1-
Variables and frequency distribution

1.1 Basic definitions

1.1.1 Variables

In engineering systems, whether we deal with raw materials, equipment, utilities or products, a lot of variables are encountered, such as mass, level of purity, flow rate, temperature, pressure, etc. These variables can be classified into two types:

Discrete variables: These can take only certain specific numerical values. For example, the maximum ambient temperature, as recorded in ^{o}C in five consecutive days whose values belong to a finite set:{25, 26, 26, 24, 27}.

Continuous variables: Here, the variable does not take specific values but can vary within a real interval taking all possible values within this interval. For example, , the mole fraction of a volatile component in a distillate fraction will belong to the real interval (0.6, 0.64).

1.1.2 Population and samples

Consider the total production per day of a factory producing sacks of fertilizers. The daily produced sacks represent a **population**. Now, if we choose 50 of these sacks to test them for their weight (For example), this set is called a **sample**. Each of the 50 sacks of this sample is called a **specimen**.

A characteristic variable of a population is called a **parameter** while that of a sample is called a **statistic**.

In the following sections we will be dealing with samples.

1.2 Frequency distribution

Consider the following data, describing the daily productivity of an oil well (bbl) over 30 days:

800	850	790	940	1000	740	820	940	960	940
970	920	850	870	800	760	1030	1010	980	890
960	900	920	990	850	930	860	840	800	780

Given this way, these data are termed **ungrouped**. If, however, a table can be drawn, that shows the different classes of concentration figures in 50 bbl intervals and the frequency of their occurrence, then this table represents **grouped data**. This is shown in Table (1.1) where the mid value of each class is shown in the third row.

Table (1.1): Class intervals of daily production

Class	**700-750**	**750-800**	**800-850**	**850-900**	**900-950**	**950-1000**	**1000-1050**
Frequency	1	3	5	6	7	5	3
Mid-class	725	775	825	875	925	975	1025

These data can be represented graphically in the form of **histogram** (Figure 1.1).

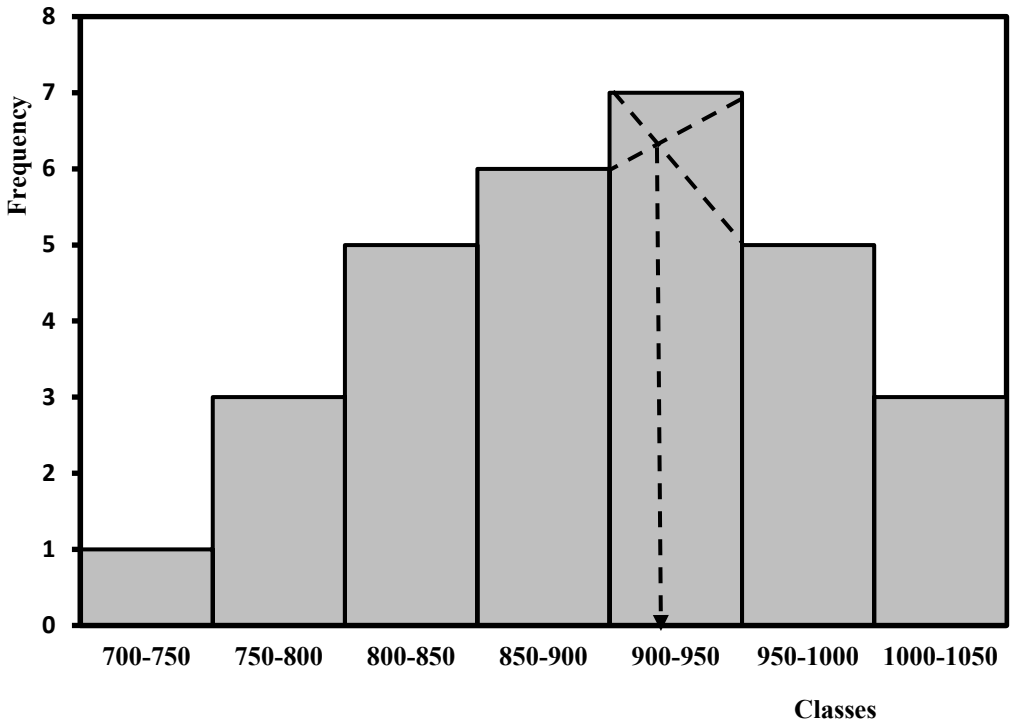


Fig (1.1): Histogram for S.M. Concentration

Plotting the frequency against mid-class values produces a broken line frequency plot, shown in Figure (1.2). It is sometimes referred to as **Frequency polygon**.

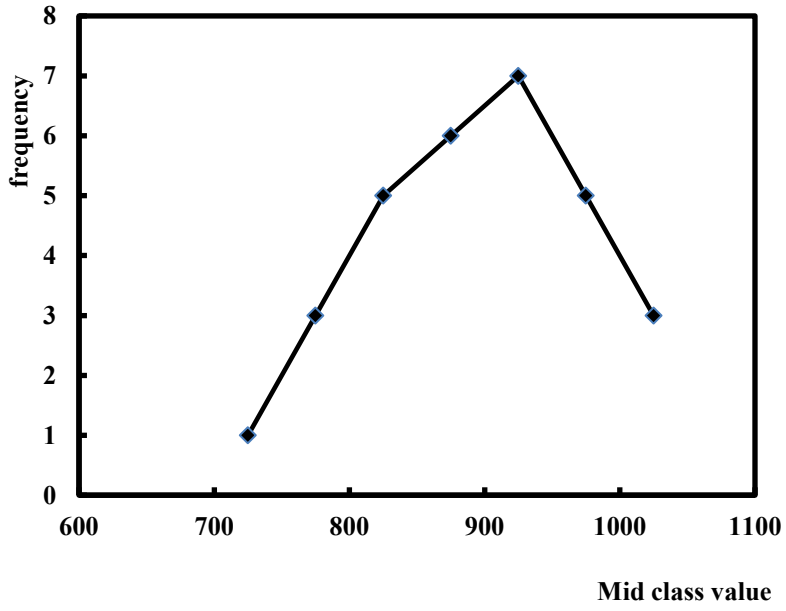


Fig (1.2): Frequency polygon

The above plot shows only one maximum value, which represents the most encountered pollutant concentration. This value is usually known as the **mode** of

distribution. This curve is thus known as a **monomoda**l curve. Bimodal curves will have two maximum values, whereas skewed curves will have their mode shifted towards one end of the class values interval.

1.3 Statistical averages

Often, we need to quote a single value that represents some typical average of the statistical distribution. There are several such values, commonly known as central tendencies or averages.

1.3.1 The mean value (Arithmetic average)

The mean (or average) is the most popular and well-known measure of central tendency. For a sample of **ungrouped** values such as $x_1, x_2, x_3 \ldots x_i, \ldots x_n$, the mean value is simply defined as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.1}$$

A similar definition for a population of N ungrouped values can be written:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{1.2}$$

Using EXCEL, it is possible to obtain the mean value by simply highlighting the numbers using the function **AVERAGE**.

One obtains the value $\bar{x}$ **= 889.7**

For a sample of **grouped** values, the definition becomes:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i \bar{x}_i}{n} \tag{1.3}$$

Whereas for a population, the corresponding expression is:

$$\mu = \frac{\sum_{i=1}^{N} f_i \bar{x}_i}{N} \tag{1.4}$$

In both cases, $\bar{x}_i$ represents the mid-class value. The following table shows the calculations using grouped data from Table (1.1):

Table (1.2): Calculation of mean value

$\bar{x}_i$	725	775	825	875	925	975	1025	**Sum**
f_i	1	3	5	6	7	5	3	**30**
$f_i\bar{x}_i$	725	2325	4125	5250	6475	4875	3075	**26850**

$$\bar{x} = \frac{26850}{30} = \mathbf{895}$$

This value is slightly different from the exact value obtained from raw data (889.7)

1.3.2 The median of a distribution

In an array of **ungrouped data** arranged by ascending order of magnitude, the median is the middle value. For example, the following data shows the maximum temperature recorded over one week period in Cairo (in ^{o}C):
26, 28, 29, 27, 25, 24, 24.

The data are then grouped in ascending order.
24, 24, 25, **26**, 27, 28, 29.
As these data are grouped in an ascending order, the median = 26.

If the number of data points is even, then the median value will consist of the arithmetic average of the two middle values. In any case, the EXCEL function **MEDIAN** directly discloses the median value of any set of ungrouped data. For the data in Table (1.1), the median value is $\boldsymbol{M = 895}$.

For **grouped data**, a cumulative plot is made, and the median is the value corresponding to one half the sum of all frequencies.

For example, in the preceding example, Table (1.2) can be re-written in terms of "number of days where the production is less than…" as shown in Table (1.3):

Table (1.3): Cumulative distribution of data in Table (1.1)

End value of class	750	800	850	900	950	1000	1050
No of days with production less than	1	4	9	15	22	27	30

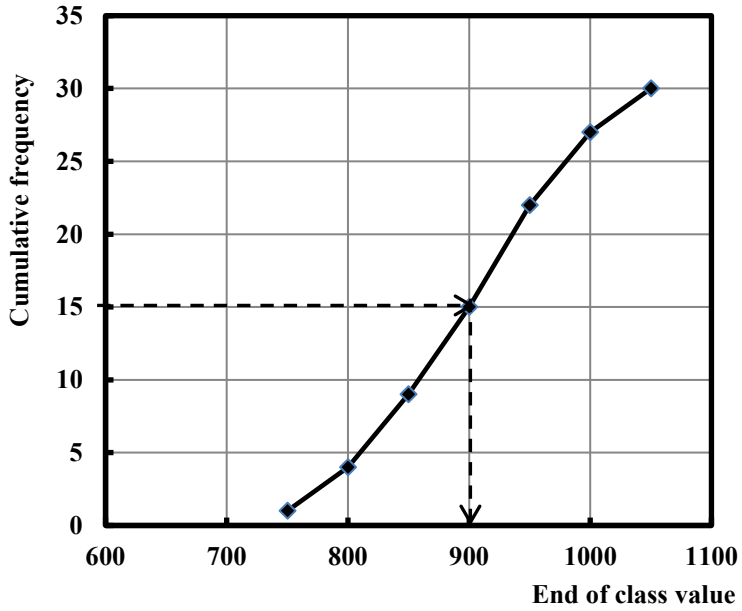


Fig (1.3): Cumulative frequency curve

The cumulative plot represents the cumulative frequency of values less than the entry in the table. This cumulative plot is called an **ogive**. The plot is performed for cumulative frequency against end of class values.

This cumulative plot is shown in Figure (1.3) and the median value is obtained at a frequency = $0.5 \times 30 = 15$. Its value is about **900**.

1.3.3 The mode of a distribution

In its simplest form, the mode can be defined as the value which occurs at the highest frequency among the data. In Table (1.2), this value is **940.** However, this value is only approximate, since it corresponds to the mid – value of the interval.

To get a more accurate value, the following interpolation can be made within the class corresponding to the maximum frequency (modal class):

$$Mo = L_1 + \frac{D_1 \times i}{D_1 + D_2} \tag{1.5}$$

Where:

L_1 represents the lower value of modal class
D_1 is the difference in the frequency of modal class and the previous one
D_2 is the difference in the frequency of modal class and the next one
i is the size of class interval.

For data of Table (1.1), the modal class is 900 –950, hence $L_1 = 900$
$D_1 = 7 - 6 = 1$ and $D_2 = 7 - 5 = 2$ and $i = 50$

Substituting in equation (1.6), one gets:

Mo ≈ 916.7

A simpler graphical method relying on the above equation can be applied to get the value of the mode with reasonable accuracy. This can be followed in Figure (1.1) where the interpolation is made graphically to give: Mode ≈ 920.

1.3.4 Level of measurements of variables

In some cases, we are confronted with non – numerical data. In this respect the following represents the classical **level of measurement** of data that is treated statistically:

These are categorized as **nominal**, **ordinal, interval, or ratio** variables.

Nominal variables are those variables that have no specific numerical value. They may be numbered only for convenience. For example, chemicals in a warehouse are categorized as: 1. Inorganic solids, 2. inorganic solutions, 3. Organic solids, 4. Organic liquids, etc. Note that the numbering system is extremely arbitrary, and the value of the number holds no indication about its magnitude. Nominal variables do not have a median or mean value. They may, however, have a mode. For example, if we consider a class list the most frequent surname will be the mode.

Ordinal variables, on the other hand, possess numerical values that can only help in ordering them in an ascending or descending way. For example, the Moh's scale of hardness sorts the minerals according to their hardness in a 1 to 10 scale. Talc, which is assigned as the ordinal 1 is the mineral of least hardness while quartz is much harder (7). On this scale, diamond is the hardest mineral with an ordinal of

10. Such numbers bear only relative numerical significance since quartz is not 7 times as hard as talc. For such variables a median and a mode can both be defined but not a mean value.

Interval variables: These are variables that can be either added or subtracted or multiplied by a constant number. A familiar example of interval scale measurement is temperature with the Celsius scale. In this scale, the unit of measurement is 1/100 of the temperature difference between the freezing and boiling points of water under a pressure of 1 atmosphere. The "zero point" on an interval scale is arbitrary; and negative values can be used. Variables measured at the interval level are called "interval variables" or sometimes "scaled variables" as they have units of measurement.

Ratios between numbers on the scale are not meaningful, so operations such as multiplication and division cannot be carried out directly. But ratios of differences can be expressed.

The central tendency of a variable measured at the interval level can be represented by its mode, its median, or its arithmetic average. Statistical dispersion can be measured in most of the usual ways, such as range and standard deviation. Since one cannot divide, one cannot define measures that require a ratio, such as coefficient of variation.

Ratio variables: This represents the highest level of measurement and includes common variables that can be subjected to all numerical operations.

1.4 Measures of dispersion

1.4.1 The concept of dispersion

The following data represents the daily consumption of electrical energy of a small factory over a one-week period (in kWh): 310, 330, 400, 290, 320, 290, 300. Their mean value = 320.

On the other hand, another factory yields the following figures: 310, 330, 300, 360, 290, 350, 300. The mean value = 320.

Although both distributions have the same mean value, we can see that the difference between the highest and the lowest values in the first is 110, while it is 70 in the second. This suggests that the first distribution is **more dispersed** than the second.

The difference between the highest and lowest values is known as **the range** and represents the crudest measure of dispersion.

This measure can be misleading: For, if we discard the number 400 from the first set of data, we get a range of 40.

Another commonly used measure of dispersion is called the interquartile range (IQR). This will not, however, be discussed in this work.

The most reliable measure of dispersion used in statistical analyses is the **standard deviation**.

1.4.2 The Standard Deviation

(a) For **ungrouped data**, the standard deviation of **sample** data is defined by:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{1.6}$$

For a **population** the corresponding expression is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \tag{1.7}$$

Equation (1.7) can be simplified as follows:

$$\sum_{i=1}^{N}(x_i - \mu)^2 = \sum_{i=1}^{N}x_i^2 - 2\mu.x_i + \mu^2 = \sum_{i=1}^{N}x_i^2 - 2\mu.\sum_{i=1}^{N}x_i + N.\mu^2$$

Since: $\sum_{i=1}^{N}x_i = N.\mu$, the previous equation simplifies to:

$$\sum_{i=1}^{N}(x_i - \mu)^2 = \sum_{i=1}^{N}x_i^2 - N.\mu^2$$

Hence equation (1.7) simplifies to:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}x_i^2}{N} - \mu^2} \tag{1.8}$$

For a sample size n the equation is slightly different:

$$s = \sqrt{\frac{\sum_{i=1}^{n}x_i^2 - n.\bar{x}^2}{n-1}} \tag{1.9}$$

Using EXCEL, the function **STDEV.S** is applied for a sample and **STDEV.P** for a population. For the sample data in Table (1.1), one obtains: ***s* = 81.1**

(b) For **grouped data**, the standard deviation of **sample** data is defined by:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\bar{x}_i^2.f_i - n.\bar{x}^2}{n-1}} \tag{1.10}$$

For a **population** the corresponding expression is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\bar{x}_i^2.f_i}{N} - \mu^2} \tag{1.11}$$

Calculations of the standard deviation using the grouped data in Table (1.2) are displayed in Table (1.4)

Table (1.4): Calculation of standard deviation from grouped data

$\bar{x}_i$	725	775	825	875	925	975	1025	Sum
f_i	1	3	5	6	7	5	3	30
$f_i\bar{x}_i$	725	2325	4125	5250	6475	4875	3075	26850
$f_i\bar{x}_i^2$	525625	1801875	3403125	4593750	5989375	4753125	3151875	24218750

Hence $s = \sqrt{\frac{24218750 - 30\times 895^2}{30-1}} =$ **80.5**

This value is close to the exact value of 81.1 obtained from raw data.

It is to be finally noted that the units of standard deviation are identical to those of the variable under consideration.

1.4.3 The Variance

The variance can be thought of as the square of the standard deviation, implying its being a measure of dispersion as well.

For a population and a sample of ungrouped data respectively:

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} \qquad (1.12)$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \bar{x}^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \qquad (1.13)$$

These can be directly obtained by using the EXCEL function **VAR.P** or **VAR.S** respectively.

The corresponding formulas for grouped data are:

$$\sigma^2 = \frac{\sum_{i=1}^{N} \bar{x}_i^2 . f_i}{N} - \mu^2 \qquad (1.14)$$

And

$$s^2 = \frac{\sum_{i=1}^{n} \bar{x}_i^2 . f_i - n.\bar{x}^2}{n-1} \qquad (1.15)$$

1.4.4 The coefficient of variation (CV)

When comparing dispersions in two distributions using their standard deviations as measure, the variables must have the same units, and their mean values should be equal. Since this is not usually the case, the comparison is rather based on the **coefficient of variation**, defined for populations and samples respectively as:

$$\text{CV} = \frac{\sigma}{\mu} \times 100\% \tag{1.16}$$

$$\text{CV} = \frac{s}{\bar{x}} \times 100\% \tag{1.17}$$

For example, for the ungrouped data of Table (1.1), since $\bar{x}$ = 889.7 and s = 81.1, the coefficient of variation is:

$$\text{CV} = \frac{81.1}{889.7} \times 100\% = \mathbf{9.11\%}$$

The coefficient of variation allows investors to determine how much risk is assumed in comparison to the amount of expected return. The lower the coefficient of variation, the better is the risk-return trade-off.

1.5 Skewness of a distribution

An ideal distribution where the mean, medium and mode are equal, is a symmetrical distribution. Figure (1.4)

On the other hand, some distributions are not symmetrical as can be seen from the same figure. These are called **skewed distributions**. This is calculated from the following formulas of ungrouped data:

$$S_k = \frac{\sum_{i=1}^{N}(x_i - \mu)^3}{N.\sigma^3} \text{ (For a population)} \tag{1.18}$$

As for samples:

$$S_k = \frac{\sum_{i=1}^{n} n.(x_i - \bar{x})^3}{(n-1)(n-2).s^3} \text{ (For a sample)} \tag{1.19}$$

For ungrouped data, skewness is readily obtained from the EXCEL function: **SKEW.P** or **SKEW** that uses equations (1.18) or (1.19) respectively. For the data in Table (1.1), this function yields a low skewness of $\mathbf{-0.094}$. This is emphasized by the high symmetry observed in Figure (1.2).

For grouped data, similar formulas can be applied:

$$S_k = \frac{\sum_{i=1}^{N} f_i(x_i - \mu)^3}{N.\sigma^3} \text{ (For a population)} \tag{1.20}$$

$$S_k = \frac{\sum_{i=1}^{n} nf_i.(x_i - \bar{x})^3}{(n-1)(n-2).s^3} \text{ (For a sample)} \tag{1.21}$$

A positive value of skewness means that the distribution possesses a "tail" to the right while the tail is at the left for negative values of skewness.

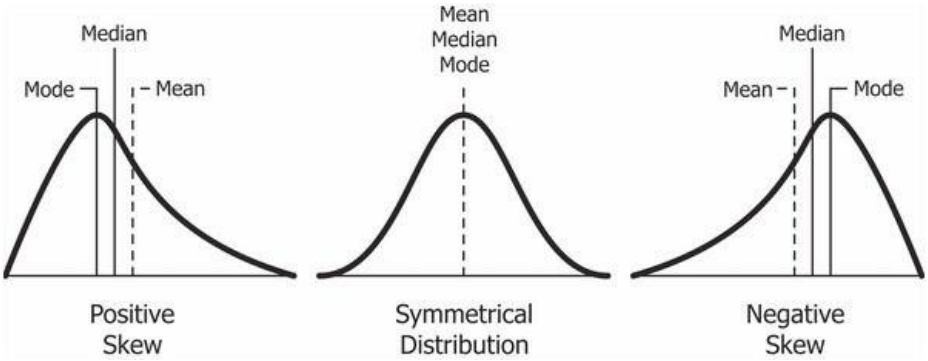


Fig (1.4): Symmetrical and skewed distributions

1.6 Kurtosis

Kurtosis defines the flatness of a distribution rather than its being symmetrical or not. It is defined by a formula using ungrouped data. It shows how far the distribution is "tailed". A positive value of kurtosis indicates an elongated distribution (Leptokurtic) while a negative value will denote a flat distribution (Platykurtic) (Figure 1.5).

The value of kurtosis can be obtained for ungrouped data from the EXCEL function KURT. For the data of Table (1.1), $= \mathbf{-1.094}$, which is considered to denote a moderate platycurtic distribution (Figure 1.5). If its value < -2, the distribution is considered to be extremely flat.

For ungrouped data, kurtosis is defined as follows:

$$K_r = \frac{\sum_{i=1}^{N}(x_i - \mu)^4}{N.\sigma^4} \qquad (1.22)$$

This formula has been challenged by many statisticians since it does not allow for negative values. It has been replaced by more sophisticated ones.

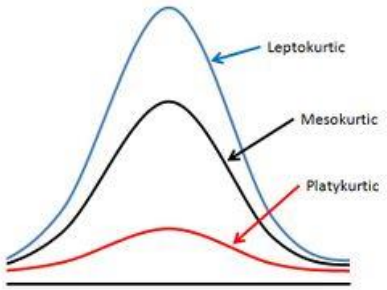


Fig (1.5): Kurtosis as related to curve shapes

1.7 Statistical moments

Statistical moments are parameters that define the distribution of a population. The n^{th} moment of a distribution is defined by:

$$M_n = \sum_{i=1}^{N}(x_i - \mu)^n \qquad (1.23)$$

They are related to the various statistical parameters as follows:

1.7.1 First moment

The First Moment of a distribution is related to its arithmetic average which for ungrouped data takes the form:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Hence, $\sum_{i=1}^{N}(x_i - \mu) = 0$ (First moment) (1.2)

1.7.2 Second moment

The second moment measures the dispersion of the distribution through its variance, and is defined by:

$\sum_{i=1}^{N}(x_i - \mu)^2 = N.\sigma^2$ (Second moment) (1.12)

1.7.3 Third moment

The third moment relates to skewness as follows:

$\sum_{i=1}^{N}(x_i - \mu)^3 = S_k.N.\sigma^3$ (Third moment) (1.18)

1.7.4 Fourth moment

The fourth moment describes the flatness of the distribution, being related to kurtosis

$\sum_{i=1}^{N}(x_i - \mu)^4 = K_r.N.\sigma^4$ (Fourth moment) (1.22)

1.8 Exercise problems

(1) The following data were collected over a two-week period as a sample for COD of wastewater ($mg.L^{-1}$) of a food processing plant.

1550 2070 1800 2020 1560 2700 2530 2100 1890 2050 2450 1720 1050 1400

Estimate the mean value of COD, its standard deviation, median value, the skewness of the distribution and its kurtosis.

(2) The following table represents the ages of a random sample of 130 people:

Age	[0 –10)	[10 – 20)	[20 – 30)	[30 – 40)	[40 – 50)	[50 – 60)	[60 – 70)	[70 – 80)
Frequency	2	18	26	37	22	14	9	2

Calculate the average sample age, its standard deviation, median value and the mode of that distribution.

(3) The following table shows the incidence of stoppage occurring in a production line during a 320-day year

Nº of stoppages	0	1	2	3	4	5	6	7
Days	84	88	64	38	25	11	8	2

Calculate the average number of stoppages and the variance.

(4) The following data represents the scores of a class of 40 students in a test.

16	14	15	9	3	16	11	19	5	13
20	5	6	17	13	14	14	17	11	10
4	6	18	20	16	17	11	8	9	1
20	14	19	18	12	19	7	13	7	6

Obtain the average score and the standard deviation out of these raw data. Also determine the median and skewness of the distribution.

(5) The following table summarizes the chemical analyses related to the purity of samples of an ore obtained from two different quarries A and B. Compare the mean values of ore purity and their COV. In your opinion which quarry is more reliable?

Sample number	1	2	3	4	5	6	7	8	9	10
% Purity (A)	39	44	52	37	40	45	32	55	48	43
% Purity (B)	40	42	46	39	37	41	45	44	44	39

(6) Determine the values of the first four moments from the following data: $2, 5, 12, 8, 6, 11, 17, 3.$

-2-
Probabilities

2.1 The statistical definition of probability

Consider 10 samples taken at random from a production line for insulating panels. Following standards, the density of the product should not exceed 1000 kg/m^3. Let the measured densities be represented by the set: S = {995, 985, 995, 1025, 995, 985, 1010, 975, 990, 1030}.
The probability of getting a defective sample will simply be 3 out of 10, that is 0.3.

In general, the set of all samples under consideration is called the **sample space *S***. Any number of samples belonging to S is called a sub-set of S. This sub-set is called an **event**. For example, in the above case, the event: "The chosen item is defective" corresponds to a sub-set:

A = {1025, 1010, 1030}

Let the number of elements of the sample space = $N(S)$, and that of the event A = $N(A)$. Then the probability of occurrence of A is:

$$P(A) = \frac{N(A)}{N(S)} \tag{2.1}$$

2.2 Calculation of *N*(*A*)

2.2.1 Elements of combinatory algebra

To calculate $N(S)$ or $N(A)$, we recall some basic definitions:

(a) The **factorial *n*!** of a positive integer n

$$n! = 1 \times 2 \times 3 \times 4 \ldots n \tag{2.2}$$

This represents the number of ways one can choose n items out of n, the choice being done **one by one without replacement.**

(b) A **combination C_r^n** represents the number of ways one can choose r items out of n, the choice being done **once at a time.**

The value of C_r^n can be calculated from the following formula:

$$C_r^n = \frac{n!}{r!\,n-r!} \tag{2.3}$$

For example, out of 8 different brands of paint, by how many methods can one choose 3 brands?

$$C_3^8 = \frac{8!}{3!\,8-3!} = \frac{40320}{6\times 720} = 56$$

2.2.2 Types of choice

Consider event A consisting of choosing r items out of n. There are several methods by which this choice can be accomplished:

(a) Items are chosen **one by one with replacement**:

$$N(A) = n^r \tag{2.4}$$

(b) Items are chosen **one by one without replacement**:

$$N(A) = \frac{n!}{n-r!} \tag{2.5}$$

(c) Items are chosen **once at a time**

$$N(A) = C_r^n \tag{2.6}$$

For example, we need to choose 3 samples out of 10.

If drawing of samples was done one by one with replacement, then:

$N(A) = 10^3 = 1000$

If drawing of samples was done one by one without replacement, then:

$$N(A) = \frac{10!}{10-3!} = 720$$

If drawing of samples was done once, at a time, then:

$$N(A) = \frac{10!}{3!10-3!} = 120$$

2.3 Elements of event algebra

Consider the sample space $S = \{a_1, a_2, a_3, \ldots, a_n\}$. This consists of n single events. Any sub-set of S, like A usually consists of either one **elementary event** such as $\{a_3\}$ or a **compound event** such as $\{a_2, a_3\}$.

For example, let $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and let a random number be chosen from that set.

Then event A = "An odd number is obtained", $A = \{1, 3, 5, 7, 9\}$

Event B: "the number obtained is ≥ 6", $B = \{6, 7, 8, 9, 10\}$

The event: "A multiple of 6 is obtained" is $C = \{6\}$

These events are represented in Figure (2.1) by the **Venn diagram**.

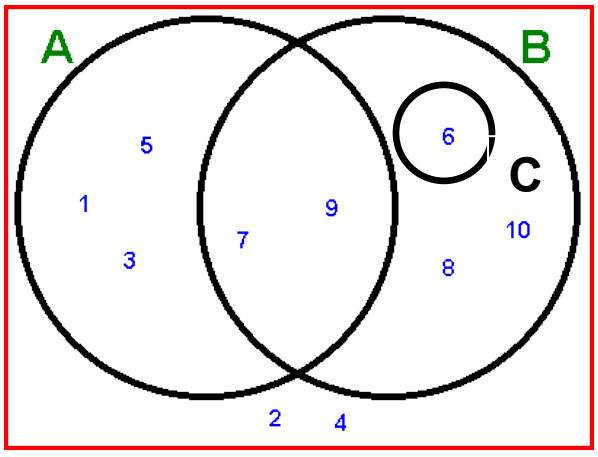


Fig (2.1) Venn diagram

- A **complementary event** A' is the set of single events (elements) that do not belong to A. In the present case: $A' = \{2, 4, 6, 8, 10\}$
- An **impossible event** is an event containing zero elements, like throwing a die and obtaining the number 7. It is written $\emptyset$.
- The **intersection** of two events is the set of elements belonging to both A and B. It is written $A \cap B$. In the present case: $A \cap B = \{7, 9\}$, and $B \cap C = \{6\}$
- The **union** of two events is the set of elements belonging to either A or B or both. It is written $A \cup B$. In the present case: $A \cup B = \{1, 3, 5, 6, 7, 8, 9, 10\}$ and $A \cup C = \{1, 3, 5, 6, 7, 9\}$
- The **difference** between two events A and B is the set of elements present in A but not in B: $A - B = A \cap \bar{B} = \{1, 5, 3\}$
- Two events A and C are said to be **mutually exclusive** if $A \cap C = \emptyset$
- Finally, if C is **sub-set** of B, $C \subseteq B$, then $B \cap C = C$ and $B \cup C = B$

Example 2.1
A pump house contains 4 rotary and 6 centrifugal pumps. If 2 pumps are chosen at random, what is the probability that both will be of the latter type?

Solution
$N(S) = C_2^{10} = 45$
Let A be the event: The two chosen pumps are of the centrifugal type.
Hence, $N(A) = C_2^6 = 15$
From equation (2.1): $P(A) = \frac{15}{45} = \frac{1}{3}$

2.4 The axiomatic definition of probability

Consider the sample space S consisting of a finite number of elementary events $\{A_1, A_2, A_3, \dots, A_i, \dots, A_n\}$.

If A is a sub-set of S, then the probability function $P(A)$ can be defined according to the following axioms:

(1) $P(A) > 0 \quad (A \neq \emptyset)$
(2) $P(S) = 1$
(3) If A and B are two mutually exclusive events, then:
$P(A \cup B) = P(A) + P(B)$

Applying these axioms on the elementary mutually exclusive events of S, the following two conditions are obtained:

(1) $P(A_i) > 0$
(2) $\sum P(A_i) = 1$

For example, the following table shows the probability that out of 100items chosen from a production line, there will be (n) defective items.

n	0	1	2	3	4	5	6
Probability	0.2	0.25	0.3	0.15	0.06	0.03	0.01

It is easy to notice that the two previous conditions are fulfilled.

2.5 Basic laws of probability

The following laws represent the basic laws of probability. They are given here without proof.

(1) $P(\emptyset) = 0$ (2.7)

(2) $P(A') = 1 - P(A)$ (2.8)

(3) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (2.9)

Note that if A and B are mutually exclusive, then $A \cap B = \emptyset$ and this law reduces to the third axiom.

Example 2.2

In a statistic involving the origin of the main raw material used by 30 factories, it was found that 18 of them draw their raw materials from source A, 16 from source B, and 7 from both sources. If a factory is chosen at random, what is the probability that:

(1) The factory gets its raw materials from either A or B

(2) The factory gets its raw materials from a totally different source.

Solution:

$P(A) = \frac{18}{30}$, $P(B) = \frac{16}{30}$ and $P(A \cap B) = \frac{7}{30}$

(1) This is the event $A \cup B$ the probability of which can be obtained from equation (2.9):

$$P(A \cup B) = \frac{18}{30} + \frac{16}{30} - \frac{7}{30} = \mathbf{\frac{9}{10}}$$

(2) This is the complementary event $(A \cup B)'$

Its probability is calculated from equation (2.8):

$$P(A \cup B)' = 1 - P(A \cup B) = \mathbf{\frac{1}{10}}$$

Example 2.3

In a research facility there are 8 chemists and 4 engineers. Two people were chosen at random.

What is the probability that both are chemists? Both are engineers. One is a chemist and the other an engineer.

Solution:

$N(S) = C_2^{12} = 66$

Let event A be: Both people are chemists, then $N(A) = C_2^8 = 28$

Hence $P(A) = \frac{\mathbf{28}}{\mathbf{66}} = \mathbf{0.424}$

Let event B be: Both people are engineers, then $N(B) = C_2^4 = 6$.

Hence $P(B) = \frac{\mathbf{6}}{\mathbf{66}} = \mathbf{0.091}$

The remaining situation (One chemist and one engineer) is the complementary event of both A and B:

Its probability = 1 – (0.424 + 0.091) = **0.485**

2.6 Conditional probability

2.6.1 Basic concept

Consider the following experiment: Out of a bag containing 7 white balls and 3 black balls are drawn 2 balls, one after the other, without replacement. $N(S) = 10 \times 9 = 90$.

Let event A be: the two balls are black. There will be 3 chances in the first draw and 2 in the second **if the first was black**. The probability of getting two successive black draws will be: $\frac{\mathbf{3}}{\mathbf{10}} \times \frac{\mathbf{2}}{\mathbf{9}} = \frac{\mathbf{1}}{\mathbf{15}}$

The first term in the above product is the probability of getting a black ball in the first draw $P(A_1)$ while the second term represents the probability of drawing a second black ball if the first one was black. This is written: $P(A_2/A_1)$ and is termed **conditional probability**. The general rule governing such situation is:

$$P(A_1 \cap A_2) = P(A_1).P(A_2/A_1) \qquad (2.10)$$

2.6.2 The total probability – Bayes theorem

Consider the following situation: In a factory there are three production lines. The first, A, produces 30% of the total output stock, B produces 20% and C produces 50%. It is known from past practice that 5% of products from line A are defective, while the percentage is 4% from B and 6% from C. We wish to calculate the probability that a sample selected at random will be defective $P(D)$.

This probability will be:

$$P[(A \cap D)U(B \cap D)U(C \cap D)] = P(A \cap D) + P(B \cap D) + P(C \cap D)$$

Following equation (2.10), this can be written as:

$$P(D) = P(A).P(D/A) + P(B).P(D/B) + P(C).P(D/C)$$

$$P(D) = 0.3 \times 0.05 + 0.2 \times 0.04 + 0.5 \times 0.06 = 0.053$$

Now, one may ask: If an item was found to be defective, what is the probability that it would have come from line C? It will be required to calculate $P(C/D)$.

From equation (2.10), we may write: $P(C \cap D) = P(D).P(C/D)$

Hence, $0.5 \times 0.06 = 0.053 \times P(C/D)$, from which $P(C/D) = 0.566$

The steps undertaken to solve the latter problem can be written in a more general form as follows:

$$P(B) = \sum_{i=1}^{n} P(A_i).P(B/A_i) \tag{2.11}$$

This is **the law of total probability.**

And,

$$P(A_i/B) = \frac{P(A_i).P(B/A_i)}{\sum_{i=1}^{n} P(A_i).P(B/A_i)} \tag{2.12}$$

This formula is known as **Bayes theorem**.

Example 2.4

In a plant, there are four major departments: Technical (A), sales (B), financial (C) and R&D (D). The percentage of personnel in these departments represent 40%, 15%, 30% and 15% of the plant working power respectively. The percent of women in these departments is: 25%, 20%, 45% and 30% respectively.

An employee was chosen at random. What is the probability that it was a woman? And if it turned out to be a woman, what is the probability that it would have come from department C?

Solution:

The following probabilities are readily calculated:

$P(A) = 0.4, P(B) = 0.15, P(C) = 0.3$ and $P(D) = 0.15$.

$P(W/A) = 0.25, P(W/B) = 0.2, P(W/C) = 0.45, P(W/D) = 0.3$

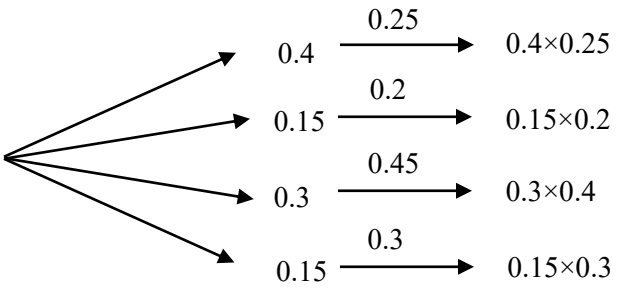
The law of total probability is applied:

$P(W) = \sum P(A_i).P(W/A_i)$

$= 0.4 \times 0.25 + 0.15 \times 0.2 + 0.3 \times 0.45 + 0.15 \times 0.3 = \mathbf{0.31}$

Bayes theorem can then be written as:

$P(C/W) = P(C \cap W)/P(W) = \frac{0.3 \times 0.45}{0.31} = \mathbf{0.435}$



Total = 0.31

2.7 Independent events

Consider once more the following experiment: Out of a bag containing 7 white balls and 3 black balls are drawn 2 balls, one after the other, but this time with replacement. $N(S) = 10 \times 10 = 100$.

Let event A be the two balls are black. There will be 3 chances in the first draw and 3 also in the second. The probability of getting two successive black draws will be: $\frac{3}{10} \times \frac{3}{10} = \frac{9}{100}$.

In this case, the probability of the second draw does not depend on the outcome of the first. Such events are said to be independent.

In this case, $P(A_2/A_1)=P(A_2)$, and equation (2.10) can be written as:

$$P(A_1 \cap A_2) = P(A_1).P(A_2) \qquad (2.13)$$

This is the law of **independent events.**

Example 2.5

An oil company is bidding for the rights to drill a well in field A and a well in field B. The probability it will drill an oil well in field A is 40%. If it does, the probability the well will be successful is 45%. The probability it will drill a well in field B is 30%. If it does, the probability the well will be successful is 55%. Calculate each of the following probabilities:

(a) Probability of a successful well in field A,
(b) Probability of a successful well in field B,
(c) Probability of both a successful well in field A and a successful well in field B,
(d) Probability of at least one successful well in the two fields together,
(e) Probability of no successful well in the two fields together
(f) Probability of exactly one successful well in the two fields together.

Solution:

$P(A) = 0.4 \quad P(S/A) = 0.45 \qquad\qquad P(B) = 0.3 \quad P(S/B) = 0.55$

(a) $P(A \cap S) = P(A).P(S/A) = 0.4 \times 0.45 = \mathbf{0.18}$

(b) $P(B \cap S) = P(B).P(S/B) = 0.3 \times 0.55 = \mathbf{0.165}$

(c) $P[(A \cap S) \cap (B \cap S)] = 0.18 \times 0.165 = \mathbf{0.0297}$ (Independent events)

(d) $P[(A \cap S) \cup (B \cap S)] = 0.18 + 0.165 - 0.0297 = \mathbf{0.3153}$

(e) $P[(A \cap S)' \cap (B \cap S)'] = P[(A \cap S) \cup (B \cap S)]' = 1 - 0.3153 = \mathbf{0.6847}$

(f) $P[(A \cap S) \cup (B \cap S)] - P[(A \cap S) \cap (B \cap S)] = 0.3153 - 0.0297 = \mathbf{0.2856}$

Example 2.6

The probability that an experiment will produce a positive result is 0.8. How many times do we have to repeat that experiment so that the probability of obtaining at least one positive result exceeds 0.998?

Solution:
The sequence of experiments is assumed to represent independent events. Let the probability of obtaining a positive result = $P(A) = 0.8$. Hence the probability of failure = $P(A') = 0.2$.

The probability of n consecutive experiments to fail = 0.2^n.

Therefore, the probability of having at least one positive result will be $1 - 0.2^n$.

This means solving the inequality $1 - 0.2^n > 0.999 \rightarrow 0.2^n < 0.001$.

$n.\ln 0.2 > \ln 0.001 \rightarrow n > 4.292 \rightarrow \boldsymbol{n = 5}$

Example 2.7
In a lot containing 100 samples of a certain commercial product, 5 are known to be defective. If 3 samples are drawn, one after one with replacement, calculate the following probabilities:
(1) The three samples are defective
(2) At least one sample is not defective
(3) The three samples are not defective

Solution:
Since drawing has been done one after the other with replacement, then we are in presence of independent events.

(1) $P(D \cap D \cap D) = (0.05)^3 = \boldsymbol{0.000125}$

(2) This is the complementary event of $D \cap D \cap D$: Its probability is therefore:

$1 - P(D \cap D \cap D) = 1 - 0.000125 = \boldsymbol{0.999875}$

(3) $P(D' \cap D' \cap D') = (1 - 0.05)^3 = (0.95)^3 = \boldsymbol{0.875}$

2.8 Markov Chains

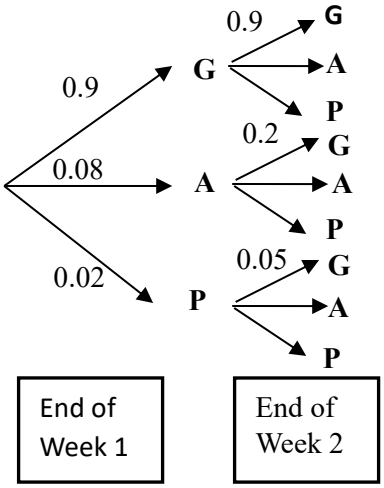
Consider the following situation:
A production line includes a set of control valves regulating its operation. Their performance is rated as Good (G), average (A) or poor (P). The probability that any of such performances as established during the first week of operation will affect the second week's performance is given in the following table by following quality control charts for the first year of operation.

	G	**A**	**P**
G	0.9	0.08	0.02
A	0.2	0.7	0.1
P	0.05	0.25	0.7

The matrix *M* so defined is called the **transition matrix**:

$$\begin{pmatrix} 0.9 & 0.08 & 0.02 \\ 0.2 & 0.7 & 0.1 \\ 0.05 & 0.25 & 0.7 \end{pmatrix}$$

Assume that we began in the first week with a control system rated G then, the first column in the following "tree diagram" shows the probabilities of the corresponding performance at the end of this first week.



The probability of a good performance at the end of the second week is the sum of the following products:

$0.9\times0.9 + 0.08\times0.2 + 0.02\times0.05 = 0.827$

Had we ended with an average performance, this product would have been:

$0.9\times0.08 + 0.08\times0.7 + 0.02\times0.25 = 0.133$

And, finally, in case of ending with poor performance:

$0.9\times0.02 + 0.08\times0.1 + 0.02\times0.7 = 0.04$

Note that the sum of these probabilities is 1.

These represent the first row of the matrix $M^2 = M*M$
Upon matrix multiplication we get the following expression for M^2:

$$\begin{pmatrix} 0.827 & 0.133 & 0.04 \\ 0.325 & 0.531 & 0.144 \\ 0.13 & 0.354 & 0.516 \end{pmatrix}$$

To follow up the performance after any number of weeks we keep on finding the expressions of the matrices M^n for increasing values of n:

For example, M^4 and M^8 will respectively show as:

$$\begin{pmatrix} 0.7324 & 0.1948 & 0.4601 \\ 0.4601 & 0.3762 & 0.1638 \\ 0.2896 & 0.3879 & 0.3224 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0.6471 & 0.2442 & 0.1088 \\ 0.5574 & 0.2946 & 0.1479 \\ 0.484 & 0.3274 & 0.1886 \end{pmatrix}$$

The following matrices are M^{16}, M^{32} and M^{64} respectively:

$$\begin{pmatrix} 0.6074 & 0.2656 & 0.127 \\ 0.5965 & 0.2714 & 0.1321 \\ 0.587 & 0.2764 & 0.1366 \end{pmatrix}$$

$$\begin{pmatrix} 0.6025 & 0.2682 & 0.1293 \\ 0.6012 & 0.2689 & 0.1299 \\ 0.6 & 0.2695 & 0.1305 \end{pmatrix}$$

$$\begin{pmatrix} 0.6019 & 0.2685 & 0.1296 \\ 0.6018 & 0.2685 & 0.1296 \\ 0.6018 & 0.2685 & 0.1296 \end{pmatrix}$$

We note that as $n \to \infty$, the entries of the columns stabilize to constant values, meaning that the performance after enough time does not depend on the initial state.

In the present example, let the initial performance of purchased valves be as follows: 95% *G*, 4% *A*, 1% *P*. These data are represented by the row matrix $A_0 = (0.95, 0.04, 0.01)$.
The performance after n weeks is calculated from the row matrix:

$$A_n = A_0 . M^n \qquad (2.14)$$

Then:

$$A_n = (0.95, 0.04, 0.01) \times \begin{pmatrix} 0.6019 & 0.2685 & 0.1296 \\ 0.6018 & 0.2685 & 0.1296 \\ 0.6018 & 0.2685 & 0.1296 \end{pmatrix}$$

Hence, $A_n = (0.6019, 0.2685, 0.1296)$

This means that whether the initial performance was good or average or poor, then after about one year, the probability of having a good performance will equal 0.6019, whereas an average performance will have a probability of 0.2685 and there will be 0.1296 probability of having a poor performance.

This chain of matrices is called a **regular Markov chain** and establishes an extremely important result regarding the performance of different equipment in a production line. That is: **The performance after enough time does not depend on the initial state.**

Generally, let the steady state Markov matrix = $M^n \times \begin{pmatrix} a & b & c \\ a & b & c \\ a & b & c \end{pmatrix}$

And the initial status be $A_0 = (x_0, y_0, z_0)$.

Then, from equation (2.14):

$$A_n = a.(x_0 + y_0 + z_0) + b.(x_0 + y_0 + z_0) + c.(x_0 + y_0 + z_0)$$

Since $x_0 + y_0 + z_0 = 1$, then $A_n = (a, b, c)$, regardless of the values of x_0, y_0, z_0.

This result is made use of in predicting long-term performance based on statistical field data gathered during the initial period of operation or from similar production lines that have been operating for some time.

Here is another interesting example:

Let us now assume that there was no initial probability that poor performance on a certain day would give way to good performance the next day. However, there is a very small probability that good performance might result in subsequent poor performance.

The transition matrix reads:

$M =$

	G	**P**
G	0.999	0.001
P	0	1

The values of matrix entries will then stabilize very slowly owing to the high probability of the a_{11} cell. For example:

$M^{8192} =$

	G	***P***
G	0.0003	0.9997
P	0	1

This result is very interesting: It means that if there is even an infinitesimal probability that a good performance might lead to a poor one, the long-term probability of having a good performance will approach zero!!

This result is used to classify the performance of some critical pieces of equipment. The value of n in M^n must exceed the lifetime of the piece for a predetermined probability of accepted performance.

For example, if the accepted probability of good performance in this example is 0.99, then we calculate M^n until the a_{11} term decreases below 0.99. This will take place for M^{11} as M^{10} shows as:

$$\begin{pmatrix} 0.99005 & 0.00995 \\ 0 & 1 \end{pmatrix}$$

This shows that the performance will get below the acceptable level after only 10 days, which is obviously unacceptable.

To decide about the initial level of performance we set the general form of the transition matrix as:

$$M = \begin{pmatrix} a & 1-a \\ 0 & 1 \end{pmatrix}$$

It can easily be shown that:

$$M^n = \begin{pmatrix} a^n & 1-a^n \\ 0 & 1 \end{pmatrix}$$

If the expected lifetime of the equipment is, say, 5 years (1500 working days), then the condition for good performance is:

$a^{1500} > 0.99$

Yielding: $a = 0.999993$

This represents the safe limit for the initial equipment performance that would guarantee an acceptable performance for five years. In practice, this corresponds to an initial failure probability of $1 - a = 0.0000067$.

The **logarithm of the reciprocal** of this figure is known as the **SIL value** (Safety Integrity Level). In this example it is in the range of 5 although in practice **its maximum value is 4.** This is the highest level, representing the most stringent requirements and a high level of risk reduction. (Like in nuclear power stations, for example).

Using EXCEL:

To perform matrix multiplication using excel:

- First: determine the order of the product matrix
- Second: Choose a region of cells having the above order
- Third: in this region, write: = MMULT (Range 1st matrix, range 2nd matrix), then CTRL SHIFT =
- The result will be displayed within the chosen region

Example 2.8

A company purchases several pumps. After 2 years of operation, their performance was recorded as follows:

Excellent: 38% - Good: 32% - Acceptable 20% and Poor 10%.

Long practice with this type of pump has made it possible to establish the following annual performance matrix:

$$\begin{pmatrix} 0.52 & 0.21 & 0.21 & 0 \\ 0.40 & 0.50 & 0.08 & 0.02 \\ 0.10 & 0.35 & 0.45 & 0.10 \\ 0 & 0.07 & 0.23 & 0.70 \end{pmatrix}$$

Find the initial performance row vector as well as after 5 years of operation.

Solution:

$A_2 = (0.38 \quad 0.32 \quad 0.2 \quad 0.1)$

Since $A_2 = A_0.M^2$ therfore $A_0 = A_2.M^{-2}$

We first obtain M^2 =

$$\begin{pmatrix} 0.3844 & 0.3132 & 0.2064 & 0.096 \\ 0.416 & 0.3758 & 0.1522 & 0.056 \\ 0.237 & 0.3655 & 0.2695 & 0.128 \\ 0.057 & 0.1815 & 0.2487 & 0.5128 \end{pmatrix}$$

We then get M^{-2} =

$$\begin{pmatrix} 6.7945 & -1.354 & -4.42 & -0.021 \\ -11.488 & 8.4261 & 3.7732 & 0.2886 \\ 10.4355 & -11.55 & 3.7438 & -1.626 \\ -1.7504 & 2.7712 & -2.66 & 2.6391 \end{pmatrix}$$

$$A_0 = (0.38 \quad 0.32 \quad 0.2 \quad 0.1) \times \begin{pmatrix} 6.7945 & -1.354 & -4.42 & -0.021 \\ -11.488 & 8.4261 & 3.7732 & 0.2886 \\ 10.4355 & -11.55 & 3.7438 & -1.626 \\ -1.7504 & 2.7712 & -2.66 & 2.6391 \end{pmatrix}$$

$$\boldsymbol{A_0 = (0.8182 \quad 0.14828 \quad 0.0105 \quad 0.023)}$$

To get A_5, we need to get M^3 since: $A_5 = A_2 \times M^3$ and $M^3 = M^2 \times M$

$$M^3 = \begin{pmatrix} 0.3844 & 0.3132 & 0.2064 & 0.096 \\ 0.416 & 0.3758 & 0.1522 & 0.056 \\ 0.237 & 0.3655 & 0.2695 & 0.128 \\ 0.057 & 0.1815 & 0.2487 & 0.5128 \end{pmatrix} \times \begin{pmatrix} 0.52 & 0.21 & 0.21 & 0 \\ 0.40 & 0.50 & 0.08 & 0.02 \\ 0.10 & 0.35 & 0.45 & 0.10 \\ 0 & 0.07 & 0.23 & 0.70 \end{pmatrix}$$

$$M^3 = \begin{pmatrix} 0.3458 & 0.3297 & 0.2073 & 0.1172 \\ 0.3819 & 0.3461 & 0.1852 & 0.0868 \\ 0.2964 & 0.3455 & 0.2201 & 0.138 \\ 0.1271 & 0.2376 & 0.2444 & 0.3909 \end{pmatrix}$$

$$A_5 = (0.38 \quad 0.32 \quad 0.2 \quad 0.1) \times \begin{pmatrix} 0.3458 & 0.3297 & 0.2073 & 0.1172 \\ 0.3819 & 0.3461 & 0.1852 & 0.0868 \\ 0.2964 & 0.3455 & 0.2201 & 0.138 \\ 0.1271 & 0.2376 & 0.2444 & 0.3909 \end{pmatrix}$$

$$\boldsymbol{A_5 = (0.3256 \quad 0.3289 \quad 0.2065 \quad 0.139)}$$

2.9 Exercise problems

(1) Polypropylene is mainly produced by two processes: process (A) and process (B). Out of 40 factories it was found that 30 adopt the first process, 8 the second process and 5 factories contain production lines of both types. If a factory is randomly chosen, calculate the following probabilities:
   (a) The factory chosen adopts either process
   (b) All the production lines of the chosen factory are type (A)
   (c) The factory uses a totally different process than either (A) or (B).

(2) A plant engineer looks for a certain type of fuse (A). He knows that 80 fuses have been purchased of which 30 are of the A type. All fuses were mixed up in

a drawer. How many fuses should he choose at a time to ensure that he will get at least one proper fuse with at least 99.9% chance?

(3) An importing company purchases chemicals from 3 different sources: A, B and C. The annual purchased amounts from these three sources are 150 tons, 200 tons and 250 tons respectively. Of these, the amounts of grade "ANALAR" reagents are 5 tons, 8 tons and 6 tons respectively. A product is chosen at random. Calculate the probability that it is of the ANALAR type. And if it turns out to be from that type, what is the probability that it originates from source B?

(4) A certain product was found to have two types of minor defects. The probability that an item of the product has only a type A defect is 0.2, and the probability that it has only a type B defect is 0.15. The probability that it has both defects is 0.1 Find the probabilities of the following events:

(a) An item has either a type A or type B defect.
(b) An item does not have any of these defects.
(c) An item has defect A, but not defect B.
(d) An item has exactly one of the two defects.

(5) Companies manufacturing LWPE get their raw materials from two suppliers only: X and Y; out of 100 factories, 55 get it from X and 65 from Y. What is the probability that a factory chosen at random would get its raw materials from both sources?

(6) A plant installs temperature controllers that were purchased with the following initial performance: (E = excellent, G = good, F = fair) corresponding to the row vector $(0.94 \quad 0.05 \quad 0.01)$. Long time experience has permitted to set the following Markov transition matrix based on weekly performance

$$\begin{pmatrix} 0.92 & 0.07 & 0.01 \\ 0.05 & 0.93 & 0.02 \\ 0.01 & 0.06 & 0.93 \end{pmatrix}$$

Estimate the performance:

After 8 weeks, after 10 weeks and the ultimate performance.

(7) The grades transition matrix of students attending a certain program is assumed to remain constant as follows:

	A	**B**	**C**	**D**	**F**
A	0.7	0.2	0.08	0.02	0
B	0.2	0.65	0.05	0.08	0.02
C	0.02	0.18	0.67	0.1	0.03
D	0	0.02	0.22	0.66	0.1
F	0	0	0.05	0.8	0.15

Find their status after 6 semesters if their original status was as follows:
$[0.15 \quad 0.25 \quad 0.35 \quad 0.25 \quad 0]$

(8) A company performs a semiannual assessment of the performance of its employees. They are graded as: E (excellent), G (good), and U (unsatisfactory). The biannual transition matrix is assumed to be constant:

$$\begin{pmatrix} 0.7 & 0.25 & 0.05 \\ 0.3 & 0.6 & 0.1 \\ 0.1 & 0.6 & 0.3 \end{pmatrix}$$

Determine the ultimate performance status.

(9) In January 2020, a company purchased several electric bulbs of different brands. Experience has shown that brand A bulbs have a superior lifetime than those of brands B or C. As time elapses, the lifetime of each type generally decreases to obtain the following semi-annual transition matrix:

$$\begin{pmatrix} 0.3 & 0.65 & 0.05 \\ 0.1 & 0.55 & 0.35 \\ 0.0 & 0.1 & 0.9 \end{pmatrix}$$

In January 2023, the status of the bulbs was given by the row vector:
$[0.044378 \quad 0.25732 \quad 0.6983]$. Determine the original status of the bulbs.

-3-
Distributions of discrete random variables

3.1 Basic concepts

Consider choosing 2 people from a plant out of 8 chemists (C) and 4 engineers (E).

The probability of choosing 2 chemists = $P(2C) = \frac{C_8^2}{C_{12}^2} = \frac{14}{33}$

The probability of choosing 2 engineers = $P(2E) = \frac{C_4^2}{C_{12}^2} = \frac{1}{11}$

The probability of choosing 1chemist and 1engineer = $P(2C \cap E) = \frac{8\times4}{C_{12}^2} = \frac{16}{33}$

If we define a variable x: "Number of engineers chosen", then:

$P(x = 0) = \frac{14}{33}$ $P(x = 1) = \frac{16}{33}$ and $P(x = 2) = \frac{1}{11}$

Usually, $P(x = x_i)$ is simply written $p(x_i)$. So that the previous results can be written in a tabulated from as follows:

x_i	0	1	2
$p(x_i)$	$\frac{14}{33}$	$\frac{16}{33}$	$\frac{1}{11}$

In this table, *x* is called a **discrete random variable**. To each event belonging to S, such a number can be assigned. The shown table expresses the **probability distribution (or probability law)** of the discrete random variable x.
It can be readily seen that:

$$\sum_{i=1}^{N} P(x_i) = 1 \quad (3.1)$$

The **mean value (Expectation)** of the random variable of a population is given by:

$$\mu = \sum_{i=1}^{N} x_i P(x_i) \quad (3.2)$$

While its **standard deviation** in the population can be calculated from:

$$\sigma = \sqrt{\sum_i^N x_i^2 . P(x_i) - \mu^2} \quad (3.3)$$

Details of such calculations for the previous example can be tabulated as follows:

x_i	0	1	2	Total
$P(x_i)$	14/33	16/33	1/11	**1**
$x_i . P(x_i)$	0	16/33	2/11	**2/3**
$x_i^2 . P(x_i)$	0	16/33	4/11	**28/33**

So that: $\mu = \frac{2}{3}$

And $\sigma^2 = \frac{28}{33} - \left(\frac{2}{3}\right)^2 = 0.404$ From which: $\boldsymbol{\sigma = 0.636}$

Example 3.1

A game consists of throwing three distinct dice once. The points associated with the possible results are as follows:

Three of a kind 500

A pair 100

All different 50

a) Find the probability distribution of the number of points.

b) Find the expected value of the number of points.

c) Find the standard deviation of the number of points.

Solution:

$N(S) = 6^3 = 216$

To get three of the same kind, the probability = $\frac{6}{216} = \frac{1}{36}$

To get two of the same kind with a different third outcome, then, considering that each element of the event can occur in three ways: (For example, $1, 1, 2$ can occur as $112, 121, 211$), the probability of that event is multiplied by 3. The probability is therefore: $\frac{6\times5\times3}{216} = \frac{15}{36}$

To obtain 3 different outcomes, the probability = $\frac{6\times5\times4}{216} = \frac{20}{36}$

$\boldsymbol{X}$	500	100	50
$\boldsymbol{P(X)}$	$\frac{1}{36}$	$\frac{15}{36}$	$\frac{20}{36}$

$\boldsymbol{X}$	500	100	50
$\boldsymbol{P(X)}$	$\frac{1}{36}$	$\frac{15}{36}$	$\frac{20}{36}$
$\boldsymbol{X.P(X)}$	$\frac{500}{36}$	$\frac{100\times15}{36}$	$\frac{50\times20}{36}$
$\boldsymbol{X^2.P(X)}$	$\frac{500^2}{36}$	$\frac{100\times1500}{36}$	$\frac{50\times1000}{36}$

$\mu = \frac{500}{36} + \frac{1500}{36} + \frac{1000}{36} = \boldsymbol{\frac{250}{3}}$

$\sigma^2 = \frac{500^2}{36} + \frac{100\times1500}{36} + \frac{50\times1000}{36} - \left(\frac{250}{3}\right)^2 = \frac{50000}{9}$

From which: $\boldsymbol{\sigma = 74.54}$

3.2 The binomial distribution

3.2.1 Law of probability of a binomial distribution

Consider the following situation: There are 3 independent reactors in a plant. The probability that at any time any of them will be operating is 0.85. Therefore, if we denote by x: "The number of operating reactors", we can calculate the probability distribution of x as follows:

$P(0)$ = The probability that none is operating. Since the three reactors are independent then: $p(0) = (1 - 0.85)^3 =$ **0.003375**

The probability that one reactor in particular is operating is the product.

$0.85\times(1 - 0.85)^2$, but since we have three reactors, then the probability that only one of them be operating will be $3\times0.85\times(1 - 0.85)^2 =$ **0.057375**

Now, to calculate the probability that 2 of them will be operating we first need to state that these two can be chosen out of three by $C_2^3 = 3$ methods. Hence $f(2) = 3\times(0.85)^2\times(1 - 0.85) =$ **0.325125**

Finally, the probability that all three are operating is $(0.85)^3 =$ **0.614125**
The probability distribution is therefore:

x_i	0	1	2	3
$p(x_i)$	0.003375	0.057375	0.325125	0.614125

The average value, as calculated from equation (3.2) = **2.55**
While the standard deviation, as calculated from equation (3.3) = **0.618**

The above situation is an example of a distribution that is encountered in many engineering applications, known as **the binomial distribution.**

This distribution is characterized by three main features:
(1) The number of trials must be fixed beforehand
(2) Consecutive elementary trials must be independent
(3) There exist exactly two outcomes for each trial:
Success of probability p and failure of probability $q = 1 - p$

The law of probability of the binomial distribution is given by:

$$p(x_r) = C_r^n . p^r . q^{n-r} \tag{3.4}$$

This is the probability of getting (r) successes out of (n) trials. ($0 \le r \le n$)

We note that:

$$\sum_{i=0}^{n} p(x_i) = \sum_{i=0}^{n} C_r^n . p^r . q^{n-r} = q^n + C_1^n . q^{n-1} . p + C_2^n . q^{n-2} . p^2 + \ldots + p^n = (p+q)^n = 1$$

3.3.2 Characteristics of a binomial distribution

Mean value
From equation (3.2), the average value is obtained from:

$$\mu = \sum_{i=0}^{n} x_i.p(x_i) = 0.p(0) + 1.p(1) + 2.p(2) + ..n.p(n)$$

$$= 0.q^n + 1.C_1^n.q^{n-1}.p + 2.C_2^n.q^{n-2}.p^2 + ..n.p^n$$

$$= n.p.(q^{n-1} + 2.\frac{(n-1)}{2}.q^{n-2}.p + 3.\frac{(n-1).(n-2)}{3.2}.q^{n-3}.p^3 + ...p^{n-1})$$

$$= n.p.(q^{n-1} + C_1^{n-1}.q^{n-2}.p + C_2^{n-1}.q^{n-3}.p^3 + ...p^{n-1}) = n.p.(n+p)^{n-1} = n.p$$

Hence

$$\mu = n.p \tag{3.5}$$

Standard deviation

The value of standard deviation is obtained from the following equation, given here without proof.

$$\sigma = \sqrt{n.p.q} \tag{3.6}$$

If we apply the above equations to the example given in section (3.1):
$n = 3$, $p = 0.85$ and $q = 0.15$
Hence, $\mu = 3 \times 0.85 =$ **0.255**
And $\sigma = \sqrt{3 \times 0.85 \times 0.15} =$ **0.618**
These are the same results obtained under (3.3.1)

3.2.3 Calculations of binomial probabilities using EXCEL

Although equation (3.7) can be used to predict the probability of *r* occurrences out of *n*, it is more practical to use the EXCEL function BINOM.DIST as follows:

= BINOM.DIST (number_s, trials, probability_s, cumulative)

The value of (*r*) is put in the first cell, followed by the value of (*n*) in the second, then (*p*) in the third. The fourth cell consists of a logical variable:
If *P(r*) is required, then write FALSE.

If $\sum_{r=0}^{k} P(r)$ is required, then write TRUE. This will calculate $P(x \leq k)$

Example 3.1

There are 5 pumps in the pumping house of a factory. At any time, the probability that any of them will be non-operating is 0.2. If a random variable is defined as "*x* = Number of non-operating pumps", find the mean value of *x*, its standard deviation, and its law of probability. Then deduce the median and the mode of this distribution.

Solution:

Here, $p = 0.2$, $q = 0.8$ and $n = 5$
Hence, from equations (3.5) and (3.6), we get:
$\mu = 5 \times 0.2 =$ **1** and $\sigma = \sqrt{5 \times 0.2 \times 0.85} =$ **0.894**
This means that, on average, there will be **one** non-operating pump

As for the law of probability, the BINOM.DIST function is used with logical variable = FALSE.

$P(0) = 0.32768$
$P(1) = 0.4096$
$P(2) = 0.2048$
$P(3) = 0.0512$
$P(4) = 0.0064$
$P(5) = 0.00032$

$\boldsymbol{r}$	0	1	2	3	4	5
$\boldsymbol{P(r)}$	0.3277	0.4096	0.2048	0.0512	0.0064	0.00032

To get the median, we use the cumulative probability $P(r)$ by setting the logical variable TRUE as follows:

$P(0) = 0.32768$
$P(1) = 0.73728$
$P(2) = 0.94208$
$P(3) = 0.99328$
$P(4) = 0.99968$
$P(5) = 1$

$\boldsymbol{r}$	0	1	2	3	4	5
$\boldsymbol{P(r)}$	0.3277	0.7373	0.9421	0.9933	0.9997	1

A plot of P against r reveals that one gets $P = 0.5$ at a value $\approx$ **0.4**

As for the mode, we take the value corresponding to the maximum probability, that is, $P(1) = 0.4096$, which would simply correspond to **1**.

Example 3.2
A company is considering drilling four oil wells. The probability of success for each well is 0.40, independent of the results for any other well. The cost of each well is $200,000. Each well that is successful will be worth $600,000.

(a) What is the probability that one or more wells will be successful?
(b) What is the expected number of successes?
(c) What is the standard deviation of the number of successes?
(d) What is the expected gain?
(e) What will be the gain if only one well is successful? Less than 4 wells?
(f) Considering all possible results, what is the probability of a loss rather than a gain?

Solution:
$p = 0.4, q = 0.6$

(a) This is the complementary event of "No wells are successful"
$= 1 - 0.6^4 = \mathbf{0.8704}$

(b) The success probability table has been set using EXCEL as follows:

$\boldsymbol{r}$	0	1	2	3	4
$\boldsymbol{P(r)}$	0.1296	0.3456	0.3456	0.1536	0.0256

Applying Equation (3.5): $\mu = 4 \times 0.4 = \mathbf{1.6}$

(c) Applying Equation (3.6): $\sigma = \sqrt{4 \times 0.4 \times 0.6} = \mathbf{3.1}$

(d) If r wells are successful then the gain = $r \times 600000 - 4 \times 200000$

So, the following table was set:

$\boldsymbol{r}$	0	1	2	3	4
$\boldsymbol{P(r)}$	0.1296	0.3456	0.3456	0.1536	0.0256
Gain	-800000	-200000	400000	1000000	1600000

The expected gain = $1.6 \times 600000 - 800000 = \$\mathbf{160000}$

(e) If one well only is successful, then the company looses $\$\mathbf{200000}$

If less than 4 wells are successful, then the gain = $-800000 - 200000 + 400000 + 1000000 = \$\mathbf{400000}$

(f) The probability of a loss = $P(0) + P(1) = \mathbf{0.475}$

3.3 The Poisson distribution

3.3.1 Law of probability of the Poisson distribution

This distribution is like the binomial distribution in that it is characterized by the same three conditions cited under (3.2.1). However, it is usually used whenever the value of the single probability of success p is very small. The probability of getting (r) successes out of (n) defines the following law of probability:

$$P(r) = \frac{\lambda^r . e^{-\lambda}}{r!} \qquad (3.7)$$

Where: $\lambda = n.p$

This distribution is commonly used in engineering applications dealing with events of low probability, such as equipment failure in a plant, defective items in a high-quality product, etc.

It is generally restricted to the cases where: $\boldsymbol{n > 50}$ ($n \to \infty$) and $\boldsymbol{\lambda = n.p < 6}$, although it is common to use it whenever the probability of occurrence is very low, even if $\lambda > 6$, and the number of possible outcomes n is moderately elevated.

We note that:

$$\sum_{r=0}^{\infty} P(r) = \sum_{r=0}^{\infty} \frac{\lambda^r . e^{-\lambda}}{r!}$$

$$= \frac{\lambda^0 . e^{-\lambda}}{0!} + \frac{\lambda^1 . e^{-\lambda}}{1!} + \frac{\lambda^2 . e^{-\lambda}}{2!} + \frac{\lambda^3 . e^{-\lambda}}{3!} + ..\infty = e^{-\lambda}[1 + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + ..\infty]$$

$$= e^{-\lambda} . e^{\lambda} = 1$$

3.4.2 Characteristics of the Poisson distribution

Mean value

$$\mu = \sum_{i=0}^{\infty} x_i . p(x_i) = 0.p(0) + 1.p(1) + 2.p(2) + ..\infty$$

$$= 0.\frac{\lambda^0.e^{-\lambda}}{0!} + 1.\frac{\lambda^1.e^{-\lambda}}{1!} + 2.\frac{\lambda^2.e^{-\lambda}}{2!} + 3.\frac{\lambda^3.e^{-\lambda}}{3!} + ..\infty$$

$$= e^{-\lambda}[1.\frac{\lambda^1}{1!} + 2.\frac{\lambda^2}{2!} + 3.\frac{\lambda^3}{3!} + ..\infty] = e^{-\lambda}.\lambda[1 + \frac{\lambda}{1!} + .\frac{\lambda 2}{2!} + ..\infty] = e^{-\lambda}.e^{\lambda}.\lambda$$

$$\mu = \lambda = n.p \tag{3.8}$$

Standard deviation

The following formula gives the value of the standard deviation of a variable following the Poisson distribution:

$$\sigma = \sqrt{\lambda} \tag{3.9}$$

3.4.3 Calculation of Poisson probabilities using EXCEL

The probability values in a binomial distribution can be readily obtained using the function POISSON.DIST as follows:

POISSON.DIST(*x,* mean, cumulative)

If *p(r*) is required, the probability shall be calculated from:

POISSON.DIST(*r,* λ, false)

If the cumulative probability $P(r) = P(x \leq r)$ is required, then the function is:

POISSON.DIST(*r,* λ, true)

Example 3.3

It is known that 3% of a certain batch of detergent packages is defective. Out of a sample of 100 items, calculate the probability that the number of defective items will be: None, 1, 2, 3, 4. Deduce the median value.

Solution:

To apply the Poisson distribution, we usually check the following:

$p = 0.03 \approx 0$

$n = 100 > 50$

$\lambda = n.p = 100 \times 0.03 = 3 < 6$.

Therefore:

$p\,(0)$ **= 0.04978**

$p\,(1)$ = **0.1493**

$p\,(2)$ = **0.2240**

$p\,(3)$ = **0.2240**

$p\,(4)$ = **0.1680**

We note that the values of probability first increase to reach a maximum **modal value** at $r \approx \lambda$. It then decreases for higher values.

Cumulative values are obtained using the TRUE logical variable and the **median** value was deduced as equal to **2.3** corresponding to a cumulative probability of 0.5 as follows:

$P(0) = 0.04978$
$P(1) = 0.19908$
$P(2) = 0.42308$
$P(3) = 0.64708$…….

Example 3.4
It was found that, on average, a production line is stopped five times a year. Calculate the probability that in a certain year:
(1) Exactly four times
(2) At most four times
(3) At least four times

Solution:
(1) Using EXCEL with mean = 5 and $x = 4$, we get for a FALSE cumulative input: $P(4) =$ **0.1755**

(2) Using EXCEL with mean = 5 and $x = 4$, we get for a TRUE cumulative input: $P(4) =$ **0.4405**

(3) This means calculating $P(x \geq 4) = 1 - P(x < 4) = 1 - P(x \leq 3) = 1 - 0.2650 =$ **0.7350**

Example 3.5
It was found from previous practice that there is a 2% probability that 2 accidents take place yearly for company cars. Assuming a Poisson distribution, estimate the mean number of annual accidents.

Solution:
In this example $P(2) = 0.02$ and it is required to calculate the value of λ.

Start assuming any value of λ, say $\lambda = 4$. The probability of getting a probability of 0.02 is 0.14.

Using the Goal – Seek technique, we get $\boldsymbol{\lambda = 7.15}$

Example 3.6
A company's cars are known to have on average 5 accidents per year. In a lifetime of 5 years, what is the probability that a car will have had 15 accidents? At least 15 accidents? More than 25 accidents? Knowing that the occurrence of accidents follows Poisson distribution.

Solution:
The average number of accidents in 5 years = $5 \times 5 = 25$

$P(X = 15) = \mathbf{0.0156}$

$P(X \geq 15) = 1 - P(X < 15) = 1 - P(X \leq 14) = \mathbf{0.895}$

$P(X > 25) = 1 - P(X \leq 25) = \mathbf{0.112}$

3.6 Exercise problems

(1) X is a random variable that can take the values: $\{1, 2, 3, 4\}$. The probability that $X = r$ is given by:

$$P(X = r) = 0.021\, e^{A.X}$$

Find the value of A then obtain the expectation and the variance of X.

(2) Given a random variable with the following distribution:

X_i	1.34	2.55	3.67	4.88	5.21	m
$P(X_i)$	n	0.3	0.25	0.1	0.15	0.1

Given that the standard deviation of X is 1.66, find the values of m and n.

(3) The nominal weight of a detergent box is 5 kg. It was found that 25% of boxes have slightly lower weights. In a sample of 50 randomly chosen boxes, find the probability that 10 of them will have weights less than 5 kg.

(4) There are 80 light bulbs in an indoor sector in a chemical plant. On an average day, 3 of them will be defective and will have to be replaced. Use a suitable distribution to calculate the probability that on a given day, 5 bulbs will be out of order.

(5) The proportion of students passing a certain exam is 68%. Out of a group of 120 students, calculate the following probabilities:
(a) Exactly 80 students will pass.
(b) At most 80 students will pass.

(6) In a lot of 1000 items, the probability of finding at least 5 defective items is 0.02, what is the mean number of defective items?

(7) Long term practice has shown that on average, there are 2 fires a year on a certain plant site. What is the probability that in a certain year there will be no fire at all? Three fires? At least three fires?

(8) The nominal production of a company is 2000 ton "gypsum" per day of which, on average, 25 tons are lost by dusting. In a 30-days month, what is the probability that the company will lose by dusting more than 700 tons? On how many days will the loss exceed 30 tons?

-4-
Distributions of continuous random variables

4.1 Basic concepts

If the number of values taken by a variable is extremely high, then the histogram can be approximated by a continuous curve.
In that case, the probability cannot be calculated by the ordinate of $P(x)$ versus (x) plot. This is since the values of x are infinite.
It is customary to define a function, known as the density function $f(x)$, which is the derivative of the probability function $P(x)$. Accordingly, the probability that x would lie between two values x_1 and x_2 is expressed as:

$$P(x_1<x<x_2) = \int_{x_1}^{x_2} f(x).dx \qquad (4.1)$$

This situation is shown in Figure (4.1), where a plot of the density function $f(x)$ is performed against x.
If x is defined over the interval $[a, b]$, then the conditions for $f(x)$ to represent a density function are:
(1) The function should be positive and single valued over $[a, b]$
(2) The total probability should equal one.
This can be written as:

$$P(a < x < b) = \int_a^b f(x).\,dx = 1 \qquad (4.2)$$

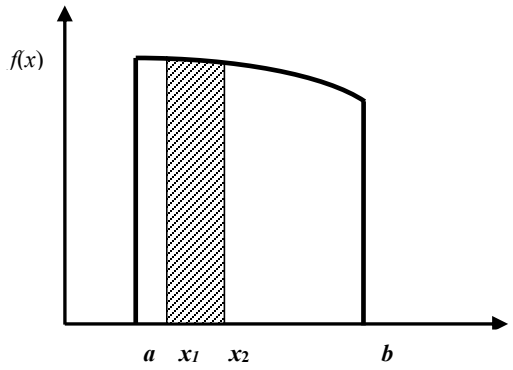


Fig (4.1): Calculation of probability out of a density function

The mean value of a continuous random variable is calculated from:

$$\mu = \int_a^b x.f(x).dx \qquad (4.3)$$

The median value *m* of a continuous random variable is calculated from:

$$\int_a^m f(x).dx = 0.5 \qquad (4.4)$$

The mode is obtained by getting the **maximum value** of $f(x)$ on [a, b]

The standard deviation is calculated from the variance as follows:

$$\sigma^2 = \int_a^b x^2 . f(x).dx - \mu^2 \qquad (4.5)$$

Example 4.1
A continuous variable x is defined in the range $0 < x < 3$. Its density function over this range is: $f(x) = \frac{1}{15}.(2 + 4x - x^2)$
Prove that this function fulfills the conditions of a density function, then calculate the mean value of x, the median value, the modal value, and the standard deviation.

Solution:
As defined, f(x) is positive and single valued.

$$\int_0^3 f(x).dx = \int_0^3 \frac{1}{15}.(2 + 4x - x^2).dx = \frac{1}{15}.\left[2x + 2x^2 - \frac{x^3}{3}\right]_0^3 = 1$$

Then, f(x) is a density function.

Mean value:

$$\mu = \int_a^b x.f(x).dx = \int_0^3 \frac{x}{15}(2 + 4x - x^2).dx = \frac{1}{15}.\left[x^2 + \frac{4x^3}{3} - \frac{x^4}{4}\right]_0^3 = \mathbf{1.65}$$

Median value:

$$\int_a^m f(x).dx = \int_0^m \frac{1}{15}.(2 + 4x - x^2).dx = \frac{1}{15}.\left[2x + 2x^2 - \frac{x^3}{3}\right]_0^m$$

$$= \frac{1}{15}.(2m + 2m^2 - \frac{m^3}{3}) = 0.5$$

Solving we get ***m* =1.69**

(The solution $m = 6.38$ is discarded)

Modal value:

We first get any extreme values in the interval [0, 3] by setting $f'(x) = 0$:

$f'(x) = \frac{1}{15}.(4 - 2x) = 0$ giving $x = 2$

The value of x corresponding to a maximum value is then obtained by calculating $f(x)$ at the borders of the interval and at $x = 2$:

x	0	2	3
$f(x)$	0.133	0.4	0.333

So, the **modal value is 2**

Standard deviation:

$$\sigma^2 = \int_a^b x^2 . f(x).dx - \mu^2 = \int_0^3 \frac{x^2}{15}(2+4x-x^2).dx - 1.65^2$$

$$= \frac{1}{15}.\left[\frac{2x^3}{3}+\frac{4x^4}{4}-\frac{x^5}{5}\right]_0^3 - 1.65^2 = 0.6375$$

Hence $\boldsymbol{\sigma \approx 0.8}$

4.2 The continuous uniform distribution

4.2.1 The density function of a uniform distribution

The density function of a continuous uniform distribution defined over an interval [a; b] is simply $f(x) = C$ (constant). As shown in Figure (4.2), the density function is a simple segment line // to the X – axis.

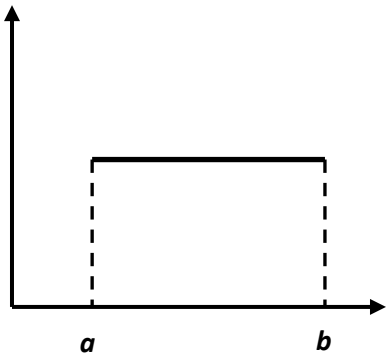


Fig (4.2): Uniform distribution

Following equation (4.1):

$\int_a^b C.dx = 1$

Hence $C.(b-a) = 1$

And:

$$C = \frac{1}{b-a} \qquad (4.6)$$

4.2.2 Characteristics of a continuous uniform distribution

Mean value

This is calculated from equation (4.2) as follows:

$$\mu = \int_a^b C.x.dx = \frac{1}{b-a} \int_a^b x.dx = \frac{b^2-a^2}{2(b-a)}$$

$$\mu = \frac{a+b}{2} \tag{4.7}$$

Standard deviation

This is calculated from equation (4.3) as follows:

$$\sigma^2 = \int_a^b C.x^2.dx - \mu^2 = \frac{1}{b-a} \int_a^b x^2.dx - \left(\frac{a+b}{2}\right)^2$$

$$\sigma^2 = \frac{b^3-a^3}{3(b-a)} - \left(\frac{a+b}{2}\right)^2 = \frac{a^2+b^2+ab}{3} - \frac{a^2+b^2+2ab}{4} = \frac{a^2+b^2-2ab}{12}$$

$$\sigma^2 = \frac{(a-b)^2}{12}$$

$$\sigma = \frac{\sqrt{3}}{6}(a-b) \tag{4.8}$$

4.3 The normal distribution

4.3.1 Defining density function

This represents a continuous distribution that is commonly encountered in engineering applications.

$$f(x) = \frac{1}{\sigma.\sqrt{2\pi}}.e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{4.9}$$

It is usually easier to deal with this distribution by defining the dimensionless variable:

$$z = (x-\mu)/\sigma \tag{4.10}$$

so that the above function reads:

$$f(z) = \frac{1}{\sqrt{2\pi}}.e^{\frac{-z^2}{2}} \tag{4.11}$$

This is called *standard normal distribution*. The shape of this function is shown in Figure (4.3). Its average (mean) value = 1 and its standard deviation = 1.

This function has the following properties:

1- Its curve is symmetrical with respect to the vertical axis (even function)

2- Its average (mean) value = 0 and its standard deviation = 1.

3- Its two branches are asymptotic with respect to the z – axis. i.e., as $z \rightarrow \pm \infty$ $f(z) \rightarrow 0$. However, in practice, the value of $f(z)$ becomes negligible at value of $|z| > 3$.

4- The probability that z lies within any two values can be calculated by equation (4.2). However, $f(z)$ cannot be integrated by analytical means, so numerical methods have been used. They show that:

* $P(-1 < z < 1) = 0.6826$
* $P(-2 < z < 2) = 0.9544$
* $P(-3 < z < 3) = 0.9974 \approx 1$, showing that most values of z lie within the interval $(-3, 3)$

In practice, however, we seldom deal with the variable z, but rather with physical variables such as pressure, temperature, time, length, although we need sometimes to use the transformation: $z = (x - \mu) / \sigma$, as will be shown in a coming Chapter.

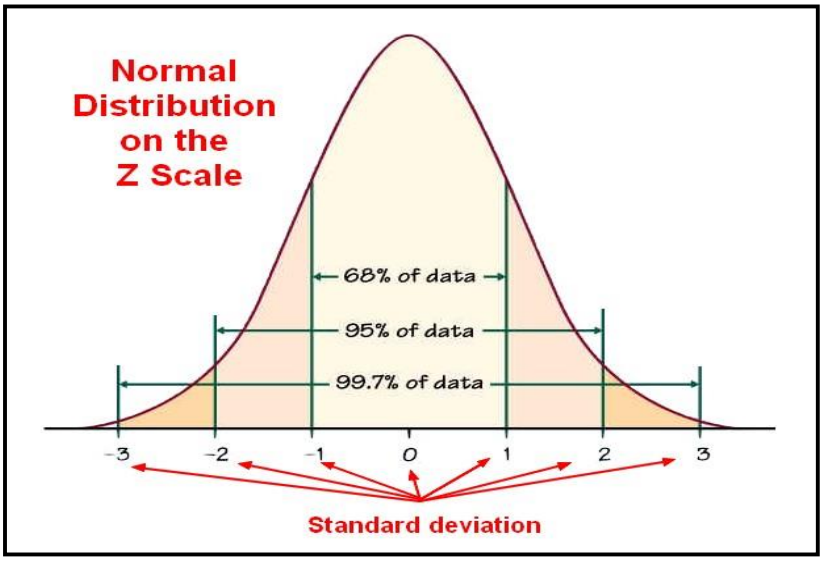


Fig (4.3): Standard normal distribution

4.3.2 Using EXCEL

It is easier to deal with cases of normal distribution using the EXCEL function NORM.DIST. This function gives the probability that $x < a$. Its use requires four entries:

NORM.DIST (x, MEAN, STANDARD_DEV, TRUE)

Other probabilities than $P(x < a)$ can be readily calculated from:

$$P(x > a) = 1 - P(x < a)$$

$$P(a < x < b) = P(x < b) - P(x < a)$$

Example 4.2
PVC pipes of nominal diameter 24" are produced by a factory. Specifications require that the diameter should not deviate from the nominal value by more than 0.5". If the diameters follow a normal distribution of standard deviation = 0.2", calculate the percentage of non-conforming pipes.

Solution:
The pipes whose diameters conform to specifications should have their diameters ranging from 23.5 to 24.5".
That is, the probability that a pipe chosen at random will conform to norms is:
$P\ (23.5 < x < 24.5)$.
Since $\mu = 24$ and $\sigma = 0.2$, then:
$P\ (x < 23.5) = 0.00621$ and by symmetry $P\ (x > 24.5) = 0.00621$
This corresponds to $P\ (x < 23.5) + P\ (x > 24.5) =$ **1.24%**

Example 4.3
Values of dust concentration in air in an industrial area have been monitored over a 300-day period. The average value was found to be 50 mg/Nm^3 and the standard deviation10 mg/Nm^3.If the dust concentrations follow a normal distribution find the number of days on which the concentration was:

(1) Lower than 40 mg/Nm^3
(2) Lower than 70 mg/Nm^3
(3) Lying between 60 and 75 mg/Nm^3
(4) Lying between 35 and 65 mg/Nm^3

Solution:
(1) $P\ (x < 40) = 0.1587$
Number of days = 300×0.1587 ≈ **47 days**
(2) $P\ (x < 70) = 0.9772$
Number of days = 300×0.9772 ≈ **292 days**
(3) $P\ (60 < x < 75) = P\ (x < 75) - P\ (\mathrm{x} < 65) = 0.4938 - 0.3413 = 0.1525$
Number of days = 300×0.1525 ≈ **46 days**
(4) $P\ (35 < x < 65) = P\ (x < 65) - P\ (x < 35) = 0.\ 933 - 0.0668 = 0.8664$
Number of days = 300×0.8664 ≈ **260 days**

Example 4.4
The percent sulfur in some crude specimens is assumed to follow a normal distribution of mean value 1.2%. If it is found that 20% of these specimens have sulfur content above 2.3%, what is the standard deviation of this distribution?

Solution:
We introduce in the icon of the NORMDIST function the entry 2.3, mean = 1.2 and TRUE as fourth entry by assuming a certain value for standard deviation. We then use the goal – seek module to correct that value. We get: $\boldsymbol{\sigma = 1.307}$

Example 4.5

There are 300 employees in a company. It was found that 95 of them had their salaries less than LE 8000 and 25 of them exceeded LE 15000. If the salaries are normally distributed across the population, find the average salary and the standard deviation.

Solution:

$P(x < 8000) = \frac{95}{300} = 0.3167$

$P(X > 15000) = \frac{25}{300} = 0.0833$

The difficulty of this problem is that both μ and σ are unknown. This necessitates using a different approach than that used in the previous example.

Let z_1 be the standard normal variable corresponding to 60 and z_2 the one corresponding to 90, so that:

$z_1 = \frac{8000 - \mu}{\sigma}$ and $z_2 = \frac{15000 - \mu}{\sigma}$

Hence: $P(z < z_1) = 0.3167$ and $P(z > z_2) = 0.0833$

Or: $P(z < z_1) = 0.3167$ and $P(z < z_2) = 0.9167$

We use the NORM.S.INV function to obtain the values of z_1 and z_2: Enter the probability 0.3167 to get $z_1 = -0.477$

Repeat with 0.9167 to get $z_2 = 1.383$

Hence: $-0.477 = \frac{8000 - \mu}{\sigma}$ and $1.383 = \frac{15000 - \mu}{\sigma}$

$\mu - 0.477\sigma = 8000$ $\mu + 1.383\sigma = 15000$

Solving: $\boldsymbol{\mu = 9795}$ $\boldsymbol{\sigma = 3763}$

4.4 Exercise problems

(1) The density function of a continuous random variable (x) is defined as follows:

$$f(x) = \begin{cases} \frac{1}{7}(x^3 + x) & (0 < x < a) \\ 0 & \text{(otherwise)} \end{cases}$$

(a) Calculate the value of the constant (a)

(b) Find the mean, median, mode and the standard deviation of x.

(2) A continuous distribution encountered in risk and reliability analysis is the exponential distribution. Its density function is defined by:

$f(x) = \lambda e^{-\lambda x} (x \geq 0)$

(a) Use the integral $\int_0^\infty x^n . e^{-\lambda x} . dx = n! \lambda^{-(n+1)}$ to prove the following:

- $\int_0^\infty f(x) . dx = 1$

- $\mu = 1/\lambda$
- $\sigma = 1/\lambda$

(b) Prove that $P(a < x < b) = e^{-\lambda a} - e^{-\lambda b}$

(3) Standard specifications require that the weights of polyethylene sacks do not deviate by more than 2% of their mean value of 50 kg. If the weights of sacks follow a normal distribution of standard deviation = 0.5 kg, what is the percentage of sacks that cannot be accepted?

(4) A factory produces steel pipes of nominal diameter 12″. Specifications require that the deviation from this value does not exceed ± 1%. If 94% of the production abides by specifications, calculate the standard deviation of pipe diameters assuming they are normally distributed.

(5) The density of an organic solvent produced by a plant has a nominal value of 850 kg/m^3. Samples drawn from the production line have their densities following a normal distribution of standard deviation 6 kg/m^3. Find the percentage of product that would have a density:

(a) Ranging from 840 to 860 kg/m^3 (b) More than 855 kg/m^3

(6) The scores of a large number of students are normally distributed with a mean value of 55 and a variance of 530. If 10% of students get grade A, what is the minimum score required to get A?

(7) The standard deviation of the scores of students in a test is 16.6. If the scores are normally distributed among the population, find the mean score of the whole student population if it is known that the score of 20% of this population exceeds 75.

(8) In an exam attended by 245 students, marked from 100, it was found that 62 of them failed to get 60 and scored (F), while 16 of them scored (A) by getting over 90. Estimate the number of students who got a (C) (between 70 and 80).

-5-

Sampling and distribution of sample means

5.1 Introduction

Let us consider the following process: Taking underground water samples from different locations in a field and analyzing for total dissolved salts content. To this aim, the location must be divided into several sections having the same area (say 4 m^2). If we have 500 such sections (Population), we will choose 50 of them (Sample) to take underground water specimens. The choice is made by giving each section a serial number: 1, 2, 3, etc. Then 50 computer-generated random numbers between 1 and 500 are produced: 224, 23, 28, 326, 489, 450, 238, 105, 167, 469, etc. Water specimens will be taken out of sections corresponding to these numbers. Random integers between A and B are generated by the EXCEL function RANDBETWEEN (A, B).

The question to be answered in the present chapter is the following: To what extent do the results obtained for the samples represent the population?

5.2 The distribution of sample means

5.2.1 Basic concepts

In the preceding example, it is obvious that each time, a different computer run will generate a new set of random numbers. Hence each time, the sample obtained will have a different mean and a different standard deviation. The sample mean will be denoted by $\bar{x}$; it represents a random variable of mean value μ_x and standard deviation $\sigma_{\bar{x}}$. This concept is best understood by the following example.

Consider a processing unit containing 6 reactors.
The following table shows the time elapsed a reactor must be revamped.

Converter	A	B	C	D	E	F
***x* days**	240	300	255	270	264	270

From these data, the mean $\mu = 265.5$ and the standard deviation $\sigma = 19.94$

This is a small population such that no sampling is necessary. Let us assume, however, that a sample of 3 reactors is going to be chosen. There are $C_3^6 = 20$ such samples. The following table shows these samples together with the mean value of each.

Sample	ABC	ABD	ABE	ABF	ACD	ACE	ACF	ADE	ADF	AEF
Mean	265	270	268	270	255	253	255	258	260	258
Sample	BCD	BCE	BCF	BDE	BDF	BEF	CDE	CDF	CEF	DEF
Mean	275	273	275	278	280	278	263	265	263	268

The mean value of all sample means can be calculated. $\mu_{\bar{x}} = 265.5 = \mu$

While the standard deviation of sample means is $\sigma_{\bar{x}} = 8.35 < 19.94\ (\sigma)$

This example sets a very important principle: **"The mean value of sample means is equal to the mean value of population"**; that is:

$$\mu_{\bar{x}} = \mu \tag{5.1}$$

As for the standard deviation of sample means, it is always smaller than the standard deviation of population. This means that the values of sample means are less scattered about their mean value than the values of *x* for population.

If the size of sample $n > 30$, then the following relation holds for $\sigma_{\bar{x}}$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{5.2}$$

In the above example, since $n = 3$ only, the latter relation cannot be applied. The value of calculated by this relation would have been:

$\frac{19.94}{\sqrt{3}} = 11.5 \neq$ the calculated value of 8.35

Example 5.1

400 sacks of PE beads are chosen from a factory outlet. The average weight = 50 kg and standard deviation = 0.5 kg. Calculate the mean and standard deviation of sample means.

Solution:

$\mu_{\bar{x}} = \mu = 50$ Hence, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{400}} = 0.025$ kg

5.2.2 Distribution of sample means in a population that is normally distributed

If the values of the random variable *x* are normally distributed in the population, then the mean values of sample means will also be normally distributed with the same mean value but less scattered since $\sigma_{\bar{x}} < \sigma$ (Equations 5.1 and 5.2). Figure (5.1) depicts this situation.

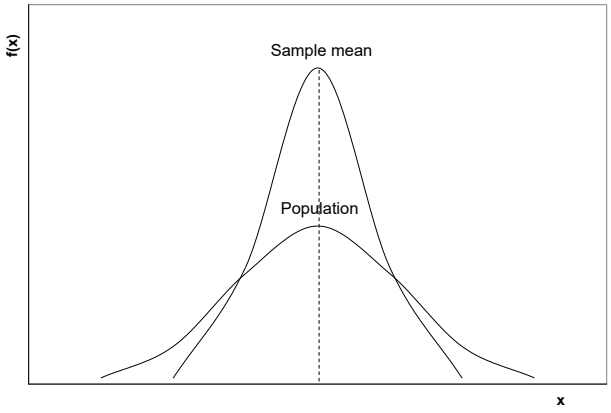


Fig (5.1): Distribution of population and sample means

5.2.3 Distribution of sample means in a population that is not normally distributed

Even if the random variable is not normally distributed over the population, it can be proved that, provided $n > 30$, the sample means will still be normally distributed and equations (5.1) and (5.2) will still hold. The following figure illustrates this principle.

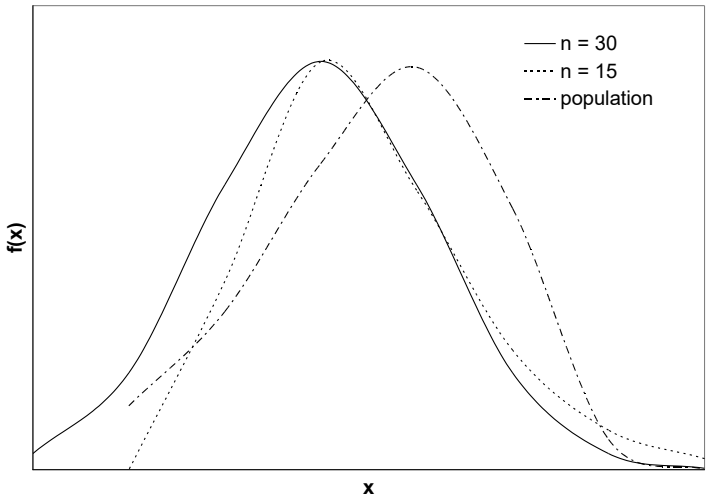


Fig (5.2): Effect of size on distribution of means

5.3 Estimation of the population mean: Levels of confidence

5.3.1 Point estimate of the population mean: Large samples

Consider a population of unknown mean μ with $n > 30$. If we choose a random sample of mean value $\bar{x}$, this value will not necessarily be equal to the mean of the population. As previously stated, the mean values of samples will be normally distributed about μ. It is said that $\bar{x}$ is a **point estimate** of the population mean μ.

The question to be asked is the following: To what extent does the sample mean differ from the population mean?

Let the difference between the two means be *a*, then the following inequality holds:

$-a < \bar{x} - \mu < a$

The probability of such an occurrence is: $P\,(-a < \bar{x} - \mu < a\,)$.

Dividing all sides by the standard deviation of sample means $\sigma_{\bar{x}}$, this probability becomes:

$$P\left(\frac{-a}{\sigma_{\bar{x}}} < \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} < \frac{a}{\sigma_{\bar{x}}}\right)$$

Since the means are normally distributed about μ, then this can be written in the following form:

$$P\left(\frac{-a}{\sigma_{\bar{x}}} < z < \frac{a}{\sigma_{\bar{x}}}\right) \quad \text{or} \quad P\left(-z_{crit} < z < z_{crit}\right) \quad \text{or}$$

$$P\left(-z_{crit} < \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} < z_{crit}\right) \tag{5.3}$$

The value of this probability is termed the **level of confidence L**. This is an arbitrary level that expresses the extent to which the sample mean will differ from the population mean. The usual values of L are 0.9, 0.95 and 0.99. Each of these levels corresponds to a definite value of $z_{crit} = a / \sigma_{\bar{x}}$ that can be obtained from the normal tables. The following table shows these values.

Table 5.1: Values of z_{crit} corresponding to different confidence levels

L	0.99	0.95	0.90
z_{crit}	2.58	1.96	1.64

The inequality in equation (5.3) can be written as: $-z_{crit}.\sigma_{\bar{x}} < \bar{x} - \mu < z_{crit}.\sigma_{\bar{x}}$

Or, from equation (5.2):

$$-z_{crit}.\frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < z_{crit}.\frac{\sigma}{\sqrt{n}} \tag{5.4}$$

So, if the value of σ is known, we can estimate, to any level of confidence, the deviation of sample mean from population mean.

Since it is seldom possible to know the value of σ, the standard deviation of population, it is safe enough to use the standard deviation of sample instead (s). This way, the above equation becomes:

$$-z_{crit}.\frac{s}{\sqrt{n}} < \bar{x} - \mu < z_{crit}.\frac{s}{\sqrt{n}}$$

Better written as: $$\bar{x} - z_{crit}.\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{crit}.\frac{s}{\sqrt{n}} \tag{5.5}$$

The interval defined by equations (5.4) or (5.5) is **a confidence interval for mean value of population.**

Finally, it is worth mentioning that the value $\alpha = 1 - L$ is called, in statistical terminology: **Level of significance**.

For values of L different from Table (5.1), the value of z_{crit} is obtained from the EXCEL function NORM.S.INV with entry $\frac{1+L}{2}$

For example, for a confidence level of 0.98, z_{crit} is obtained by introducing the entry 1.98/2 = 0.99 into the function NORM.S.INV. We get $z_{crit} = 2.326$

Example 5.2

Data belonging to a sample of 35 mortar cubes tested for compressive strength (MPa) showed that the mean value is 24.2 MPa and the standard deviation = 3.75 MPa. Estimate the value of the mean strength of the batch using a 95% confidence level**.**

Solution:

$\bar{x}$ = 24.2 and s = 4.75 MPa

Applying equation (5.5), we get for L = 0.95:

$$24.2 - \frac{1.96 \times 4.75}{\sqrt{35}} < \mu < 24.2 + \frac{1.96 \times 4.75}{\sqrt{35}}$$

or: **$22.63 < \mu < 25.77$**

This result means that **we are 95% sure that the mean value of population lies between 22.63 and 25.77 MPa.**

Note that if we increase the level of confidence to 99%, then the value of z_{crit} = 2.58, and application of equation (5.5) will give $22.13 < \mu < 26.27$, which is wider a range.

5.3.2 Effect of confidence level on width of confidence interval of mean

In general, the higher the confidence level used, the higher will be the error in predicting the. This error = $\frac{z_{crit} \cdot s}{\sqrt{n}}$

Consider for example a sample of size = 30 drawn from a population. Let the standard deviation of sample = 3. Figure (5.3) clearly shows that increasing the value of L will increase the confidence interval width, yielding more uncertain estimations for the population mean. That is why it is common to use 0.95 as reasonable confidence level.

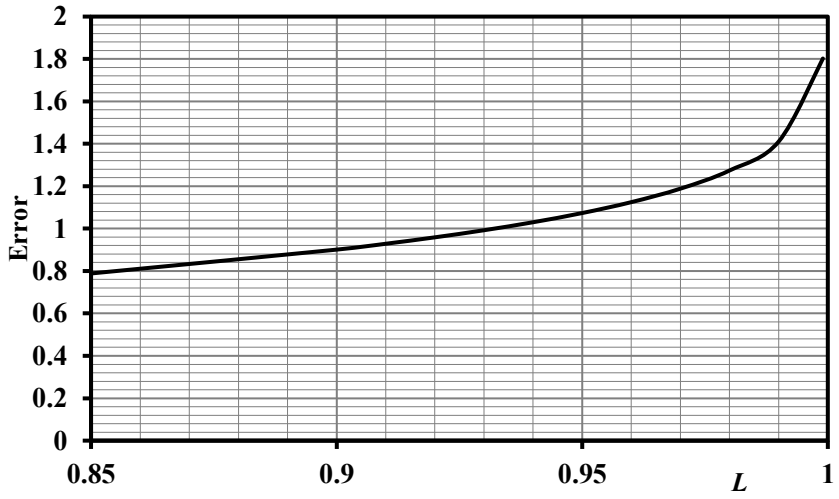


Fig (5.3): Effect of confidence level on the accuracy of predicting the population mean

5.3.3 The case of small samples

If $n < 30$, then the normal approximation for sample means is no longer valid; we must use another distribution. This is known as **the t – distribution**. This is a symmetrical distribution like the normal distribution. Its density function is given by the following law:

$$f(t) = \frac{\Gamma\left(\frac{n}{2}\right).\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}}{\sqrt{(n-1).\pi.\Gamma(\frac{n-1}{2})}} \tag{5.6}$$

This distribution gives values of a parameter t instead of z_{crit} in equation (5.4), which is a function of both α and n in the form: $t(\alpha, n-1)$, where $n-1$ is the number of degrees of freedom (*d.f.*)

In case of small samples, equation (5.5) is written as follows:

$$\bar{x} - t.\frac{s}{\sqrt{n-1}} < \mu < \bar{x} + t.\frac{s}{\sqrt{n-1}} \tag{5.7}$$

The EXCEL function T.INV2t can be readily used to display the required value of t for a known value of α and $d.f. = n-1$

There is however one major assumption related to the use of the t – distribution; namely, that the population must be normally distributed.

Example 5.3

Six specimens of lubricating oil were chosen from the daily production of a factory and tested for viscosity, the values were: 105, 112, 97, 102, 107, 107 cP. At 95% confidence level, set a confidence interval for the mean value of viscosity in the produced batch.

Solution:

The average value of sample was calculated as 105 cP

The standard deviation was calculated as 5.1 cP

For a level of significance = 1 – 0.95 = 0.05 and *d.f.* = 6 – 1 = 5, we get from the T.INV2t function: $t = 2.57$.

Replacing in equation (5.7), we get:

$$105 - \frac{2.57 \times 5.1}{\sqrt{5}} < \mu < 105 + \frac{2.57 \times 5.1}{\sqrt{5}} \text{ or } \mathbf{99.13 < \mu < 110.86}$$

5.3.4 Sample size

Suppose that we wish to choose a sample from a population and ensure, at a given confidence level, that the population mean will not differ from the sample mean by more than a certain amount.

Let the maximum deviation be D, hence: $|\ \bar{x} - \mu\ | = D$

Hence, $D = \frac{z_{crit}\sigma}{\sqrt{n}}$

$$n = \frac{z_{crit}^2 . \sigma^2}{D^2} \tag{5.8}$$

Example 5.4

It is required to choose a sample from a stream of industrial wastewater. The standard deviation of the percent TDS is known to be 3. At a confidence level of 95%, how many specimens should we take so that the sample mean percent would not deviate from the population mean by more than 1?

Solution:

Here: $\sigma = 3$, $D = 1$ and $z_{crit} = 1.96$

Substituting in equation (5.7), we get:

$n = \frac{1.96^2 \times 3^2}{1^2} = 34.57 \approx$ **35 specimens**

5.4 Sampling of proportions

In many cases, the most important factor is the proportion of successes in any case at hand rather than their number. For example, it is known that factories can tolerate a certain number of defective items in their production line, provided it does not exceed a certain ratio. In general, the proportion of "success" in the population will be denoted π whereas the proportion of failure $\tau = 1 - \pi$.

For example, in a production line of porcelain ware the proportion of first grade products is 0.4. So, $\pi = 0.4$ and $\tau = 0.5$.

If now we take a sample of say, 5 specimens, then all odds are possible. That is, the number of first grade products can take any value from 0 to 5. We denote the proportion of success (first grade) in sample by p and failure by q. The likelihood of any proportions of successes will follow a binomial distribution as shown by the following table:

Table 5.2: Binomial distribution of success probability in a sample

No of successes	No of failures	p	q	Probability $f(x_i)$	$p.f(x_i)$	$x_i^2.f(x_i)$
0	5	0	1	$(0.6)^5$	0	0
1	4	0.2	0.8	${}^5C_1 \times 0.6^4 \times 0.4$	0.05184	0.010368
2	3	0.4	0.6	${}^5C_2 \times 0.6^3 \times 0.4^2$	0.13824	0.055296
3	2	0.6	0.4	${}^5C_3 \times 0.6^2 \times 0.4^3$	0.13824	0.082944
4	1	0.8	0.2	${}^5C_4 \times 0.6 \times 0.4^4$	0.06144	0.049152
5	0	1	0	0.4^5	0.01024	0.01024
Total				**1**	**0.4**	**0.208**

Note that the total sum of probabilities is 1 and the mean value of proportions of success in different samples is 0.4, which is equal to the proportion of successes in population (π).
On the other hand, the standard deviation of success proportions in samples can be obtained from equation (1.8):

$$\sqrt{\Sigma\, x_i^2 . f_i - \bar{x}^2} = \sqrt{0.208 - 0.4^2} = 0.219$$

Now, the standard deviation of population σ_p of a random variable for a binomial distribution was obtained from $\sqrt{npq} = \sqrt{\mu q}$. In case the variable is a proportion, this equation has to be rewritten to include the mean value of population proportion π (instead of μ) and the complementary proportion τ instead of q.

$$\sigma_p = \sqrt{\pi . \tau} \tag{5.9}$$

Now, from equation (5.2), the standard deviation of sample means is:

$$\sigma_s = \sqrt{\pi . \tau / n} \tag{5.10}$$

In the example at hand, $\sigma_s = \sqrt{\dfrac{0.6 \times 0.4}{5}} = 0.219$

The foregoing example yields two important results that coincide with those previously enunciated through equations (5.1) and (5.2). That is, the mean value of sample proportion means is equal to the proportion of successes in population, and their standard deviation can be obtained by an equation analogous to (5.2), namely (5.10).

The standard deviation of sample proportion means is also termed the **standard error of sample proportion.**

If now we select a sample of n items out of a large population, the proportion of population success will not be known. A **point estimate of population mean** will be the proportion of successes in sample (p) and a **point estimate for standard error** of sample proportion can be obtained from the following equation:

$$s_p = \sqrt{\frac{p.q}{n-1}} \tag{5.11}$$

For a given confidence interval, the range of values of the mean population proportion of successes can be obtained from:

$$p - z_{crit} . s_p < \pi < p + z_{crit} . s_p \tag{5.12}$$

Example 5.5
A sample of 100 specimens was taken out of many products. It was found that 7 of these specimens were defective. What is the estimate for the mean proportion of defective items in the production line at a 95% confidence level?

Solution:
The point estimate of the proportion of defective items $p = 0.07$ and the standard error of sample is:

$$s_p = \sqrt{\frac{0.07 \times 0.93}{100 - 1}} = 0.0256$$

Hence, from equation (5.12):

$0.07 - 0.0256 \times 1.96 < \pi < 0.07 + 0.0256 \times 1.96$ or $\mathbf{0.02 < \pi < 0.12}$

This means that, based on the sample chosen, we are 95% sure that the proportion of defective items in the production line will range from 0.02 to 0.12, that is 2% to 12%.
We note that this range is somewhat wide. Had we taken a sample of 500 specimens, the previous result would have been: $0.048 < \pi < 0.092$, which represents a better estimate.

In case of proportions, we must take into consideration the effect of sample size as previously done under (5.3.4). In that case, the equation corresponding to equation (5.8) that determines the sample size required obtaining a maximum deviation $|p - \pi|$ is approximately:

$$n \approx z_{crit}^2 \pi\tau / D^2 \qquad (5.13)$$

If π and τ are not known, the best approximation is to use a sample estimate for p and q.

Example 5.6
In the preceding example, what is the size sample required to get a maximum deviation between the mean proportion of success in population and its sample estimate of 0.01?

Solution:
π and τ being unknown, we take $p = 0.07$ and $q = 0.93$, whereas $D = 0.01$ and for a 95% confidence level, $z_{crit} = 1.95$.

Replacing in equation (5.13), we get:
$n = 1.96^2 \times 0.07 \times 0.93 / 0.01^2 =$ **2500**

In that case, we get: $0.06 < \pi < 0.08$

5.5 Exercise problems

(1) The concentration of salt in the effluent stream of a reactor was measured for 35 samples. The mean value was 0.4 g mole/dm^3 and the standard deviation = 0.02. At a 95% confidence level, what is the probable mean value of the salt in the effluent stream?

(2) 100 PVC pipes were tested for bending strength. The mean value of samples was found to be 8.5 MPa and the standard deviation = 0.5 MPa. What are the lower and upper limits of the population mean, at 90% and at 95% confidence level?

(3) The following table shows the marks of a sample of 30 students in an exam (attended by 2500 students). At a 95% confidence level, what would be the probable limits of the mean value of population?

Class interval	[0 – 5)	[5 – 10)	[10 – 15)	[15 – 20)	[20 – 25)	[25 – 30]
Number of	2	4	11	7	5	1

(4) The following data represents the level of solid dust concentration (mg/m^3) gathered from 10 different locations inside a plant. At a 95% confidence level, what would be the probable mean value of dust concentration inside this plant?

56 73 66 52 60 58 44 44 50 61

(5) It is required to choose a sample from the production of an oil well. The standard deviation of the percent sulfur in the crude is known to be 0.3. At a confidence level of 95%, how many samples should we take so that the mean percent sulfur in sample would not deviate from that of population mean by more than 0.05?

(6) A sample of 200 specimens was taken out of a production line. It was found that 20 of these specimens were defective. What is the estimate for the mean proportion of defective items in the production line at a 95% confidence level?

(7) A factory produces articles for exportation. However, due to process problems, the proportion of non-conforming items in a 300-specimen sample was found to be 15%. What would be your estimate for the proportion of non-conforming articles in the whole production at a 90% confidence level?

(8) A sample of bottled edible oil taken from the market showed that a proportion of 0.12 of the products are outdated. What would be a reasonable sample size such that the sample proportion represents the population with a maximum error of 0.01? Use a 95% confidence level.

(9) A batch storage tank dispenses on the average batches of volume 0.8 m^3 and standard deviation of 0.1 m^3. The volumes dispensed are assumed to follow a normal distribution of means. How many batches should be chosen to ensure with 90% confidence that its mean value would not deviate by more than 1% of the true mean? If it is intended to take 50 samples, what significance level would guarantee that the samples mean would not deviate from the true mean by more than 1%?

-6-

Testing of hypotheses

6.1 Introduction

Consider a factory producing PE sacks. The manufacturer claims that the mean mass of produced sacks exceeds 10 kg. This claim is known as the **null hypothesis** and is written: H_0: $\mu > 10$.

The hypothesis that does not support the claim is known as the **alternate hypothesis** and is written as follows: H_a: $\mu \leq 10$.

This defines two regions on a line indicating the mean mass of population: An **acceptance region** if $\mu > 10$ and a **rejection region** if $\mu \leq 10$. This situation in shown in Figure (6.1).

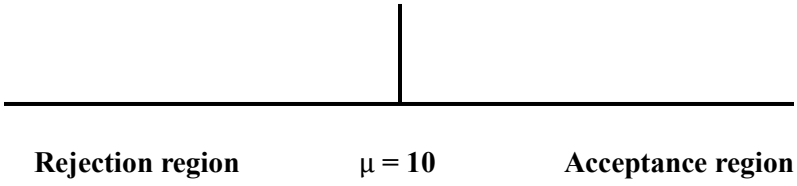


Fig (6.1): Acceptance and rejection regions for H_0: $\mu > 10$

In practice it is common to allow for a certain deviation from the theoretical case discussed above. This means accepting the null hypothesis for values of mean $= \mu - a$. The determination of the allowed error (a) will be explained later.

A similar case may show up if a manufacturer of a certain chemical claims that, on average, the percentage of impurities in his product is less than 0.5%. In that case: The null hypothesis is H_0: $\mu < 0.5$ and the alternate hypothesis: H_a: $\mu \geq 0.5$. In that case the acceptance and rejection regions show as in Figure (6.2).

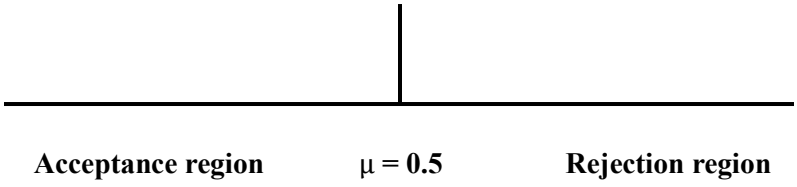


Fig (6.2): Acceptance and rejection regions for H_0: $\mu < 0.5$

Here also, we may accept the null hypothesis for values of mean $= \mu + a$.

In the two previous examples, the acceptance or rejection of the null hypothesis depends on whether the mean of population is higher or lower than a certain control value. Such cases represent a **one – tailed hypothesis.**

Consider now the production of several tons per day of distilled water of mean pH = 7. Any deviation from that figure would be considered in principle inadequate so that the null hypothesis takes the form: H_0: $\mu = 7$ while the alternate hypothesis is H_a: $\mu \neq 7$. This way, there will appear two rejection regions from both sides of the hypothetical value 7 (Figure 6.3). This is called a **two – tailed hypothesis**.

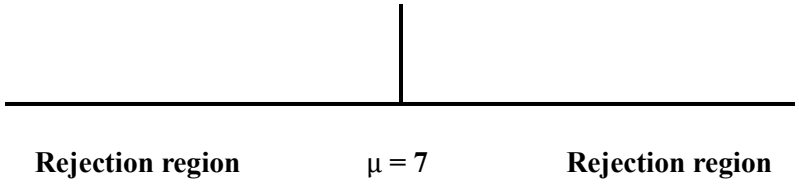


Fig (6.3): Rejection regions for H_0: $\mu = 7$

In that case, it is common practice to allow for deviations from the mean from both sides of its hypothetical values, that is in the range $]\mu - a\ ;\ \mu + a[$.

6.2 Hypotheses concerning the mean of a population:

6.2.1 The case of large samples $(n > 30)$

To check the veracity of the null hypothesis we choose a sample from the population of size n and determine the mean value of the parameter of interest $\bar{x}$ and its standard deviation s. In that case, the null hypothesis to be tested will consist of one of two possibilities, assuming that the studied parameter is normally distributed among the population.

(a) One – tailed hypotheses

This is the case when the null hypothesis takes the form: H_0: $\mu > k$ or H_0: $\mu < k$. We first calculate the **test statistic** z form the expression:

$$|z| = \frac{|\bar{x} - k|}{\frac{s}{\sqrt{n}}} \qquad (6.1)$$

The next step is to compare the value of the calculated statistic to a critical z – value (z_{crit}) obtained from the function NORM.S.INV (L). The null hypothesis will be accepted if $|z| < |z_{crit}|$. Figure (6.4) explains the rationale of this criterion for the two cases where the null hypothesis is H_0: $\mu > k$ or H_0: $\mu < k$. The figures show that the alternate hypothesis H_a is more probable if $|z| > |z_{crit}|$.

Example 6.1

In the preparation of concrete slabs, concrete cubes are tested for compressive strength, and the average strength should exceed 30MPa. If 35 cubes are tested and the mean strength was found to be 28.7MPa with a standard deviation of 2.66MPa, would you accept the null hypothesis H_0: $\mu > 30$ at 0.05 significance level?

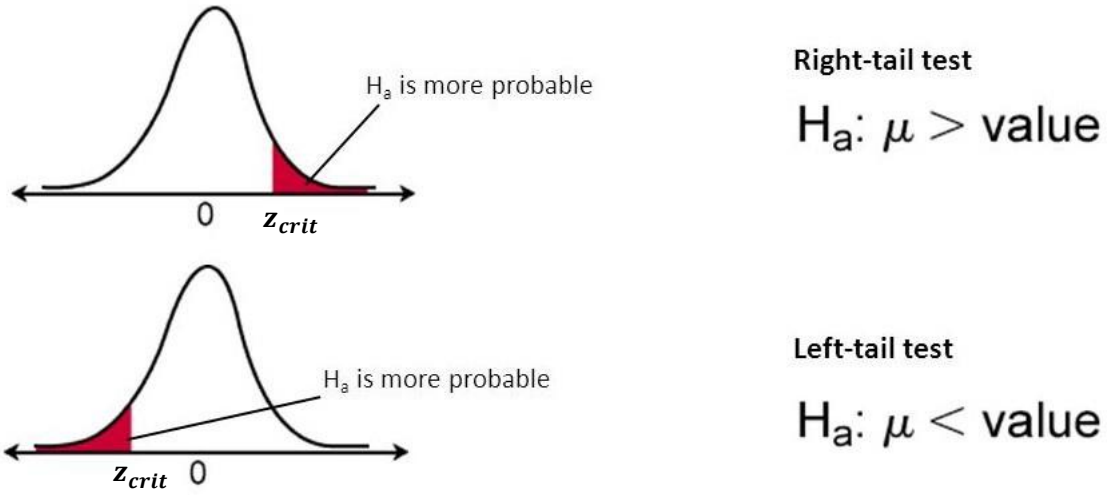


Fig (6.4): Alternate hypotheses for one – tailed tests

Solution:

The mean value of sample $\bar{x}$ = 28.7 and the standard deviation s = 2.66. The null hypothesis H_0: $\mu > 30$ corresponds to a value of $k = 30$.

$$|z| = \frac{|28.7-30|}{\frac{2.66}{\sqrt{35}}} = 2.89$$

At $\alpha = 0.05, L = 0.95, z_{crit} = 1.645$

Since $2.89 > 1.645$, the null hypothesis cannot be accepted.

Example 6.2

The daily amount of rejected items from a production line is normally distributed. The company's policy requires that this should not exceed 7% of the production. Data gathered over a one-month period (30 days) show that the daily mean percentage of rejected items = 7.62% with a standard deviation of 1.41%. What maximum significance level should be taken in order not to reject the null hypothesis: H_0: $\mu \leq 7$?

Solution:

The mean value of sample $\bar{x}$ = 7.62 and the standard deviation s = 1.41 while $n = 30$ and $k = 7$.

$$|z| = \frac{|7.62-7|}{\frac{1.41}{\sqrt{30}}} = 2.408$$

The L value corresponding to 2.408 is obtained from NORM.S.DIST $(2.408, \text{TRUE})$. This gives $L = 0.992$. This means that the null hypothesis cannot be rejected if $L > 0.992$ or $\alpha < 1 - 0.992 \rightarrow \alpha < 0.008$.

Hence the maximum significance level that would not reject the null hypothesis is $\boldsymbol{\alpha = 0.008}$

(b) Two – tailed hypotheses

When the null hypothesis takes the form H_0: $\mu = k$, the z – statistic is also determined from equation (6.1). However, the two rejection regions where the alternate hypothesis is accepted are now form both sides of the mean as revealed from Figure (6.5). In that case, the critical vlaue of z is obtained from the function NORM.S.INV $\left(\frac{1+L}{2}\right)$

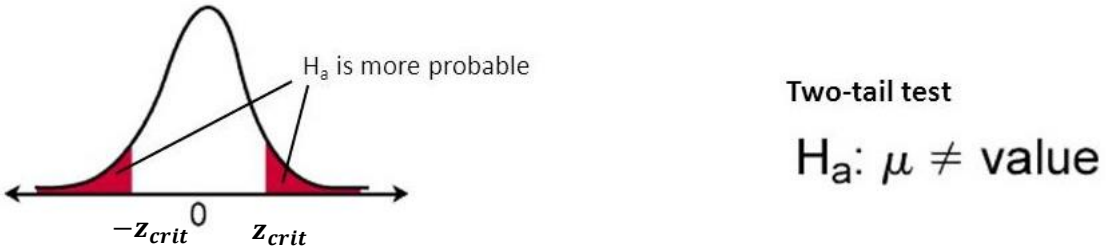


Fig (6.5): Alternate hypotheses for two – tailed tests

Example 6.3

For proper operation, the length of a particular rod in an engine should equal 645 mm. When 40 pieces from the daily production were chosen, the mean length was 645.5 mm with a standard deviation of 0.84 mm. At 0.05 significance level, would you consider that the manufacturing process of these parts needs some adjustment?

Solution:

This is a two – tailed hypothesis with H_0: $\mu = 645$ and $\bar{x}$ = 645.5, s = 0.84, n = 40 and $L = 0.95$. The test statistic is:

$$|z| = \frac{|645.5-645|}{\frac{0.84}{\sqrt{40}}} = 3.7646$$

The critical value of z is obtained from NORM.S.INV $\left(\frac{1+0.95}{2}\right)$ = 1.96

Since $3.7646 > 1.96$, then the null hypothesis H_0: $\mu = 645$ cannot be accepted and the manufacturing process effectively needs some adjustment.

6.2.2 The case of small samples $(n < 30)$

Usually, it is not possible to test large number of specimens and the need for a modification of the last criterion arises. In case of $n < 30$, the t – distribution is used whereby the test statistic takes the form:

$$|t| = \frac{|\bar{x} - k|}{\frac{s}{\sqrt{n-1}}} \quad (6.2)$$

The null hypothesis is accepted if $|t| < |t_{crit}|$. The critical value of t is obtained as follows:

- **For one – tailed tests**: Use the function T.INV $(\alpha\,, n-1)$
- **For two – tailed tests**: Use the function T.INV.2T $(\alpha\,, n-1)$

Example 6.4

Six specimens were chosen from the production line of HDPE bottles, and their density was determined. The values were as follows (g.cm^{-3}): $0.962\,, 0.949\,, 0.958\,, 0.961\,, 0.960\,, 0.944$. The factory claims that the mean density does not exceed 0.95 g.cm^{-3}. At a significance level = 0.05, would you accept that claim?

Solution:

The null hypothesis to be tested is: H_0: $\mu \leq 0.95$.

First the mean value and standard deviation of sample density is determined: $\bar{x}$ = 0.9557, s = 0.0074 with n = 6.

$$|t| = \frac{|0.9557 - 0.95|}{\frac{0.0074}{\sqrt{6-1}}} = 1.72$$

The critical value of $|t|$ = $|\text{T. INV}\,(0.05\,, 6-1)|$= 2.015

Since $1.72 < 2.015$, then the null hypothesis cannot be rejected.

Example 6.5

The pH in a certain chemical reaction must be fixed at pH = 8.3 for proper results. Specimens from the reaction mixture were drawn at regular hourly intervals for 8 hours and their pH determined. The results obtained were as follows: $8.29\,, 8.18\,, 8.09\,, 8.12\,, 8.31\,, 8.34\,, 8.22\,, 8.19$. Show that the alternate hypothesis H_a: $\mu \neq 8.3$ cannot be rejected and determine the maximum significance level necessary for the null hypothesis H_0: $\mu = 8.3$ not to be rejected.

Solution:

The mean value and standard deviation of sample density is determined: $\bar{x}$ = 8.2175, s = 0.09 with n = 8.

$$|t| = \frac{|8.2175 - 8.3|}{\frac{0.09}{\sqrt{8-1}}} = 2.424$$

The critical value of $|t|$ = $|\text{T. INV. 2T}\,(0.05\,, 8-1)|$ = $2.364 < 2.424$. Hence the null hypothesis cannot be accepted and consequently we do not reject the alternate hypothesis.

Using Goal – Seek, we can get the maximum value of α which would not reject the null hypothesis. This is found to equal $\boldsymbol{\alpha = 0.0458}$.

6.3 Sample size

When it is required for the sample's mean to be as close as possible to the population's mean, one must choose a large sample. If D is the difference between sample mean and population's mean, $D = |\bar{x} - \mu|$ then from equation (6.1):

$$z_{crit} = \frac{D.\sqrt{n}}{\sigma}$$

$$\text{Or } n = \left(\frac{z_{crit}\sigma}{D}\right)^2 \qquad (6.3)$$

6.3 Hypotheses concerning the mean of a proportion

Sometimes we are more interested in testing a proportion rather than a parameter. The null hypothesis in that case takes the form:

- H_0: $\pi < k$ or $\pi > k$ for **one tailed hypotheses**
- H_0: $\pi = k$ for **two tailed hypotheses**

In case of large samples ($n > 30$), the test statistic is:

$$|z| = \frac{|p-k|}{\sqrt{\frac{\pi.\tau}{n-1}}} \qquad (6.4)$$

Here, p is the proportion of sample while π is that of population ($\pi = k$) and $\tau = 1 - \pi$

The critical value of z is determined either from NORM.S.INV (L) or NORM.S.INV $\left(\frac{1+L}{2}\right)$ depending on whether the test is one – tailed or two – tailed respectively.

In case of small samples ($n < 30$), the test statistic is also obtained from equation (6.1) and the critical value of t obtained from $|\text{T. INV}(\alpha, n-1)|$or T.INV.2T $(\alpha, n-1)$ for one- tailed and two – tailed tests respectively.

Example 6.6

An instructor claims that at least 60% of his students have passed the final exam. When a sample of 30 students was chosen, 15 out of them turned out to be passing. At 0.05 significance level, would you accept the instructor's claim?

Solution:

H_0: $\pi \geq 0.6$ Hence $\pi = 0.6, \tau = 0.4$

The sample proportion $p = \frac{15}{30} = 0.5$

$|z| = \frac{|0.5 - 0.6|}{\sqrt{\frac{0.6 \times 0.4}{30-1}}} = 1.1$ At $\alpha = 0.05, z_{crit} = 1.65 > 1.1$ Hence H_0 can be accepted.

Example 6.7

A coin was tested for uniformity (That is the probability of both heads and tails are equal). This coin was thrown 26 times and the number of tails obtained was 9. At a 95% confidence level, test the null hypothesis H_0: $\pi = 0.5$ at significance level = 0.5.

In 26 throws, how many times must tails appear at least to accept the hypothesis of coin uniformity?

Solution:

H_0: $\pi = 0.5$ Hence $\pi = \tau = 0.5$

The sample proportion $p = \frac{9}{26} = 0.346$

$|t| = \frac{|0.346 - 0.5|}{\sqrt{\frac{0.5 \times 0.5}{26-1}}} = 1.54$

At $\alpha = 0.05, t_{crit} = 2.06 > 1.54$ Hence H_0 can be accepted.

To accept coin uniformity, one must have $t < t_{crit}$:

Assume any value for the number of tails (Obviously less than 9) and calculate the corresponding value of p, then obtain the value of t. Then use the Goal – Seek method for this value to equal 2.06.

For example, setting the number of tails = 8 will yield $p = 0.3077$, and $t = 1.923$. Using the Goal – Seek method, by equating t to 2.05, we get the number of tails = 7.67. Hence, the appearance of at least 8 tails is necessary to accept the hypothesis of uniformity at 0.95 confidence level.

6.4 Hypotheses concerning the difference between two means

6.4.1 Introduction

Suppose that two catalysts are tested to enhance the rate of a certain reaction. The reaction is repeated several times using each time the two catalysts individually and conversion is determined each time. The null hypothesis to be tested is whether the mean values of conversion using the two catalysts are comparable. That is:

H_0: $\mu_1 = \mu_2$

6.4.2 The case of large samples

Let the mean values of the two set of samples be $\bar{x}_1$ and $\bar{x}_2$ and their standard deviations s_1 and s_2 respectively. The statistic used takes the form:

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (6.5)$$

Example 6.8

Two different mixers were proposed to blend plastic powders. The uniformity of the blend was assessed by the time taken for the standard deviation of the mix density to stabilize. When 30 runs were carried out on the first mixer, the average time required to reach homogeneity was 23.5 min. with a standard deviation of 2.5 min. The corresponding figures for the second mixer were 25.2 and 3.8 min. respectively for a 35-size sample. Determine at 0.05 significance level, whether the performance of the two mixers can be considered comparable.

Solution:

H_0: $\mu_1 = \mu_2$

$\bar{x}_1 = 23.5 \quad s_1 = 2.5 \quad n_1 = 30$

$\bar{x}_2 = 25.2 \quad s_2 = 3.8 \quad n_2 = 35$

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{|23.5 - 25.2|}{\sqrt{\frac{2.5^2}{30} + \frac{3.8^2}{35}}} = 2.21$$

At $\alpha = 0.05 \rightarrow L = 0.95, z_{crit} = 1.96 < 2.21$ (The function NORM.INV $\left(\frac{1+0.95}{2}\right)$ was used.

Hence the null hypothesis cannot be accepted, and the performance of the two mixers cannot be considered comparable.

6.4.3 The case of small samples

In that case, the null hypothesis H_0: $\mu_1 = \mu_2$ is tested in a different way:

First a "pooled" standard deviation is used using the expression:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (6.6)$$

The statistic of the test is:

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (6.7)$$

This is then compared to the critical value of t for two – tailed test at the required significance level for a number of degrees of freedom = $n_1 + n_2 - 2$

Example 6.9

Two different methods were used to determine the nickel content in a steel alloy using 4 specimens each time. The results were as follows:

$\bar{x}_1 = 3.285 \quad s_1 = 0.00774 \quad n_1 = 4$

$\bar{x}_2 = 3.258 \quad s_2 = 0.00960 \quad n_2 = 4$

Is the difference between the two methods significant at 0.05 significance level?

Solution:

$$s_p = \sqrt{\frac{(4-1)\times 0.00774^2 + (4-1)\times 0.0096^2}{4+4-2}} = 0.00872$$

$$t = \frac{|3.285 - 3.258|}{0.00872 \times \sqrt{\frac{1}{4}+\frac{1}{4}}} = 4.379$$

The critical value of t is obtained from T.INV.2T $(0.05, 4+4-2) = 2.447 <$ 4.379. Hence, the difference between the two methods of testing is significant.

6.5 Hypotheses concerning paired samples

In paired samples, we compare two sets of matched specimens. For example, consider a group of students having failed an exam that are to attend extra sessions. To test whether these sessions have had any effect on their performance, we compare the scores of these students before and after the sessions. The null hypothesis concerns the mean difference in scores $\overline{D}$:

Null hypothesis: $\overline{D} = 0$ (The extra sessions did not improve the students' status)

Alternate hypothesis: $\overline{D} > 0$ (The extra sessions improved the students' status).

Whether the sample size is large or small, such problem is dealt with one – tailed statistics. The test statistic is:

$$t = \frac{\overline{D}}{s_D}\sqrt{n-1} \qquad (6.8)$$

Example 6.10

A nano-oxide is claimed to stabilize the viscosity index (V.I.) of lube oils when added at 0.5% level. Seven specimens of the same oil were tested for V.I. before and after the addition. Determine at 0.05 significance level whether the addition has had any significant effect in improving the viscosity index.

V.I. Before	101.3	111	103.5	107.6	105.5	108.2	104.1
V.I. After	103.2	111.5	106.5	108.2	109.3	111	106.3

Solution:

H_0: $\overline{D} = 0$

H_a: $\overline{D} > 0$

We first determine the difference:

V.I. Before	101.3	111	103.5	107.6	105.5	108.2	104.1
V.I. After	103.2	111.5	106.5	108.2	109.3	111	106.3
D	1.9	0.5	3	0.6	3.8	2.8	2.2

The mean value and standard deviation of differences are: $\overline{D} = 2.114, s_D = 1.228$

$t = \frac{\overline{D}\sqrt{7-1}}{s_D} = 4.217$

The critical value of t as calculated from $|\text{T. INV}\,(0.05\,, 7-1)| = 1.943$

Since $4.21 > 1.94$, then the null hypothesis cannot be accepted, and the alternate hypothesis cannot be rejected. Hence, we conclude that the addition significantly improved the viscosity index.

6.6 Type I and Type II errors

Consider the following case: When a sample of 6 foamed polystyrene slabs were tested for density, the mean value of density was 0.78 g.cm^{-3} with a standard deviation of 0.066 g.cm^{-3}. To check the null hypothesis:H_0: Mean density < 0.7 at 0.05 significance level, we calculate the statistic from equation (6.2) to obtain:

$|t| = \frac{|0.78 - 0.7|}{\frac{0.066}{\sqrt{6-1}}} = 2.71$ while $|t_{crit}| = 2.015$.

This means that the null hypothesis cannot be accepted. What if the null hypothesis was correct but the choice of the sample resulted in its rejection? In that case, we are in the presence of a **type I error in which a hypothesis is rejected while it is correct.** The error here is simply the significance level α (0.05). Decreasing the error means choosing a lower significance level. Effectively, if the significance level is reduced to 0.021, then $|t_{crit}| = 2.015$ and the null hypothesis can no more be rejected. Obviously, this means that we need to ensure to a probability of $1 - 0.021 = 0.979$ that our judgment is correct, which justifies not rejecting the null hypothesis in that case.

On the other hand, if in the previous example, the mean density of the sample turned out to be 0. 75 g.cm^{-3} with the same standard deviation, then we would had obtained $|t| = 1.69 < 2.015$ and the null hypothesis would have been accepted. Again, if another sample was chosen, what if its average density or standard deviation would have resulted in a value of $|t_{crit}| > 2.015$? here we are in presence of **a type II error where a hypothesis has been accepted while it is false.** This type of error is denoted by β and can be obtained using the Goal – Seek technique by determining the significance level at which $|t_{crit}| = 1.69$.This results in $\beta = 0.076$.

6.7 Exercise problems

In all forthcoming problems, assume the properties to be normally distributed among the population.

1. The presence of nitrogen oxides in the ambient atmosphere in the premises of a chemical plant is limited to a maximum of 600 $mg.L^{-1}$. On 30 consecutive days in a certain month, samples were drawn that resulted in a mean value of 624 $mg.L^{-1}$ and a standard deviation of 42.5 $mg.L^{-1}$. At 0.05 significance level, would you accept the null hypothesis that the mean nitrogen oxides concentration does not exceed 600 $mg.L^{-1}$? What maximum significance level is required so as not to reject this hypothesis?

2. High sulfur content in fuel oil is undesirable, and a refinery claims that its production of fuel oil contains no more than 2% sulfur. The standard deviation of sulfur content is known to be about 0.22%. What sample size would ensure that a mean sulfur content of 2.1% in that sample would not result in rejecting the refinery's claim at 0.05 significance level?

3. Design calculations have shown that the compressive strength of concrete slabs should exceed 32MPa. When 35 core concrete cubes were tested, they resulted in a mean strength of 30.6MPa with a standard deviation of 1.8MPa. At 0.05 significance level, would you accept or reject the null hypothesis: H_0: Strength > 32?

4. The diameters of flywheels produced by a factory should equal 280mm. When a 40 sized sample was chosen, the distribution of diameters was as shown below.

***D* mm**	277	278	279	280	281	282	283
f	3	11	7	8	5	4	2

   At 0.05 significance level, would you consider the null hypothesis H_0: $D = 280$ valid?

5. The yield strength of steel bars should exceed 125MPa. When 6 bars were tested, they produced the following results for yield strength: $117, 121, 122, 119, 117, 118$. What maximum significance level would not reject the null hypothesis: H_0: $\sigma > 125$?

6. According to a research center, a polymerization reaction reached a yield of 90% after 2 hours under certain experimental conditions. To confirm these results, the reaction was repeated five times under the same conditions and the yield after one hour determined. The results obtained for yield were as follows: $87.6, 88.8, 90.0, 86.5, 87.9$. At 0.05 significance level, would you consider the null hypothesis H_0: Yield $= 90$ valid?

7. The company that manufactures a certain type of mixer stated that the time required reaching a homogeneous mixture on agitating two specific immiscible liquids does not exceed 10 min. Mixing runs were carried out twelve times and the time required to homogenize the mixture was assessed by reaching a steady mean density. The results for mixing time were as follows (min.):

   10.6 , 11.2 , 9.8 , 10.7 , 11.0 , 10.5 , 9.7 , 11.5 , 11.2 , 10.0 , 9.5 , 10.2.

   Test the null hypothesis: H_0: $t \leq 10$ at significance level = 0.05.

8. Two sets of students enrolled in two different programs (A and B) were examined in the same subject. The results are summarized as follows:

	n	$\bar{x}$	s
A	41	64.3	15.6
B	51	59.5	17.2

   Use $\alpha = 0.05$ to decide whether the differences in their scores are significant.

9. The yields of a chemical reaction carried out in two different reactors produced the following results:
   - Reactor I: N° of tests = 5, average yield = 96.3%, standard deviation = 2.75
   - Reactor II: N° of tests = 6, average yield = 93.3%, standard deviation = 3.35

   Using 0.05 significance level, determine whether the difference in yield between the two reactors is significant.

10. A manufacturer of PE bottles claims that the percentage of defective production does not exceed 4%. 80 bottles were chosen at random and 6 of them were found to be defective. What maximum significance level would not reject the claim?

11. Two different techniques were used to prepare a certain catalyst and the specific surface area ($cm^2.g^{-1}$) is determined by nitrogen adsorption (BET analysis). The results of samples obtained using the two techniques were as follows:

A	115.6	127.8	100.9	134.6	122.7	119.8	123.2	116.9	109.8
B	103.8	110.6	118.2	101.4	104.7	110.0	103.2		

    Determine at a significance level = 0.05 whether there is any significant difference in applying the two techniques.

12. A researcher wants to confirm whether the use of his own method in the preparation of an elastomer (A) yields better yield than a conventional method (B). To that aim, he carries out two sets of tests and obtains the following results for the elastomer yield:

A	0.91	0.88	0.85	0.86	0.83	0.91
B	0.92	0.89	0.82	0.81	0.8	0.86

Prove that at significance level = 0.05, the difference between the two methods is statistically insignificant.

13. Seven students who failed to pass an exam were given extra tutorials and re-examined. Their scores (before and after the sessions) were as shown. Determine whether these sessions have significantly improved their performance at significance level = 0.05.

Before	9	6	4	11	10	2
After	11	6	7	12	10	5

14. A researcher wishes to prove that a certain doping additive has improved the viscosity index of a specific lubricating oil. He performs 5 tests on five specimens before and after the addition and obtains the following results:

V.I. Before	122	119	120	127	122
V.I. After	129	121	118	129	127

Prove that at significance level = 0.05, the proposed addition has not significantly affected the V.I.

Now, the researcher, who is unscrupulous, decides to doctor his data to prove his point of view by changing the third entry in the second row. What figure should he put instead of (118) to prove his point?

-7-

The χ^2 distribution

7.1 Introduction

This is a continuous distribution that is widely used to ascertain whether different observations are independent or related, through what is known as **independence test**. This test is also used to check whether a random variable follows a certain distribution in what is known as **goodness of fit test.**

The parameter included in this distribution is K = number of degrees of freedom, to be defined later.

The density function of this distribution is:

$$f(x) = \frac{\left(\frac{1}{2}\right)^{\frac{K}{2}} . x^{\frac{K}{2}-1} . e^{-\frac{x}{2}}}{\Gamma\left(\frac{K}{2}\right)} \tag{7.1}$$

This function, when plotted against x shows different curves corresponding to different values of the parameter K.

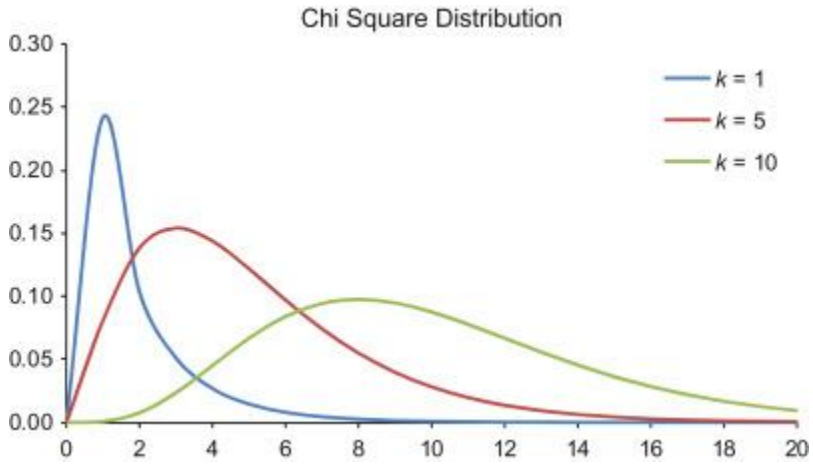


Fig (7.1): χ^2 density function for different values of K

As seen from Figure (7.1), as K increases, the distribution approaches a normal curve.

7.2 Independence tests

Consider the following simple situation: A dice is thrown 60 times. The following table shows the outcomes. The hypothesis to be tested is: the dice is uniform (that is the probability of getting any of its 6 numbers is alike). We compare in the table the observed frequency of occurrence with the hypothetical frequency.

If f_o is the value of observed frequency and f_h, the value of hypothetical frequency, then the value of χ^2 is obtained by the formula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_o - f_h)^2}{f_h} \tag{7.2}$$

In the EXCEL function CHIINV, there are two entries: K, the number of degrees of freedom and the level of significance $\alpha = 1 - L$.

In the present case, $K = 6 - 1 = 5$. This is since if we know the number of outcomes for any five numbers, we can get the sixth since the total number is known to be 60.

Number	1	2	3	4	5	6
Frequency (Obs.)	13	6	8	12	11	10
Frequency (Hyp.)	10	10	10	10	10	10

$$\chi^2 = \frac{(13-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(10-10)^2}{10}$$
$$= 3.4$$

For $K = 5$, the value of 3.4 is less than the critical value of 11.05 at a 0.05 significance level. **This means that there is a 95% probability that differences between observed values and predicted ones are due to chance**. We thus accept the hypothesis that the coin is uniform.

In the previous example, the hypothetical value of frequencies was easy to calculate as well as the number of degrees of freedom. When the hypothetical values are not known, the problem gets more complicated as follows:

It is claimed that the sales of a certain brand of detergent are independent of the geographic sector where it is sold. To test this hypothesis, we choose samples from three districts A, B and C, and obtain the number of customers using this brand.

The following table summarizes the results obtained. Such tables are called **Contingency tables**

Table (7.1): Observed frequencies of users in different districts

	Number of users	**Number of non-users**	**TOTAL**
A	125	75	200
B	143	107	250
C	89	61	150
Total	**357**	**243**	**600**

To test the validity of the hypothesis, we calculate the "theoretical frequency" of occurrence. This is done by dividing the total number of users and non – users by

the total number of customers then multiplying this figure by the total number of customers in each district.

The ratio of users is 357/600 = 0.595 and that of non – users: 243/600 = 0.405

The hypothetical number of users in zone A should be: $200 \times 0.595 = 119$, in zone B: $250 \times 0.595 \approx 149$ and in zone C: $150 \times 0.595 = 89$

Similarly, the corresponding figures for non – users are: A: $200 \times 0.405 = 81$, B: $250 \times 0.405 = 101$ and C: $150 \times 0.405 = 61$

The following contingency table shows the actual frequency and the hypothetical frequency of users.

Table 7.2: Observed and hypothetical frequencies of users

	Observed Nº of users	**Hyp. Nº of users**	**Observed Nº of non-users**	**Hyp. Nº of non-users**	**TOTAL**
A	125	119	75	81	200
B	143	148.75	107	101.25	250
C	89	89.25	61	60.75	150
Total	**357**		**243**		**600**

In such contingency tables, K, the number of degrees of freedom is obtained by the following equation:

$$K = (C - 1) \times (R - 1) \quad (7.3)$$

Where: C is the number of columns in Table 7.1 (disregarding totals) and R the number of rows.

In the current case, $K = (2 - 1) \times (3 - 1) = 2$

And from equation (7.2), calculation of χ^2 on EXCEL, yields $\chi^2 = 1.2975$

From the χ^2 function for $K = 2$, at a 0.05 level of significance, the value of 1.2975 is less than the critical value of 5.99

This means that the differences between the observed and the hypothetical values are most probably due to chance. The hypothesis is thus accepted.

Example 7.1

A large company owns three factories A, B and C. Each factory produces four brands of chemicals (C_1, C_2, C_3 and C_4). The following table shows the distribution of their products (in tons per year). Check the null hypothesis: The number of brands produced does not depend on a specific factory.

	C_1	C_2	C_3	C_4
A	16	18	15	10
B	50	8	8	10
C	30	20	10	5

Solution:

The following table summarizes the values of observed and hypothetical frequencies.

Note for example that the total number of C_1 brands produced by the three factories is 96, while the total number of products is 200. So, the C_1 brand figures must be multiplied by 96/200 = 0.48 etc…

	C_1		C_2		C_3		C_4		**Total**
	Obs.	**Hyp.**	**Obs.**	**Hyp.**	**Obs.**	**Hyp.**	**Obs.**	**Hyp.**	
A	16	28.32	18	13.57	15	9.74	10	7.38	**59**
B	50	36.48	8	17.48	8	12.54	10	9.35	**76**
C	30	31.20	20	14.95	10	10.73	5	8.13	**65**
Total	**96**		**46**		**33**		**25**		**200**

From equation (7.2), calculation of χ^2 on EXCEL, yields $\chi^2 = 25.41$

The number of degrees of freedom is $(4-1) \times (3-1) = 6$

From the χ^2 function for $K = 6$, the value of 25.41 is greater than 12.59, the value for α = 0.05. **The hypothesis cannot therefore be accepted**.

7.3 Goodness of fit tests

The χ^2 distribution can also be used to test how good a fit of data as compared to some theoretical model. The following example explains the method used.

Example 7.2

A common empirical rule, known as Trouton's rule, states that the entropy of vaporization of a solid is 88 J/mole.K. The following table gives experimental values for the latent heat of vaporization and the boiling point of some solids. From these data test the suitability of this rule at a significance level = 0.05

Solid	**Cd**	**Na**	**Mg**	**Zn**	**KCl**	**NaCl**
ΔH_{vap}**.kJ.mol^{-1}**	100	99.2	127.5	112.5	159	166
T **K**	1038	1155	1378	1180	1680	1738

Solution:

First, we recall that: $\Delta S_{vap} = \frac{\Delta H_{vap}}{T}$

The hypothesis under test is therefore: ΔS_{vap} = 88 J.mole^{-1}.K^{-1}

So, we build a table where the observed and hypothetical values of ΔS_{vap} are represented. The number of degrees of freedom is (2 – 1)×(6 – 1) = 5

Solid	$\Delta S_{vap\ obs.}$	$\Delta S_{vap\ hyp.}$
Cd	96.339	88
Na	85.887	88
Mg	92.525	88
Zn	95.339	88
KCl	94.643	88
NaCl	95.512	88

We get: $\chi^2 = 2.828$

From the χ^2 function, for K = 5, the value of 2.828 is less than 11.07, the value corresponding to α = 0.05. **The hypothesis is thus accepted.**

Example 7.3

A company produces computer chips of uniform power. To check for that uniformity, eight specimens were tested for power. Check at a significance level = 0.05 whether the power can effectively be considered uniform.

105 115 107 102 103 95 110 108

Solution:

The mean value of the sample is first determined:

$\bar{x} = 105.63$

The null hypothesis to be tested is H_0: $\mu = 105.63$

The following table shows the χ^2 calculations:

P	105	115	107	102	103	95	116	108
P_{th}	105.63	105.63	105.63	105.63	105.63	105.63	105.63	105.63
$\frac{(P - P_{th})^2}{P_{th}}$	0.0038	0.8312	0.0178	0.1247	0.0655	1.0697	0.1808	0.0532

The number of degrees of freedom = $8 - 1 - 1 = 6$ since one restriction was added by assuming the constant value as the mean value of sample. $\chi^2 = 2.347$

Critical value from CHIINV = 12.59

Since $2.347 < 12.59$, the hypothesis can be accepted, and the distribution can be assumed to be uniform.

Example 7.4

The following data belongs to the population distribution of student marks in an exam. Check, at a 0.05 level of significance, whether these data can be fitted by a normal distribution model.

Class	0 to < 5	5 to < 10	10 to < 15	15 to < 20	20 to< 25	≤ 30
Frequency	2	15	8	13	16	3

Solution:

Calculation of mean value and standard deviation: $\mu = 15.57$, $\sigma = 6.8026$

The number of degrees of freedom must be lessened by 2 since we have added a new constraint regarding the mean and standard deviation of the normal assumption = 6 – 1 – 2 = 3

Normal simulation	$x < 5$	$x < 10$	$x < 15$	$x < 20$	$x < 25$	$x < 30$
Probability	0.06011	0.20645	0.4666	0.7425	0.91716	0.9835
Calc. cum *f*	3.42649	11.7676	26.5968	42.3253	52.2783	56.0338
Obs. cum *f*	2	17	25	38	54	57
$(f_{calc} - f_{obs})^2/f_{calc}$	0.5939	2.3266	0.09587	0.4420	0.05670	0.01666

Calculated value of χ^2:
0.5939 + 3.3266 + 0.09587 + 0.442 + 0.0567 + 0.01666 = **3.532**

Critical value from CHIINV: **7.81**

Since 3.532 < 7.81, **the hypothesis is accepted.**

7.4 The use of χ^2 values to evaluate the confidence interval for variance

χ^2 values can be used to construct a confidence interval for standard deviations or variances and alternatively test hypotheses about their values assuming normally distributed population.

For a significance level α, we first determine two values for χ^2: $\chi^2_{\frac{\alpha}{2}}$ and $\chi^2_{1-\frac{\alpha}{2}}$. For a sample size n, let the variance of sample be s^2, then the confidence interval for variance of population σ^2 will be:

$$\frac{(n-1).s^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1).s^2}{\chi^2_{1-\frac{\alpha}{2}}} \qquad (7.4)$$

Example 7.4

A 10 – sized sample of crude oil was analyzed for sulfur content. The results are given in the following table.

Specimen Nº	1	2	3	4	5	6	7	8	9	10
% sulfur	1.56	1.8	1.22	1.78	2.11	1.95	1.45	2.05	2.0	1.85

Use this information to construct a confidence interval for the standard deviation of the whole production at a 90% confidence level.

Solution:

The standard deviation of sample is calculated as: $s = 0.286$

The values of $\chi^2_{0.05}$ and $\chi^2_{0.95}$ are obtained from the CHINV function as:

$\chi^2_{0.05} = 16.918$ and $\chi^2_{0.95} = 3.325$

Substituting in equation (7.4):

$$\frac{(10-1)\times 0.286^2}{16.918} < \sigma^2 < \frac{(10-1).\times 0.286^2}{3.325} \quad \text{or } 0.0435 < \sigma^2 < 0.2214$$

Hence: $\mathbf{0.2086 < \sigma < 0.4705}$

7.5 The use of χ^2 values to test hypotheses about the variance

χ^2 values can also be used to test hypotheses about the variance or alternatively the standard deviation of a population along which the variable of interest is normally distributed.

For example, let a population have an unknown variance σ^2. A sample with n values is selected, and its standard deviation (s) determined. To test the null hypothesis: H_0: $\sigma^2 = k^2$, the following statistic is used, where $d.f. = n - 1$:

$$\chi^2 = \frac{(d.f.)s^2}{k^2} \qquad (7.5)$$

The acceptance region will range from $\chi^2_{1-\frac{\alpha}{2}}$ to $\chi^2_{\frac{\alpha}{2}}$ for a significance level α

If, however, the tested hypothesis is **one – tailed**, the acceptance region will be from $\chi^2_{1-\alpha}$ to χ^2_{α}

Example 7.5

The following data show the scores of a sample of 9 students (out of 20) sampled from a large population. It will be assumed that the scores are normally distributed along this population. Test the hypothesis that the standard deviation of the population is 4 at a 5% significance level.

16 12 13 8 17 3 15 18 11

Solution:

The value of s is obtained = 4.77

For $\alpha = 0.05$: $\frac{\alpha}{2} = 0.025$ and $1 - \frac{\alpha}{2} = 0.975$

From the CHINV function, the corresponding values of χ^2, for a number of degrees of freedom = 8, are: $\chi^2_{0.025} = 17.5$ and $\chi^2_{0.975} = 2.17$.

From equation (7.5), the value of the test statistic for the variance is:

$$\chi^2 = \frac{8 \times 4.77^2}{4^2} = 11.37$$

Since this value lies within the interval (2.18; 17.5), the hypothesis is accepted at a 0.05 significance level.

7.6 Exercise problems

In all forthcoming problems take the level of significance as 0.05

(1) A dice was thrown 60 times. The following table was obtained.

1	2	3	4	5	6
10	14	9	11	9	7

Check the following hypothesis: The coin is uniform.

(2) The following table shows the pattern of sales of a brand of lube oils on different oil stations in four different governorates A, B, C and D. Test the hypothesis that the proportion of sales does not depend on the selling zone.

Zone	Users	Non – users
A	22	15
B	37	19
C	15	9
D	62	39

(3) The following represents the results obtained on monitoring the BOD of wastewater as function of the distance from the source of pollution:

Distance km	1	2	3	4	5	10	20	50	100
BOD mg.L^{-1}	61	53	48	44	41.5	38	36.5	36	35.5

Conduct a goodness of fit test to evaluate the possibility of using the following empirical equation to predict the BOD level as function of distance (D):

$$BOD = 33 + \frac{30}{D}$$

(4) The following equation represents an empirical relation between the 28-days compressive strength of cement mortar cubes (σ MPa) and the water to cement ratio used (x) for a fixed sand to cement ratio. Check the validity of that relation using the Chi squared test.

$\sigma = 8.5x^{-0.94}$

x	0.32	0.305	0.28	0.265	0.245	0.23	0.21
σ	24	28	31	32	33.5	37	42

(5) Wall tiles can contain 5 types of defects. The following table shows the number of defects found in a sample of 100 tiles.

Nº of defects	**0**	**1**	**2**	**3**	**4**	**5**
Nº of samples	15	26	25	15	12	7

Calculate the mean value and show that this distribution approximately conforms to a Poisson distribution.

(6) Ten specimens were tested from a wastewater stream for suspended solids content (ppm). At 0.05 significance level would you consider the distribution to be uniform?

450 485 460 425 460 470 430 425 450 475

(7) Construct a confidence interval for the standard deviation of the specific gravity of 8 crude specimens:

Specimen Nº	1	2	3	4	5	6	7	8
Sp. Gr.	0.82	0.76	0.77	0.80	0.75	0.76	0.82	0.84

(8) The following table shows the tensile strength (in MPa) of a 11 – sized samples of PE fibers drawn from a large population. Test the hypothesis that the standard deviation of population = 120

410 620 590 840 490 600 820 380 760 540 720

(9) The producer of a lube oil brand claims that its production is extremely uniform in properties. To this aim a client chooses 12 specimens in twelve days and tests them for viscosity at some fixed temperature. Verify the producer claim that the standard deviation does not exceed 2 cP.

32 21 27 24 30 26 28 22 25 27 32 30

-8-

Correlation and Regression

8.1 Nature of correlation

In many aspects of engineering applications, it is important to decide about the presence of some correlation between two or more variables. This is particularly true when there is an intuitive feeling of the presence of a correlation. For example, a field engineer notices that the incidence of failure of electrical insulators fitted to high voltage wire feeding an electrostatic precipitator increases with an increase in inlet gas temperature. He "feels" that there is some correlation between the two factors. The purpose of the present section is to introduce the concept of correlation and the method of its estimation in case of two or more variables. The simplest case is that of linear correlation between two variables.

The nature of correlation between two variables x and y can follow any of the five schemes shown in Figure (8.1) known as **scatter diagrams**. In the first case (a), a perfect linear correlation is shown, while in (b), data suggest an increasing character of correlation. In (c), the correlation is very poor, while in (d), there is an inverse correlation, and it becomes a perfect linear decreasing correlation in (e).

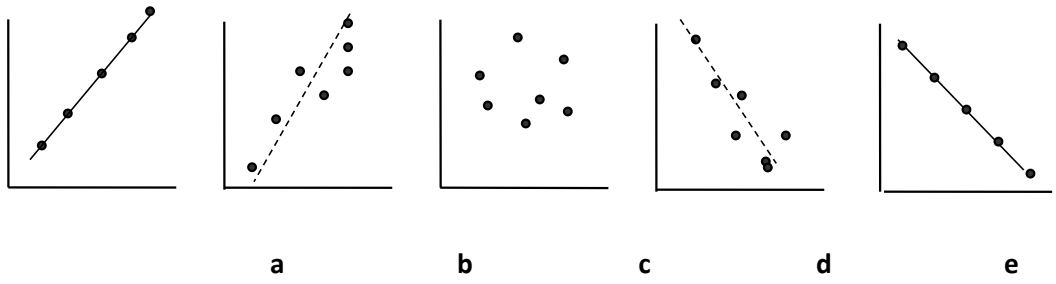


Fig (8.1): Types of linear correlations

8.1.1 The Pearson correlation coefficient

The extent to which the observed data fit to a linear correlation has been quantified by **Pearson**. He first stated that the type of correlation should be independent of the origin chosen and the scale used. So, he suggested that all x and y values be normalized in a way like that done in the normal distribution by defining:

$$X_i = \frac{x_i - \bar{x}_i}{s_x} \text{ and } Y_i = \frac{y_i - \bar{y}_i}{s_y}$$

He then suggested that the extent of correlation be calculated from the rule:

$$R = \frac{\Sigma\, Xi.Yi}{n}$$

This factor (R) is termed the **linear correlation coefficient**.

An easier way of computing this coefficient is by expanding the terms of the above definition.

$$R = \frac{\sum_{i=1}^{n} (x_i - \bar{x}).(y_i - \bar{y})}{s_x.s_y} = \frac{\sum_{i=1}^{n} (x_i.y_i - x_i.\bar{y} - \bar{x}y_i + \bar{x}.\bar{y})}{s_x.s_y}$$

Recalling that $\sum_{i=1}^{n} x_i = n.\bar{x}$ and that $\sum_{i=1}^{n} y_i = n.\bar{y}$ as well as the definition of the standard deviation from equation (1.7), we get the following form for the linear correlation coefficient:

$$R = \frac{n.\sum_{i=1}^{n} x_i.y_i - \sum_{i=1}^{n} x_i.\sum_{i=1}^{n} y_i}{\sqrt{n.\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \times \sqrt{n.\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \quad (8.1)$$

It can be generally proved that: $-1 \leq R \leq 1$.

The value of R is 1 for perfectly increasing linear correlation (Case a in Figure (8.1), and –1 for a perfectly linear decreasing correlation (Case e). The more the value of $|R|$ approaches unity, the higher is the extent to which points gather about a straight line. The case of poor correlation (c) would correspond to a value of R close to zero.

Currently, values of correlation coefficient (R) are readily obtained using the EXCEL function "correlation" (= CORELL).

Example 8.1

The following table relates the scores obtained by 16 students in two different exams A and B (Out of 20). Plot the scatter diagram then estimate the correlation coefficient.

A	17	4	8	12	15	13	3	10	18	9	12	5	19	16	12	14
B	14	7	6	16	14	10	3	14	17	5	10	6	20	17	9	15

Solution:

The scatter diagram is shown in Figure (8.2)

The value of R can be directly obtained from the CORELL function that yields ***R* = 0.8715**

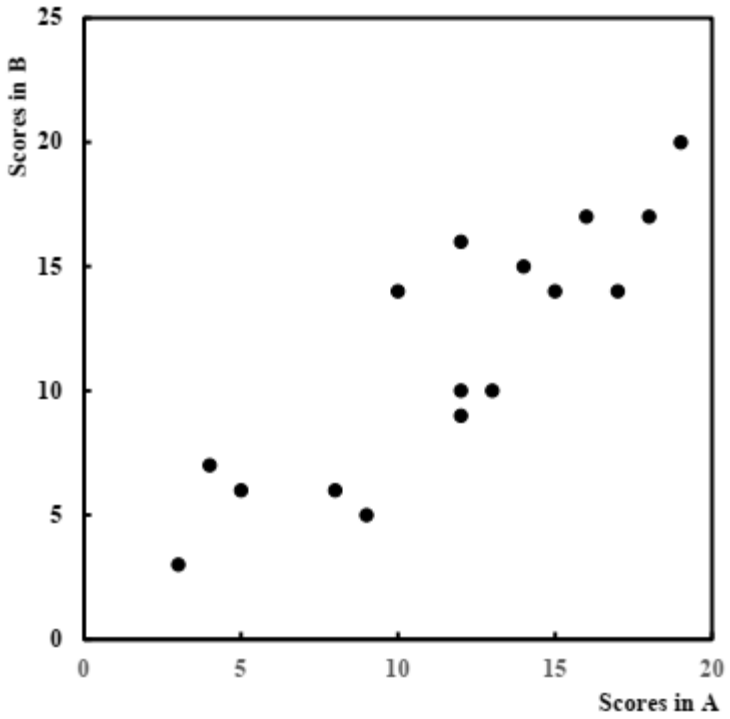


Fig (8.2): Scatter diagram of example (8.1)

8.1.2 Confidence interval of *R*: The Fisher method

The value of *R* calculated in the previous example using equation (8.1) has been obtained from sample data. The question that arises is: How far does this value represent the correlation between *x* and *y* for the whole population. The correlation coefficient for population is designated as ρ. For relatively large sample size ($n > 15$), it may be assumed that errors are normally distributed along samples taken from the population. The confidence interval of ρ is:

$$z_1 = 0.5 \ln \frac{1+R}{1-R} - \frac{z_{crit}}{\sqrt{n-3}} \quad (8.2)$$

$$z_1 = 0.5 \ln \frac{1+R}{1-R} + \frac{z_{crit}}{\sqrt{n-3}} \quad (8.3)$$

Where, z_{crit} is the critical z – value obtained from the function NORM.S.INV at a confidence level $0.5(1 + L)$.

The limits of ρ are then obtained from the following expression:

$$\tanh z_1 < \rho < \tanh z_2 \quad (8.4)$$

For example, if this method is applied on example (8.1) case where $R = 0.826$ and $n = 16$, at significance level = 0.05, we get from equations (8.2) and (8.3):

$$z_1 = 0.5 \ln \frac{1 + 0.8715}{1 - 0.8715} - \frac{1.96}{\sqrt{16 - 3}} = 0.7956$$

$$z_2 = 0.5 \ln \frac{1 + 0.8715}{1 - 0.8715} + \frac{1.96}{\sqrt{16 - 3}} = 1.883$$

From equation (8.4):

tanh 0.7975 < ρ < tanh 1.883

0.662 < ρ < 0.955

8.1.3 Testing hypothesis for correlation coefficient

When the numerical value of R is close to 1 or to zero, then it is easy to make inference about the strength of correlation; that is, if R = 0.95, for example, then one can safely say that there is a strong correlation between the two variables. The same is true if R = 0.1, where practically no correlation exists.

In the event of obtaining inconclusive values of R (such as 0.5 for example), it is necessary to find a way of telling whether the correlation exists or is absent. To this aim, a test is conducted where the null hypothesis involves assuming no correlation at all, that is: H_0: $\rho = 0$

For relatively large sample size ($n > 15$), the test statistic is:

$$z = \frac{|R|.\sqrt{n}}{\sqrt{1 - R^2}} \tag{8.5}$$

While for smaller sample size, it is:

$$t = \frac{|R|.\sqrt{n-2}}{\sqrt{1 - R^2}} \tag{8.6}$$

With $n - 2$ degrees of freedom.

For instance, in Example (8.1), n = 16 and R = 0.8715.
The null hypothesis is H_0: $\rho = 0$

From equation (8.6):

$$t = \frac{0.8715 \times \sqrt{16 - 2}}{\sqrt{1 - 0.8715^2}} = 6.649$$

The critical value of t at α = 0.05 and *d.f.* = 8 as obtained from the function T.INV.2T = 2.306.

Since 6.649 > 2.306 then the null hypothesis is not accepted meaning that the presence of a correlation between the scores in the two exams for the whole population cannot be rejected.

8.2 Linear regression

8.2.1 Equation of the regression line

The interpretation of the value of R obtained is better understood by obtaining the equation of the best straight line passing between the points.
As can be seen from Figure (8.2), the points seem to point out an increasing relation between the two variables. A straight line can be fitted to pass between these points. The best fit is obtained when the sum of squares of differences between the actual and the calculated value of y is a minimum value. The straight line obtained is called **the regression line.**

Let the equation of that line be: $y = a.x + b$

The value of y corresponding to an entry x_i calculated from the above equation is $y_i = a.x_i + b$

If the actual value of y corresponding to x_i is y_i, then the deviation between the actual and the calculated value is:

$D_i = y_i - (a.x_i + b)$

The best line is obtained when:

$\Sigma D_i^2 = \Sigma [y_i - (a.x_i + b)]^2$ is a minimum value. (8.7)

To obtain the values of the constants a and b, the following sets of equations must be solved together:

$\frac{\partial \Sigma D_i^2}{\partial a} = 0$ and $\frac{\partial \Sigma D_i^2}{\partial b} = 0$

Developing the RHS of equation (8.7) we get:

$\Sigma D_i^2 = \Sigma [y_i^2 + (a.x_i + b)^2 - 2.y_i .(a.x_i + b)]$

$= \Sigma y_i^2 + a^2. \Sigma x_i^2 + 2.a.b.\Sigma x_i + n.b^2 - 2.a. \Sigma x_i .y_i - 2.b.\Sigma y_i$

Performing partial differentiation with respect to a, we get;

$2.a.\Sigma x_i^2 + 2.b.\Sigma x_i - 2.\Sigma x_i .y_i = 0$

Also, performing partial differentiation wrt b we get:

$2.a.\Sigma x_i + 2.n.b - 2.\Sigma y_i = 0$

Solving the above two equations for a and b, we get:

$$a = \frac{n.\sum_{i=1}^{n} x_i.y_i - \sum_{i=1}^{n} x_i . \sum_{i=1}^{n} y_i}{n.\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \quad (8.8)$$

$$b = \frac{\sum_{i=1}^{n} y_i - a.\sum_{i=1}^{n} x_i}{n} \qquad (8.9)$$

The EXCEL program readily displays the regression equation through the command: "Add trend line" by right clicking on any point on the scatter diagram of Figure (8.2).

Example 8.3
Find the equation of the regression line of the data in example (8.1)

Solution:
The equation relating the marks obtained in exam A (x_A) to those of exam B (x_B) is directly obtained by the EXCEL command: "Add trend line"

$$\boldsymbol{x_B = 0.9052 x_A + 0.8575}$$

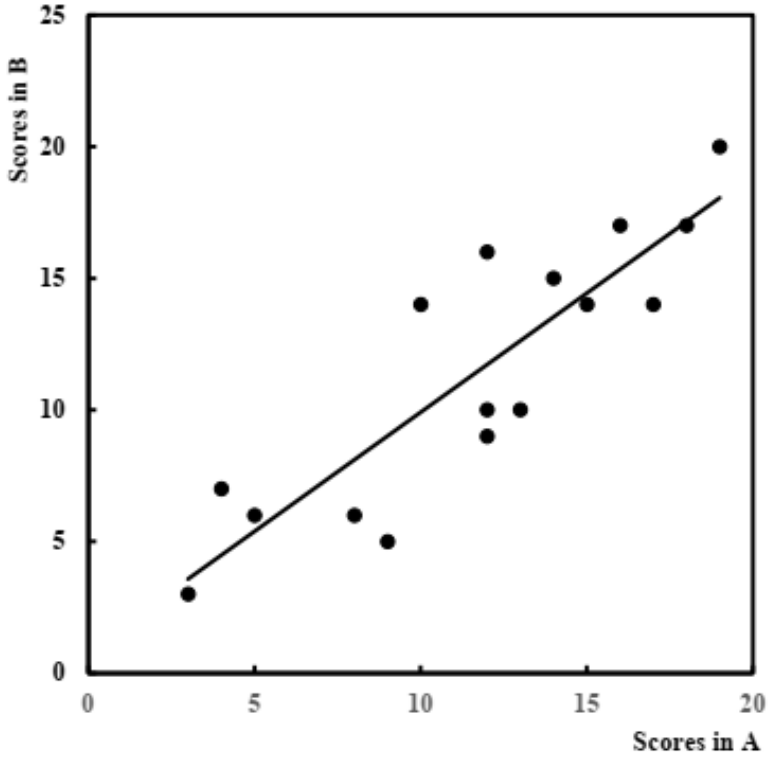


Fig (8.3): Regression line with its equation

8.2.2 The standard error of estimate

The regression line gives an estimate about the average relationship between the two variables for the set of given points. Had another set of points been chosen, the equation would have been different.

In the following section a method is shown to estimate how far does this relation represents the actual relation between variables. In other words, the errors associated

with using the regression line equation, rather than the actual values, will be computed.

Let the observed values of y corresponding to values of x_i be y_i, and the calculated values $y_{ci} = a.x_i + b$

Let $\bar{y}$ be the mean value of observed value = $\Sigma\, y_i / n$

There are three types of deviations that can be considered:

- The deviation of the mean from any observed value: $D_i = \bar{y} - y_i$
- The deviation between observed and calculated values, known as **unexplained deviation** or **residual**: $D_{si} = y_i - y_c$
- The deviation between calculated values and the mean value, known as **explained deviation**: $D_{ci} = y_c - \bar{y}$

The following definitions are related to the above differences.

- **The total variation**: $$\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (8.10)$$
- **The unexplained variation**: $$\sum_{i=1}^{n} D_{si}^2 = \sum_{i=1}^{n} (y_i - y_c)^2 \qquad (8.11)$$
- **The explained variation**: $$\sum_{i=1}^{n} D_{ci}^2 = \sum_{i=1}^{n} (y_c - \bar{y})^2 \qquad (8.12)$$

Finally, it can be shown that:

$$\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} D_{si}^2 + \sum_{i=1}^{n} D_c^2 \qquad (8.13)$$

The above definitions can be understood in the light of the following discussion:

When the value of the independent variable x is higher than the mean values $\bar{x}$ of observations, we expect in an increasing relation that the value of y will also be higher than that of the mean $\bar{y}$. This expected difference between y_c and $\bar{y}$ is hence called explained difference. The difference between observed and calculated values is termed unexplained since it will probably be associated with other factors than the independent variable under consideration.

The square root of the unexplained variance is a type of standard deviation known as the **standard error of estimate**. This is computed from the definition:

$$S_{yx} = \sqrt{\frac{\Sigma(y_c - y_i)^2}{n-2}} \qquad (8.14)$$

The calculation of this quantity is necessary to set the limits of accuracy of the calculated values of y to any desired confidence level. It is assumed that the errors are normally distributed about their mean value.

Let the required significance level be α corresponding to a value of z_{crit} as obtained from Table (6.1) (or from NORMINVS), then the expected limits of the values of y corresponding to x_i, can be determined from:

$$y_c - z_{crit}.s_{xy} < y < y_c + z_{crit}.s_{xy} \tag{8.15}$$

Note that the standard error of estimate can be directly obtained in EXCEL from the function STEYX.

8.2.3 The coefficient of determination

The extent to which the variables x and y are correlated can be estimated by obtaining the ratio of explained variation to the total variation: This is a positive number known as the determination coefficient.

$$R^2 = \frac{\sum_{i=1}^{n}(y_c - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8.16}$$

For linear regression this value is directly obtained after the line is plotted by clicking on the command "Display R^2 value on chart".

In example (8.4), its value = R^2 = 0.631.

This means that 63.1% of the variation in strength is due to the variation in water to cement ratio while the remaining 36.9% are due to other factors including experimental errors.

Example 8.4

The following data was obtained on investigating the effect temperature (T K) on the thermal conductivity of an insulating material (k W.m^{-1}.K^{-1})

T K	300	350	400	450	500	550	600
k	0.019	0.025	0.03	0.045	0.056	0.07	0.089

Plot the relation between the two variables. Then, obtain the probable limits of accuracy of the forecast at significance level = 0.05.

Solution:

First, we draw the regression line and obtain its equation. This takes the form:

$k = 0.00021T - 0.05011$

We then calculate the value of k for each value of T using that equation to obtain the third row in the table.

To obtain the probable limits of accuracy, we obtain the standard error of estimate s_{xy} using the function STEYX. This value = 0.003206. Then, equation (8.15) is applied to obtain the maximum and minimum expected values of k at each temperature, revealed in the 4th and 5th rows of the table.

Figure (8.4) summarizes the previous results.

T K	300	350	400	450	500	550	600
k	0.015	0.025	0.03	0.045	0.052	0.07	0.076
k_{calc}	0.01289	0.02339	0.0339	0.0444	0.0549	0.0654	0.0759
k_{min}	0.00661	0.017106	0.0276	0.0381	0.0486	0.0591	0.0696
k_{max}	0.01917	0.029674	0.0402	0.0507	0.0612	0.0717	0.0822

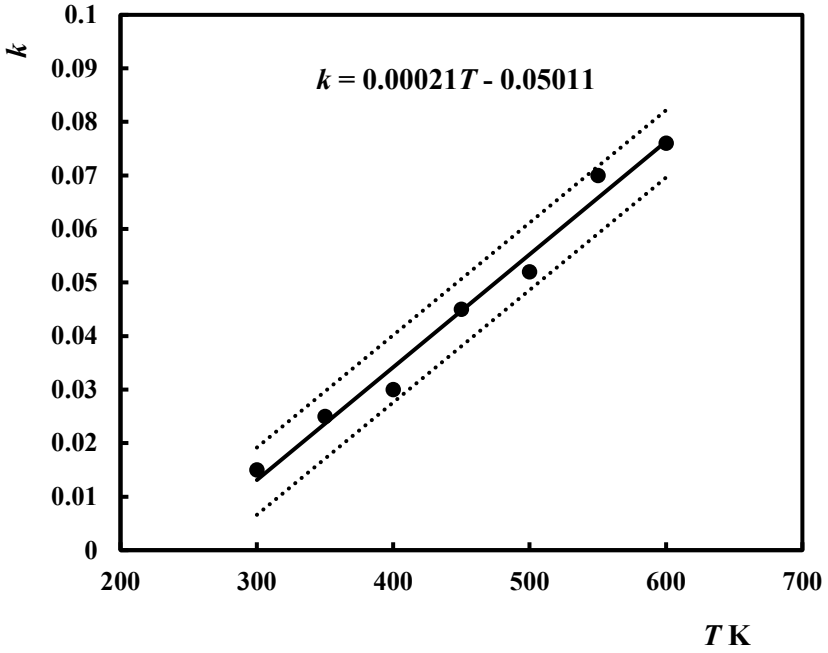


Fig (8.4): Regression line of Example (8.4)

8.3 Non – linear regression

Regressions obtained through experimentation are not necessarily linear. In that case, the methods explained in the previous section are not valid.

- In general, it is always possible to assume **a polynomial relation** between the variables. This takes the form:

$$y_c = a_0 + a_1.x + a_2.x^2 + \ldots + a_n.x^n = \Sigma\, a_k.x_i^k \qquad (8.17)$$

The values of the coefficients a_0 , a_1 , a_2, ..., a_n are obtained by setting n conditions in the form:
For k = 1 to n:

$$\frac{\partial \Sigma D_i^2}{\partial a_k} = 0 \qquad \text{Where } D_i = y_i - \Sigma\, a_k.x_i^k$$

This yields a set of n linear equations that can be written in the following matrix form:

$$\boldsymbol{Z = N.A} \qquad (8.18)$$

For example, if the suggested regression equation is a second degree polynomial in the form: $y_c = a_0 + a_1.x_i + a_2.x_i^2$, the form of the matrix N is: $\boldsymbol{N} = \begin{pmatrix} n & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \end{pmatrix}$

Where $\boldsymbol{A}$ is the column matrix $[a_k] = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}$ and $\boldsymbol{Z}$ is the column vector $\begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \\ \Sigma x_i^2 y_i \end{bmatrix}$

Hence the coefficients can be obtained from: $\boldsymbol{A} = \boldsymbol{N}^{-1}\boldsymbol{Z}$

The value of the determination coefficient is calculated from the basic definition (8.14)

- If the suggested regression is of **the exponential type** in the form:

$y_c = a.e^{kx_i}$

This can be transformed to a linear form by taking logarithms of both sides:

$\ln y_c = \ln a + kx_i$

A linear regression will thus be performed between $\ln y_c$ and x_i

- If the suggested regression is of **the power type** in the form:

$y_c = a.x_i^n$

This can be transformed to a linear form by taking logarithms of both sides to get:

$\ln y_c = \ln a + n.\ln x_i$

A linear regression will thus be performed between $\mathrm{lln}\, y_c$ and $\ln x_i$.

Currently, all common types of non – linear regressions are available through the EXCEL program using the curve fitting module (Insert chart). This gives the best fit between experimental data for any assumed regression, as well as the coefficient of determination.

Example 8.5

A distribution that is often used to relate the mass fraction of suspended solid particles in water to their size is the Rosin – Rammler distribution:

$\varphi = e^{-\left(\frac{D}{D_m}\right)^n}$

Where: φ is the fraction having particle size > D, n an empirical constant and D_m a characteristic particle diameter.

The following table illustrates experimental data obtained in this respect using a "sedigraph" analyzer.

$D\ \mu m$	80	60	50	30	15	10	8	6	5	4	3	2	1
φ	0.005	0.017	0.047	0.277	0.627	0.767	0.887	0.917	0.941	0.963	0.98	0.995	0.999

From a suitable plot, determine the values of the parameters D_m and n.

Solution:

First, the relation should be linearized. Taking logarithms of both sides:

$$\ln \varphi = -\left(\frac{D}{D_m}\right)^n$$

$$\ln(-\ln \varphi) = n \ln D - n \ln D_m$$

Therefore, a plot of $\ln(-\ln \varphi)$ against $\ln D$ should produce a straight line of slope n and intercept $n.\ln D_m$

The table of calculations is shown below together with the corresponding plot (Figure 8.5).

$D\ \mu m$	80	60	50	30	15	10	8	6	5	4	3	2	1
φ	0.002	0.018	0.048	0.248	0.638	0.798	0.858	0.908	0.932	0.952	0.971	0.987	0.999
$\ln D$	4.382	4.094	3.912	3.401	2.708	2.303	2.079	1.792	1.609	1.386	1.099	0.693	0.000
$\ln \ln -\varphi$	1.667	1.405	1.118	0.250	-0.762	-1.327	-2.121	-2.446	-2.800	-3.278	-3.902	-5.296	1.667

The regression equation is:

$$\ln(-\ln \varphi) = 1.794 \ln D - 5.862 \qquad (R^2 = 0.984)$$

Hence n = **1.794** and $n.\ln D_m = 5.862$, which finally yields D_m = **26.25 µm**

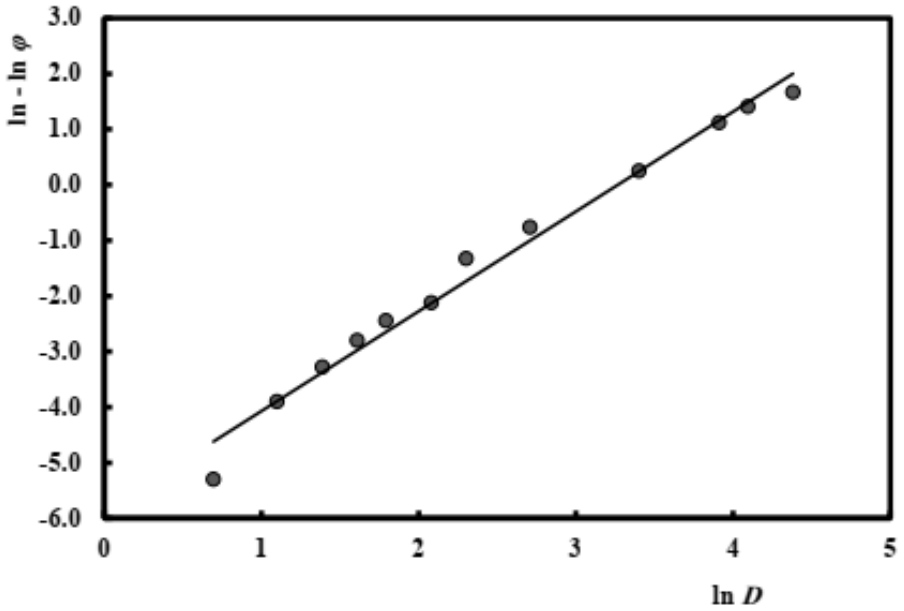


Fig (8.5): Linear relation for data of example (8.4)

Example 8.6

An experiment was conducted to follow up the effect of time on the molecular weight of a polymer. The data obtained were as follows:

Time h	60	120	180	240	300	360
MW	576	12870	45900	112000	225000	1275000

Fit an exponential to represent this correlation and calculate the coefficient of determination and deduce the expected value of the molecular weight after 420 min.

Solution:

An exponential function was fitted using EXCEL module: Insert chart: "Exponential". The plot is shown in Fig (8.6) together with the determination coefficient R^2.

The equation is: $y = 388.20e^{0.02x} \qquad R^2 = 0.99$

At $t = 420, y = 388.20e^{0.02\times420} = 1726350$

Note that both the exponent (0.02) and the determination coefficient (0.99) appear with 2 decimals only. This is not necessarily a suitable approximation. Right click on the equation and R^2 icon, go to "format trendline labels" and choose the option "Number". The default number of decimals is 2. Set it to 4 (or 5). You will get the following results for the fitted exponential equation and R^2:

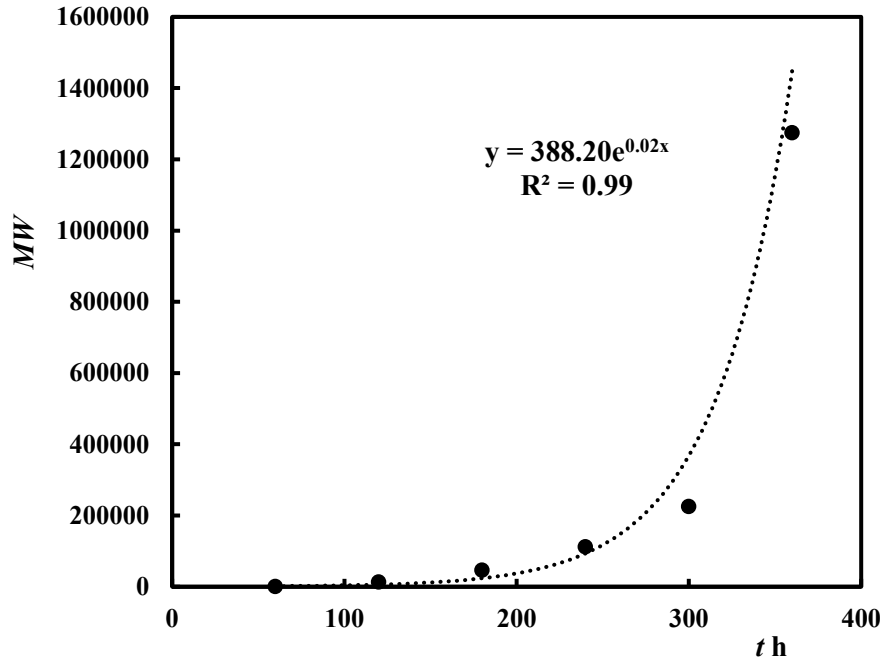


Fig (8.6): Exponential regression of example (8.6)

$y = 388.1997e^{0.0229x} \qquad R^2 = 0.9912.$

The exact molecular weight after 420 min is:

$y = 388.1997e^{0.0229\times420} = 5835788.$

Comparing with the previous figure (1726350) obtained by the two-decimals exponent shows the necessity of getting precise values for the equation constants.

Example 8.6

The following data were obtained relating the mean bending strength of hardened gypsum board samples to the particle size of gypsum, under a restriction that the maximum particle size that can be used is 1 mm (1000 μm)

Size *D* μm	74	104	147	208	294	416	590
Strength σ MPa	7.2	6.4	5.6	4.9	4.5	4.3	4.1

Optimize the relation between strength and particle size by choosing the most suitable regression equation.

Solution:

The relation is plotted, and the following trials are performed:

Power function

The regression equation is $\sigma = 22.95\ D^{-0.28}$ with $R^2 = 0.961$ (Fig 8.7)

Polynomial function

A third-degree polynomial has been fitted in Fig (8.8) and although it yields a value of $R^2 = 0.998$, it cannot be accepted as an extrapolation of the trend indicates that the strength can reach negative values if $D > 800$ μm, which is illogical. (Note that the relation should be valid for up to $D = 1000\mu m$).

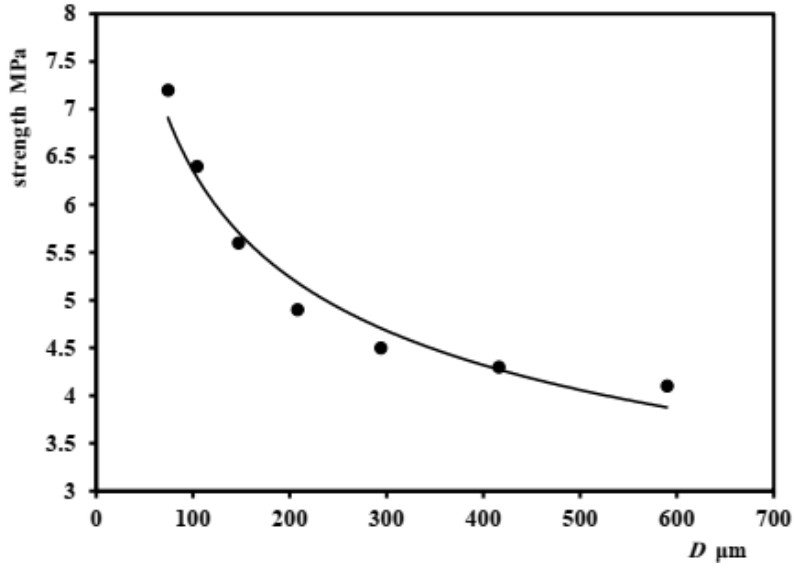


Fig (8.7): Fitted power function for data of Example (8.6)

Exponential function

A direct exponential function fitting of data does not yield a reasonable value of R^2, as exponential functions tend to approach 0 as D approaches ∞ (R^2 = 0.787). However, a look at the table reveals that an asymptotic value of about 4 MPa may be assumed as D approaches 1000.

That is why, a plot of $\sigma - 4$ was performed against D as shown in Fig (8.9).

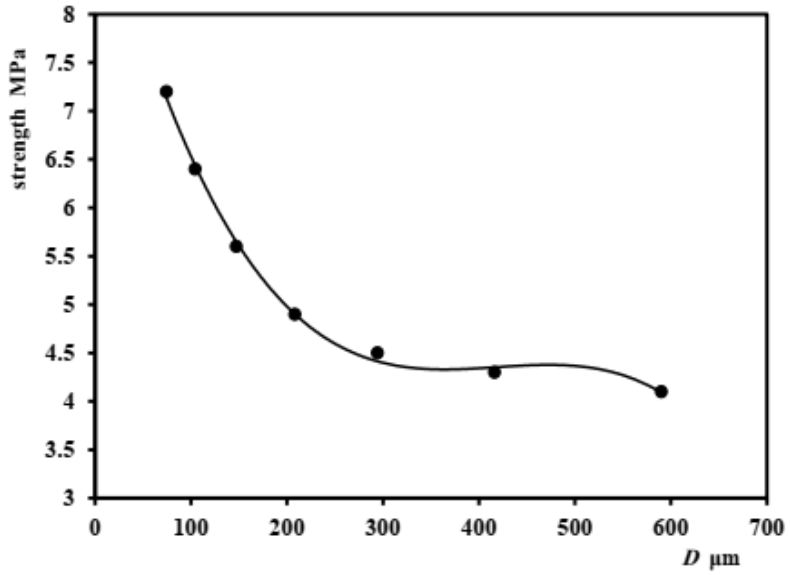


Fig (8.8): Fitted 3rd degree polynomial for data of Example (8.6)

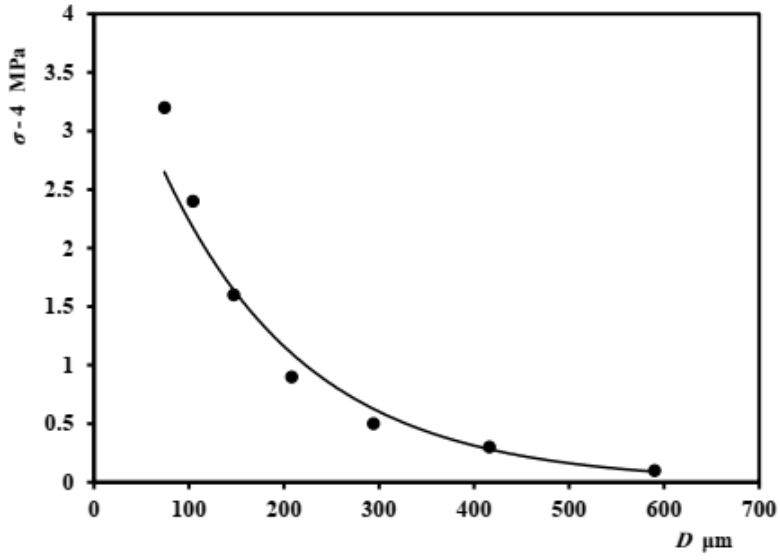


Fig (8.9): Fitted modified exponential function for data of Example (8.6)

That is why, a plot of $\sigma - 4$ was performed against D as shown in Fig (8.9).

The deduced function takes the form: $\sigma = 4 + 4.305e^{-0.0066D}$ with $R^2 = 0.983$

8.4 Exercise problems

(1) During a test with a thermocouple the e.m.f. (mV) was related to temperature through the following table:

T °C	100	200	300	400	500	600	700	800	900	1000
E mV	5	5.5	10.5	13.6	18	20.2	26.5	27.3	28.6	35

Obtain a linear regression for temperature as function of e.m.f. and determine the correlation coefficient. Then construct a confidence interval for the population coefficient at significance level = 0.05

(2) The following table shows the results obtained in a poll covering 24 randomly chosen graduate students to relate their scores in the midterm exam in a certain subject A and another subject B. The results were as follows:

A	16	12	11	13	8	14	19	4	5	14	13	20
B	14	10	12	12	10	14	20	7	4	12	13	20
A	12	13	17	11	9	15	12	11	8	20	18	14
B	12	12	15	12	11	14	10	9	5	17	17	12

Prepare a scatter diagram and deduce the regression equation and the correlation coefficient.

(3) The following table shows experimental values for thermal conductivity of insulating boards K at different values of porosity. Express K as a linear function of porosity and deduce the correlation coefficient. At a confidence level of 0.95, construct a confidence interval for the population determination coefficient using the Fisher method.

Porosity	0.35	0.43	0.28	0.49	0.36	0.50	0.38	0.44	0.52	0.33
K **W/m.K**	0.89	0.77	0.98	0.78	0.85	0.65	0.85	0.80	0.60	0.81

(4) Compressive strength tests are performed on samples of concrete mortar cubes after 28 days curing as function of cement content per cubic meter concrete. Each sample consists of 3 specimens. The results are as follows:

Cement content	Specimen 1	Specimen 2	Specimen 3
230	28	27	30.2
260	32.3	30.9	33.4
320	36	34.8	33.9
345	38.7	40	38.2
390	41.3	42.5	42.7
420	43.5	42.8	43

Obtain a linear regression equation relating the mean sample strength to the cement content. For a confidence level of 0.95, draw error bars as well as two lines representing the lower and upper boundaries of expected strength values.

(5) The following equation has been suggested to relate specific heat of a solid (J/mol.K) to temperature (K): $c_p = a + b.T + c.T^2$

Using the given data, find the values of the constants *a, b* and *c*. Obtain the determination coefficient.

T K	300	400	500	600	700	800	900
c_p	25	27.5	27.7	28.5	28.8	30.1	30.4

(6) The rate constant of a chemical reaction is known to be related to temperature (K) by the relation: $k = Ae^{-\frac{E}{RT}}$

From the following data relating *k* to temperature, estimate the values of A and E (J/mole) and estimate the coefficient of determination.

T K	300	340	385	420	455	500
k	0.016	0.02	0.022	0.023	0.024	0.026

(7) The compressibility factor *Z* of a real gas has been related to its molar volume by the relation: $Z = a + b/V + c/V^2$. Using a suitable regression find the values of the constants *a*, *b* and *c* and estimate the coefficient of determination.

V m^3/mole	0.02	0.03	0.04	0.05	0.06	0.07
Z	1.23	1.15	1.14	1.09	1.07	1.05

(8) The following data were obtained on following the sedimentation of fine silt in water. The height represents the level of interface between clear liquid and suspension.

Time min	0	15	30	45	60	90	120	150	360
Height mm	430	405	385	380	355	345	335	325	290

Find an equation describing the sedimentation operation in the form:
$h = h_\infty + ke^{-c.t}$
Write down the coefficient of determination.

(9) Agricultural waste is used for the adsorption of heavy metal ions from wastewater. The equilibrium concentration of ions (q_e mg.L^{-1}) follows the Langmuir model:

$$q_e = \frac{k.c}{1+bc}$$

Where, c is the concentration of the adsorbed phase ($mg.L^{-1}$).

Prove that the following data are compatible with the above expression using a linear plot:

c	0.0045	0.0087	0.021	0.026	0.092	0.195
q_e	0.026	0.053	0.075	0.082	0.123	0.129

Obtain a correlation coefficient for that relation.

(10) The relation between the mole fraction of a volatile component in vapor phase (y) and its mole fraction in liquid phase (x) is often obtained by the following expression where α is the relative volatility:

$$y = \frac{\alpha.x}{1+(\alpha-1).x}$$

Express the following data in linearized form, then find the value of α from two different parameters obtained from the regression equation.

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
y	0.128	0.231	0.36	0.432	0.56	0.66	0.729	0.844	0.9

(11) The following relation shows the variation of fractional conversion α with time $(t \text{ min})$ in the decomposition of aluminum hydroxide ($Al(OH)_3$):

$$\alpha = 1 - e^{-k.t^n}$$

From the following data, deduce the values of the constants k and n.

t	1	2	3	4	5	6	7	8	9	10
α	0.09	0.22	0.35	0.46	0.56	0.66	0.71	0.76	0.8	0.83