स्वाध्याय स्वमन्थन स्वावलम्बन

UTTAR PRADESH RAJARSHI TANDON OPEN UNIVERSITY

(Established vide U.P. Govt. Act No. 10, of 1999)

MCA-5.3
NUMERICAL AND STATISTICAL COMPUTING

THIRD - BLOCK
Statistical Computing

ira Gandhi National Open University

UP Rajarshi Tandon Open University

hantipuram (Sector-F), Phaphamau, Allahabad - 211013

Uttar Pradesh
Rajarshi Tandon Open University

Block

3

STATISTICAL COMPUTING

BLOCK INTRODUCTION

Statistics is in fact a science related to analysis of gathered data and hence strengthens us to conclude and comment on the data gathered by conducting an experiment, this experiment could belong to any stream of education, science, advertising, etc. and could be any software too. It is the statistical analysis of the customers requirement which make any industry (could be software industry too) to develop a variety of products, and hence facilitate their users. You may agree with this, because it was the statistical need analysis of GUI platforms which make software companies like Microsoft and many more to switch from CUI to GUI mode. Thus, the study of the topics discussed in this block will definitely strengthen your analytical skill in a scientific way.

This block is composed of three units: Probability distribution, Pseudo random number generation, Regression. Since, Computers is a discrete science so we should study the mathematical concepts, which are discrete in nature, that's why we talked about probability and its distribution. It is to be noted that in any experiment the events are not a sure shot, their occurrence is quite random and this random behaviour analysis is handled by various distributions discussed in this block. We have further extended the topic of randomness and also discussed the algorithms related to Pseudo random number generation. Finally, in this block we have talked about the technique of regression, which contributes to the determination of exactness out of randomness. Regression techniques will help you to analyse the random results of any experiment and hence bring you in the position that you can comment on the results obtained, these results could be the outcome produced by any software too.

UNIT 1 PROBABILITY DISTRIBUTIONS

Structure **Page Nos.**

1.0 INTRODUCTION

The discipline of statistics deals with the collection, analysis and interpretation of data. Outcomes may vary, even when measurements are taken under conditions that appear to be the same. Variation is a fact of life. Proper statistical methods can help us understand the inherent variability in the collected data, and facilitate the appropriate analysis of the same. Because of this variability, uncertainty is a part of many of our decisions. In medical research, for example, interest may center on the effectiveness of a new vaccine for AIDS; an agronomist may want to decide, if an increase in yield can be attributed to a new strain of wheat; a meteorologist may be interested in predicting, whether it is going to rain on a particular day; an environmentalist may be interested in testing, whether new controls can lead to a decrease in the pollution level; an economist's interest may lie in estimating the unemployment rate, etc. Statistics, and probabilistic foundations on which statistical methods are based, can provide the models that may be used to make decisions in these and many other situations involving uncertainties.

Any realistic model of a real world phenomenon must take into account the possibilities of randomness i.e., more often, the quantities we are interested in are not predicted in advance, but rather will exhibit inherent variation and that should be taken into account by the model. Such a model is, naturally enough, referred to as a probability model.

In this unit, we shall see what a random variable is, and define it for a particular random experiment. We shall see that there are two major types of probability distribution. We shall look into their properties and study the different applications.

1.1 OBJECTIVES

After going through this unit, you should be able to:

- describe the events and sample spaces associated with an experiment;
- define a random variable associated with an experiment;

- decide the whether a random variable is discrete or continuous;
- describe the following distributions:
  a) Binomial distribution
  b) Poisson distribution
  c) Uniform distribution
  d) Exponential distribution
  e) Normal distribution, and
  f) Chi-square distribution.

1.2 RANDOM VARIABLES

Definition: A "Random experiment" or a "Statistical experiment" is any act whose outcome cannot be predicted in advance. Any outcome of a random experiment is known as "event".

We will start with the following illustrations:

1) The number of telephone calls received by Monica, a telephone operator in a call center in Delhi, between 1:00 am and 3:00 am in the early morning.
2) The amount of rainfall in Mumbai on August 1st.
3) The number of misprints on a randomly chosen page of a particular book.
4) The final results of the 5 one-day matches between India-Pakistan.
5) The outcome of rolling dice.
6) The volume of sales of a certain item in a given year.
7) Time to failure of a machine.

In all the above cases there is one common feature. These experiments describe the process of associating a number to an outcome of the experiment (i.e. to an event). A function which associates a number to each possible outcome of the experiment is called a "random variable". It is often the case that our primary interest is in the numerical value of the random variable rather than the outcome itself. The following examples will help to make this idea clear.

Example 1: Suppose we are interested in the number of heads, say X, obtained in three tosses of a coin.

Solution: If we toss a coin three times, then the experiment has a total of eight possible outcomes, and they are as follows;

a_1={HHH}, a_2={HHT}, a_3={HTH}, a_4={HTT}
a_5={THH}, a_6={THT}, a_7={TTH}, a_8={TTT}

Denoting the event corresponding to getting k heads, k=0,1,2,3,.. as{X=k}, observe that {X=0} => {a_8} ; {X=1} => {a_4, a_6,a_7} ; {X=2} =>{a_2, a_3,a_5} ; {X=1} =>{a_1}

In above expressions, each value in the support of X corresponds to some element (or set of elements) in the sample space S. For example, the value 0 corresponds to the element {a_8}, while the value 1 corresponds to the set of elements { a_4, a_6,a_7}.

Therefore, the sample space S, the set of all possible outcomes of this experiment can be expressed as

$$S=\{a_1, a_2, \ldots, a_8\}$$

Since, X is the characteristics, which denotes the number of heads out of the three tosses, it is associated with each outcome of this experiment. Therefore, X is a function defined on the elements of S and the possible values of X are {0,1,2,3}. The

set of possible values that X can take is called the support of X which may be denoted as χ. Observe, that X can be explicitly expressed as follows;

$X(a_1)=3, X(a_2)=X(a_3)=X(a_5)=2, X(a_4)=X(a_6)=X(a_7)=1, X(a_8)=0$

It is interesting to observe that to each value there is always some elements in the sample space or a set of element in the sample spaces. For, example, the set of element, in the sample spaces corresponding to the value '0' is the point $\{a_8\}$; for 1, the set is $\{a_4, a_6, a_7\}$, for 2, the set is $\{a_2, a_3, a_5\}$ and for 3 the point is $\{a_1\}$.

Therefore, we can easily make the following identification of events corresponding to the values associated by X. Denoting the event corresponding to '0', as {X=0}, similarly for other values, observe that

$\{X=0\}=\{a_8\}; \{X=1\}=\{a_4, a_6, a_7\}; \{X=2\}=\{a_2, a_3, a_5\}\ \{X=3\}=\{a_1\}$

If we assume that the coin is unbiased and the tosses have been performed independently, the probabilities of all the outcomes of the sample space are equal, that is $P(a_1)=P(a_2)=\ldots=P(a_8)=\frac{1}{8}$. Therefore, using the probability law of disjoint events we can easily obtain

$$P(\{X=0\}) = P(\{a_8\}) = \frac{1}{8}$$

$$P(\{X=1\}) = P(\{a_4, a_6, a_7\}) = P(\{a_4\})+P(\{a_6\})+P(\{a_7\}) = \frac{3}{8}$$

$$P(\{X=2\}) = P(\{a_2, a3, a5\}) = P(\{a_2\})+P(\{a_3\})+P(\{a_5\}) = \frac{3}{8}$$

$$P(\{X=3\}) = P(\{a_1\}) = \frac{1}{8}$$

Therefore, in this case, the random variable X takes four values 0,1, 2,3 with the probabilities 1/8,3/8,3/8,1/8 respectively. It is also important to observe that

$P(\{X=0\})+P(\{X=1\})+P(\{X=2\})+P(\{X=3\})=1$

It is not a coincidence, it is always true. If we add the probabilities of all possible values of a random variable it is always one.

To sum up, we say that the random variable X, is a real valued function defined on all the elements of a sample space of a random experiment. The random variable takes different values, and for each value there is a probability associated with it. The sum of all the probabilities of all the points in the support χ of the random variable X adds up to one.

The *Figure 1* will demonstrate the random variable X.

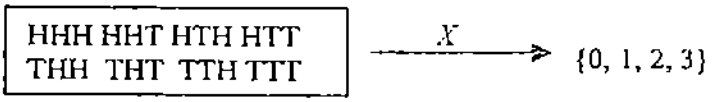


Figure 1: Schematic representation of a random variable

Because of this representation, one can define probabilities for the set of numbers (depending on the random variable) rather than working with arbitrary space and this simplifies the problem considerably. Now, let us consider the following example.

Example 2: You have purchased a new battery operated wristwatch and you have inserted one new battery into it. Suppose you are interested in the following:

a) How long will it be before the first battery needs replacement?
b) How many battery will have to be replaced during a one year period?

Solution: Note that both (a) and (b) are random variables. Let us discuss them one by one. In case (a), we want to find the duration of time before the battery needs to be replaced. Note that the variable takes values continuously along a line say from the time duration A to time duration B. No values in between A and B are left out. In other words there is no break in the values assumed by this random variable.

In case (b), the random variable is the number of batteries. This variable can take values 0 or 1 or 2 etc. There is no continuity, since only non-negative integer values can be assumed. So, the range of this variable is a discrete set of points. From this discussion it is clear that the random variable X defined in Example 1 is also a discrete random variable. The above examples show that the random variables can be of two types. We will distinguish between the following two types of random variables;

1) Discrete Random Variable , and
2) Continuous Random Variable.

☞ Check Your Progress 1

1) Suppose you take a 50-question multiple-choice examination, guessing your answer, and are interested in the number of correct answers obtained. Then

   (a) What is the random variable X that you will consider for this situation?
   (b) What is the set of possible values of X in this example?
   (c) What does $P(X=10)$ mean in this context? .

   .....

   .....

   .....

Now in the next two sections we will describe the discrete and continuous random variables in detail.

1.3 DISCRETE RANDOM VARIABLE

In this section, we define a discrete random variable and mention some of its basic properties.

Definition: A random variable X is said to be discrete, if the total number of values X can take is finite or countably infinite (i.e the support of X is either finite or countable).

The support χ of X may be listed as $\{a_0, a_1, a_2, \ldots\}$. Moreover, for each value of a_i, there is a probability associated with it. Suppose we denote them as $\{p_0, p_1, p_2, \ldots\}$, therefore, we have $P(X= a_i)=p_i$ for $I=0,1,\ldots$. From the properties of a random variable and from the probability law, we have

(a) $pi \geq 0$ for all $i \geq 0$

(b) $\sum_{i=0}^{\infty} p_i = p_0 + p_1 + p_2 + \ldots = 1$

From the above discussions, it is follows that there exists a function p: $\chi \rightarrow \mathbf{R}$ as follows;

$$p(a) = \begin{cases} p_i & if \quad a = a_i; \quad i = 0,1,2..... \\ 0 & \text{otherwise} \end{cases}$$

This function p is called the probability mass function (p.m.f.) of the discrete random variable X. The collection of the pairs $\{(a_i, p_i); i=0,1,...\}$ is called the probability distribution of X.

Another function which plays a very important role for any random variable is known as the cumulative distribution function (c.d.f.) or simply the distribution function of the random variable. The c.d.f. $F : \mathbf{R} \rightarrow [0,1]$ of the random variable X is defined as

$F(b) = P(X \leq b)$, for $-\infty < b < \infty$.

In other words, F(b) denotes the probability that the random variable X takes on a value which will be less than or equal to b. Some important properties of the c.d.f. F (.) are

(a) F (b) is a non-decreasing function of b.
(b) $\lim_{b \to \infty} F(b) = 1$
(c) $\lim_{b \to -\infty} F(b) = 0$

Now, we clarify these concepts with the same example discussed in the previous section. Suppose X is the random variable denoting the number of heads obtained in three independent tosses of a fair coin, then the probability mass function (p.m.f) p is the function, $p: \chi \rightarrow \mathbf{R}$, such that

$$p(0) = \frac{1}{8}, p(1) = p(2) = \frac{3}{8}, p(3) = \frac{1}{8}$$

Therefore, $p(a_i) = p_i \geq 0$, for all a_i and

$$\sum_{i=0}^{3} p_i = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1$$

In this case the p.m.f of the random variable by the function p and the corresponding probability distribution is the set $\left\{\left(0,\frac{1}{8}\right),\left(1,\frac{3}{8}\right),\left(2,\frac{3}{8}\right),\left(3,\frac{1}{8}\right)\right\}$. This can also be expressed in a tabular form as follows:

Table 1: Probability distribution of the number of heads in three independent tosses of a fair coin

The number of heads (X value)	Probability
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

Now, let us see the graphical representation of the distribution.

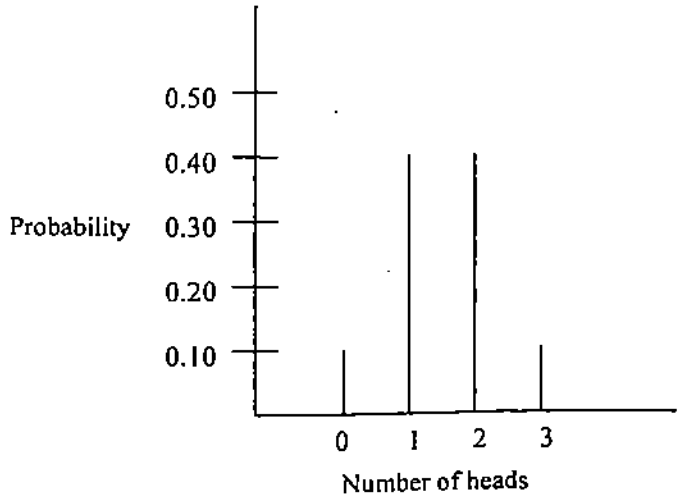


Figure 2: Graphical representation of the distribution of X

Graphically along the horizontal axis, plot the various possible values a_i of a random variable and on each value erect a vertical line with height proportional to the corresponding probability p_i.

Now, let us consider the c.d.f of the random variable X. Note that if b<0, clearly $F(b)=P(X \le b)=0$, because X takes values only {0,1,2,3}. If b=0, that is $F(0)=P(X \le 0)=P(X=0)=1/8$. If $0 \le b \le 1$, then $P(X \le b) = P(X=0)+P(0<X<b) = 1/8+0 = 1/8$. Similarly, if b=1, $F(1)= P(X \le 1) = P(X=0) + P(X=1)=1/8+3/8=4/8$ and so on. Therefore, the c.d.f. F(.) has the following form;

$$F(b)=\begin{cases} 0 & if \quad b<0 \\ \frac{1}{8} & if \quad 0 \le b<1 \\ \frac{4}{8} & if \quad 1 \le b<2 \\ \frac{7}{8} & if \quad 2 \le b<3 \\ 1 & if \quad b \le 3 \end{cases}$$

Note: Mathematical expectation or Expected values or Expectations forms, the fundamental idea in the study of probability distribution of any discrete random variable X, the expected value (or mean), denoted as $E(X)$ is defined as

$$E(X) = x_0p_0+ x_1p_1+ x_2p_2+.....=\Sigma x_ip_i$$

Where x_0, x_1, x_2 etc., are the values assumed by X and p_0, p_1, p_2 etc are probabilities of these values. Under special conditions (like all probabilities are equal) then

$$E(X) = \text{mean of } x_0, x_1, x_2.....$$

Similarly for continuous variables X having density function p(x) where $P[X=x] = p(x)$, the Expectation $E(X)$ will be given by integral of $x_ip(x_i)$ w.r.t x. This concept of Expectation also contributes to the definition of **Moment Generating Function of X i.e $M_x(t)= E(e^{tx})$.**

Example 3: A box contains twice as many red marbles as green marbles. One marble is drawn at random from the box and is replaced; then a second marble is drawn at random from the box. If both marbles are green you win Rs. 50; if both marbles are red you loose Rs. 10; and if the marbles are of different colours, then you neither loose nor win. Determine the probability distribution for the amount you win or loose?

Solution: Say X denote the amount you win (+) or loose (–) ; i.e X= +50 or –10

The probability that both marbles are green is 1/9 i.e. P[X= +50] =1/9
The probability that both marbles are red is 4/9 i.e. P[X= –10] =4/9
The probability that marbles are of different colours is 4/9 i.e. P[X=0] =4/9

Thus, the probability distribution is given by following table

Amount(in Rs won(+) or lost(–))	Probability
+50	1/9
0	4/9
–10	4/9

☞ Check Your Progress 2

1) Which of the variables given below are discrete? Give reasons for your answer.

   (a) The daily measurement of snowfall at Shimla.
   (b) The number of industrial accidents in each month in West Bengal.
   (c) The number of defective goods in a shipment of goods from a manufacturer.

   .....

   .....

   .....

1.3.1 Binomial Distribution

One very important discrete random variable (or discrete distribution) is the binomial distribution. In this sub-section, we shall discuss this random variable and its probability distribution.

Quite often we have to deal with the experiments where there are only two possible outcomes. For example, when a coin is tossed either a head or a tail will come up, a newborn is either a girl or a boy, a seed either germinates or fails to germinate. Let us consider such an experiment. For example consider the same experiment of tossing a coin independently three times. Note that, the coin need not necessarily be a fair one, that is P(Head) may not be equal to P(Tail).

This particular experiment has a certain characteristics. First of all, it involves repetition of three identical experiments (trials). Each trial has only two possible outcomes: a Head or a Tail. We refer to the outcome 'Head' as success and the outcome 'Tail' as failure. All trials are independent of each other. We also know that the probability of getting a 'Head' in a trial is p and probability of getting a 'tail' in a trial is 1 – p, that is

P(Head) = P(success) = p and P(Tail) = P(failure) = q = 1 – p

This shows that the probability of getting a 'success' or a 'failure' does not change from one trial to another. If X denotes the total number of 'Heads', obtained in three

trials, then X is a random variable, which takes values {0,1,2,3}. Then regarding the above experiment, we have observed the following;

1) It involves a repetition of n identical trials (Here n=3).
2) The trials are independent of each other.
3) Each trial has two possible outcomes.
4) The probability of success (p) and the probability of failure (q=1–p) remain constant and do not change from trial to trial.

Now, let us try to compute the probabilities $P\{X=0\}$, $P\{X=1\}$, $P\{X=2\}$ and $P\{X=3\}$ in this case. Note that

$P(X=0)$ = P(getting tails in all three trials)
$= P(\{TTT\}) = (1-p)^3 = q^3$.

Similarly,
$P(X=1)$ = P(getting one Tail and two Heads in three trials)
$= P(\{THH,HTH,HHT\}) = P(\{THH\}) + P(\{HTH\}) + (\{HHT\})$
$= (1-p)^2p+(1-p)^2p+(1-p)^2p = 3(1-p)^2p = 3q^2p$.

Similarly,
$P(X=2)$ = (getting two Tails and one Head in three trials)
$= P(\{HTT,THT,TTH\}) = P(\{HTT\}) + P(\{THT\}) + (\{TTH\})$
$= (1-p)p^2 + (1-p)p^2 + (1-p)p^2 = 3(1-p)p^2 = 3q^2p$

Finally
$P(X=3)$ = P(getting Heads in three trials)
$= P(\{HHH\}) = p^3$

Now, observe that instead of n=3, in the above example, we can easily compute the probability for any general n. Suppose we compute $P(X = r)$, for $0 \le r \le n$, then note that

$$P(X = r) = C(n,r)p^r(1-p)^{n-r} = C(n,r)p^rq^{n-r},$$

Where C(n,r) denotes the number of ways n places can be filled with r Heads and $n-r$ Tails. From your school mathematics, recall that it is the number of combination of n objects taken r at a time and it can be calculated by the following formula:

$$C(n,r) = \frac{n!}{r!(n-r)!}$$

Therefore, for $r = 0,1,\ldots,n$,

$$P(X=r) = \frac{n!}{r!(n-r)!}p^rq^{n-r},$$

where

n = the number of trials made
r = the number of successes
p = the probability of success in a trial
$q = 1 - p$ = the probability of a failure.

Now, we define the binomial distribution formally.

Let X represents the number of successes in the set of n independent identical trials.

Then X is a discrete random variable taking values $0,1,\ldots,n$. The probability of the event P(X= r) is given by

$$P(X=r)=\frac{n!}{r!(n-r)!}p^r q^{n-r},\ \text{r=0,1,2,.....,n}$$

where n, r, p, q are same as defined before. Such a random variable X is called a binomial random variable and its probability distribution is called the binomial distribution. A Binomial distribution has two parameters n and p.

☞ Check Your Progress 3

1) A farmer buys a quantity of cabbage seeds from a company that claims that approximately 90% of the seeds will germinate if planted properly. If four seeds are planted, what is the probability that exactly two will germinate?

.....

.....

.....

1.3.2 Poisson Distribution

In this sub-section, we will introduce another discrete distribution called 'Poisson Distribution'. First, we shall describe the different situations where we can apply this Poisson Distribution.

Suppose it is the first hour in a bank on a busy Monday morning, and we are interested in the number of customers who might arrive during that hour, or during a 5-minute or a 10-minute interval in that hour. In statistical terms, we want to find the probabilities for the number of arrivals in a certain time interval frame.

To find this probability, we shall make some assumptions similar to the binomial distribution.

a) The average arrival rate at any time, remains the same over the entire first hour.
b) The number of arrivals in a time interval does not depend on what has happened in the previous time intervals.
c) It is extremely unlikely that more than one customer will arrive at the same time.

Under these assumptions, we can find the required probabilities. Suppose X is the random variable denoting the number of customers that arrive in the first hour, then

$$P(X=i)=\frac{e^{-\lambda}\lambda^i}{i!},\qquad i=0,1,2,3,...$$

Where λ (the Greek letter Lambda) denotes the average arrival rate per hour. For example, suppose we know that the average number of customers that arrive in that bank during the first hour is 60 and we want to find whether there will be no more than 3 customers in the first 10 minutes. Since, we know that the average arrival rate per hour is 60, if we denote λ to be the average arrival rate per 10 minutes, then $\lambda=\frac{60\times10}{60}=10$. Therefore, we can use the above formula and get

$$P(X=i)=\frac{e^{-10}10^i}{i!},\qquad i=0,1,2,3.....$$

But we want to find the probability that no more than 3 customers will be there in the first ten minutes and that is

$$P(X \le 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3) \quad (1)$$

$$= e^{-10} + e^{-10}10 + \frac{e^{-10}10^2}{2!} + \frac{e^{-10}10^3}{3!} \quad (2)$$

$$\approx 0.00005 + 0.00050 + 0.00226 + 0.00757 = 0.01038 \quad (3)$$

What does this value 0.01038 indicate? It tells us that if the arrival rates are uniform then there is only 1% chance that less than three customers will be there, or in other words, there is a 99% chance that there will be more then 3 customers in the first 10 minutes.

Similarly, if we want to find whether there will be no more than 3 customers in the first 5 minutes, then similarly, as above we can see that in this case $\lambda = \frac{60 \times 5}{60} = 5$. Therefore, if Y denotes the number of customers presents in the first 5 minutes, then

$$P(Y \le 3) = P(Y=0) + P(Y=1) + P(Y=2) + P(Y=3) \quad (4)$$

$$= e^{-5} + 5e^{-5} + \frac{e^{-5}5^2}{2!} + \frac{e^{-5}5^3}{3!} \quad (5)$$

$$\approx 0.00674 + 0.03369 + 0.08422 + 0.14037 = 0.26502 \quad (6)$$

From the above two examples it is clear that if we change the time unit (and hence the value of λ), the probabilities will change. The probability mass function (p.m.f) given by

$$p(i) = P(X=i) = \frac{e^{-\lambda}\lambda^i}{i!}, \quad i = 0,1,2,3,\ldots$$

represents the Poisson probability distribution. From the series expansion of e^{λ}, it easily follows that as it should be.

$$\sum_{i=0}^{\infty} P(X=i) = \sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!} = 1$$

One point that should always be kept in mind is that a random variable denoting the number of occurrences in an interval of time will follow a Poisson distribution, if the occurrences have the following characteristics:

a) The average occurrence rate per unit time is constant.
b) Occurrence in an interval is independent of what has happened previously.
c) The chance that more than one occurrence will happen at the same time is negligible.

Now, let us look at some situations where we can apply the Poisson distribution. Here is an example

Example 4: Calls at a particular call center occur at an average rate of 8 calls per 10 minutes. Suppose that the operator leaves his position for a 5 minute coffee break. What is the chance that exactly one call comes in while the operator is away?

Solution: In this case the conditions (a), (b) and (c) are satisfied. Therefore, if X denotes the number of calls during a 5 minute interval, then X is a Poisson random variable with $\lambda = \frac{8 \times 5}{10} = 4$. Therefore,

$$P(X=1) = \frac{e^{-4}4^1}{1!} = 4e^{-4} \approx 0.073$$

That means the chance is 7.3% that the operator misses exactly one call.

☞ **Check Your Progress 4**

1) If a bank receives on an average $\lambda = 6$ bad Cheques per day, what is the probability that it will receive 4 bad checks on any given day.

.....

.....

.....

.....

.....

.....

1.4 CONTINUOUS RANDOM VARIABLE

So far we have discussed, discrete random variables in details and we have provided two important discrete distributions namely, binomial and Poisson distributions. Now, in this section, we will be discussing another type of random variables namely, continuous random variables.

Let us look at the part (a) of Example 2. Note that we want to find the time of occurrence rather than the number of occurrences. Therefore, if the random variable X denotes the time of occurrence of a particular event, then the random variable X can take any value on the positive real line or may be any value on a fixed interval say (A,B). Therefore, the random variable can take uncountably many values. This type of a random variable which can take uncountably many values is called a continuous random variable. For a continuous random variable X, the probability that X takes a particular value is always zero, but we can always specify the probability of X of any interval through a probability density function (p.d.f.). The exact details are given below.

Definition: Let X be a continuous random variable which takes values in the interval (A,B). A real valued function $f(x)$: $\mathbf{R} \rightarrow \mathbf{R}$ is called the p.d.f of X, if

a) $f(x) \geq 0$ and $f(x) = 0$, if $x < A$ or $x > B$.

b) $\int_A^B f(x)dx = 1$

c) $P(c < X < d) = \int_c^d f(x)dx$

Now, we shall see how we can use the graph of the p.d.f. of a continuous random variable to study real life problems.

Example 5: Suppose the Director of a training programme wants to conduct a programme to upgrade the supervisory skills of the production line supervisors. Because the programme is self-administered, supervisors require different number of hours to complete the programme. Based on a past study, it is known that the following p.d.f. shows the distribution of time spent by a candidate to complete the programme.

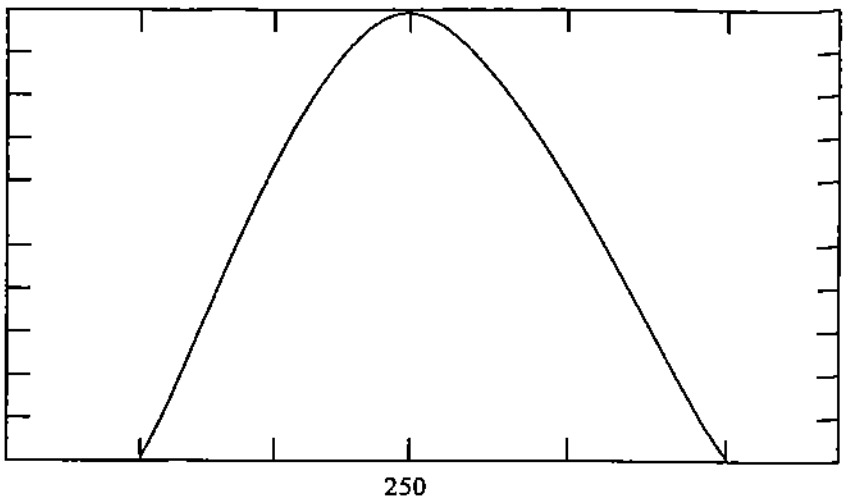


Figure 3: The p.d.f. of the time spent by a candidate to complete the program

In *Figure 3*, it is clear that the average time spent by the candidate is 250 and it is symmetrically distributed around 250. How can the Director use this graph to find the following? What is the chance that a participant selected at random will require:

a) more than 250 hours to complete the program
b) less than 250 hours to complete the program

Solution: Since the graph is symmetric, therefore, it is clear that area under the curve above 250 is half. Therefore, the probability that the random variable takes values higher than 250 is $\frac{1}{2}$. Similarly, the random variable takes value lower than 250 is also $\frac{1}{2}$.

In the following sub-sections, we will consider different continuous distributions.

1.4.1 Uniform Random Variable

The uniform distribution is the simplest of a few well-known continuous distributions, which occur quite often. It can be explained intuitively very easily. Suppose X is a continuous random variable such that if we take any subinterval of the sample space, then the probability that X belongs to this subinterval is the same as the probability that X belongs to any other subintervals of the same length. The distribution corresponding to this random variable is known as a uniform distribution and this random variable is called a uniform random variable.

Formally, we define the uniform random variable as follows: The random variable X is a uniform random variable between (A, B), if its p.d.f. is given by

$$f(x) = \begin{cases} \frac{1}{B-A} & for \quad A < x < B \\ 0 & \text{otherwise} \end{cases}$$

From the p.d.f. it is clear that if $A < a_1 < b_1 < B$, $A < a_2 < b_2 < B$ and $b_1 - a_1 = b_2 - a_2$, then $P(a_1 < X < b_1) = P(a_2 < X < b_2)$. Therefore, if the length of the intervals are the

same then, the corresponding probabilities will be also equal. Let us see some examples of such random variables:

Example 6: A train is likely to arrive at a station at any time uniformly between 6:15 am and 6:30 am. Suppose X denotes the time the train reaches, measured in minutes, after 6:00 am.

Solution: In this case X is a uniform random variable that takes value between (15,30). Note that in this $P(20 < X < 25)$ is same $P(18 < x < 23)$ and that is equal to

$$\frac{25-30}{30-15}=\frac{23-18}{30-15}=\frac{1}{3}$$

☞ Check Your Progress 5

1) An office fire drill is scheduled for a particular day, and the fire alarm is likely to ring uniformly at any time between 10:00 am to 1:00 pm.

.....

.....

.....

1.4.2 Exponential Random Variable

In making of mathematical model for a real world phenomenon, it is always necessary to make certain simplifying assumptions so as to render the mathematical tractability. On the other hand, however, we cannot make too many simplifying assumptions, for then our conclusions obtained from the mathematical model, would not be applicable to the real world problem. Thus, in short, we must take enough simplifying assumptions to enable us to handle the mathematics but not so many that the mathematical model no longer resembles the real world problem. One simplifying assumption that is often made is to assume that certain random variables are exponentially distributed. The reason for this is that the exponential distribution is both easy to work and is often a good approximation to the actual distribution.

We use exponential distribution to model lifetime data that is the data, which are mainly non-negative. Although, with proper modifications, it can be used to analyse any type of data (not necessarily non-negative only). The property of the exponential distribution, which makes it easy to analyse, is that it does not deteriorate with time. By this we mean that if the lifetime of an item is exponentially distributed, then an item which has been in use for say ten hours, is as good as a new item in regards to the amount of time remaining until the item fails.

Now, we define the exponential distribution formally: A continuous random variable X is said to be an exponential random variable if the p.d.f of X is given for some $X>0$, by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x>0 \\ 0 & \text{otherwise} \end{cases}$$

Here λ is known as the rate constant. It can be shown mathematically that the average value or the mean values of X is $\frac{1}{\lambda}$. Shapes of f (x) for different values of λ are provided in the *Figure 4*.

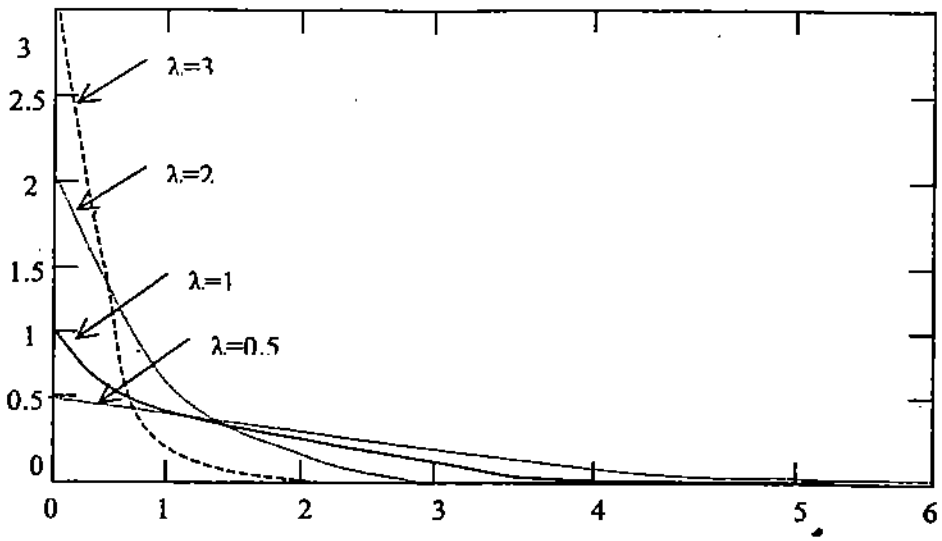


Figure 4: The p.d.f. of the exponential distribution for different values of λ

It is clear that f(x) is a decreasing function for all values of λ and f(x) tends to 0 as x tends to ∞. Now consider the following example.

Example 7: Suppose that the amount of time one spends in a bank to withdraw cash from an evening counter is exponentially distributed with mean ten minutes, that is $\lambda = 1/10$. What is the probability that the customer will spend more than fifteen minutes in the counter?

Solution: If X represents the amount of time that the customer spends in the counter than we need to find P(X>15). Therefore,

$$P(X > 15) = \int_{15}^{\infty} \lambda e^{-\lambda x} = e^{-15\lambda} = e^{-\frac{3}{2}} \approx 0.223$$

P(X>15)= .223 represents that there is a 22.3 % chance that the customer has to wait more than 15 minutes.

1.4.3 Normal Distribution

Normal distribution is undoubtedly the most used continuous distribution in different areas such as, astronomy, biology, psychology and of course in probability and statistics too. Because of its practical as well as theoretical importance it has received considerable attentions in different fields. The normal distribution has a unique position in probability theory, and it can be used as an approximation to other distributions. In practice, 'normal theory' can frequently be applied with a small risk of serious error, when substantially non-normal distributions correspond more closely to the observed value. The work of **Gauss** in 1809 and 1816 established techniques based on normal distribution, which became standard methods used during the nineteenth century. Because of **Gauss's** enormous contribution, it is also popularly known as Gaussian distribution.

We will now state the normal distribution formally: The random variable X is said to be normally distributed with parameters μ and σ , if the p.d.f f(x) of X is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}, \quad where \quad -\infty < x < \infty$$

Here μ is a real number lying between -∞ and ∞ and σ is a real number lying between 0 and ∞.

The function f(x) may look a bit strange, but do not let it bother you. Just notice the following important things. Note that it involves two parameters μ and σ, that means corresponding to each μ and σ we get a distribution function. Moreover, it can be seen that for $-\infty < \mu < \infty$ and $0 < \sigma < \infty$, the function f(x) is symmetric about μ and is a 'bell shaped' one. Both μ and σ have nice interpretation. It can be easily checked that μ is the average value or mean of the distribution and σ provides the measure of spread. The p.d.f. of two different normal distributions are provided in *Figure 5*.

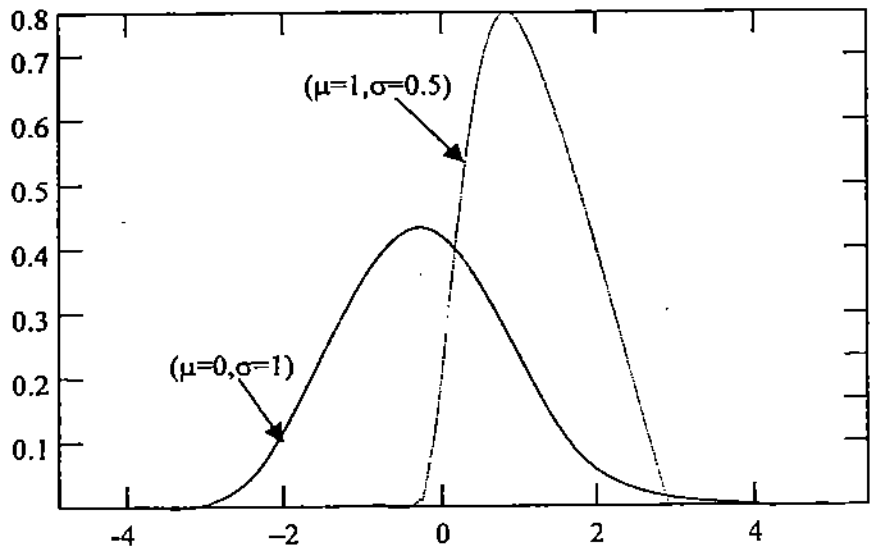


Figure 5: The p.d.f of the normal distribution for two different values of (μ, σ).

It is clear in *Figure 5*, that the p.d.f. is symmetric about μ and the shape depends on σ. The spread of the distribution is more if σ is large.

Now, let us find the $P(a < X < b)$ for any a and b, when X follows normal distribution with parameters μ and σ, note that,

$$P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz.$$

The last equality follows by simply making the transformation $z = \frac{x-\mu}{\sigma}$.

Therefore it follows

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right),$$

Where Z follows normal distribution with parameters 0 and 1. Although the probability cannot be calculated in a compact form, extensive tables are available for P (Z < z) for different values of z. The table values can be used to compute P (a < X < b) for any μ and σ.

Say we denote F(a) = P[Z<= a], the probability that the standard normal variable Z takes values less than or equal to 'a'. The values of F for different values of a are calculated and listed in the table. One such table is given at the end of this unit.

Note that, the entries in the table are values of z for z=0.00 0.01,0.02 , ..0.09. To calculate the probability that a random variable having standard normal distribution will take on a value between a and b, we use the equation

$$P[a < Z < b] = F(b) - F(a)$$

And if either a or b is negative then we can make use of the identity

$$F(-z) = 1 - F(z)$$

Example 8: Use the table to find the following probabilities

a) P[0.87 < Z < 1.28]
b) P[–0.34 < Z < 0.62]
c) P[Z >= 0.85]
d) P[Z >= –0.65]

Solution: a) P[0.87 < Z < 1.28] : Find F(1.28 from the table of Normal distribution given at the end of this block/unit). In the table in the row for Z=1.2 find the value under column 0.08 it will be 0.8997 . Similarly find F(0.87) =0.8078

so, P[0.87 < Z < 1.28] = 0.8997 – 0.8078 = 0.0919

b) Similarly, P[–0.34 < Z < 0.62] = F(0.62) – F(0.34) = F(0.62) – [1– F(0.34)]
= 0.7324 – (1 – 0.6331) = 0.3655

c) Similarly, calculate P[Z > 0.85] = 1– P[Z <= 0.85] = 1–F(0.85) = 0.1977

d) P[Z > –0.65] = 1 – P [Z <= – 0.65]
= 1 – F (–0.65)
= 1– F(1–F(0.65))
= 0.7422

Next, we shall see that how to use the standard normal probability table to calculate probability of any normal distribution.

Standardising

Any normal random variable X, which has mean μ and variance σ^2 can be standardised as follows.

Take a variable X, and
i) subtract its mean (m or μ) and then,
ii) divide by its standard deviation(s or σ).

We will call the result, Z, so

$$Z = \frac{X - \mu}{\sigma}$$

For example, suppose, as earlier, that X is an individual's IQ score and that it has a normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 15$. To standardize

and individuals IQ score, X, we subtract $\mu = 100$ and divide the result by $\sigma = 15$ to give,

$$Z = \frac{X - 100}{15}$$

In this way every value of X, has a corresponding value of Z. For instance, when $X = 130$, $Z = \frac{130 - 100}{15} = 2$ and when $X = 90$, $Z = \frac{90 - 100}{15} = -0.67$.

The distribution of standardised normal random variables

The reason for standardizing a normal random variable in this way is that a standardised normal random variable $Z = \frac{X - \mu}{\sigma}$ has a standard normal distribution.

That is, Z is $N(0,1)$. So, if we take any normal random variable, subtract its mean and then divide by its standard deviation, the resulting random variable will have standard normal distribution. We are going to use this fact to calculate (non-standard) normal probabilities.

Calculating probabilities

With reference to the problem of IQ score, suppose we want to calculate the probability that an individual's IQ score is less than 85, i.e. P[X<85]. The corresponding area under the pdf $N(100,15^2)$ is shown in *Figure 6*.

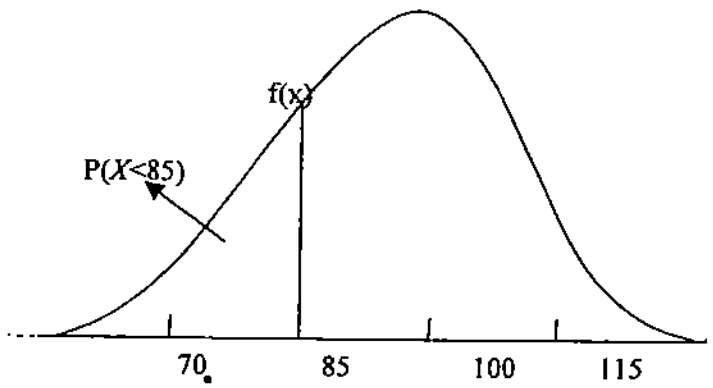


Figure 6: Area under the pdf $N(100,15^2)$

We cannot use normal tables directly because these give $N(0,1)$ probabilities. Instead, we will convert the statement $X<85$ into an equivalent statement which involves the standardised score, $Z = \frac{X - 100}{15}$ because we know it has a standard normal distribution.

We start with $X=85$. To turn X into Z we must standardise the X, but to ensure that we preserve the meaning of the statement we must treat the other side of the inequality in

exactly the same way. (Otherwise we will end up calculating the probability of another statement, not $X<85$). 'Standardising' both sides gives, $\frac{X-100}{15} < \frac{85-100}{15}$.

The left hand side is now a standard normal random variable and so we can call it Z, and we have,

$$Z < \frac{85-100}{15}$$

which is

$$Z < -1.$$

So, we have established that the statement we started with, $X < 85$ is equivalent to $Z < -1$. This means that whenever an IQ score, X is less than 85 the corresponding standardised score, Z will be less than -1 and so the probability we are seeking, P[X<85] is the same P[$Z < -1$].

P[$Z < -1$] is just a standard normal probability and so we can look it up in given Table at end of block/unit, in the usual way, which gives 0.1587. We get that P[$X < 85$] = 0.1587.

This process of rewriting a probability statement about X, in terms of Z, is not difficult if you are systematically writing down what you are doing at each stage. We would lay out the working we have just done for P[$X < 85$] as follows:

X has a normal distribution with mean 100 and standard deviation 15. Let us find the probability that X is less than 85.

$$P[X < 85] = P\left[\frac{X-100}{15} < \frac{85-100}{15}\right]$$

$$= P[Z - 1] = 0.1587$$

Let us do some problems now.

Example 9: For each of these write down the equivalent standard normal probability.

a) The number of people who visit a historical monument in a week is normally distributed with a mean of 10,500 and a standard deviation of 600. Consider the probability that fewer than 9000 people visit in a week.

b) The number of cheques processed by a bank each day is normally distributed with a mean of 30,100 and a standard deviation of 2450. Consider the probability that the bank processes more than 32,000 cheques in a day.

Solution: Here, we want to find the standard normal probability corresponding to the probability P[X < 9000].

a) We have $P[X < 9000] = P\left[\frac{X-10500}{600} < \frac{9000-10500}{600}\right] = P[Z < -2.5]$.

b) Here, we want to find the standard normal probability corresponding to the probability P[X > 32000].

$$P[X < 32000] = P\left[\frac{X-30100}{2450} < \frac{32000-30100}{2450}\right] = P[Z < -0.78]$$

Note: Probabilities like P[a < X < b] can be calculated in the same way. The only difference is that when X is standardised, similar operations must be applied to both a and b i.e., $a < X < b$ becomes,

$$\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}$$

which is

$$\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}$$

Example 10: An individual's IQ score has a $N(100, 15^2)$ distribution. Find the probability that an individual's IQ score is between 91 and 121.

Solution: We require P[91 < X < 121]. Standardising gives

$$P\left[\frac{91-100}{15} < \frac{X-100}{15} < \frac{121-100}{15}\right]$$

The middle term is standardised normal random variable and so we have,

$$P\left[\frac{-9}{15} < Z < \frac{21}{15}\right] = P[-0.6 < Z < 1.4] = 0.9192 - 0.2743 = 0.6449.$$

☛ Check Your Progress 6

1) If a random variable has the standard normal distribution, find the probability that it will take on a value

   a) Less than 1.50
   b) Less than −1.20
   c) Greater than −1.75

   .....

   .....

   .....

2) A filling machine is set to pour 952 ml of oil into bottles. The amount to fill are normally distributed with a mean of 952 ml and a standard deviation of 4 ml. Use the standard normal table to find the probability that the bottle contains oil between 952 and 956 ml?

   .....

   .....

   .....

1.4.4 Chi-Square Distribution

In the last sub-section, we discussed normal distribution. The chi-square distribution can be obtained from the normal distribution as follows. Suppose $Z_1,\ldots,Z_n$ are n independent identically distributed normal random variables with parameters 0 and 1, then $Z_1^2 + \ldots + Z_n^2$ is said to have chi-square distribution with n degrees of freedom. The degrees of freedom here basically indicates the number of independent components which constitute the chi-square distribution. It has received a great deal of attention because of its appearance in the constructing analysis of variable tables,

contingency tables and for obtaining the critical values of different testing procedure. Now, we formally provide the p.d.f of a chi-square random variable with n degrees of freedom.

If the random variable X has chi-square distribution with n-degrees of freedom, then the p.d.f. of X is $f(x)$

$$f(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{(n/2)-1} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

here $\Gamma(.)$ is a gamma function and it is defined as

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

Although, the p.d.f of chi-square random variable is not a very nice looking one, do not bother about that. Keep in mind that the shapes of density functions are always skewed. In this case also if we want to compute $P(a < X < b)$ for any a, b and n, explicitly it will not be possible to do so. Numerical integration is needed to compute this probability. Even for chi-square distribution extensive tables of $P(a < X < b)$ are available for different values of a, b and n.

Note: We have a standard table corresponding to Chi-Square Distribution vary often you may need to refer to the values from the Table given at end of block/unit. So, the same is given at the end, the method of usage is similar to that discussed under Normal distribution.

Example 11: Show that the moment generating function of a random variable X which is chi-square distributed with v degrees of freedom is $M(t) = (1 - 2t)^{-v/2}$.

Solution: $M(t) = E(e^{tx}) = \dfrac{1}{2^{v/2}\Gamma(v/2)} \int_0^\infty e^{tx} x^{(v-2)/2} e^{-x/2} dx$

$$= \frac{1}{2^{v/2}\Gamma(v/2)} \int_0^\infty x^{(v-2)/2} e^{-x(1-2t)/2} dx$$

Letting $(1-2t)^{x/2} = u$ in the last integral we find

$$M(t) = \frac{1}{2^{v/2}\Gamma(v/2)} \int_0^\infty \left(\frac{2u}{1-2t}\right)^{(v-2)/2} e^{-u} \frac{2du}{1-2t}$$

$$= \frac{(1-2t)^{-v/2}}{\Gamma(v/2)} \int_0^\infty u^{(v/2)-1} e^{-u} du = (1-2t)^{-v/2}$$

☞ Check Your Progress 7

1) Let X_1 and X_2 be independent random variables, which are chi-square distributed with v_1 and v_2 degrees of freedom respectively. Show that the moment generating function of $Z = X_1 + X_2$ is $(1-2t)^{-v(v_1+v_2)/2}$

.....

.....

.....

.....

.....

2) Find the values of x^2 for which the area of the right-hand tail of the x^2 distribution is 0.05, if the number of degrees of freedom ν is equal to (a) 15, (b) 21, (c) 50.

.....

.....

.....

.....

.....

1.5 SUMMARY

In this unit we have covered the following points:

a) A random variable is a variable that takes different values according to the chance outcome

b) Types of random variables: Discrete and Continuous

c) Probability distribution gives the probabilities with which the random variables takes various values in their range

d) Discussed probability distributions:

a. Binomial Distribution: The probability of an event P[X=r] in this distribution is given by

$$P(X = r) = C(n,r)p^r(1-p)^{n-r} = C(n,r)p^rq^{n-r},$$

b. Poisson Distribution: The probability of an event P[X=i] in this distribution is given by

$$P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!}, \qquad i = 0,1,2,3,...$$

c. Uniform Distribution: The probability density function is defined by

$$f(x) = \begin{cases} \frac{1}{B-A} & for \quad A < x < B \\ 0 & \text{otherwise} \end{cases}$$

d. Normal Distribution: The probability for this distribution is determined by calculating the area under the curve of probability density function defined by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}, \quad where \quad -\infty < x < \infty$$

e. Chi-Square Distribution: If the random variable X has chi-square distribution with n-degrees of freedom, then the probability density function of X is given by

$$f(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{(n/2)-1} & \text{if } x > 0 \\ 0 & otherwise \end{cases}$$

here $\Gamma(.)$ is a gamma function and it is defined as

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$$

f. Mathematical expectation or Expected values or Expectations

$E(X)$ is defined as $E(X) = x_0p_0 + x_1p_1 + x_2p_2 + \ldots\ldots\ldots = \Sigma x_ip_i$ when all probabilities are equal then $E(X)$ = mean of $x_0, x_1, x_2 \ldots\ldots$

Similarly for continuous variables X having density function p(x) where $P[X=x] = p(x)$, the Expectation $E(X)$ will be given by integral of $x_ip(x_i)$ w.r.t x .

This concept of Expectation also contributes to the definition of Moment Generating Function of X *i.e* $M_x(t) = E(e^{tx})$.

1.6 SOLUTIONS/ANSWERS

Check Your Progress 1

1) a) If X denotes the number of correct answers, then X is the random variable for this situation
b) X can take the values 0,1,2,3.....up to 50
c) $P[X=10]$ means the probability that the number of correct answers is 10

Check Your Progress 2

1) Case (a) is not discrete where as case (b) and (c) are discrete because in case (a) we are taking values in an interval but in case(b) the number of accident is finite , similarly you argue for case (c)

Check Your Progress 3

1) This situation follows the binomial distribution with n=4 and p=90/100=9/10 The random variable X is the number of seeds that germinate. We have to calculate the probability that exactly two of the four seeds will germinate. That is P[X=2]. By applying Binomial formula , we get

$$P[X=2] = ({}^4C_2) * (9/10)^2 * (1/10)^2$$

$$= 6 * (81/100) * (1/100) = 486/10000 = 0.0486$$

So, the required probability is 0.0486

Check Your Progress 4

1) Here, we are dealing with the problem related to the receipt of bad Cheques, which is an event with rare occurrence over an interval of time (which is a day In this case). So, we can apply Poisson distribution

Average bad Cheques received per day = 6
Thus, by substituting $\lambda = 6$ and x=4 in Poisson formula we get

$$P[X=4] = (6^4 e^{-6})/4! = 0.135$$

Check Your Progress 5

1) Suppose X denotes the time, the fire alarm starts measured in minutes after 10:00 am. Then clearly X is a uniform random variable between (0,180). If we want to calculate the probability that fire alarm will ring before noon, then

$$P(X \leq 12:00 \text{ noon}) = \frac{(12-10)\times 60}{180} = \frac{2}{3}.$$

Check Your Progress 6

1) a) 0.9332
   b) 0.1151
   c) 0.9599

2) The standard normal probability corresponding to this probability is given by

$$\begin{aligned} P[952 < Z < 956] &= P[((952-952)/4) < ((X-952)/4) < ((952-956)/4)] \\ &= P[0 < Z < 1] \\ &= F(1) - F(0) \\ &= 0.8413 - 0.5 = 0.343 \end{aligned}$$

Check Your Progress 7

1) The moment generating function of $Z = X_1 + X_2$ is
$M(t) = E\left[e^{t(X_1+X_2)}\right] = E(e^{tX_1})E(e^{tX_2}) = (1-2t)^{-v_1/2}(1-2t)^{-v_2/2} = (1-2t)^{-(v_1+v_2)/2}$
using **Example 9.**

2) Using the table in for Chi Square distribution we find in the column headed $x^2_{.95}$ the values: (a) 25.0 corresponding to $v = 15$; (b) 32.7 corresponding to $v = 21$; (c) 67.5 corresponding to $v = 50$.

UNIT 2 PSEUDO – RANDOM NUMBER GENERATION

Structure **Page Nos.**

2.0 INTRODUCTION

A pseudo-random number generation is the methodology to develop algorithms and programs that can be used in, probability and statistics applications when large quantities of random digits are needed. Most of these programs produce endless strings of single-digit numbers, usually in base 10, known as the decimal system. When large samples of pseudo-random numbers are taken, each of the 10 digits in the set {0,1,2,3,4,5,6,7,8,9} occurs with equal frequency, even though they are not evenly distributed in the sequence.

Many algorithms have been developed in an attempt to produce truly random sequences of numbers, endless strings of digits in which it is theoretically impossible to predict the next digit in the sequence based on the digits upto a given point. But the very existence of the algorithm, no matter how sophisticated, means that the next digit can be predicted! This has given rise to the term pseudo-random for such machine-generated strings of digits. They are equivalent to random-number sequences for most applications, but they are not truly random according to the rigorous definition.

A simulation that has any random aspects at all, must involve sampling or generating random variables from different probability distributions. These distributions are often specified, that is the form of the distribution functions is explicitly known, for example it can be exponential, gamma, normal or Poisson as discussed in Unit 1.

Random number generation has intrigued scientists for several years and a lot of effort has gone into examining the creation of randomness on a deterministic (non-random) machine, that is to design computer algorithms that are able to produce 'random' sequences of integers. This is not a trivial task. Such algorithms are called generators and all generators have flaws because all of them construct the n-th number in the sequence as a function of the $(n-1)$ – th number, initialised with a non-random seed value. Numerous techniques have been invented over the years that measure just how random a sequence is, and the most well known generator, have been subjected to rigorous testing. The mathematical tools that are required to design such an algorithm are largely number theoretic and combinatorial in nature. These tools differ drastically from those needed to generate sequences of integers with certain non-uniform distributions given that a perfect uniform random number generator is available.

The methodology of generating random numbers has a long and interesting history. The earliest methods were essentially carried out by hand, such as casting lots, throwing dice, dealing out cards or drawing numbered balls from a well-stirred urn. Many lotteries still operate in this fashion. In the early twentieth century, statisticians

joined gamblers in generating random numbers and mechanised devices were built to generate random numbers more quickly. Sometime later, electric circuits based on randomly pulsating vacuum tubes were developed that delivered random digits at rates up to 50 per second. One such random number generator machine the Electronic Random Number Indicator Equipment (ERNIE) was used by the British General Post Office to pick the winners in the Premium Savings Bond lottery. Another electronic device was used by the Rand Corporation to generate a table of million random digits. In India, Indian Statistical Institute has published a book with a collection of a million random numbers in the mid twentieth century which was used for sample survey planning.

As computers and simulation became more widely used, attention was paid to methods of random number generation compatible with the way computers work. A good uniform between 0 and 1, random generator should possess certain properties as listed below:

- Above all, the numbers produced should appear to be distributed uniformly on [0, 1] and should not exhibit any correlation with each other; otherwise, the simulation's results may be completely invalid.
- From a practical point of view, a generator should be fast and avoid the need for a lot of storage.
- We should be able to produce a given stream of random numbers from a given initial (seed) value for at least two reasons. First, this can sometimes make debugging or verification of the computer program easier or we might want to use identical random numbers in simulating different systems in order to obtain a more precise comparison.

In this unit, we will describe how to generate $U(0,1)$ (uniform between 0 and 1, see unit 1 for the actual definition) in a computer. Once we have a $U(0, 1)$ random number, we will use that to generate several other deviates from different discrete and continuous distributions.

2.1 OBJECTIVES

After going through this unit, you should be able to:

- how to generate $U(0, 1)$ random number in a computer;
- how to generate random deviates from any discrete distribution, and
- how to generate random numbers from many continuous distributions like, exponential, Weibull, gamma, normal, chi-square etc.

2.2 UNIFORM RANDOM NUMBER GENERATORS

You may recall that we had mentioned in the previous section that we need the uniform random number of generator for generating random numbers from any other distributions. Therefore, it is very important to have a very good uniform random number generator. There are several methods available for generating uniform random numbers. But currently the most popular one is the linear congruential generator (LCG). Most of the existing software's today use this LCGs proposed by **Lehmer** in the early 50's. The LCGs provides an algorithm that generates uniform random number between (0,1). It can be simply described as follows.

A sequence of integers $Z_1, Z_2, \ldots$ is defined by the recursive formula

$$Z_i = (aZ_{i-1} + c) \pmod{m}, \qquad (1)$$

where m, the modulus, a, the multiplier, c, the increment and Z_0, the seed or the initial value, are all non-negative integers. Those who are not familiar with the definition of modules, note that for non-negative integers, x, y and z, $x = y \pmod{z}$ means x is the remainder when the integer y divides the integer z. For example if $y = 10$ and $z = 3$, then $x = 1$, or if $y = 10$ and $z = 2$, then $x = 0$. Therefore, from (1) it is clear that to obtain Z_i, first divide $aZ_{i-1} + c$ by m and Z_i is the corresponding remainder of this division. It is clear that $1 \le Z_i \le m - 1$ and to obtain the desired random numbers U_i, for $i = 1, 2....$ On [0, 1], we let $U_i = \frac{Zi}{m}$. The choice of the non-negative integers, a, c and m are the most crucial steps, and they come from the theoretical considerations. In addition to non-negatively, they also should satisfy $0 < m$, $a < m$ and $c < m$. Moreover, the initial value $Z_0 < m$.

Immediately, two objections could be raised against LCGs. The first objection is one which is common to all random number generators, namely, the Z_i's defined by (1) are not really random. It can be easily seen that for $i = 1, 2, ...$,

$$Zi = \left[a^i Z_0 + \frac{c(a^i - 1)}{a - 1}\right] \pmod{m},$$

so that every Z_i is completely determined by m, a, c and Z_0. However, by careful choice of these four parameters the aim is to induce a behavior in the Z_i's that makes the corresponding U_i's appear to be independent identically distributed $U(0, 1)$ random variates when subjected to a variety of statistical tests.

The second objection to LCGs might be that the U_i's can take on only the rational numbers $0, \frac{1}{m}, \frac{2}{m}, ..., \frac{(m-1)}{m}$; in fact the U_i's might take only a fraction of these values, depending on the specifications of the four parameters. Therefore, there is no possibility of getting a value of U_i between, say $\frac{0.1}{m}$ and $\frac{0.9}{m}$, whereas this should occur with probability $\frac{0.8}{m} > 0$. We will see later that the modulus m is usually chosen to be very large, say 10^9 or more, so that the points in [0, 1] where U_i's can fall are very dense. This provides an accurate approximation to the true continuous $U(0, 1)$ distribution, sufficient for most purposes.

Let us consider the following example:

Example 1 : Consider the LCG defined by $m = 16$, $a = 5$, $c = 3$ and $Z_0 = 7$. The following table gives Z_i and U_i (up to three decimal places) for $i = 1, ..., 19$. Note that $Z_{17} = Z_1 = 6$, $Z_{18} = Z_2 = 1$ and so on. Therefore, from $i = 17$, the sequence repeats itself. Naturally, we would not seriously suggest to anybody to use this generator. The main reason being that in this case m is too small. We are also presenting this for illustrative purposes.

Table 1: The LCG $Z_i = (5Z_{i-1} + 3) \pmod{16}$ with $Z_0 = 7$

i	Z_i	U_i	i	Z_i	U_i	i	Z_i	U_i	i	Z_i	U_i
0	7	-	5	10	0.625	10	9	0.563	15	4	0.250
1	6	0.375	6	5	0.313	11	0	0.000	16	7	0.438
2	1	0.063	7	12	0.750	12	3	0.188	17	6	0.375
3	8	0.500	8	15	0.938	13	2	0.125	18	1	0.063
4	11	0.688	9	14	0.875	14	13	0.813	19	8	0.500

Note that, the repeating behaviour of LCG is inevitable. By the definition of Z_i, whenever it takes on a value it had taken previously, from that point onward the sequence will repeat itself endlessly. The length of a cycle is called the period of a generator. For LCG, Z_i depends only on the previous value Z_{i-1} and since $0 \le Z_i \le m-1$, it is clear that the period is at most m. If it is m, the LCG is said to have full period. Clearly, if a generator is full period, any choice of the initial seed Z_0 from $\{0,\ldots,m-1\}$ will produce the entire cycle in some other order.

Since for large scale simulation, projects may require hundreds of thousands of random numbers, it is desirable to have LCGs with long periods. Moreover, it is desirable to have full period LCGs, because then it is assured that every integer between 0 and $m-1$ will occur exactly once in every cycle. Thus, it is very important to know how to choose a, m and c so that the corresponding LCG will have full period. We should also keep in mind that obtaining full period is just one desirable property for a good LCG. It should also have good statistical properties, such as apparent independence, computational and storage efficiency and reproducibility. Reproducibility is simple. For reproducibility, we must only remember that the initial seed Z_0 initiates the generator with this value again to obtain the same sequence of U_i exactly. Some of the standard LCGs with different values of a, m and c are presented below. These LCG have been observed to perform very well in several machines and passed the standard statistical tests also.

Generator 1: $a = 16807$, $m = 2^{31}-1$, $c = 0$.
Generator 2: $a = 1664525$, $m = 2^{32}$, $c = 1664525$.

Fortunately today, most of the simulation packages and even simple scientific calculators have reliable $U(0\ 1,)$ generator available.

☞ Check Your Progress 1

1) What do you mean by Pseudo-random number generation? What is the practical advantage of the concept of random number generation? Do you know any algorithm which works in designing the software for Random number generation?

.....

.....

.....

.....

.....

2.3 GENERATING RANDOM VARIATES FROM ARBITRARY DISTRIBUTIONS

A simulation that has any random aspects at all must involve generating random variates from different distributions. We usually use the phrase generating a random variate to refer to the activity of obtaining an observation or a realisation on a random variable from the desired distribution. These distributions are often specified as a result of fitting some appropriate distributional form. They are often specified in advance, for example exponential, gamma or Poisson etc. In this section, we assume that the distributional form has already been specified including the values of the parameters and we address the issue of how to generate random variate from this specified distribution.

We will see in this section that the basic ingredient needed for every method of generating random variates from any distribution is a source of independent identically distributed $U(0, 1)$ random variates. Therefore, it is essential that a statistically reliable $U(0, 1)$ generator is available for generating random deviate correctly from any other distribution. Therefore, from now on, we assume that we have a reliable sequence of U(0,1) variates available to us.

There are several issues, which should be kept in mind before using any particular generator. The most important issue is of course the exactness. It is important that one should use an algorithm that results in random variates with exactly the desired distribution, within the unavoidable external limitations of machine accuracy and the exactness of $U(0,1)$ random number generator. The second important issue is efficiency. Given that we have several choices, we should choose that algorithm which is efficient in terms of both storage space and execution time. Now, we provide some of the most popular and standard techniques to generator non-uniform random deviates, it may be both discrete or continuous as well as even mixed distribution.

2.4 INVERSE TRANSFORM

Suppose, we want to generate a random variate X that has a continuous and strictly increasing distribution function F, when $0 < F(x) < 1$, i.e., whenever $x_1 < x_2$ then $F(x_1) < F(x_2)$. We draw a curve of F in the *Figure 1*. Let F^{-1} denote the inverse of the function F. Then an algorithm for generating a random variate X having distribution function F is as follow.

Algorithm

- Step 1: Generate U_I from $U_I(0,1)$
- Step 2: Return $X_I = F^{-1}(U_I)$.

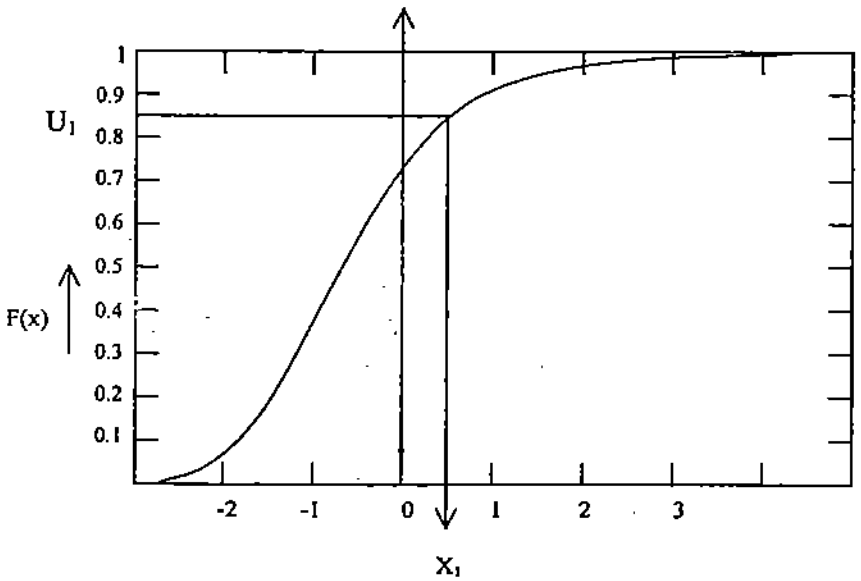


Figure 1: Distribution function of a continuous a random variable.

Note that when F is a strictly increasing function, $F^{-1}(U)$ will be always defined, since $0 \leq U \leq 1$ and the range of F is $[0, 1]$. *Figure 1* illustrates the algorithm graphically. According to the figure it is clear that the uniform random variable U_1

results the random variable X_1 with the distribution function F. To show that the value X_1 returned by the above algorithm has the desired distribution function F, note that

$$P(X_1 \le x) = P(F^{-1}(U_1) \le x) = P(U_1 \le F(x)) = F(x).$$

The first equality follows from the definition of X_1, the second equality follows because F is invertible and the third equality follows because U_1 follows $U(0,1)$.

Example 2: Let X have the Weibull distribution with the following probability density function:

$$f(x) = \begin{cases} \alpha\lambda e^{-\lambda x^{\alpha}} x^{\alpha-1} & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x < 0 \end{cases} \quad \text{Find } F^{-1}$$

Solution: Here α and λ both are known constants and both of them are strictly greater than 0.

Therefore, X has the distribution function

$$F(x) = \begin{cases} 1 - e^{-\lambda x^{\alpha}} & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \le 0 \end{cases}$$

Therefore, to compute $F^{-1}(u)$, let us equate $u = F(x)$ and we solve for x to obtain

$$F^{-1}(u) = \left[\frac{1}{\lambda}\{-\ln(1-u)\}\right]^{\frac{1}{\alpha}}$$

Therefore, to generate X from a Weibull distribution with $\alpha = 2$ and $\lambda = 1$, generate U from $U(0, 1)$ and then set

$$X = [\{-\ln(1-u)\}]^{\frac{1}{2}}$$

In this case also as before, it is possible to replace U by $1 - U$, therefore we can use

$$X = [\{-\ln u\}]^{\frac{1}{2}}, \quad \text{to avoid one subtraction.}$$

The inverse-transform method can be used also when the random variable X is discrete. In this case the distribution function is

$$F(x) = P(X \le x) = \sum_{x_i \le x} p(x_i),$$

Where $p(x_i)$ is the probability mass function of X, *i.e.*,

$$p(x_i) = P(X = x_i).$$

We can assume that X can take values $x_1, x_2, \ldots$, where $x_1 < x_2 < \ldots$. Then the algorithm is as follows:

Algorithm

- Step 1: Generate U from $U(0,1)$
- Step 2: Determine the smallest positive integer I such that $U \le F(x_I)$ and return $X = x_I$. The *Figure 2* illustrates the method. In that case we generate $X = x_4$

Now to show that the discrete inverse transform method is valid, we need to show that $P(X = x_i) = p(x_i)$ for all $i = 1$. we get $X = x_1$, if and only if $U \le F(x_1) = p(x_1)$, since x_i's are in the increasing order and U follows $U(0,1)$. For $i \ge 2$, the algorithm sets $X = x_i$ if and only if $F(x_{i-1}) < U \le F(x_i)$, since the i chosen by the algorithm is the smallest positive integer such that $U \le F(x_i)$. Further, since U follows $U(0,1)$ and $0 \le F(x_{i-1}) < F(x_i) \le 1$,

$$P(X = x_i) = P\{F(x_{i-1} < U \le F(x_i)\} = F(x_i) - F(x_{i-1}) = p(x_i).$$

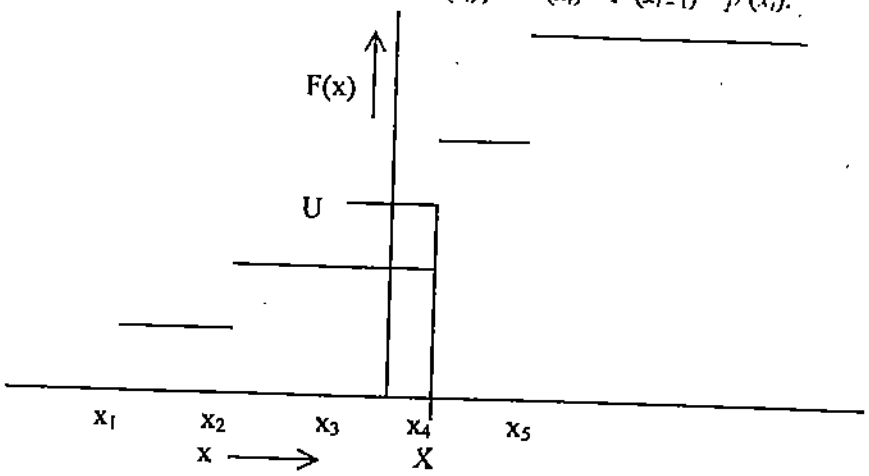


Figure 2: Distribution function of a discrete random variable

Now consider the following example.

Example 3: Suppose we want to generate a random sample from the following discrete probability distribution.

$P(X=1) = \frac{1}{2}$, $P(X=2) = \frac{1}{4}$, $P(X=4) = \frac{1}{4}$. Generate a random sample from X?

Solution: The distribution function of the random variable X is

$$F(x) = \begin{cases} 0 & if \quad x < 1 \\ \frac{1}{2} & if \quad 1 \le x < 2 \\ \frac{3}{4} & if \quad 2 \le x < 4 \\ 1 & if \quad x \ge 4 \end{cases}$$

The distribution function X is presented in *Figure 3*. If we want to generate a random sample from X, first generate a random variable U from $U(0, 1)$. If $U \le \frac{1}{2}$, then assign $X = 1$. If $\frac{1}{2} < U \le \frac{3}{4}$, then assign $X = 2$, otherwise assign $X = 4$.

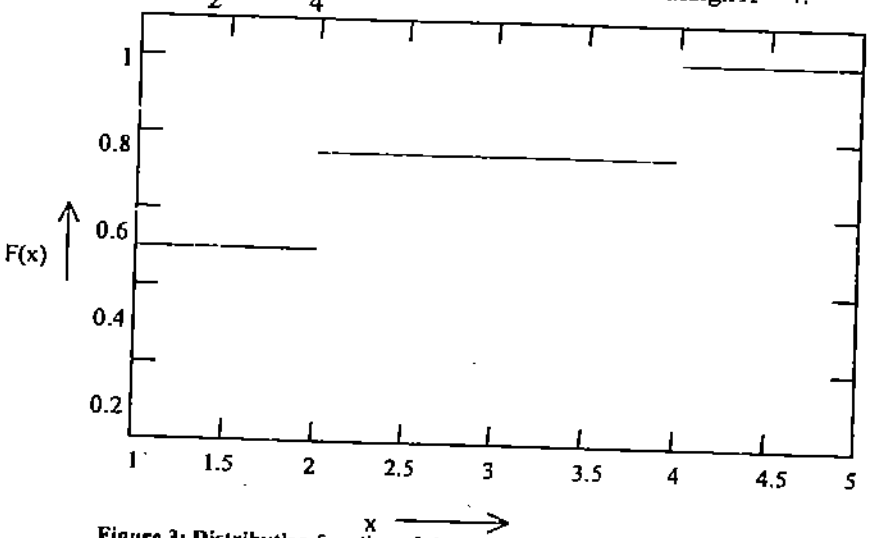


Figure 3: Distribution function of the random variable X of example 3.

☞ **Check Your Progress 2**

1) Let X have the exponential distribution with mean 1. The distribution function is

$$F(x) = \begin{cases} 1-e^{-x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$ Find F^{-1}

.....

.....

.....

2) Consider another discrete random variable which may take infinitely many values. Suppose the random variable X has the following probability mass function.

$P(X = i) = p_i = \frac{1}{2^i}$, i = 1, 2, 3, ... Generate a random sample from X?

.....

.....

.....

2.5 ACCEPTANCE-REJECTION METHOD

In the last section, we have discussed the inverse transformation method to generate random number from different non-uniform distributions. Note that apparently, the inverse transformation method seems to be the most general method for generating random deviates from any distribution function functions. In fact, it can be used provided the distribution function can be written in an explicit form, or more precisely the inverse of the distribution function can be computed analytically. For example, in case of exponential, Weibull, Cauchy distributions the distribution function and also their inverses can be constructed analytically. Unfortunately that is not the case with the general method. Suppose the random variable X follows gamma with the shape and scale parameters as α and λ respectively. Then the density function of X, say $f(x|\alpha, \lambda)$, is

$$f(x|\alpha, \lambda) = \begin{cases} \frac{\lambda^\alpha}{\Gamma\alpha} x^{\alpha-1} e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

The distribution function X, say $F(x|\alpha,\lambda) = \int_0^x f(y|\alpha,\lambda)dy$ cannot be expressed in explicit form. Therefore, $F^{-1}(x|\alpha,\lambda)$ also cannot be calculated explicitly. Exactly the same problem arises if X is a normal random variable. Suppose X is a normal random variable with mean μ and variance σ^2, then the probability density function of X, say $f(x|\mu, \sigma)$ is

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-\mu)^2}{2\sigma^2}}, \text{ for } -\infty < x < \infty.$$

In this case also, the distribution function cannot be computed analytically and similarly it's inverse. Therefore, in these cases, we cannot apply the inverse transformation method to generate the corresponding random deviates. The acceptance-rejection method can be used quite effectively to generate these random deviates. It can be described as follows.

Supnose we have a density function $f(x)$ and we want to generate a random deviate from the density function $f(x)$. The distribution function of $f(x)$ can not be expressed in explicit form. The acceptance-rejection method requires that we specify a function $g(x)$ that majorises the function $f(x)$, that is, $f(x) \leq g(x)$ for all x. Naturally, $g(x)$ will not be a density function always, since

$$c = \int_{-\infty}^{\infty} g(x)dx \geq \int_{-\infty}^{\infty} f(x) = 1,$$

but the function $h(x) = \frac{1}{c} g(x)$ is clearly a density function provided $c < \infty$. Now for any given $f(x)$, we choose the function $g(x)$, such that $c < \infty$ and it is possible to generate random deviate from $g(x)$ by a inverse transformation method. Then we can generate the random deviate X from $f(x)$ as follows:

Algorithm

- Step 1: Generate Y with density function $g(x)$.
- Step 2: Generate U from $U(0,1)$ which is independent of Y.
- Step 3: If $U \leq \frac{f(Y)}{g(Y)}$, $X = Y$, otherwise go back to Step 1 and try again.

Note that the algorithm is looping back to Step 1 until we generate a pair (Y, U) pairs in Steps 1 and 2 for which $U \leq \frac{f(Y)}{g(Y)}$, when we accept the value Y for X.

Theoretically, it can be shown that the random deviate X generated by the above algorithm has indeed the probability density function $f(x)$. Since it is not very easy to prove the result we do not provide it.

Example 4: Consider the following density function

$$f(x) = \begin{cases} 60x^3(1-x)^2 & if \quad 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

generate random deviates from given f(x) by using Acceptance Rejection method.

Solution: In this case the distribution function, it is presented in Figure 4, of $f(x)$ is

$$F(x) = \begin{cases} 0 & if \quad x < 0 \\ 10x^6 + 15x^4 - 24x^5 & if \quad 0 < x < 1 \\ 1 & if \quad x > 1 \end{cases}$$

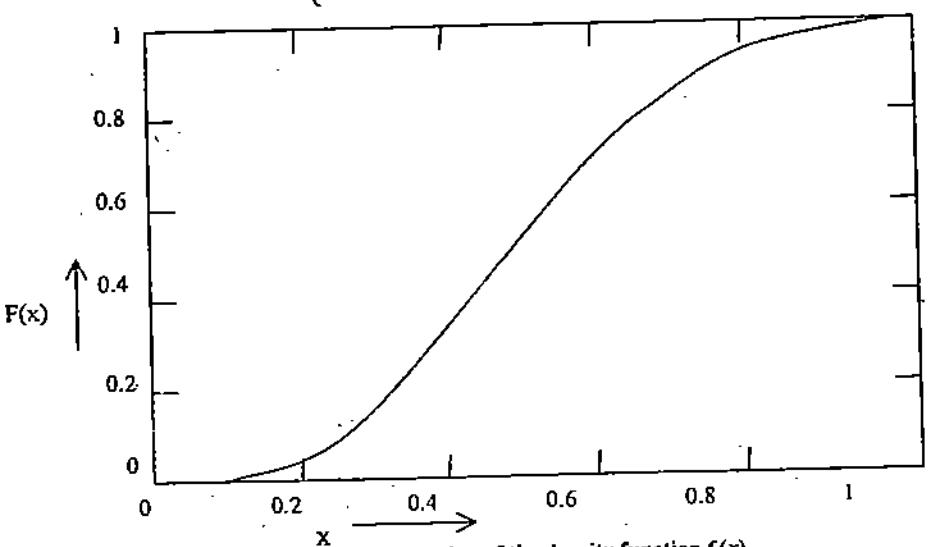


Figure 4 : Distribution function of the density function $f(x)$

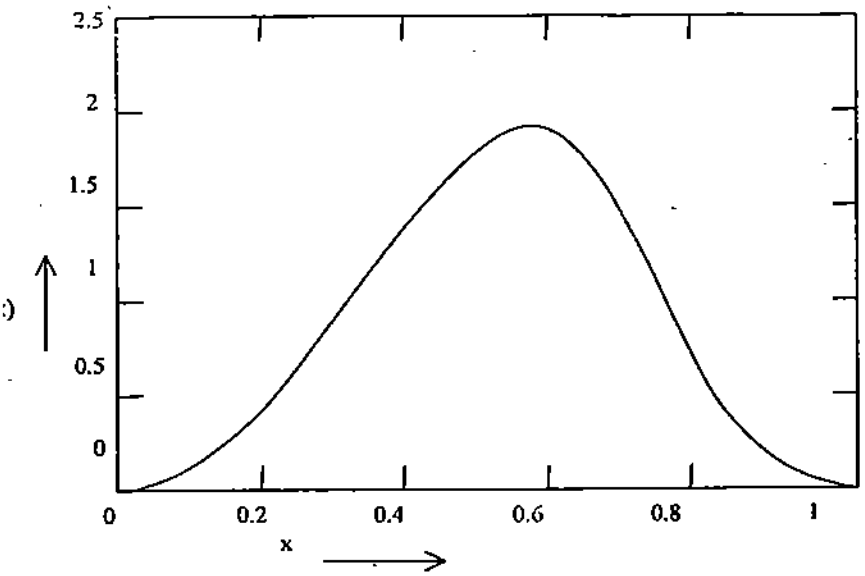


Figure 5: Density function $f(x)$

From the distribution function $F(x)$ is provided in *Figure 4*. It is clear that the distribution function $F(x)$ is a strictly increasing function of x in [0, 1]. Therefore $F^{-1}(x)$ exists, but unfortunately to find $F^{-1}(x)$ we need to solve a six degree polynomial, which cannot be obtained analytically. We need to solve this numerically only. Therefore, we cannot generate random deviate from the density function $f(x)$ using the inversion method. But we will be able to generate random deviate from $f(x)$ using the acceptance-rejection method.

First, let us look at the graph of the density function $f(x)$. It is presented in *Figure 5*. From *Figure 5* it is clear that $f(x)$ is an unimodal function with a unique maximum. The maximum can be easily obtained by the standard differential calculus, that is by setting $\frac{df(x)}{dx} = 0$. We see that the maximum of $f(x)$ occurs at $x = 0.6$ and the maximum value at 0.6, that is $f(0.6) = 2.0736$. Therefore, if we define

$$g(x) = \begin{cases} 2.0736 & if \quad 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

then clearly $f(x) \leq g(x)$ for all x. Now to calculate $h(x)$, first we need to calculate c. Since,

$c = \int_0^1 2.0735 = 2.0736$, therefore,

$$h(x) = \begin{cases} \frac{2.0736}{c} = 1 & if \quad 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

It is immediate that $h(x)$ is just the U(0,1) density function. Now the algorithm takes the following form.

Algorithm

- Step 1: Generate Y from $U(0, 1)$
- Step 2: Generate U from $U(0,1)$ which is independent of Y.

- Step 3: If $U \leq \frac{60Y^3(1-Y)^2}{2.0736}$ then return with $X = Y$, otherwise go back to Step 1.

In this case X has the desired density function $f(x)$.

☞ Check Your Progress 3

1) Use acceptance-rejection method to generate a random deviate from gamma density function. The gamma density function with the shape parameter α can be written as

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} & \text{if} \quad x > 0 \\ 0 & \text{otherwise} \end{cases}$$

.....

.....

.....

2.6 SUMMARY

In this unit, we have discussed the meaning of pseudo random number generation and along with that we have described uniform random number generator and arbitrary random number generator. Uniform random number generator we emphasised on **LCG (Linear Congruential Generator)** and some objections related to LCG. Actually LCGs provides an algorithm how to generate uniform random number between (0,1). It can be simply described as follows.

A sequence of integers $Z_1, Z_2, \ldots$ is defined by the recursive formula

$$Z_i = (aZ_{i-1} + c) \pmod{m}, \qquad (1)$$

where m, the modulus, a, the multiplier, c, the increment and Z_0, the seed or the initial value, are all non-negative integers

Then we have discussed the concept, algorithm and application of **Inverse transforms** for random number generation, in brief. Suppose, we want to generate a random variate X that has a continuous and strictly increasing distribution function F, when $0 < F(x) < 1$, i.e., whenever $x_1 < x_2$ then $F(x_1) < F(x_2)$. Let F^{-1} denote the inverse of the function F. Then an algorithm for generating a random variate of X with distribution function F is as follows:

Algorithm

- Step 1: Generate U_i from U_i (0,1)
- Step 2: Return $X_i = F^{-1}(U_i)$.

The inverse-transform method can be used when the random variable X is discrete.

In this case the distribution function is

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i),$$

Where $p(x_i)$ is the probability mass function of X, *i.e.*,

$$p(x_i) = P(X = x_i).$$

We can assume that X can take values $x_1, x_2, \ldots$, where $x_1 < x_2 < \ldots$. Then the algorithm is as follows:

Algorithm

- Step 1: Generate U from $U(0,1)$
- Step 2: Determine the smallest positive integer I such that $U \le F(x_I)$ and return $X = x_I$.

Note: The inverse transformation method to generate random number from different non-uniform distributions. Note that apparently, the inverse transformation method seems to be the most general method to generate random deviate from any distribution function. In fact, it can be used provided the distribution function can be written in an explicit form.

Finally, we had the **Acceptance Rejection method** of random number generation, this method is quite important especially, when the distribution function cannot be computed analytically and similarly it's inverse also. Then in such cases, we cannot apply the inverse transformation method to generate the corresponding random deviates. The acceptance-rejection method can be used quite effectively to generate these random deviates.

In brief, suppose we have a density function $f(x)$ and we want to generate a random deviate from the density function $f(x)$. The distribution function of $f(x)$ cannot be expressed in explicit form. The acceptance-rejection method requires that we specify a function $g(x)$ that majorises the function $f(x)$, that is, $f(x) \le g(x)$ for all x. Naturally, $g(x)$ will not be a density function always, since

$$c = \int_{-\infty}^{\infty} g(x)dx \ge \int_{-\infty}^{\infty} f(x) = 1,$$

but the function $h(x) = \frac{1}{c} g(x)$ is clearly a density function provided $c < \infty$. Now for any given $f(x)$, we choose the function $g(x)$, such that $c < \infty$ and it is possible to generate random deviate from $g(x)$ by a inverse transformation method. Then we can generate the random deviate X from $f(x)$ as follows:

Algorithm

- Step 1: Generate Y having density function $g(x)$.
- Step 2: Generate U from $U(0,1)$ which is independent of Y.
- Step 3: If $U \le \frac{f(Y)}{g(Y)}$, $X = Y$, otherwise go back to Step 1 and try again.

Note that the algorithm is looping back to Step 1 until we generate a pair (Y, U) pairs in Steps 1 and 2 for which $U \le \frac{f(Y)}{g(Y)}$, when we accept the value Y for X.

Theoretically, it can be shown that the random deviate Y generated by the above algorithm has indeed the probability density function $f(x)$. Since, it is not very easy to prove the result we do not provide it.

2.7 SOLUTIONS/ANSWERS

Check Your Progress 1

1) A pseudo-random number generation is the methodology to develop algorithms and programs that can be used in, probability and statistics applications when large quantities of random digits are needed. Most of these programs produce endless strings of single-digit numbers, usually in base 10, known as the decimal system. When large samples of pseudo-random numbers are taken, each of the 10 digits in the set {0,1,2,3,4,5,6,7,8,9} occurs with equal frequency, even though they are not evenly distributed in the sequence.

   Many algorithms have been developed in an attempt to produce truly random sequences of numbers, endless strings of digits in which it is theoretically impossible to predict the next digit in the sequence based on the digits up to a given point. But the very existence of the algorithm, no matter how sophisticated, means that the next digit can be predicted! This has given rise to the term pseudo-random for such machine-generated strings of digits. They are equivalent to random-number sequences for most applications, but they are not truly random according to the rigorous definition.

   There are several methods available to generate uniform random numbers. But currently the most popular one is the linear congruential generator (LCG). Most of the existing software's today use this LCGs proposed by Lehmer in the early 50's.

Check Your Progress 2

1) To find F^{-1}, we set $u = F(x)$ and solve for x to obtain

$$x = F^{-1}(u) = -\ln(1-u).$$

   Therefore, to generate random variate X from exponential distribution with mean 1, first generate U from a $U(0,1)$ and then let $X = -\ln(1-U)$. Therefore X will have exponential distribution with mean 1. Since U and $1-U$ have the same $U(0,1)$ distribution, we can also use $X = \ln U$. This saves a subtraction.

2) Note that $\sum_{i=1}^{\infty} \frac{1}{2^i} = 1$, therefore p_i denotes the probability mass function of a discrete random variable. The corresponding distribution function of the discrete random variable X can be written as

$$F(x) = \begin{cases} 0 & if \quad x < 1 \\ \sum_{i=1}^{m} \frac{1}{2^i} & if \quad m \le x < m+1, \end{cases}$$

   where m is any positive integer. Now to generate a random deviate from the random variable X, first draw a random sample U from $U(0,1)$, since $0 \le U \le 1$, there exits a positive integer m such that $\sum_{i=1}^{m-1} \frac{1}{2^i} \le U < \sum_{i=1}^{m} \frac{1}{2^i}$, where $\sum_{i=1}^{0} \frac{1}{2^i} = 0$, then $X = m$.

Check Your Progress 3

1) Our problem is to generate random deviate from $f(x)$ for a given $0 < \alpha < 1$. Note that, we cannot use the acceptance-rejection method in this case. It is easily observed if we take

$$g(x) = \begin{cases} 0 & \text{if} \quad x \le 0 \\ \dfrac{x^{\alpha-1}}{\Gamma(\alpha)} & \text{if} \quad 0 < x < 1 \\ \dfrac{e^{-x}}{\Gamma(\alpha)} & \text{if} \quad x > 1, \end{cases}$$

then $f(x) \le g(x)$ for all x. In this case

$$c = \int_0^\infty g(x) = \int_0^1 \frac{x^{\alpha-1}}{\Gamma(\alpha)} dx + \int_1^\infty \frac{e^{-x}}{\Gamma(\alpha)} dx = \frac{1}{\Gamma(\alpha)}\left[\frac{(e+a)}{ae}\right].$$

Therefore, $h(x) = \dfrac{1}{c} g(x)$ is

$$h(x) = \begin{cases} 0 & \text{if} \quad x \le 0 \\ \dfrac{\alpha x^{\alpha-1}}{b} & \text{if} \quad 0 \le x \le 1 \\ \dfrac{\alpha e^{-x}}{b} & \text{if} \quad x > 1, \end{cases}$$

where $b = \dfrac{e+a}{e}$. The distribution function $H(x)$ corresponds to the density function $h(x)$ is

$$H(x) = \int_0^x h(y)\, dy = \begin{cases} \dfrac{x^\alpha}{b} & \text{if} \quad 0 \le x \le 1 \\ 1 - \dfrac{\alpha e^{-x}}{b} & \text{if} \quad x > 1, \end{cases}$$

Which can be easily inverted as

$$H^{-1}(u) = \begin{cases} (bu)^{\frac{1}{\alpha}} & \text{if} \quad u \le \dfrac{1}{b} \\ -\ln \dfrac{b(1-u)}{\alpha} & \text{if} \quad u > \dfrac{1}{b} \end{cases}$$

Therefore, it is possible to generate random deviate from the density function $h(x)$ using the simple inversion method. Generation of a random deviate Y from the density function $h(x)$ can be performed as follows. First generate U_1 from $U(0, 1)$, if $U_1 \le \dfrac{1}{b}$ we set $Y = (bU_1)^{\frac{1}{\alpha}}$, in this case $Y \le 1$. Otherwise $Y = -\ln \dfrac{b(1-U_1)}{\alpha}$ and in this case $Y > 1$.

Also note that

$$\frac{f(Y)}{g(Y)} = \begin{cases} e^{-Y} & if \quad 0 \le Y \le 1 \\ Y^{\alpha-1} & if \quad Y > 1 \end{cases}$$

Now, the algorithm to generate random deviate from a gamma density function with the shape parameter α, for $0 < \alpha < 1$ takes the following form:

- Step 1: Generate U_1 from $U(0,1)$ and let $P = bU_1$. If $P > 1$, go to Step 3 otherwise proceed to Step 2.
- Step 2: Let $Y = P^{\frac{1}{\alpha}}$ and generate U_2 from $U(0, 1)$. If $U_2 \le e^{-Y}$, return $X = Y$. Otherwise go back to Step 1.
- Step 3: Let $Y = -\ln \frac{(b - P)}{\alpha}$ and generate U_2 from $U(0, 1)$. If $U_2 \le Y^{\alpha-1}$, return $X = Y$, otherwise go back to Step 1.

UNIT 3 REGRESSION ANALYSIS

Structure **Page Nos.**

3.0 INTRODUCTION

In many problems, there are two or more variables that are inherently related and it may be necessary to explore the nature of their relationship. Regression analysis is a statistical technique for modeling and investigating the relationship between two or more variables. For example, in a chemical process suppose that the yield of the product is related to the process operating temperature. Regression analysis can be used to build a model that expresses yield as a function of temperature. This model can be used to predict yield at a given temperature level. It can also be used for process optimisation or process control purposes.

In general, suppose that there is a single dependent variable or response variable y and that is related to k independent or regressor variables say $x_1,.....,x_k$. The response variable y is a random variable and the regressor variables $x_1,.....,x_k$ are measured with negligible error. The relationship between y and $x_1,.....x_k$ is characterised by a mathematical model and it is known as the regression model. It is also known as the regression of y on $x_1,.....,x_k$. This regression model is fitted to a set of data. In many situations, the experimenter knows the exact form of the functional relationship between y and $x_1, ..., x_k$, say $\varphi(x_1, ..., x_k)$, except for a set of unknown parameters. When the functional form is unknown, it has to be approximated on the basis of past experience or from the existing information. Because of its tractability, a polynomial function is popular in literature.

In this unit, we will be mainly discussing the linear regression model, when $k = 1$, that is only one regressor variables. We will be discussing in detail how to estimate the regression line and how it can be used for prediction purposes from a given set of data. We will also discuss briefly how we can estimate the function ϕ, if it is not linear.

3.1 OBJECTIVES

After going through this unit, you should be able to:

- decide how two variables are related;
- measure the strength of the linear relationship between two variables;
- calculate a regression line that permits the prediction of the value of one of the variables if the value of the other variable is known;

- analyse data by the method of least squares to determine the estimated regression line to be used for prediction, and
- apply the least squares methods to fit different curves and use it for prediction purposes.

3.2 SIMPLE LINEAR REGRESSION

We wish to determine the relationship between a single regressor variable x and a response variable y*(note: The linear regression with one independent variable is referred to as a simple linear regression).* We will refer to y as the dependent variable or response and x as the independent variable or regressor. The regressor variable x is assumed to be a continuous variable controlled by the experimenter. You know that it is often easy to understand data through a graph. So, let us plot the data on *Scatter diagram* (a set of points in a 2-D graph where horizontal axis is regressor and vertical axis is response). Suppose that the true relationship between y and x is a straight line. Therefore, each observation y can be described by the following mathematical relation (model)

$$y = \beta_0 + \beta_1 x + \in \qquad (1)$$

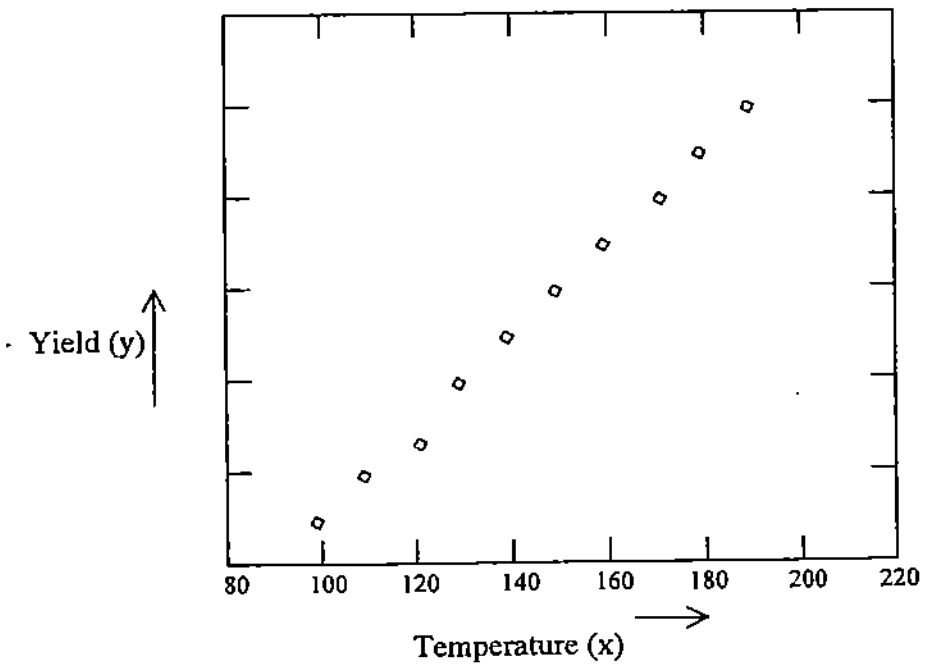


Figure 1: Scatter diagram of yield versus temperature

where $\in$ is a random variable with mean 0 and variance σ^2. The $\in$ is known as the error component and it is assumed to be small. If the error $\in$ was absent then it is a perfect relation between the variables y and x which may not be very practical. Let us look at the following example.

Example 1: A chemical engineer is investigating the effect of process operating temperature on product yield. The study results in the following data.

Temperature °C (x)	100	110	120	130	140	150	160	170	180	190
Yield,%(y)	45	51	54	61	66	70	74	78	85	89

The scatter diagram between the temperature and the yield is presented in *Figure 1*. In *Figure 1*, it is clear that there is a linear relationship between yield and temperature but clearly it is not perfect. For example, we cannot write the relationship between y and x as follows

$$y = \beta_0 + \beta_1 x$$

Clearly the presence of the error $\in$ is needed. Moreover the error $\in$ is a random variable because it is not fixed and it varies from one temperature to another. It may also vary when two observations are taken at the same temperature. If there was a perfect linear relationship between y and x we would have required just two points to calculate the relationship. Since, the relationship is not perfectly linear it is usually requires more than two data points to calculate their relationship. Our main objective is to calculate the relationship between them from the existing information (data points). Since, it is assumed that the relationship between x and y is linear therefore, the relationship can be expressed by equation (1) and finding the relationship basically boils down to finding the unknown constants β_0 and β_1 from the observations.

Let us discuss this concept of linear regression once more, through the illustration / collection of data described in *Table 1*. Table 1, encloses the data of 25 samples of cement, for each sample we have a pair of observations (x,y) where x is percentage of SO_3, a chemical and y is the setting time in minutes. These two components are strongly related; it is the percentage of SO_3 which influences the setting time of any cement sample, the recorded observations are given in *Table 1*.

Table 1: Data on SO_3 and Setting Time

Sl.No. (i)	Percentage of SO_3 (X)	Setting Time Y (in minutes)
1	1.84	190
2	1.91	192
3	1.90	210
4	1.66	194
5	1.48	170
6	1.26	160
7	1.21	143
8	1.32	164
9	2.11	200
10	0.94	136
11	2.25	206
12	0.96	138
13	1.71	185
14	2.35	210
15	1.64	178
16	1.19	170
17	1.56	160
18	1.53	160
19	0.96	140
20	1.7	168
21	1.68	152
22	1.28	160
23	1.35	116
24	1.49	145
25	1.78	170
Total Sum of Squares	39.04 64.446	4217 726539

In *Table 1*, you see that setting time y increases as percentage of SO_3 increases. Whenever you find this type of increasing (or decreasing) trend in a table, same will be reflected in the scatter diagram (Figure 5), and it indicates that there is a linear relationship between x and y. By drawing the scatter diagram you can observe that the relationship is not perfect in the sense that a straight line cannot be drawn through all the points in the scatter diagram.

Nevertheless, we may approximate it with some linear equation. What formula shall we use? Suppose, we use the formula $y = 90 + 50x$ to predict y based on x. To examine how good this formula is, we need to compare the actual values of y with the corresponding predicted values. When $x = 0.96$, the predicted y is equal to 138(= 90 + 50 x 0.96). Let (x_i, y_i) denote the values of (x, y) for the i^{th} sample. From *Table 1*, notice that $x_{12}=x_{19}=0.96$, whereas $y_{12} = 138$ and $y_{19} = 140$.

Let $\hat{y} = 90 + 50x_i$. That is, $\hat{y}_i$ **is the predicted value of y** (then using y = 90 + 50x for the i^{th} sample. Since, $x_{12}=x_{19}=0.96$, both $\hat{y}_{12}$ and $\hat{y}_{19}$ are equal to 138. Thus, the difference $\hat{e}_i = y_i - \hat{y}_i$, **the error in prediction**, also called residual is observed to be $\hat{e}_{12} = 0$ and $\hat{e}_{19} = 2$. The formula we have considered above, y = 90 + 50x, is called a **simple linear regression equation**, we will study these terms in detail in our successive sections.

3.2.1 Least Squares Estimation

Suppose that we have n pairs of observations, say $(x_1, y_1),.....(x_n, y_n)$. It is assumed that the observed y_i and x_i satisfy a linear relation as given in the model (1). These data can be used to estimate the unknown parameters β_0 and β_1. The method we are going to use is known as the method of least squares, that is, we will estimate β_0 and β_1 so that the sum of squares of the deviations from the observations to the regression line is minimum. We will try to explain it first using a graphical method in *Figure 2*. For illustrative purposes we are just taking 5 data points (x, y) = (0.5, 57), (0.75, 64), (1.00, 59), (1.25, 68), (1.50,74). The estimated regression line can be obtained as follows. For any line we have calculated the sum of the differences (vertical distances) squares between the y value and the value, which is obtained using that particular line. Now, the estimated regression line is that line for which the sum of these differences squares is minimum.

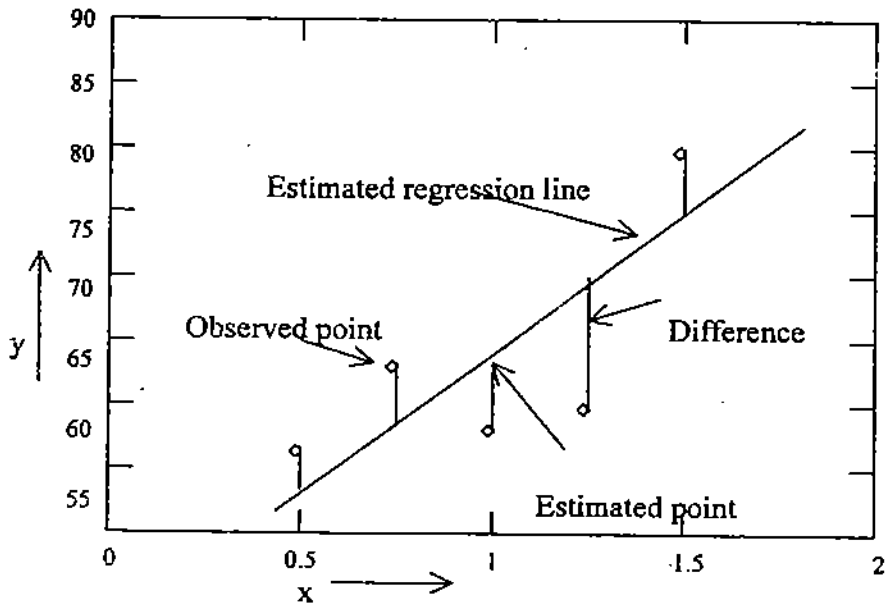


Figure 2 Differences between y values and the estimated values using regression line

Matematically the sum of squares of the deviations of the observations from the regression line is given by

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \qquad (1)$$

The least squares estimators of β_0 and β_1, are $\hat{\beta}_0$ and $\hat{\beta}_1$ which can be obtained by solving the following two linear equations.

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \qquad (1a)$$

Simplifying these two equations yields

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad (2)$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \qquad (3)$$

Solving (2) and (3) $\hat{\beta}_0$ and $\hat{\beta}_1$ can be obtained as follows:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (4)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\left(\sum_{i=1}^{n} yi\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} xi\right)^2} \qquad (5)$$

where $\bar{y} = \sum_{i=1}^{n} y_i$ and $\bar{x} = \sum_{i=1}^{n} x_i$. Therefore, the fitted simple linear regression line between these n points is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad (6)$$

Note : Linear correlation and regression are very similar. One uses the correlation coefficient to determine whether two variables are linearly related or not. The correlation coefficient measures the strength of the linear relationship. Regression on the other hand is used when we want to answer questions about the relationship between two variables.

Some Properties for Linear Correlation and Regression

1) The line of regression of y on x always passes through($\bar{x}$, $\bar{y}$) where $\bar{x}$ and $\bar{y}$ are the mean of x and y values.

2) There are always two line, of regression one of y on x and other of x on y.
   i.e., $y = a_1 + b_{yx} x$ or $x = a_2 + b_{xy} y$

   where byx = Regression coeff of y on x $= r\frac{\sigma_y}{\sigma_x}$

$$\text{bxy} = \text{Regression coeff of y on x} = r\frac{\sigma_x}{\sigma_y}$$

Correlation can be obtained by the following formula also,

$$r = \sqrt{b_{xy} * b_{yx}} \quad (-1 \le r \le 1)$$

Angle between lines of regression is given by,

$$\theta = \tan^{-1}\left\{\frac{r^2 - 1}{r}\left(\frac{\sigma_x * \sigma_y}{\sigma^2{}_x + \sigma^2{}_y}\right)\right\}$$

Where r = correlation coeff between x and y

σ_x = standard deviation of variable x

σ_y = standard deviation of variable y

So, now, Regression equation of y on x can be rewritten as

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

And Regression equation of x on y as,

$$(x - \bar{x}) = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Example 1 (contd.) Now, we will compute the estimates of β_0 and β_1 for the data points given in Example 1. In this case it is observed that

$$n = 10, \qquad \sum_{i=1}^{10} x_i = 1450, \qquad \sum_{i=1}^{10} yi = 673, \quad \bar{x} = 145, \qquad \bar{y} = 67.3 \cdot$$

$$\sum_{i=1}^{10} x_i^2 = 218{,}500, \quad \sum_{i=1}^{10} y_i^2 = 47{,}225, \quad \sum_{i=1}^{10} x_i y_i = 101{,}570$$

Therefore,

$$\hat{\beta}_1 = \frac{101{,}570 - 10 \times 1450 \times 673}{218{,}500 - 10 \times 1450^2} = 0.483,$$

and

$$\hat{\beta}_0 = 673 - 0.483 \text{ X } 145 = -2.739.$$

The fitted simple linear regression line through those 10 points is

$$\hat{y} = -2.739 + 0.483x \qquad (7)$$

The best fitted line along with the data points are plotted in *Figure 3*. Note that the best fitted line can be used effectively for prediction purposes. For example, suppose we want to calculate the expected yield when the temperature is 170°C, for which the data is not available. We can use the (7) for this purpose as follows.

$$\hat{y} = -2.739 + 0.483 \times 170 = 79.371 .$$

Therefore, the best fitted line shows that the expected yield at 170°C is 79.371.

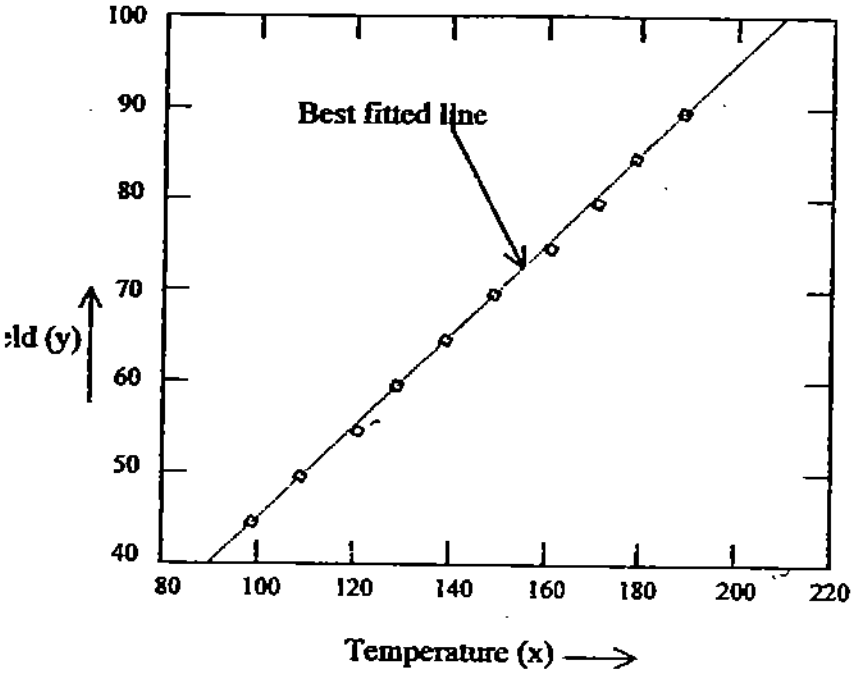


Figure 3: Data points and the best fitted regression line passing through these points

Shortly, we will discuss the technique that consists of a few steps; which can be used to fit a line in the best way, such that the error is minimum. In short, we will study the technique to determine the best equation, that can fit a line in the data such that the error is minimum. But before that let's see one more example.

Example 2: A survey was conducted to relate the time required to deliver a proper presentation on a topic, to the performance of the student with the scores he/she receives. The Table 2 shows the matched data:

Table 2

Hours (x)	Score (y)
0.5	57
0.75	64
1.00	59
1.25	68
1.50	74
1.75	76
2.00	79
2.25	83
2.50	85
2.75	86
3.00	88
3.25	89
3.50	90
3.75	94
4.00	96

1) Find the regression equation that will predict a student's score if we know how many hours the student studied.
2) If a student had studied 0.85 hours, what is the student's predicted score?

Solution. We will arrange the data in the form of a chart to enable us to perform the computations easily.

Table 3

X	y	x^2	xy
0.5	57	0.25	28.5
0.75	64	0.56	48.0
1.00	59	1.00	59.0
1.25	68	1.56	85.0
1.50	74	2.25	111.0
1.75	76	3.06	133.0
2.00	79	4.00	158.0
2.25	83	5.06	186.75
2.50	85	6.25	212.5
2.75	86	7.56	236.5
3.00	88	9.00	246.0
3.25	89	10.56	289.25
3.50	90	12.25	315.0
3.75	94	14.06	352.50
4.00	96	16.00	384.0
33.75	1188	93.44	2863

In this case $n = 15$, therefore

$$\hat{\beta}_1 \frac{15\times 2863 - 33.75\times 1188}{15\times 93.44 - 33.75^2} = 10.857\,,\quad \hat{\beta}_0 = \frac{1}{15}[1188 - 10.857\times 33.75] = 54.772$$

Therefore, the prediction line becomes:

$$\hat{y} = 54.772 + 10.857x$$

Now, the predicted score when $x = 0.85$, is

$$\hat{y} = 54.772 + 10.857\times 0.85 = 64.00$$

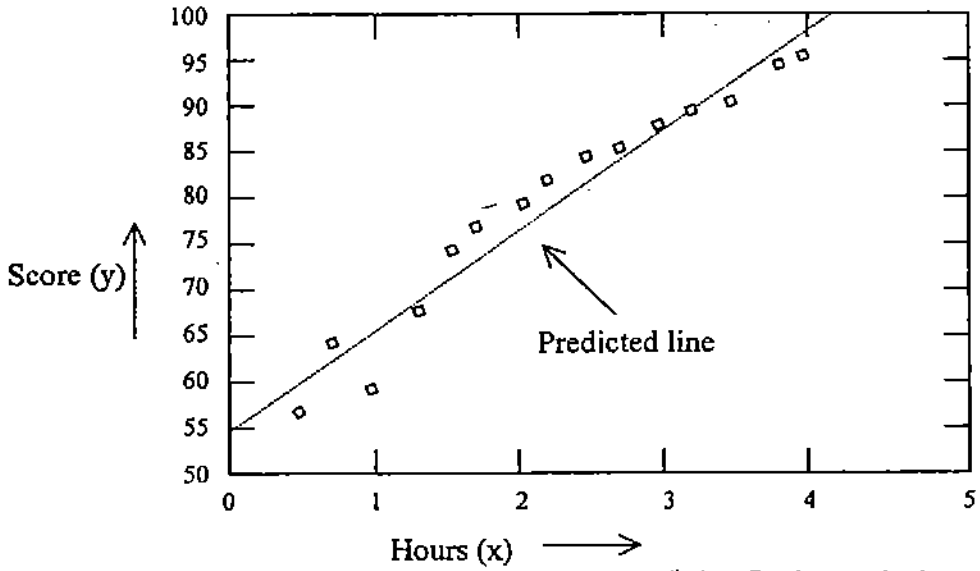


Figure 4: Hours studied and the corresponding score with the best fitted regression line passing through these points

Thus, the predicted score of the student who had studied 0.85 hours is approximately 64.00.

We have plotted the individual scores and the hours studied with the best fitted prediction line in the *Figure 4*. It shows the hours studied by the student, the corresponding score follows a linear pattern and the predicted line can be used quite effectively to predict the score of a student if we know how many hours the student had studied.

Now, it's the time to discuss the technique for determining the best equation, i.e., the equation which fits the line in a way that the overall error is minimised.

From the above illustrations and examples you might have noticed that different equations give us different residuals. What is the best equation? Obviously, the choice will be that equation for which $\hat{e}$s are small.

This means that whatever straight line we use, it is not possible to make all $\hat{e}$s zero, where $\hat{e}_i = y_i - \hat{y}_i$ (the difference). However, we would expect that the errors are positive in some cases and negative in the other cases so that, on the whole, their sum is close to zero. So, our job is to find out the best values of β_0 and β_1 in the formula $y = \beta_0 + \beta_1 x + e$ $(s.t.\ \ e \simeq 0)$. Let us see how we can do this.

Now our aim is to find the values β_0 and β_1 so that the error $\hat{e}$s are minimum. For that we state here four steps to be followed.

1) Calculate a sum S_{xx} defined by $S_{xx} = \sum_{i=1}^{n} x^2_i - n\bar{x}^2$ (8)

   where x_i's are given value of the data and $\bar{x} = \frac{\Sigma x_i}{n}$ is the mean of the observed values and n is the sample size.
   The sum S_{xx} is called the corrected sum of squares.

2) Calculate a sum S_{xy} defined by $S_{xy} = \sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}$ (9)

   where x_i's and y_i's are the x-values and y-values given by the data and $\bar{x}$ and $\bar{y}$ are their means.

3) Calculate $\frac{S_{xy}}{S_{xx}} = \beta_1$ say. That is $\beta_1 = \frac{S_{xy}}{S_{xx}}$ (10)

4) Find $\bar{y} - \beta_1 \bar{x} = \beta_0$, say.

Let us now compute these values of the data in Table 1: Data on SO_3 and Setting Time, we get

$\bar{x}$ = 1.5616, $\bar{y}$ = 168.68, S_{xx} = 3.4811, and S_{xy} = 191.2328.

Substituting these values in (10) and (11), we get

$\beta_1 = \frac{S_{xy}}{S_{xx}}$ = 54.943 and β_0 = 168.68 − 54.943 x 1.5616 = 82.88 (11)

Therefore, the best linear prediction formula is given by
y = 82.88 + 54.943x.

After drawing this line on the scatter diagram, you will discover that straight line is close to more points, and hence, it is the best linear prediction.

Example 3: A hosiery mill wants to estimate how its monthly costs are related to its monthly output rate. For that the firm collects data regarding its costs and output for a sample of nine months as given in the Table 4.

Table 4

Output (tons)	Production cost (thousands of dollars)
1	2
2	3
4	4
8	7
6	6
5	5
8	8
9	8
7	6

1) Find the scatter diagram for the data given above.
2) Find the regression equation when the monthly output is the dependent variable (x) and monthly cost is the independent variable (y).
3) Use this regression line to predict the firm's monthly cost if they decide to produce 4 tons per month.
4) Use this regression line to predict the firm's monthly cost if they decide to produce 9 tons per month.

Solution: a) Suppose that x_i denotes the output for the ith month and y_i denotes the cost for the i^{th} month. Then we can plot the graph for the pair (x_i, y_i) of the values given in Table . Then we get the scatter diagram as shown in Figure 5.

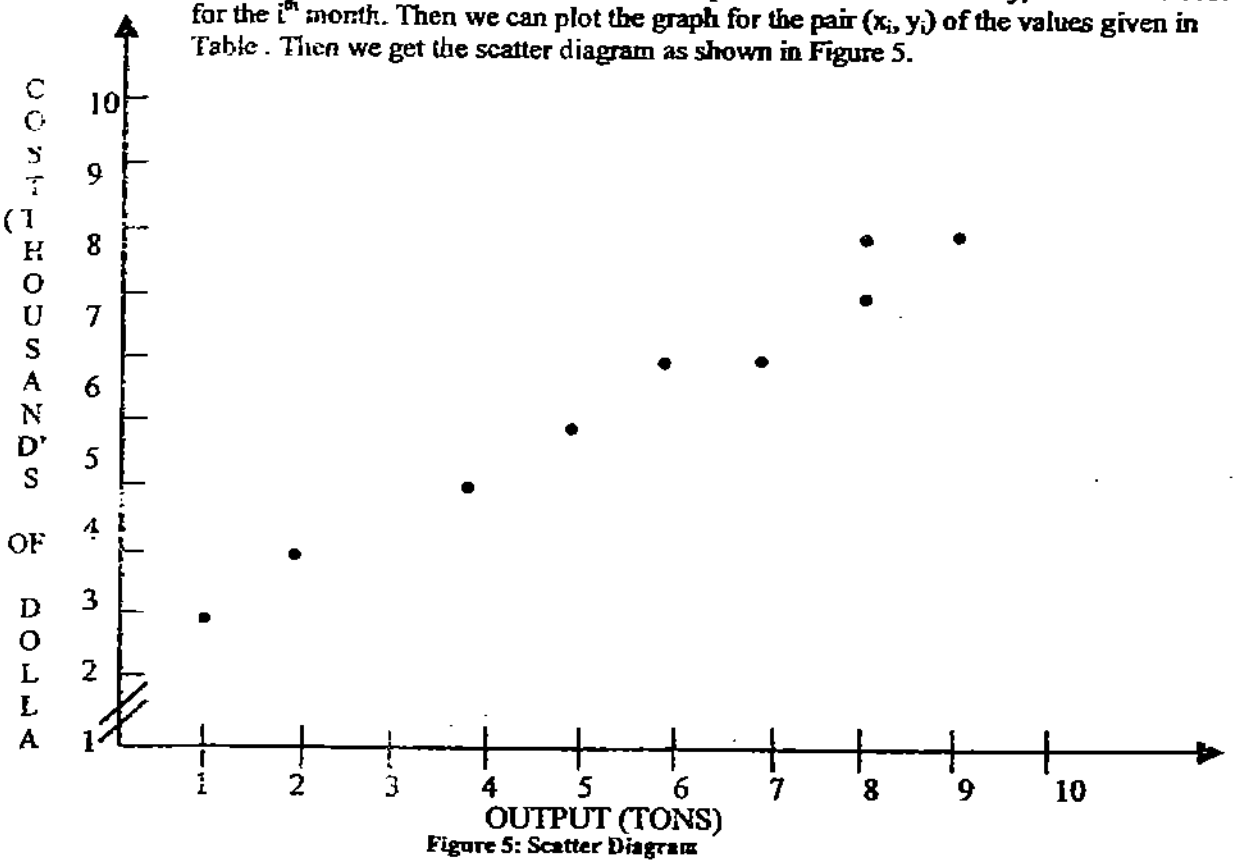


Figure 5: Scatter Diagram

a) Now to find the least square regression line, we first calculate the sums S_{xx} and S_{xy} from Eqn.(8) and (9).

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Note that from Table(4) we get that

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{50}{9}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{49}{9}$$

$$\sum x^2_i = 340$$

$$\sum y^2_i = 303$$

and $\sum x_i y_i = 319$

Therefore, we get that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$= \frac{9\ x\ 319 - 50\ x\ 49}{9\ x\ 340 - 50^2}$$

$$= \frac{421}{560} = 0.752$$

Correspondingly, we get

$$\beta_0 = \frac{49}{9} - (0.752)\ x\ \frac{50}{9}$$

$$= 1.266$$

Therefore, the best linear regression line is

$$y = 1.266 + (0.752)x$$

b) If the firms decides to produce 4 tons per month, then one can predict that its cost would be

$1.266 + (0.752) \times 4 = 4.274$
Since, the costs are measured in thousands of dollars, this means that the total cost would be expected to be \$4,274.

c) If the firms decides to produce 9 tons per month, then one can predict that its cost would be $1.266 + (0.752) \times 9 = 8.034$

Since, the costs are measured in thousands of dollars, this means that the total costs would be expected to be \$8034.

☞ Check Your Progress 1

1) In partially destroyed laboratory record of an analysis of correlation data, the following results only are legible.

Variance of $x = 9$
Regression equations : $8x - 10y + 66 = 0$
$40x - 18y - 214 = 0$
what were (1) the mean values of x and y,
(2) the correlation coeff between x and y
(3) the standard deviation of y

.....

.....

.....

2) Since humidity influences evaporation, the solvent balance of water reducible paints during sprayout is affected by humidity. A controlled study is conducted to examine the relationship between humidity (X) and the extent of evaporation (Y) is given below in Table 5. Knowledge of this relationship will be useful in the sense that it will allow the painter to adjust his or her spray gun setting to account for humidity. Estimate the simple linear regression line and predict the extent of solvent evaporation (i.e loss of solvent, by weight) when the relative humidity is 50%.

Table 5

Observation	(x) Relative humidity, (%)	(y) Solvent Evaporation, (%) wt
1	35.3	11.0
2	29.7	11.1
3	30.8	12.5
4	58.8	8.4
5	61.4	9.3
6	71.3	8.7
7	74.4	6.4
8	76.7	8.5
9	70.7	7.8
10	57.5	9.1
11	46.4	8.2
12	28.9	12.2
13	28.1	11.9
14	39.1	9.6
15	46.8	10.9
16	48.5	9.6
17	59.3	10.1
18	70.0	8.1
19	70.0	6.8
20	74.4	8.9
21	72.1	7.7
22	58.1	8.5
23	44.6	8.9
24	33.4	10.4
25	28.6	11.1

3.2.2 Goodness of Fit

We have seen in the previous subsection that the regression line provides estimates of the dependent variable for a given value of the independent variable. The regression line is called the best fitted line in the sense of minimising the sum of squared errors. The best fitted line shows the relationship between the independent (x) and dependent (y) variables better than any other line. Naturally the question arises "How good is our best fitted line?" We want a measure of this goodness of fit. More precisely, we want a numerical value which measures this goodness of fit.

For developing a measure of the goodness of fit, we will first examine the variation in y. Let us try the variation in the response y. Since, y depends on x, if we change x, then y will also change. In other words, a part of variation in y's is accounted by the variation in x's. Actually, we can mathematically show that the total variation in y's can be split up as follows:

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{S^2_{xy}}{S_{xx}} + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \qquad (12)$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2; S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \hat{y}_i)$$

Now if we divide (12) by S_{yy} on both sides, we get

$$1 = \frac{S^2_{xy}}{S_{xx}S_{yy}} + \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{S_{yy}}$$

Since, the quantities on the right hand side are both non-negative, none of them can exceed one. Also if one of them is closer to zero the other one has to be closer to one. Thus, if we denote

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

then

$$R^2 = \frac{S^2_{xy}}{S_{xx}S_{yy}}$$

Since R^2 must be between 0 and 1, R must be between – 1 and 1. It is clear that if $R^2 = 1$, then

$$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{S_{yy}} = 0 \text{ or } \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0 \quad \text{or} \quad y_i = \hat{y}_i \quad \text{for all } i.$$

Again when R^2 is close to 1, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is close to zero. When R is negative, it means that y decreases as x increase and when R is positive y increases when x increases. Thus, R gives a measure of strength of the relationship between the variables x and y.

Now let us compute the value of R for Example 1. For calculating the numerical value of R, the following formula can be used;

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2}}$$

Therefore, for Example 1, the value of R becomes;

$$R = \frac{101,570 - 10 \times 145 \times 67.3}{\sqrt{218,500 - 10 \times 145^2}\sqrt{47225 - 10 \times 67.3^2}} = \frac{3985}{\sqrt{8250}\sqrt{1932.1}} = 0.9981$$

and $R^2 = 0.9963$.

Therefore, it is clear from the value of R or from R^2 that both of them are very close to one, thus the predicted line fits the data very well.

Moreover R is a positive mean, there is a positive relation between the temperature and yield. As the temperature increases the yield also increases.

Now, the natural question is how large should this R or R^2 be in order for the fit very good. There is a formal statistical test based on F-distribution which can be used to test whether R^2 is significantly large or not. We will not go into those details here. But as a thumb rule we may say that if R^2 is greater that 0.9, the fit is very good, if it is between 0.6 to 0.8, the fit is moderate and if it is less than 0.5 it is not good.

☞ Check Your Progress 2

1) For the data given in the Table 6 compute R and R^2

Table 6: $\hat{y}_i$ and $\hat{e}_i$ For Some Selected *i*

Sample No. (i)	12	21	15	1	24
x_i	0.96	1.28	1.65	1.84	2.35
y_i	138	160	178	190	210
$\hat{y}_i$	138				
$\hat{e}_i$	0				

Note: $\hat{y}_i = 90 + 50x$ and $\hat{e}_i = y_i - \hat{y}_i$

3.2.3 Residual Analysis

Fitting a regression model to a set of data requires several assumptions. For example, estimation of the model parameters requires the assumptions that the errors are uncorrelated random variables with mean zero and constant variance. If these assumptions are not satisfied, then using the simple least squares method may not produce the efficient regression line. In fitting a linear model, it is also assumed that the order of the model is correct, that is if we fit a first order polynomial, then we are assuming that phenomenon actually behaves in a first order manner. Therefore, for a practitioner it is important to verify these assumptions and the adequacy of the model.

We define the residuals as $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$, where y_i is an observation and $\hat{y}_i$ is the corresponding estimated value from the best fitting regression line.

Analysis of the residuals is frequently helpful in checking the assumption that errors are independent and identically distributed with mean zero and finite variance and in determining whether the additional terms in the model would be required not. It is advisable to plot the residuals

a) in time sequence (if known),
b) against the $\hat{y}_i$ or
c) against the independent variable x. These graphs will usually look like one of the four general patterns as shown in the *Figures 6* to *9*.

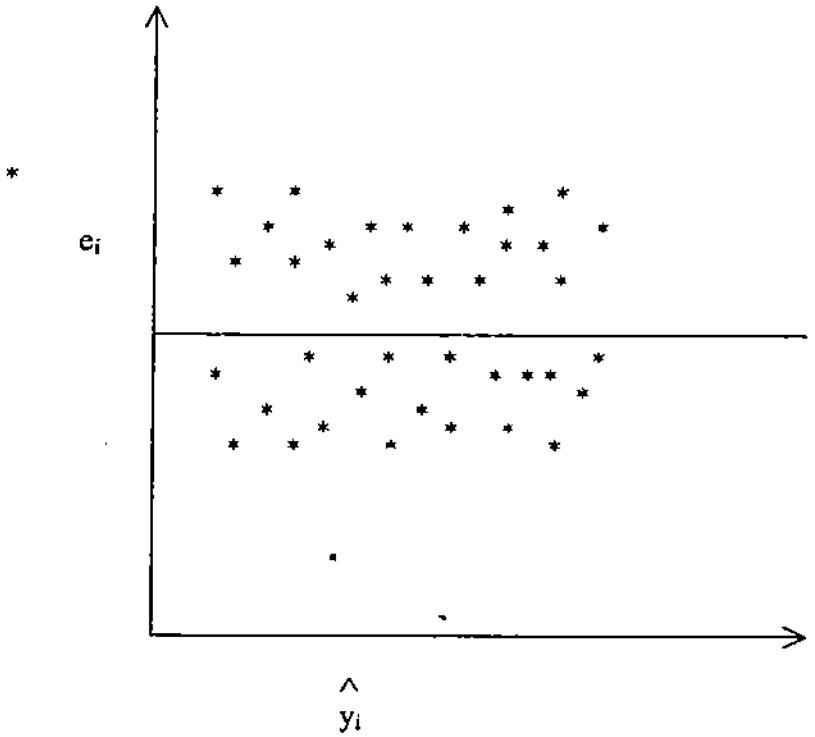


Figure 6 Pattern of the residual plot; satisfactory.

Figure 6 indicates that the residuals are behaving in satisfactory manner and the model assumptions can be assumed to be correct. *Figures 7 – 9* indicates unsatisfactory behaviour of the residuals. The *Figure 7* clearly indicates that the variances are gradually increasing. Similarly *Figure 8* indicates that the variances are not constant. If the residuals plot is like *Figure 9*, then it would seem that the model order is not correct, that means,

the first order model is not the correct assumption. We should look for higher order models.

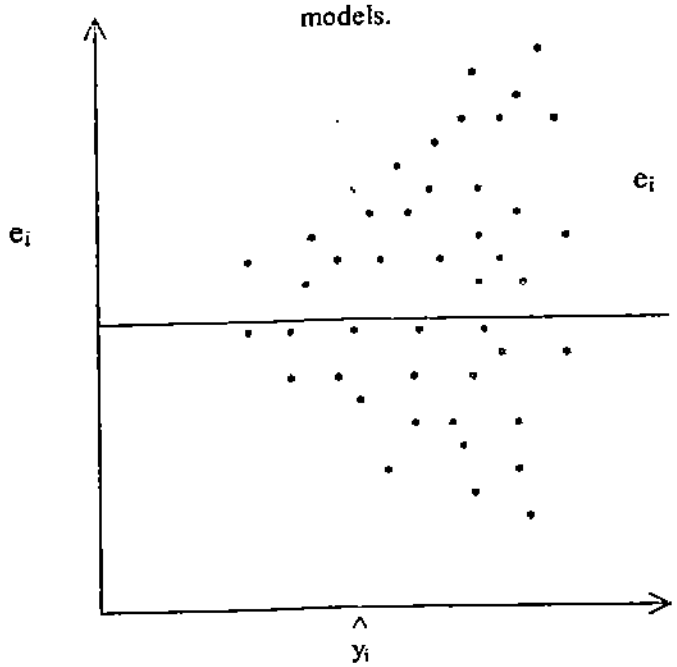


Figure 7: Pattern of the residual plot; indicates the variance is gradually increasing this case

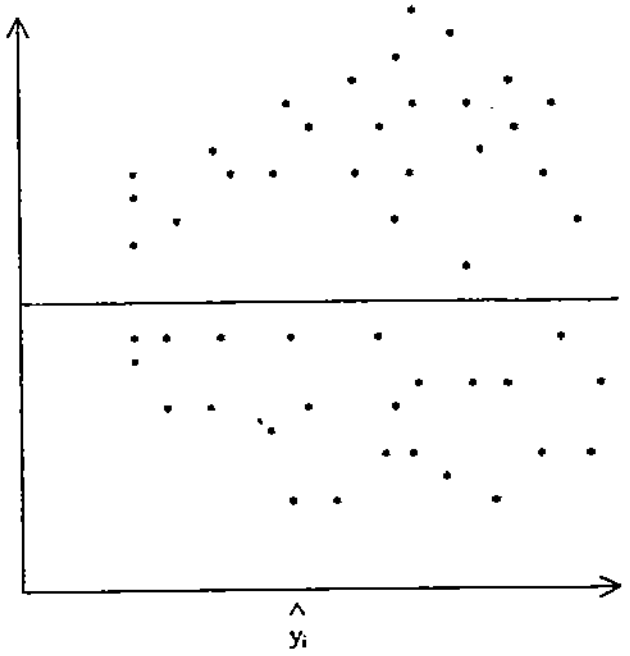


Figure 8: Pattern of the residual plot; Indicates the variance is no constant.

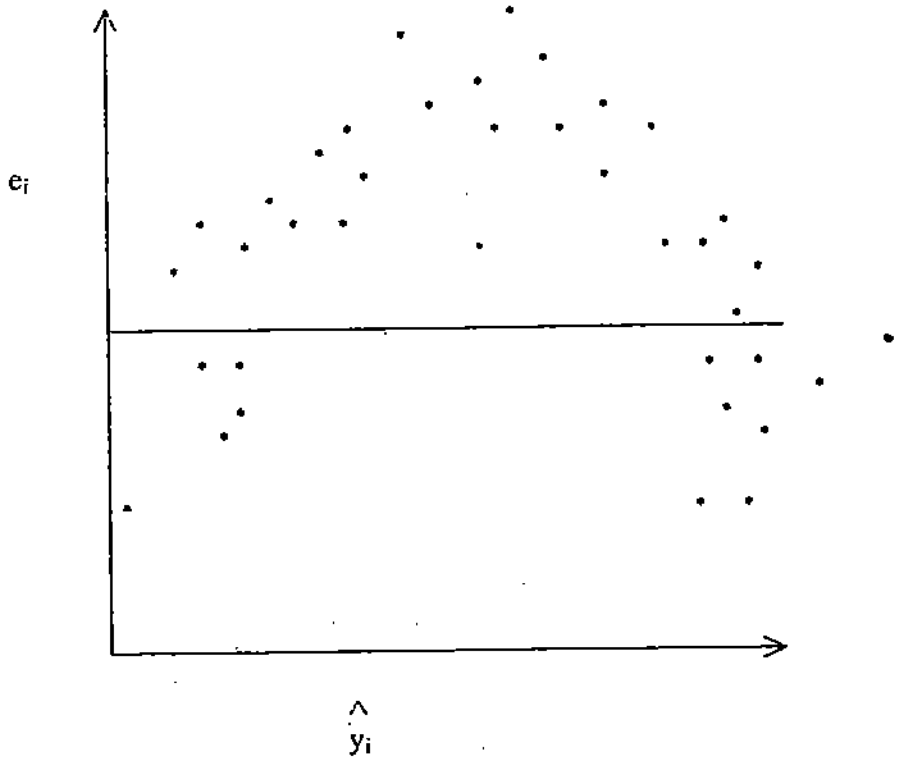


Figure 9: Patent of the residual plot; indicates the model order is not correct

Example 4: Now, we provide the residual plots of the data given in Example 1. We have plotted $\hat{y}_i$ vs. e_i. It is provided in *Figure 10*. From *Figure 10*, it is quite clear that the residuals plot is quite satisfactory and apparently all the model assumptions are satisfied in *Figure 10*.

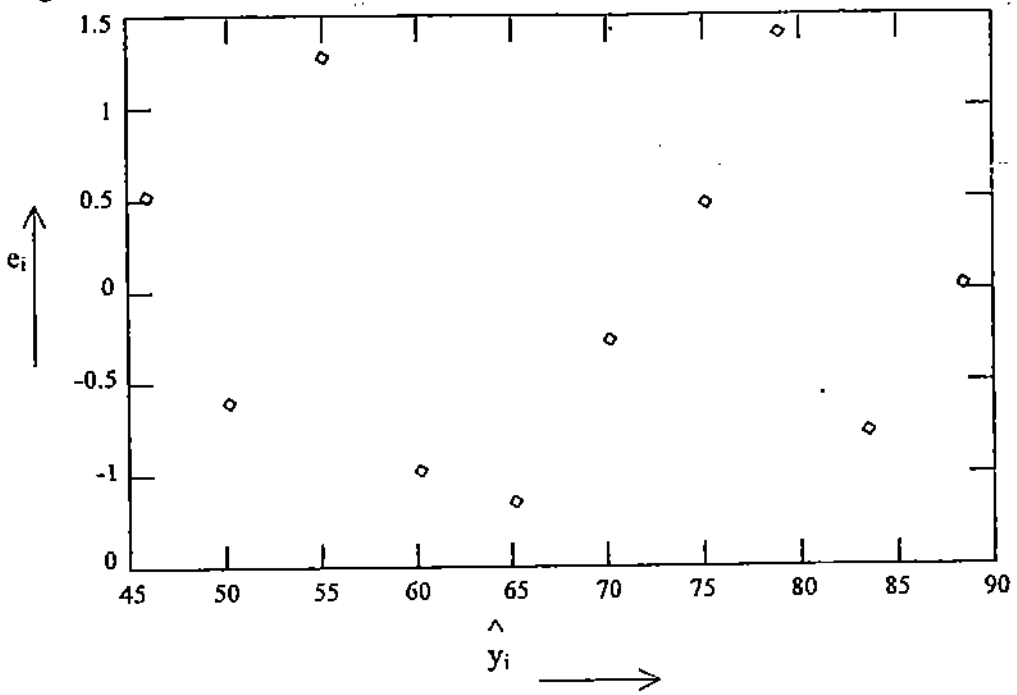


Figure10: Pattern of the residual plot; satisfactory.

☞ Check Your Progress 3

1) What is the utility of residual plots? What is the disadvantage of residual plots?

.....

.....

.....

.....

.....

.....

.....

.....

.....

3.3 NON-LINEAR REGRESSION

Linear regression is a widely used method for analysing data described by models which are linear in parameters. However, in many practical situations, people come across data where the relationship between the independent variable and the dependent variable is not linear. In that case, definitely one should not try to use a linear regression model to

represent the relationship between the independent and dependent variable. Let us consider the following example.

Example 5: Data on the amount of protein generated by a certain experiment was counted and reported over time. The results are presented below in Table 7:

Table 7

Time (min)	Protein (gm)	Time (min)	Protein (gm)	Time (min)	Protein (gm)	Time (min)	Protein (gm)
0	0.844	80	0.818	160	0.580	240	0.457
10	0.908	90	0.784	170	0.558	250	0.448
20	0.932	100	0.751	180	0.538	260	0.438
30	0.936	110	0.718	190	0.522	270	0.431
40	0.925	120	0.685	200	0.506	280	0.424
50	0.908	130	0.685	210	0.490	290	0.420
60	0.881	140	0.628	220	0.478	300	0.414
70	0.850	150	0.603	230	0.467	310	0.411

We present the Time vs. Protein generated in the *Figure 11*.

From *Figure 11* it is clear that the relationship between time taken and protein generated is not linear. Therefore, they cannot be explained by a linear equation. In a situation like this, we may often go for a non-linear model to explain the relationship between the independent and dependent variables and they are called the non-linear regression model.

A non-linear regression model can be formally written as

$$y = f(x,\theta) + \epsilon, \tag{13}$$

where $f(x,\theta)$ is a known response function of k-dimensional vector of explanatory variable x and p-dimensional vector of unknown parameter θ. Here also ϵ represents the error component and it is assumed that it has mean zero and finite variance. Therefore, it is clear that the non-linear regression model is a generalisation of the linear regression model. In case, the linear regression model $f(x, \theta)$ is a linear function, but, it can also be any non-linear function. Similar to the linear regression model, here also, our problem is the same, that is, if we observe a set of n, $\{(x_1, y_1),\ldots\ldots, (x_n, y_n)\}$, how do we estimate the unknown parameters θ, when we know the functional form of $f(x, \theta)$.

3.3.1 Least Squares Estimation

Similar to the linear regression method here also to estimate the unknown parameters, we adopt the same method. We calculate the estimate of θ by minimising the residual sums of squares, that is minimize

$$Q(\theta) = \sum_{i=1}^{n} [y_i - f(x_i,\theta)]^2, \tag{14}$$

with respect to the unknown parameters. The idea is the same as before, that is we try to calculate that particular value of θ for which the sum of squares of the distance between the points y_i and $f(x_i,\theta)$ is minimum. Unfortunately, in this case the minimum cannot be performed as easily as before. We need to adopt some numerical technique to minimise

the function $Q(\theta)$. This minimisation can be performed iteratively and one technique that can be used to accomplish this is the Gauss-Newton method.

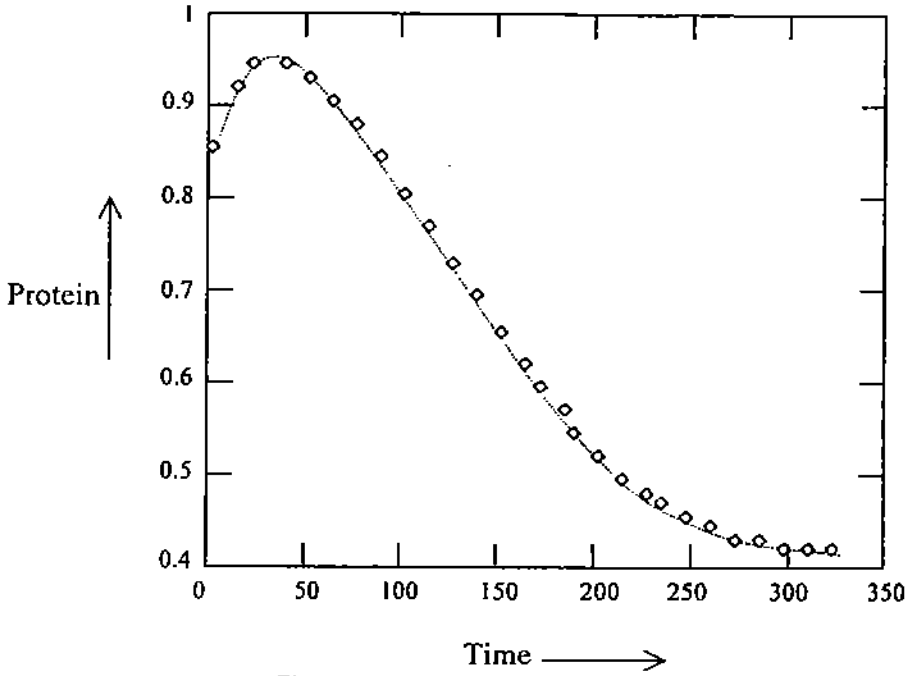


Figure 11: Time vs. Protein generated in an experiment.

You have already learned about the Gauss-Newton method in detail earlier, we just give a brief description for your ready reference. We use the following notations below:

$$\theta = (\theta_1, \ldots\ldots, \theta_p), \qquad \theta^{(k)} = (\theta_1^{(k)}, \ldots\ldots\ldots, \theta_p^{(k)}) \tag{15}$$

Expand the function $f(x,\theta)$ using a Taylor series expansion about the starting point $\theta^{(0)}$ and using only the first order expansion, we get:

$$f(x_i,\theta) \approx f(x_i,\theta^{(0)}) + v_{i1}(\theta_1 - \theta_1^{(0)}) + \ldots\ldots + v_{ip}(\theta_p - \theta_p^{(0)})$$

where

$$v_{ij} = \left.\frac{\partial f(x_i,\theta)}{\partial \theta_j}\right|_{\theta=\theta^{(0)}} \qquad \text{for} \qquad j = 1, \ldots\ldots\ldots, p.$$

Let $\eta(\theta) = (f(x_1,\theta), \ldots\ldots, f(x_n,\theta))'$ and $y = \ y = (y_1, \ldots\ldots y_n)'$ then in the matrix notation we can write (15)

$$\eta(\theta) \approx \eta(\theta^{(0)}) + V^{(0)}(\theta - \theta^{(0)}),$$

where $V^{(0)}$ is the $\eta \times p$ derivative matrix with elements v_{ij}. Therefore, to compute the first estimates beyond the starting value is to compute

$$b_0 = [V^{(0)'}V^{(0)}]^{-1}[y - \eta(\theta^{(0)})]$$

and then solve for the new estimate $\theta^{(1)}$ as

$$\theta^{(1)} = b_0 + \theta^{(0)}.$$

This procedure is repeated then with $\theta^{(0)}$ is replaced by $\theta^{(1)}$ and $V^{(0)}$ by $V^{(1)}$ and this produces a new set of estimates. This iterative procedure continues until convergence is achieved. Naturally these computations cannot be performed by hands, we need calculators or computers to perform these computations.

Example 5: (Contd). In this case it is observed (theoretically) that the following model (16) can be used to explain the relationship the time and yield generated y_t where

$$y_t = \alpha_0 + \alpha_1 e^{\beta_1 t} + \alpha_2 e^{\beta_2 t} + \epsilon_t . \qquad (16)$$

Note that as we have mentioned, for the general non-linear regression model, in this case also, the form of the non-linear function namely, $\alpha_0 + \alpha_1 e^{\beta_1 t} + e^{\beta_2 t}$ is known, but the parameters of the model, that is, $\theta = (\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2)$ is unknown. Given the data as provided in Example 5, we want to estimate the unknown parameters.

We use the following initial guess $\alpha_0 = 0.5, \alpha_1 = 1.5, \alpha_2 = -1.0, \beta_1 = -0.01, \beta_2 = -0.02,$ and finally using the Gauss-Newton algorithm we obtain the estimates of the parameters as follows:

$\hat{\alpha}_0 = 0.375, \quad \hat{\alpha}_1 = 1.936 \quad \hat{\alpha}_2 = 1.465, \quad \hat{\beta}_0 = -0.013 \quad \hat{\beta}_1 = -0.022$

We have plotted the points and also the best fitted regression line, namely in *Figure 12.*

$$\hat{y} = 0.375 + 1.936e^{-0.013t} - 1.465e^{-0.013t} \qquad (17)$$

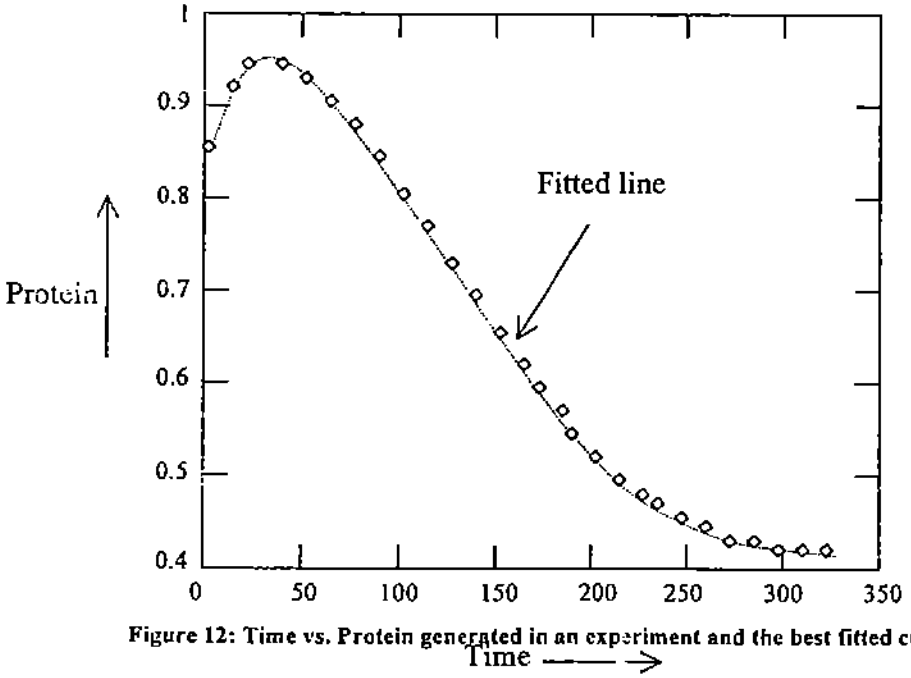


Figure 12: Time vs. Protein generated in an experiment and the best fitted curve

Figure 12 indicates that the fitted regression curve provides a very good relationship between the time and protein generated in that experiment. As before the prediction curve, that is, the curve (17) can be used easily for prediction purposes also. For example, suppose we want to estimate the expected protein generation at the time 115 minutes after the experiment, then using (17), we obtain

$$\hat{y} = 0.375 + 1.936e^{-0.013\times 115} - 1.465e^{-0.013\times 115} = 0.698.$$

Therefore, at the 115 minutes the expected protein generation is 0.698 gms.

Some points we should keep in mind about the non-linear regression model is that we have to know the functional form of the relationship, but the parameters involved are unknown. Usually the functional forms are obtained from the physical nature of the process and they are available. If they are completely unknown it is difficult to use the non-linear regression model. In that case, we need to try with some guess models but they are not very easy and they are not pursued here. Another important issue is the choice of the guess value of the iterative process. This is also a difficult problem. Usually from prior experimental results we may have to use the trial and error method to calculate the initial guess values.

☞ Check Your Progress 4

1) Data on the amount of heat generated by friction were obtained by Count Rumford in 1798. A bore was fitted into a stationary cylinder and pressed against the bottom by means of a screw. The bore was turned by a team of horses for 30 minutes and Rumford measured the temperature at small intervals of time. They are provided in the Table 8.

Table 8

Time (min)	Temperature (°F)	Time (min)	Temperature (°F)
4	126	24	115
5	125	28	114
7	123	31	113
12	120	34	112
14	119	37.5	111
16	118	41	110
20	116		

a) Plot the time versus temperature curve and convince yourself that the linear regression model is not the correct model in this case.

b) A model based on Newton's law of cooling was proposed as

$$f(t,\theta) = 60 + 70e^{-\theta t}$$

Using an initial guess of $\theta^{(0)} = 0.01$, find the least squares estimate of θ.

c) Based on the fitted least squares regression line find the expected temperature at the time 15th minute after starting the experiment.

.....

.....

.....

3.4 SUMMARY

In this unit you have seen:

- That regression analysis is an important technique, which can be used to verify the results of any experiment,
- How the relationship between a dependent and an independent variable can be determined by using the Scatter diagram,
- That by using the technique of regression you have an edge that can help you analyse the results in an organised way. Further, this analysis is smoothened by the

application of concepts such as, least square estimation, goodness to fit and residual analysis,

- That very often such as, data obtained by conducting an experiment does not follow the linear relation. So, to handle such aspects we have also discussed the concept of non linear regression and we have emphasised least the square estimation technique.
- Formulas and applications of the following topics:
  - Simple Linear Regression
    - Least Squares Estimation
    - Goodness to Fit
    - Residual Analysis
  - Non-Linear Regression
    - Least Squares Estimation

3.5 SOLUTIONS/ANSWERS

Check Your Progress 1

1) Since both the regression lines pass through the point $(\bar{x}, \bar{y})$, we have

$$8\bar{x} - 10\bar{y} + 66 = 0$$
$$40\bar{x} - 18\bar{y} - 214 = 0$$

Solving we get, $\bar{x} = 13$
$\bar{y} = 17.$

Let $8x - 10y + 66 = 0$ and $40x - 18y - 214 = 0$

Be the lines of regression of y and x and x on y respectively. Now, we put them in the following form.

$$= \frac{8}{10}x + \frac{66}{10} \text{ and } x = \frac{18}{40}y + \frac{214}{40} \qquad (4)$$

$\therefore$ byx = regression coeff of y on $x = \frac{8}{10} = \frac{4}{5}$

bxy = regression coeff of x on $y = \frac{18}{40} = \frac{9}{20}$

Hence, r^2 = bxy. byx = $\frac{4}{5} \quad \frac{9}{20} \quad = \quad \frac{9}{25}$

So r = $\pm \quad \frac{3}{5} \quad = \quad \pm \quad 0.6$

Since, both the regression coeff are +ve, we take $r = +0.6$

3) We have, byx = $r\frac{\sigma_y}{\sigma_x}$ $\Rightarrow$ $\frac{4}{5} = \frac{3}{5} \times \frac{\sigma_y}{3}$

$\sigma_y = 4$

Remarks (i) had we taken $8x - 10y + 66 = 0$ as regression equation of x on y and $40x - 18y = 214$, as regression equation of y on x.

Then bxy $= \frac{10}{8}$ and byx $= \frac{40}{18}$

or r^2 = bxy byx = $\frac{10}{8} \times \frac{40}{18} = 2.78$

so $r = \pm 1.66$

Which is wrong as r lies between $\pm$ 1.

2)

Observation	(x) Relative humidity, (%)	(y) Solvent Evaporation, (%) wt
1	35.3	11.0
2	29.7	11.1
3	30.8	12.5
4	58.8	8.4
5	61.4	9.3
6	71.3	8.7
7	74.4	6.4
8	76.7	8.5
9	70.7	7.8
10	57.5	9.1
11	46.4	8.2
12	28.9	12.2
13	28.1	11.9
14	39.1	9.6
15	46.8	10.9
16	48.5	9.6
17	59.3	10.1
18	70.0	8.1
19	70.0	6.8
20	74.4	8.9
21	72.1	7.7
22	58.1	8.5
23	44.6	8.9
24	33.4	10.4
25	28.6	11.1

Summary statistics for these data are

$n = 25$ $\Sigma x = 1314.90$ $\Sigma y = 235.70$
$\Sigma x^2 = 76.308.53$ $\Sigma y^2 = 2286.07$ $\Sigma xy = 11824.44$

To estimate the simple linear regression line, we estimate the slope β_1 and intercept β_0. these estimates are

$$=\beta_1 => \hat{\beta}_1 = b_1 = \frac{n\Sigma xy - [(\Sigma x)(\Sigma y)]}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{25(11,824.44) - [(1314.90)(235.70)]}{25(76,308.53) - (1314.90)^2}$$

$= -.08$

β_0 $=> \hat{\beta}_0 = b_0 = \bar{y} - b_1\bar{x}$

$= 9.43 - (-.08)\ (52.60) = 13.64$

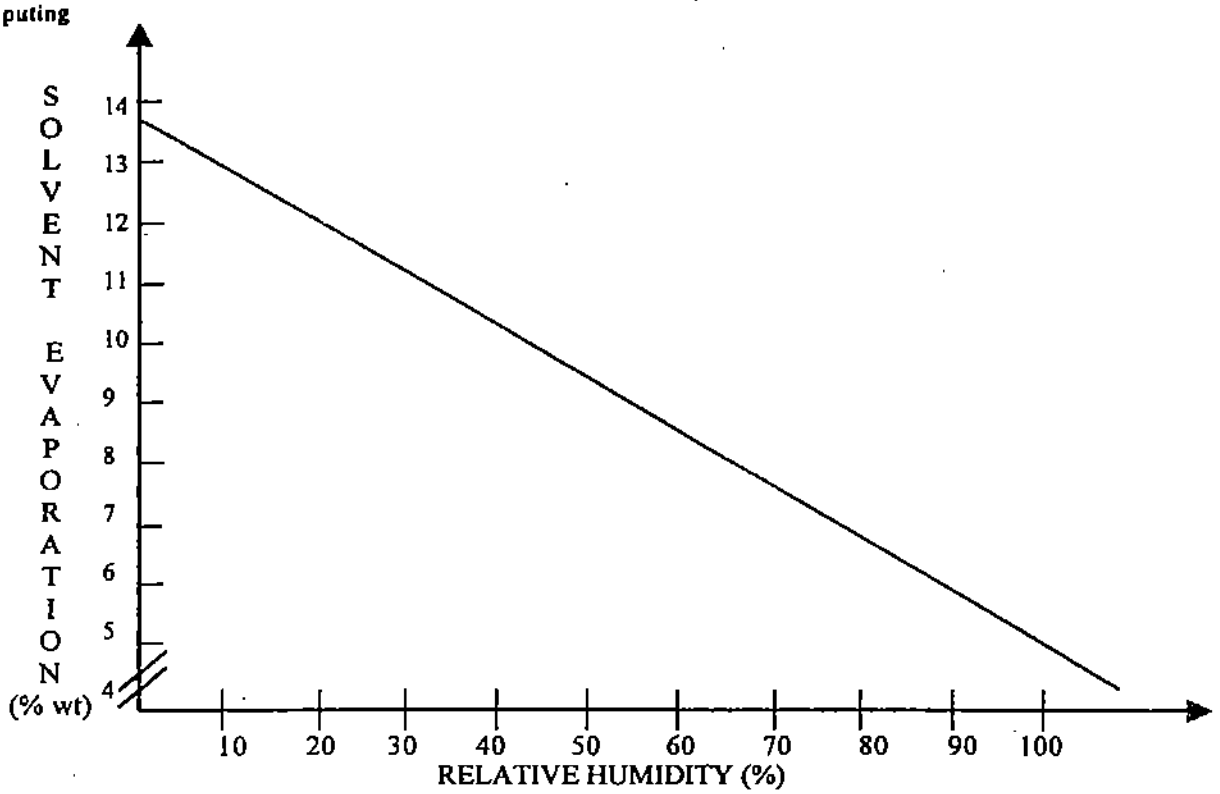


Figure: A graph of the estimated line of regression of Y, the extent of evaporation, on X, the relative humidity

Hence, the estimated regression equation is $\hat{\mu}_{Y|x} = \hat{y} = 13.64 - .08x$

The graph of this equation is shown in Figure above. To predict the extent of solvent evaporation when the relative humidity is 50 %, we substitute the value 50 for x in the equation, $\hat{y} = 13.64 - .08x$

To obtain $\hat{y} = 13.64 - .08(50) = 9.64$. That is, when there relative humidity is 50 % we predict that 9.64% of the solvent, by weight, will be lost due to evaporation.

Recall from elementary calculus that the slope of a line gives the change in y for a unit change in x. If the slope is positive, then as x increases so does y; as x decreases, so does y. If the slope is negative, things operate in reverse. An increase in x signals a decrease in y, whereas a decrease in x yields an increase in y.

Check Your Progress 2

1) $R = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = (138.076) / [(2.6929 * 7839.93)^{1/2}] = 0.9504.$

So, $R^2 = 0.9033$

Check Your Progress 3

1) Residual plots are helpful in spotting potential problems. However, they are not always easy to interpret. Residual patterns are hard to spot with small data set except in extreme cases, residual plots are most useful with fairly large collection of data.

Check Your Progress 4

1) Refer to example solved in the section 3.3.1

Table 1: The distribution function of standard normal random variable

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Table 2: The critical values of chi-square distribution. The areas given across the top are the areas to the right of the critical value. To look up an area on the left, subtract it from one, and then look it up (i.e., 0.05 on the left is 0.95 on the right).

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	-	-	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.597	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.19	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.17	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.341	46.928

Uttar Pradesh
Rajarshi Tandon Open University

MCA-5.3

Numerical and Statistical Computing

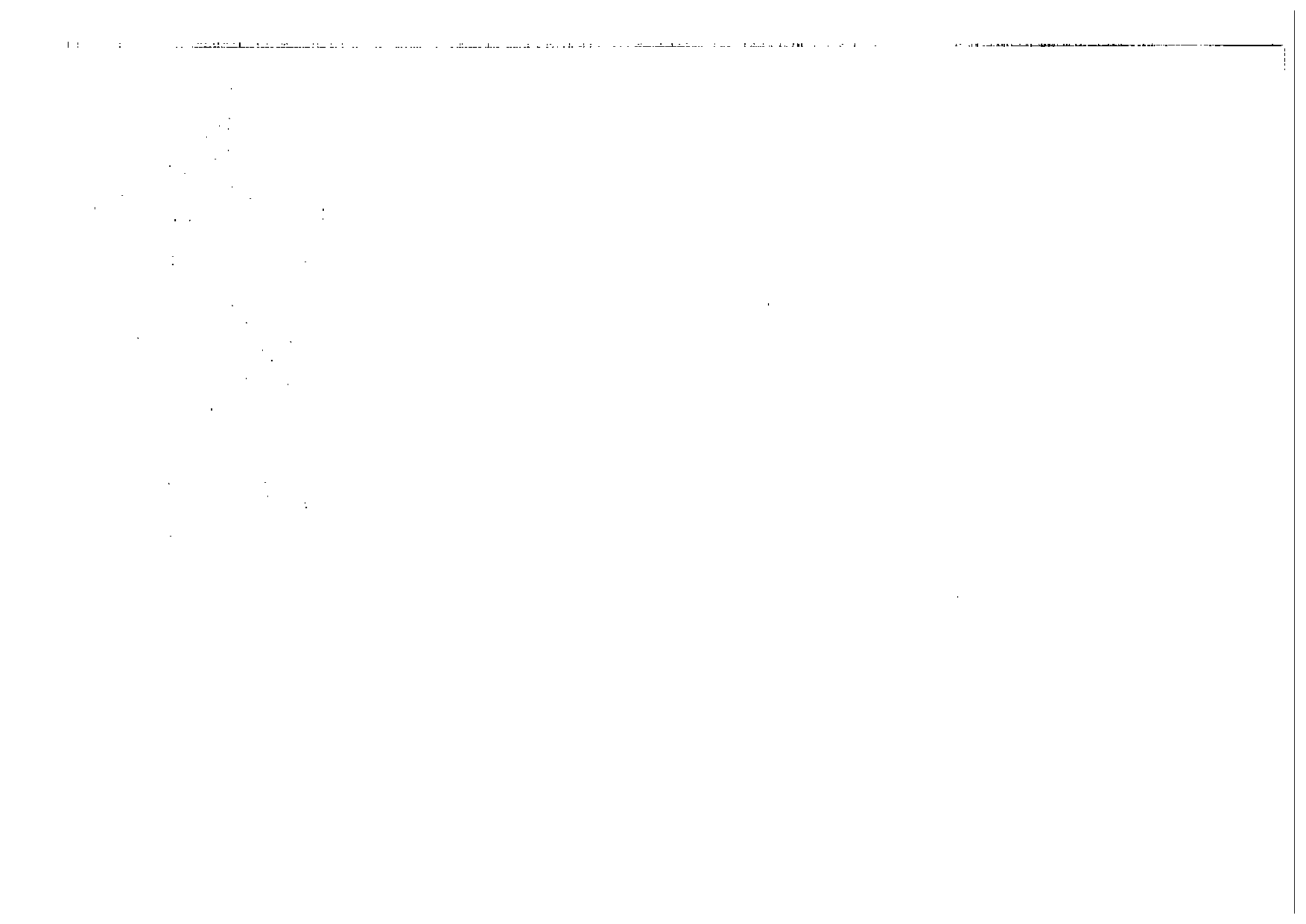
Block

2

NUMERICAL COMPUTING-II

BLOCK INTRODUCTION

This block is also composed of three units **Interpolation, Numerical Integration and Numerical Solution of ODE**. Interpolation is quite interesting because this technique will let you to learn the methods by which you can use the available data and hence determine the result of the experiment for which it was not conducted. You will understand the meaning while reading the text given in the block, secondly you will understand that determining values at quite close intervals contributes to the integration of the curve. So we had also discussed some methods for Numerical Integration and finally Numerical Solution of ODE is also discussed in this block.



UNIT 1 INTERPOLATION

Structure **Page No.**

1.0 INTRODUCTION

Interpolation is an interesting topic and has wide application in various fields. Using this concept you can analyse a problem on solution in a better way. So, as to understand and conside an example say a thermometer has a least count of 2.5°C, when it is calibrated than a correction chart is given by the concerned authority. Say the correction chart has following entries.

Actual Reading(°C)	Correction
5	+0.8
10	–0.4

i.e., when reading in device is 5°C, the actual temperature is 5.8°C, and when it is 10°C, the actual temperature is 9.6°C. Say, the device is used in an experiment and the mercury level is stopped between 5°C and 10°C. Now, what will you say about the temperature correction at 7.5°C. Since, the device is not calibrated at this point, we use the concept of interpolation to solve such tasks.

Calculation of the value of a function between the values already known is called interpolation. The problem of interpolation can be briefly stated as follows. A function f(x) is defined by a table of its values for a certain finite set of values X_i, i=0(1)n. Compute the value of the function at some point ξ not in the table. If the point ξ is included in the interval $[X_0, X_n]$, it is called the problem of interpolation. It is obvious that, for any general f(x) and any arbitrary ξ, the solution to the problem is difficult unless certain assumptions are made regarding the nature of f(x). Informally, f(x) is assumed regular or more precisely the tabulated function is amenable for approximation by some type of function y(x), preferably using polynomials or trigonometric functions. The problem of interpolation is to compute y(x). In order to reduce computational effort over a certain interval, it may be required on some occasions to resort to interpolation, even if an analytical representation of f(x) is known. The other type of problem, called inverse interpolation, consists in finding the value of the argument x corresponding to a given value of y. The problem is similar to direct interpolation since the roles of x and y can be interchanged. In particular, inverse interpolation can be used for computing the approximate root of a function, i.e., for obtaining x in the interval (x_i, x_{i+1}), which brackets the zero of f(x).

1.1 OBJECTIVES

After studying this unit, you should be able to:

- solve to Interpolation problem;
- evaluation of differences of any function;
- use forward and backward differences;
- use of Newton's forward and backward differences formulas, and
- use of Lagrange's Interpolation formula.

1.2 DIFFERENCES-FORWARD AND BACKWARD DIFFERENCES

Calculation of the value of a function between the values already known is called Interpolation.

The problem is as follows. Given a set of pairs of values (x_i,y_i), i=0(1)n, obtain an n^{th} degree polynomial y(x) that passes through all the points (x_i,y_i). This polynomial is an approximation to the function f(x), which coincides with the polynomial at (x_i,y_i), i=0(1)n.

In such a problem, the concept of differences is important.

Forward differences These are defined as

first forward differences $\Delta y_i = y_{i+1} - y_i \qquad i = 0(1)n-1$

second forward differences $\Delta^2 y_i = \Delta(\Delta y_i)$

$$= \Delta(y_{i+1} - y_i)$$
$$= \Delta y_{i+1} - \Delta y_i$$
$$= (y_{i+2} - y_{i+1}) - (y_{i+1} - y_i)$$
$$= y_{i+2} - 2y_{i+1} + y_i$$

k^{th} forward difference $\Delta^k y_i = \Delta^{k-1} y_{i+1} - \Delta^{k-1} y_i \qquad i = 0(1)n-k \qquad (1)$

These differences are usually expressed as in *Table 1*, the quantities in each column representing the differences between the quantities in the preceding column. These are usually placed midway between the quantities being subtracted so that the forward differences with like subscripts lie along the diagonals indicated in the table by arrows. It should be noted that if the r^{th} differences $\Delta^r y_i$ are constants, then all the differences of an order higher than r are zero.

If follows, from formula (1), that

$$y_1 = y_0 + \Delta y_0$$
$$y_2 = y_1 + \Delta y_1 = (y_0 + \Delta y_0) + (\Delta^2 y_0 + \Delta y_0) = y_0 + 2\Delta y_0 + \Delta^2 y_0$$

These results can be written symbolically as

$$y_1 = (1+\Delta)y_0,\ y_2 = (1+\Delta)^2 y_0, \ldots\ldots$$

in which $(1+\Delta)^k$ is an operator on y_0 with the exponent on the Δ indicating the order of the difference. The difference operator is analogous to the differential operator $D(\equiv d/dx)$.

By induction,

$$y_k = (1+\Delta)^k y_0,\ k = 1, 2, \ldots\ldots\ldots, \qquad (2)$$

or, in expanded form,

$$y_k = y_0 + k\Delta y_0 + \frac{k(k-1)}{2!}\Delta^2 y_0 + \frac{k(k-1)(k-2)}{3!}\Delta^3 y_0 + \ldots\ldots \qquad (3)$$

Formula (3) enables us to write an expression of every value y_k in terms of y_0 and the forward differences $\Delta y_0, \Delta^2 y_0, \ldots\ldots\ldots$

Table 1: Forward differences

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
x_0	y_0	Δy_0	$\Delta^2 y_0$	$\Delta^3 y_0$
x_1	y_1	Δy_1	$\Delta^2 y_1$	
x_2	y_2	Δy_2		
x_3	y_3			
.	.			
.	.			
.	.			
.	.			
x_{n-1}	y_{n-1}	Δy_{n-1}		
x_n	y_n			

Remark: $\Delta^r P_n(x) = 0$ for $r > n$, where $P_n(x)$ is a polynomial of degree n.

Check Your Progress 1

1) Construct a forward difference task for the data:

x	:	1	2	3	4
f(a)	:	7	13	18	25

.....

.....

.....

2) Estimate the missing term in the following data valid it is represents a polynomial of degree.

x	:	1	2	3	4	5
f(x)	:	3	7	?	21	31

.....

.....

.....

Backward differences The pairs of points (x_i, y_i), i = 0 (1)n, are given. Then,

first backward differences $\nabla y_i = y_i - y_{i-1}$ i = n (–1) 1

second backward differences $\nabla^2 y_i = \nabla y_i - \nabla y_{i-1}$ i = n (–1) 2

k-th backward differences $\nabla^k y_i = \nabla^{k-1} y_i - \nabla^{k-1} y_{i-1}$ i = n(–1)k (4)

The backward differences are indicated in *Table 2*.

From formula (4),

$\nabla y_n = y_n - y_{n-1}$,

$\nabla^2 y_n = \nabla y_n - \nabla y_{n-1} = y_n - 2y_{n-1} + y_{n-2}$

These results can be written symbolically as

$y_{n-1} = y_n - \nabla y_n = (1 - \nabla) y_n$

$y_{n-2} = y_n - 2\nabla y_n + \nabla^2 y_n = (1 - \nabla)^2 y_n$

$y_{n-k} = (1 - \nabla)^k y_n$ (5)

in which $(1 - \nabla)^k$ is an operator on y_n with the exponent on the ∇ indicating the order of the differences. ∇ is called the backward differences operator. Formula (5), when expanded, reads

$$y_{n-k} = y_n - k\nabla y_n + \frac{k(k-1)}{2!}\nabla^2 y_n - \frac{k(k-1)(k-2)}{3!}\nabla^3 y_n + \ldots\ldots \quad (6)$$

Formula (6) enables us to represent every value y_{n-k} in terms of y_n and the backward differences $\nabla^k y_n$.

Table 2: Backward differences

X	Y	∇y	$\nabla^2 y$	$\nabla^3 y$
x_0	y_0			
.	.			
.	.			
.	.			
.	.			
x_{n-3}	y_{n-3}			
x_{n-2}	y_{n-2}	∇y_{n-2}		
x_{n-1}	y_{n-1}	∇y_{n-1}	$\nabla^2 y_{n-1}$	
x_n	y_n	∇y_n	$\nabla^2 y_n$	$\nabla^3 y_n$

Example 1: Compute the forward differences for the following set of data:

x	1	2	3	4	5	6	7	8
y	2.105	2.808	3.614	4.604	5.857	7.451	9.467	11.985

x	Y	∇y	$\nabla^2 y$	$\nabla^3 y$	$\nabla^4 y$
1	2.105				
		0.703			
2	2.808		0.103		
		0.806		0.081	
3	3.614		0.184		−0.002
		0.990		0.079	
4	4.604		0.263		−0.001
		1.253		0.078	
5	5.857		0.341		+0.003
		1.594		0.081	
6	7.451		0.422		−0.001
		2.016		0.080	
7	9.467		0.502		
		2.518			
8	11.985				

Check Your Progress 2

1) Evaluate the missing term in the following:

x	:	100	101	102	103	104
f(x)=log x	:	2.000	2.0043	?	2.0128	2.0170

.....

.....

.....

2) Obtain the estimate of the missing figure in the following table:

x	:	1	2	3	4	5	6	7	8
f(x)	:	1	8	?	64	?	216	343	512

.....

.....

.....

1.3 NEWTON'S FORWARD DIFFERENCE INTERPOLATION FORMULA

For a given set of data (x_i, y_i), $i = 0\ (1)n$, let x_i be equally spaced and h be the interval size, i.e.,

$x_k = x_0 + kh, \qquad v_i = 0, 1, \ldots\ldots\ldots, n$ (7)

Then, $k = \frac{x_k - x_0}{h}$ (8)

On inserting this value of k in

$$y_k = y_0 + k\Delta y_0 + \frac{k(k-1)}{2!}\Delta^2 y_0 + \ldots\ldots\ldots\ldots + \frac{k(k-1)\ldots(k-n+1)}{n!}\Delta^n y_0 \quad (9)$$

we get

$$y_k = y_0 + \frac{x_k - x_0}{h}\Delta y_0 + \frac{(x_k - x_0)(x_k - x_0 - h)}{2!h^2}\Delta^2 y_0 + \ldots\ldots\ldots +$$
$$\frac{(x_k - x_0)(x_k - x_0 - h)\ldots\ldots(x_k - x_0 - nh + h)}{n!h^n}\Delta^n y_0. \quad (10)$$

This relation is satisfied by n + 1 pairs of the tabulated values. Assuming that the value of y corresponding to an arbitrary x can be obtained from formula (10) by replacing x_k by x, we get,

$$y(x) = y_0 + \frac{x - x_0}{h}\Delta y_0 + \frac{(x - x_0)(x - x_0 - h)}{2!h^2}\Delta^2 y_0 + \ldots\ldots\ldots$$
$$+ \frac{(x - x_0)(x - x_0 - h)\ldots\ldots(x - x_0 - nh + h)}{n!h^n}\Delta^n y_0$$

$$= y_0 + \left(\frac{x - x_0}{n}\right)\Delta y_0 + \frac{(x - x_0)(x - x_1)}{2!n^2}\Delta^2 y_0 + \frac{(x - x_0)(x - x_1)(x - x_2)}{3!h^3}\Delta^3 y_0 \quad (11)$$

Formula (11) is called the Newton forward difference interpolation formula.

Let $X = (x - x_0)/h$ be an undimensioned variable, which represents the distance of x from x_0 in the units of h. Then, Formula (11) becomes

$$y_x = y_0 + X\Delta y_0 + \frac{X(X-1)}{2!}\Delta^2 y_0 + \ldots\ldots + \frac{X(X-1)\ldots\ldots(X-n+1)}{n!}\Delta^n y_0. \quad (12)$$

where $y_x = y(x_0 + hX) = y(x)$. y(x) is known as the Newton-Gregory forward polynomial.

Check Your Progress 3

1) Write down the polynomial of lowest degree which satisfies the following set of numbers, using the forward difference polynomial

x	0	1	2	3	4	5	6	7
f(x)	0	7	26	63	124	215	342	511

.....

.....

.....

.....

2) What is the lowest degree polynomial, which takes the following values? Find this polynomial using the forward difference polynomial.

x	:	0	1	2	3	4	5
f(x)	:	0	3	8	15	24	35

.....

.....

.....

.....

1.4 NEWTON'S BACKWARD DIFFERENCE INTERPOLATION FORMULA

In the same way as in Section 1.3, we get from formula (6), the Newton backward difference interpolation formula

$$y(x) = y_n + \frac{x-x_n}{n}\nabla y_n + \frac{(x-x_n)(x-x_{n-1})}{2!h^2} + \ldots\ldots \frac{(x-x_n)\ldots(x-x_1)}{x!h^2}\nabla^n y_n$$

$$y_{n+x} = y_n + X\nabla y_n + \frac{X(X+1)}{2!}\nabla^2 y_n + \ldots\ldots + \frac{X(X+1)\ldots\ldots(X+n-1)}{n!}\nabla^n y_n, \quad (13)$$

where $X = (x - x_n)/h$ and $y_{n+x} = y(x_n + hX) = y(x)$. $y(x)$ is known as the **Newton-Gregory backward polynomial.**

Derivatives of Tabulated Functions

On differentiating formula (11) successively with respect to x and setting $x = x_0$ in the result, we get

$$y'(x_0) = \frac{1}{h}\left(\Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 - \frac{1}{4}\Delta^4 y_0 + \frac{1}{5}\Delta^5 y_0 - \ldots\ldots\ldots\right),$$

$$y''(x_0) = \frac{1}{h^2}\left(\Delta^2 y_0 - \Delta^3 y_0 + \frac{11}{12}\Delta^4 y_0 - \frac{5}{6}\Delta^5 y_0 + \ldots\ldots\ldots\ldots\right),$$

$$y'''(x_0) = \frac{1}{h^3}\left(\Delta^3 y_0 - \frac{3}{2}\Delta^4 y_0 - \frac{7}{4}\Delta^5 y_0 - \ldots\ldots\ldots\right),$$

$$y^{(4)}(x_0) = \frac{1}{h^4}\left(\Delta^4 y_0 - 2\Delta^5 y_0 + \ldots\ldots\ldots\ldots\right) \quad (14)$$

Example 2: Given the values

x	1	2	3	4	5.	6	7	8
y	2.105	2.808	3.614	4.604	5.857	7.451	9.467	11.985

Obtain y(2.2) using forward differencing and linear, quadratic and cubic interpolation.

Solution: We form the table of forward differences.

		Δ	Δ^2	Δ^3
1	2.105	0.703	0.103	
2	2.805	0.806		0.081
3	3.614	0.990		
4	4.604			

If two nearest points are used (n = 1), i.e., linear interpolation, then $x_0 = 2$, $y_0 = 2.808$ $x_1 = 2$, $y_1 = 2.808$, $h = 1$.

We get (note x_0 is a suitable origin).

$$y(x) = y_0 + (x - x_0)\frac{\Delta f_0}{n}$$

$$\begin{aligned} y(2.2) &= y_0 + 0.2\,\Delta f_0 \\ &= 2.808 + 0.2\,(0.806) \\ &= 2.9692 \end{aligned}$$

If three nearest points are used (n = 2), i.e., quadratic interpolation, then we have
$x_0 = 1$, $y_0 = 2.105$
$x_1 = 2$, $y_1 = 2.808$, $x_2 = 3$, y2 = 3.614.

Then, $X = (2.2 - 1)/1 = 1.2$

Hence, using (12) we get,

$$y(1.2) = 2.105 + 1.2\,(0.703) + \frac{(1.2)(0.2)}{2!}(0.103) = 2.961.$$

It should be noted that we could have as well defined $x_0 = 2$, $y_0 = 2.808$, $\Delta y_0 = 0.806$, $\Delta^2 y_0 = 0.184$, $X = (2.2 - 2)/1 = 0.2$.

If four nearest points are used (n = 3), i.e., cubic interpolation, then we take $x_0 = 1$, $x_1 = 2$, $x_2 = 3$, $x_3 = 4$. Then, $X = (2.2 - 1)/1 = 1.2$.

Using (12), we get,

$$y(2.2) = 2.105 + 1.2(0.703) + \frac{(1.2)(0.2)}{2!}(0.103) + \frac{(1.2)(0.2)(-0.8)}{6}(0.081) = 2.958.$$

Error in Newton's Interpolation polynomial

A data of n+1 values can be represented by a unique polynomial of degree $\leq$ n. Hence, the Newton forward and backward differences interpolation formulae are basically the same, and have the same error bound. Both these formulae give an n-th degree polynomial y(x) passing through (n + 1) given points (x_i, y_i), i = 0(1)n. Hence, the error involved is the same as that described for the Lagrange interpolation (to be derived in Section 1.5). The maximum absolute error is given by

$$E_1 = \max \left|\Pi_0^n (x - x_i)\right| \max \left|\frac{f^{(n+1)}(\xi)}{(n+1)!}\right|, x_0 \leq \xi \leq x_n,$$

where x is in (x_0, x_n) and $\Pi^n{}_0 = (x-x_0)(x-x_1).....(x-x_n)$.

Error in linear interpolation: The maximum absolute error in estimating f(x) by linear interpolation between x_0 and x_1 is as follows. Since, n = 1, we have

$$E_1 = \max|\Pi_0^1(x-x_i)|\max\left|\frac{f^{\cdot}(\xi)}{2!}\right|, \qquad x_0 \le \xi \le x_1,$$

$$= \max|(x-x_0)(x-x_1)|\max\left|\frac{f^{\cdot}(\xi)}{2!}\right|, \qquad x_0 \le \xi \le x_1,$$

From calculus, in stationary points of the quantity $(x-x_0)(x-x_1)$ is obtained from

$$\frac{d}{dx}\{(x-x_0)(x-x_1)\}=0 \text{ or } \qquad x=\frac{x_0+x_1}{2}$$

Hence, $\max|(x-x_0)(x-x_1)| = \left|\frac{x_1-x_0}{2}.\frac{x_0-x_1}{2}\right| = \frac{(x_1-x_0)^2}{4}$

Thus, $E_1 = \frac{(x_1-x_0)^2}{8}\max|f''(\xi)|, \quad x_0 \le \xi \le x_1.$

Error in parabolic interpolation: In this case, n = 2.

Therefore, $E_1 = \max\left|\Pi_0^2(x-x_i)\right|\max\left|\frac{f'''(\xi)}{2!}\right| \qquad , x_0 \le \xi \le x_2,$

where, $\Pi^2{}_0 = (x-x_0)(x-x_1)(x-x_2)$

We invoke the value of X = (x – x_0)/h. Thus,

$$E_1 = \max\left|X(X-1)(X-2)h^3\right|\max\left|\frac{f'''(\xi)}{3!}\right| \qquad , x_0 \le \xi \le x_2,$$

$$E_1 = h^3\max\left|X(X-1)(X-2)\right|\max\left|\frac{f'''(\xi)}{3!}\right| \qquad , x_0 \le \xi \le x_2,$$

The points of the quantity x (x – 1) (x – 2) are stationary

$$\frac{d}{dx}\{X\,(X-1)(X-2)\}=0$$

or $\frac{d}{dx}\{x^3-3x^2+2x\}=0$ or $3x^2-6x+2=0$

That is, if $X = 1 \pm (1/\sqrt{3})$. When $X = 1+(1/\sqrt{3})$ or $1-(1/\sqrt{3})$,

Then, $|X(X-1)(X-2)| = 2/(3\sqrt{3})$.

Thus, $E_1 = \frac{h^3}{9\sqrt{3}} \max|f'''(\xi)|, \qquad x_0 \le \xi \le x_2.$

Since, $|x-x_0| < |x_n-x_0|, ..., |x-x_{n-1}| < |x_n-x_0|$ for $x_0 < x < x_n$, we get the error in the (n+1) point interpolation as

$$|Error| \le \frac{|x_n-x_0|^{n+1}}{(n+1)!}\max\left|f^{(n+1)}(\xi)\right|.$$

Note that this is a very rough upper bound. For n=1, we get

$$|Error| \le \frac{1}{2}|x_1 - x_0|^2 \max|f''(\xi)|$$

and for n = 2, we get

$$|Error| \le \frac{1}{6}|x_1 - x_0|^2 \max|f'''(\xi)|.$$

The values on the right hand sides are very large compared to the values derived earlier.

Example 3: What is the interpolating polynomial for $f(x) = x^2 + \sin\Pi x$ through (0, 0), (1,1), (2,4)? What is the error when $x = ½$? What is the maximum error?

Solution: We from the following table,

x	y	Δy	$\Delta^2 y$
0	0		
		1	
1	1		2.
		3	
2	4		

The interpolating polynomial is

$$y(x) = y(0) + x\frac{\Delta y_0}{1!} + x(x-1)\frac{\Delta^2 y_0}{2!} = 0 + x + x(x-1)\frac{2}{2!} = x^2$$

Now, the exact and computed values are,

$$f\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^2 + \sin(\pi/2) = \frac{1}{4} + 1 = \frac{5}{4},$$

$$y\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^2 = \frac{1}{4}.$$

Therefore, the exact error is $f\left(\frac{1}{2}\right) - y\left(\frac{1}{2}\right) = \frac{5}{4} - \frac{1}{4} = 1$

The maximum error in quadratic interpolation is

$$E_1 = \frac{h^3}{9\sqrt{3}} \max|f'''(\xi)|, 0 \le \xi \le 2$$

(say $p \Rightarrow \pi$)

Hence, $f(x) = x^2 + \sin p\,x$

$f'(x) = 2x + p\cos p\,x$

$f''(x) = 2 - p^2 \sin p\,x$

$f'''(x) = -p^3 \cos p\,x$

Hence, $E_1 = \frac{1}{9\sqrt{3}} \max|-\pi^3 \cos\pi\xi|, \quad 0 \le \xi \le 2.$

Since, $\cos\xi = 1$, at $\xi = 0$, we have $E_1 = \frac{p^3}{\sqrt{3}}$.

Example 4: Find the largest value of h that will ensure five-place accuracy in the table of sin x.

Solution: The maximum error in linear interpolation should satisfy,

$$\frac{h^2}{8} \max |f''(\xi)| \leq .000005$$

Since, $f'' = -\sin x$ and $\max |f''| = 1$, we require, $h^2/8 \leq .000005$.
Therefore, h = 0.0063 rad.

Check Your Progress 4

From the following data estimate the value of f(2.25) using forward difference formula.

x :	0	0.5	1.0	1.5	2.0	2.5
f(x):	1.0	3.625	7.000	11.875	19.000	29.125

.....

.....

.....

2) Estimate the sale of a particular quantity for 1966 using the following table:

Year :	1931	1941	1951	1961	1971	1981
Sale in thousands :	12	15	20	27	39	52

.....

.....

.....

1.5 LAGRANGE'S INTERPOLATION FORMULA

Give (x_i, y_i), i = 0(1)n, where x_i may or may not be equally spaced, the problem is to obtain an n-th degree polynomial y(x) that passes through all the points (x_i, y_i). This polynomial is an approximation to the function f(x), which coincides with the polynomial at (x_i, y_i), i = 0(1)n,

Let the n-th degree polynomial $P_k(x)$, k = 0(1)n

$$P_k(x) = \prod_{i=0}^{n} (x - x_i), \text{ for } i \neq k, k = 0(1)n \qquad (15)$$

That is, $P_k(x) = (x - x_0).....(x - x_{k-1})(x - x_{k+1}).....(x - x_n)$

Then, the coefficients A_k of

$$y(x) = \sum_{k=0}^{n} A_k P_k(x) \qquad (16)$$

can be determined so that equation (16) is satisfied by each of n + 1 pairs (x_i, y_i).
For, if $x = x_k$, then

$$y(x_k) = y_k = A_0 P_0(x_k) + A_1 P_1(x_k) + ... + A_k P_k(x_k) + ... + A_n P_n(x_n)$$
$$= A_k P_k(x_k)$$

Therefore, $A_k = \frac{y_k}{P_k(x_k)}$ (17)

since $P_k(x_i) = 0$, if $i \neq k$.

Hence, $y(x) = \sum_{k=0}^{n} \frac{y_k P_k(x)}{P_k(x_k)}$ (18)

is the equation of the required n-th degree polynomial, which passes through the n + 1 points (x_k, y_i). Equation (18) is called the *Lagrange interpolation formula*.

Remainder and Error in Lagrange's interpolation formula

Let $P(x) = \prod_{i=0}^{n} (x - x_i)$. (19)

From (15), we have $P(x) = (x - x_k) P_k(x)$ and $P'(x) = (x - x_k) P'_k(x) + P_k(x)$.

Hence, $P'_k(x_k) = P'_k(x_k)$ (20)

Thus, we can also write y(x) from (18) as

$$y(x) = \sum_{k=0}^{n} \frac{y_k p_k(x)}{p_k(x_k)} = \sum_{k=0}^{n} \frac{y_k P(x)}{(x - x_k) P'(x_k)}. \quad (21)$$

Let α be a point in $[x_0, x_n]$ and $\alpha \neq x_k$, $k = 0(1)n$. Then, the error in interpolation is defined as $g(x) = f(x) - y(x)$.

The function $f(x) - y(x)$ vanishes for $x = x_0, x_1, \ldots x_n$. We assume that, for $x = \alpha$, we have

$f(x) = y(x) + RP(x)$ (22)

Where R is a constant. Now, define a function
$G(t) = f(t) - y(t) - RP(t)$ (23)

Where P(t) is a polynomial of degree n+1 in t and y(t) is a polynomial of degree n.

G(t) is zero for $t = x_0, x_1, \ldots x_n$. By using the Roller's theorem repeatedly, we conclude that $G^{(n+1)}(\xi) = 0$, where ξ is in $[x_0, x_n]$.

But, $G^{(n+1)}(\xi) = f^{(n+1)}(\xi) - R(n + 1)$ (24)

Since (n+1)th derivative (n+1)! of y(t) = 0 and (n+1)th derivative of P(t) = (n+1)!

Hence, $R = \frac{f^{(n+1)}(\xi)}{(n+1)!}$ (25)

Therefore, we obtain

$$f(x) = y(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!} P(x) \quad (26)$$

or $$f(x) = y(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)(x - x_1) \ldots\ldots (x - x_n) \quad (27)$$

Thus, the remainder in the Lagrange interpolation formula is given by

$$\text{Error} = f(x) - y(x) = \prod_{i=0}^{n}(x - x_i)\frac{f^{(n+1)}(\xi)}{(n+1)!}, \qquad x_0 \le \xi \le x_n \tag{28}$$

where x is in $[x_0, x_n]$. Hence, the maximum absolute error in the Lagrange interpolation formula is

$$E_1 = \max\left|\prod_{i=0}^{n}(x - x_i)\right| \max\left|\frac{f^{(n+1)}(\xi)}{(n+1)!}\right|, \quad x_0 \le \xi \le x_n \tag{29}$$

Remarks:

(a) A higher order formula does not necessarily give a better result than a lower order formula. For instance, let us consider the curve $y = x^2$. Also, suppose the values on this curve are given by

X	0	1	2	3	4
Y	0	1	8	27	64

For computing x when y = 20. Let us use the inverse Lagrange interpolation. This is obtained by interchanging the roles of x and y in equation (18). Thus,

$$x = \frac{(y-1)(y-8)(y-27)(y-64)}{(-1)(-8)(-27)(-64)}(0) + \frac{(y-0)(y-8)(y-27)(y-64)}{(1)(-7)(-26)(-63)}(1) + \frac{(y-0)(y-1)(y-27)(y-64)}{(8)(7)(-19)(-56)}(2) + \frac{(y-0)(y-1)(y-8)(y-64)}{(27)(26)(19)(-37)}(3) + \frac{(y-0)(y-1)(y-8)(y-27)}{(64)(63)(56)(37)}(4)$$

For y = 20, we obtain x = -1.31. The actual value is x = 2.71. Now, let us use linear interpolation with the data points (8,2), (27,3). We obtain

$$x = \frac{y-27}{(-19)}(2) + \frac{y-8}{(19)}(3).$$

For y = 20, we get x = 2.63, with the magnitude of error as only 0.03.

(b) The Lagrange formula is often, not used in practical computation. However, it is very useful in theoretical work within different branches in numerical analysis (for instance, deriving the Gauss – Legendre quadrature formula).

Example 4: Find the interpolating polynomial that fits the data:

x_k	0	1	2	5
f_k	2	3	12	147

Using the Lagrange interpolation formula

Solution: We have

$$f(x) = \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)}f(x_0) + \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}f(x_1) + \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)}f(x_2) + \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}f(x_3)$$

$$= \frac{(x-1)(x-2)(x-5)}{(0-1)(0-2)(0-5)}2 + \frac{(x-0)(x-2)(x-5)}{(1-0)(1-2)(1-5)}3 +$$
$$\frac{(x-0)(x-1)(x-5)}{(2-0)(2-1)(2-5)}12 + \frac{(x-0)(x-1)(x-2)}{(5-0)(5-1)(5-2)}147$$
$$= -\frac{(x-1)(x-2)(x-5)}{5} + \frac{3}{4}x(x-2)(x-5) - 2x(x-1)(x-5) + \frac{49}{20}x(x-1)(x-2)$$

Example 5: Compute f'(2.0) from the following data of values

X	1.8	1.9	2.0	2.1
f(x)	6.05	6.69	7.39	8.17

Solution:

Since, four data points are given, the data represents a cubic polynomial, P(x). We can construct the polynomial P(x)' and find P'(2.0). Alternatively, we can construct a quadratic polynomial P(x) using the data values (1.9, 6.69), (2.0, 7.39), (2.1, 8.17) and find P'(2.0). We have,

$$P(x) = \frac{(x-1.9)(x-2.0)(x-2.1)}{(-0.1)(0.2)(0.3)}(6.05) + \frac{(x-1.8)(x-2.0)(x-2.1)}{(-0.1)(0.1)(0.2)}(6.69) +$$
$$\frac{(x-1.8)(x-1.9)(x-2.1)}{(-0.2)(0.1)(-0.1)}(7.39) + \frac{(x-1.8)(x-1.9)(x-2.0)}{(0.3)(0.2)(0.1)}(8.17)$$

We have,

$$P'(x) = -\left(\frac{6.05}{0.006}\right)(3x^2 - 12x + 11.99) + \left(\frac{6.69}{0.002}\right)(3x^2 - 11.8x + 11.58)$$
$$-\left(\frac{7.39}{0.002}\right)(3x^2 - 11.6x + 11.19) + \left(\frac{8.17}{0.006}\right)(3x^2 - 11.4x + 10.82)$$

Hence, P'(2.0) = 10.08 – 66.9 + 39.95 + 27.33 = 7.46

The quadratic polynomial is given by

$$P(x) = \frac{(x-2)(x-2.1)}{(-0.1)(-0.2)}(6.69) + \frac{(x-1.9)(x-2.1)}{(0.1)(-0.1)}(7.39) + \frac{(x-1.9)(x-2.0)}{(0.2)(0.1)}(8.17)$$

$$P'(x) = +334.5(2x - 4.1) - 739(2x - 4) + 408.5(2x - 3.9)$$

Hence, P'(2.0) = – 33.45 + 40.85 = 7.40

Example 6: Obtain a second degree polynomial approximation to l_n x by expanding the function as a Taylor series about $x_0 = 1$. Calculate l_n 1.2 and obtain a bound for the truncation error.

Solution: We have, $f(x) = I_n x$, $f'(x) = \frac{1}{x}$, $f'(x) = -\frac{1}{x^2}$, $f''(x) = \frac{2}{x^3}$.

Writing Taylor series about $x_0 = 1$, we get the second degree polynomial approximation as

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x-x_0)^2}{2!}f''(x_0)$$

$$l_n x = l_n 1 + (x-1)(1) + \frac{(x-1)^2}{2}(-1) = (x-1) - \frac{1}{2}(x-1)^2$$

Now, $l_n 1.2 = 0.2 - \frac{1}{2}(0.2)^2 = 0.18$

$$|\text{Truncation error}| = \left|\frac{(x-1)^3}{6} f'''(\xi)\right| \qquad , 1 \le \xi \le 1.2$$

Since, $\max_{[1,1.2]} |f'''(\xi)| = \max_{[1,1.2]} \left|\frac{2}{x^3}\right| = 2$, and $\max_{[1,1.2]} |x-1|^3 = 0.008$

We get, $|\text{Error}| \le \frac{(0.008)^2}{6} = 0.00267$

Example 7: What is linear interpolation? Use linear interpolation to find f(0.3) for $f(x) = 5^x$.

Solution: Let a function f(x) be given at two points x_0, x_1, i.e. the data is $(x_0, f(x_0))$ and $f(x_1, f(x_1))$.Take the data at the points $x_0 = 0$, $x_1 = 1$,

Then, linear interpolation gives the value of f (x) at $x = x_a$, $x_0 < x_a < x_1$ as

$$f(x_a) = \frac{(x_a - x_1)}{x_0 - x_1} f(x_0) + \frac{(x_a - x_0)}{x_1 - x_0} f(x_1).$$

This process of finding f (x_a) is called linear interpolation.

Now, $f(x) = 5^x$. Hence, the data is (0, 1), (1, 5).

By the above formula, we get,

$$f(0.3) = \frac{0.3-1}{0-1}(1) + \frac{0.3-0}{1-0}(5) = 2.2$$

Example 8: Given f(x) = sin x, f(0.1) = 0.09983, f(0.2) = 0.19867, use the method of linear interpolation to find f(0.16). Find the error in f(0.16).

Solution

Using the Lagrange interpolation formula, we obtain

$$f(0.16) = \frac{0.16-0.2}{0.1-0.2}(0.09983) + \frac{0.16-0.1}{0.2-0.1}(0.19867)$$

$= 0.039932 + 0.119202 = 0.159134.$

The error in the Lagrange linear interpolation formula is given by

$$|Error| \le \frac{(x_1 - x_0)^2}{8} \max_{[0.1,0.2]} |f'''(\xi)| = (0.00125)(0.19867) = 0.00025$$

Example 9: Find a polynomial of degree ≤2 with the properties p(1) = 5, p(1.5) = – 3, p(3) = 0.

Solution: The data given is (1, 5), (1.5, –3), (3, 0).Using the Lagrange interpolation formula, we obtain the second degree polynomial as

$$p(x) = \frac{(x-1.5)(x-3)}{(1-1.5)(1-3)}(5) + \frac{(x-1)(x-3)}{(1.5-1)(1.5-3)}(-3) + \frac{(x-1.5)(x-1.5)}{(3-1)(3-1.5)}(0)$$

$= 5(x - 1.5)(x - 3) + 4(x - 1)(x - 3) = 9x^2 - 38.5x + 34.5$

Example 10: Sin x is to be computed from a table of values based on the points 1.0, 0.9, ..., 1.0 – 0.1 n. How big should n be to guarantee that |error| < 0.00001 ?

Solution: The truncation error is given by

$$|\text{error}| \le \frac{|x_n - x_0|^{n+1}}{(n+1)!} \max \left| f^{(n+1)}(\xi) \right|, \qquad x_0 \le \xi \le x_n.$$

From the given data, we get $x_n - x_0 = (1.0 - 0.1n) - 1.0 = 0.1n$.
Also, max |sin x| = 1 = max | (cos x) |.

Hence, we require

$$\frac{(0.1n)^{n+1}}{(n+1)!} < 0.00001$$

We find n = 6

Check Your Progress 5

1) In the following table h is the height above the sea level and p the barometric pressure. Calculate p when $h = 5280$.

h	:	0	4763	6942	10594
p	:	27	25	23	20

.....

.....

.....

2) The following table is given:

x	:	0	1	2	5
f(x)	:	2	3	12	147

Find the interpolating polynomial that fits this data.

.....

.....

.....

3) The following values of the function f(x) for values of x are given:
f(1) = 4, f(2) =5, f(7) = 5, f(8) = 4.

Find the value of f(6) and also the value of x for which f(x) is maximum or minimum.

.....

.....

.....

4) By means of Lagrange's formula, prove that
$y_1 = y_3 - 0.3(y_5 - y_{-3}) + 0.2(y_{-1} - y_{-5})$
where $y_i = y(i)$.

.....

.....

.....

1.6 SUMMARY

In this Unit on Interpolation, we have:

1) defined forward and backward differences.

2) derived the Newton-Gregory forward and backward difference formulas.

3) derived the errors in Newton forward and backward formulas.

4) derived the Lagrange's formula

$$y(x) = \sum_{k=0}^{n} \frac{y_k P_k(x)}{P_k(x_k)}$$

where $P_k(x) = (x - x_0)(x - x_1).....(x - x_{k-1})(x - x_{k+1}).....(x - x_n)$

5) derived error in Lagrange's formula

1.7 SOLUTIONS/ANSWERS

Check Your Progress 1

1) Since four values of f(x) are known, we can assume that f(x) can be represented by a polynomial of degree three

x	f(x)	Δf	Δ^2	Δ^3
1	7			
		6		
2	13		– 1	
		5		3
3	18		2	
		7		
4	25			

Since four data values are given, a polynomial of degree 3 can be passed through them. Hence, $\Delta^4 P_3(x) = 0$ and third differences have the same value.

2) Let the missing valuc be a.

X	f(x)	Af	Δ^2	Δ^3
1	3			
2	7	4	a – 11	39 – 3a
3	A	a – 7	28 – 2a	3a - 39
4	21	21 – a	a – 11	
5	31	10		

Since, third degree are to be same 39 – 3a = 3a – 39, or 6a = 78 and a = 13.

Check Your Progress 2

1) The value for x = 102. as four values of f (x) are known, for x = 1 we get, $\Delta^4 f(x) = 0$.

So f(2) = 2.0086 i.e. log 102 = 2.0086.

2) As six values of [x, f(x)] are given therefore we may assume the function to be represented by a polynomial of degree five. So that $\Delta^5 f(x)$ = constant and

$\Delta^6 f(x) = 0$ (30)

For x = 1, (1) becomes
$f(5) + f(3) = 152$ (31)

For x = 2, (1) becomes
$10f(5) + 3f(3) = 1331$.

Solving (30) and (31), we get
$f(3) = 27$ and $f(5) = 125$.

Check Your Progress 3

1) Write the forward difference table

x	f(x)	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
0	0				
1	7	7			
2	26	19	12	6	
3	63	37	18	6	0
4	124	61	24	6	0
5	215	91	30	6	0
6	342	127	36	6	0
7	511	169	42		

Since $\Delta^4 f=0$ the data represents a polynomial of degree 3 Using equation (11) with h=1 we get $y(x)=x^3+3x^2+3x$

2) Proceed as 1) above. The answer is $x^2 + 2x$.

Check Your Progress 4

1) Write the backward difference table

x	F(x)	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
0	1.0				
0.5	3.625	2.625			
1.0	7.000	3.375	0.750		
1.5	11.875	4.875	1.500	0.150	0
2.0	19.000	7.125	2.250	0.750	
2.5	29.125	10.125	3.000	0.750	0

Since $\Delta^4 f = 0$, we note that the data represents a cubic polynomial. Now,

$$X = \frac{x-n}{a} = \frac{2.25-2.5}{0.5} = -0.5$$

Newton's backward difference formula gives

$$f(2.25) = f(X_n) + X\nabla f_n + \frac{X(X+1)}{2}\nabla^2 f_n + \frac{X(X+1)(X+2)}{6}\nabla^3 f_n$$

$$= 29.125 - 0.5\,(10.125) + \frac{(-0.5)(0.5)}{2}(3) + \frac{(-0.5)(0.5)(1.5)}{6}(0.75)$$

$=23.640625$

2) Proceed as 1) above, difference formula. Here $x_n = 1981$, $x = 1966$, use backward difference Answer is 32.3437.

Check Your Progress 5

1) Use Lagrange Interpolation p(52.80) =24.8

2) By Lagrange,s formula, we get, $f(x) = x^3 + x^2 - x + 2$.

3) Applying Lagrange,s formula for the above values, we get

$$f(x) = \frac{[-x2 + 9x + 16]}{6}$$

Hence, $f(6) \approx 5.667$.

The stationary points (points of maxima/minima) are the solutions of $f'(x)=0$. We get $f'(x) = \frac{1}{6}(9 - 2x) = 0$, which gives x = 4.5. Since, $f''(x) < 0$, f(x) has a maximum at X = 4.5.

4) We are required to obtain y_1 while $y_{5}, y_{-5}, y_{-3}, y_3$, are given. By Lagranges formula, we have

$$y_1 = \frac{(1+3)(1-3)(1-5)}{(-5+3)(-5-3)(-5-5)} y_{-5} + \frac{(1+5)(1-3)(1-5)}{(-3+5)(-3-3)(-3-5)} y_{-3}$$
$$+ \frac{(1+5)(1+3)(1-5)}{(3+5)(3+3)(3-5)} y_3 + \frac{(1+5)(1+3)(1-3)}{(5+5)(5+3)(5-3)} y_5$$
$$= y_1 - 0.3(y_5 - y_{-3}) + 0.2(y_{-3} - y_{-5}).$$

UNIT 2 NUMERICAL INTEGRATION

Structure **Page No.**

2.0 INTRODUCTION

In many physical problems, we are often required to integrate a function between two specified limits. The function may be known either explicitly or as a tabulation of data (equally or unequally spaced). The numerical methods can be used to solve all such problems. The methods we describe demand the knowledge of the function either

(i) at equidistant points (e.g., all Newton – Cotes closed quadrature rules, or
(ii) at points corresponding to the zeros of some orthogonal polynomial (e.g., the Gauss – Legendre open quadrature formula).

In case at least one of these two conditions does not hold, the integrand has to be represented by some interpolating polynomial and the polynomial is then integrated.

2.1 OBJECTIVES

After studying this unit, you should be able to:

- use Newton – Cotes formula;
- use composite formula, and
- use Gaussian Quadrature formula.

2.2 NEWTON – COTES FORMULAS

The Newton – Cotes formulas are the numerical integration formulas and arise when the interpolating polynomials are integrated over the entire interval at which they match the tabulated data. These formulae are used to compute the definite integral

$$\int_a^b f(x)dx \qquad (1)$$

where $f(x)$ is known either explicitly or as a tabulation of data (at equispaced points).

Let the function f be known through its values shown in table 1, where $x_i - x_{i-1} = h$, $i = 1(1)n$ (i.e., the ordinates f_i are known at equidistant points), $x_0 = a$,

$x_1 = x_0 + h, x_2 = x_2 = x_0 + 2G, ..., b = x_n = x_0 + nh$

Table 1: Function f as tabulation of Data (at equispaced points)

x	x_0	x_1	x_2	....	x_n
$f(x)$	f_0	f_1	f_2	....	f_n

We transform the limits of integration as follows. Let

$$t = \frac{x - x_0}{h} \tag{2}$$

Then, $dx = hdt$, $f(x) = f(x_0 + ht)$ when $x = x_0, t = 0$, and when $x = x_n =, t = n$ Thus, from integral (1),

Thus, from integral (1),

$$\int_a^b f(x)dx = \int_{x_0}^{x_n} f(x)dx = h\int_0^n f(x_0 + ht)dt = h\int_0^n F(t)dt, \tag{3}$$

where $F(t) = f(x_0 + ht)$.

The values in Table 1 can now be rewritten as in Table 2.

Table 2: Function values for equidistant values of transformed variable t

x	x_0	x_1	x_2	. . . .	x_n
$t = \frac{x - x_0}{h}$	0	1	2	. . . .	n
$F(t) = f(x_0 + ht)$	f_0	f_1	f_2	. . . .	f_n

Now we write, the Newton – Georgy forward difference polynomial for $F(t)$,as

$$F(t) = f_0 + t\Delta f_0 + \frac{t(t-1)}{2!}\Delta^2 f_0 + (t) + (t-1)(t-2)\frac{\Delta^3 f_0}{3} + ... + R_n$$

Where R_n is the even term given by

$$R_n = \frac{h^{n+1}}{(n+1)!}(t)(t-1)...(t-n+1)f^{(n+1)}(\xi), x_0 \le \xi \le x_n,$$

Integrating , we get (from (3)),

$$I = \int_{x_0}^{x_n} f(x)dx = h\int_0^n F(t)dt$$

$$= h\int_0^n [f_0 + \frac{\Delta f_0}{1!}t + \frac{\Delta^2 f_0}{2!}t^{(2)} + ... + \frac{\Delta^n f_0}{n!}t^{(n)} + h^{n+1}t^{(n+1)}\frac{f^{(n+1)}(\xi)}{(n+1)!}]dt, \tag{4}$$

2.2.1 Trapezoidal Rule over $[x_0, x_1]$

Equation (4) produces the Trapezoidal rule for $n = 1$ as

$$I = \int_{x_0}^{x_1} f(x)dx = h\int [f_0 + \frac{\Delta f_0}{1!}t + \frac{h^2}{2!}t^{(t-1)}f''(\xi)]dt \tag{5}$$

Hence,

$$I = h\left[f_0 t + \Delta f_0 \cdot \frac{t^2}{2} + \frac{h^2}{2!}\left(\frac{t^3}{3} - \frac{t^2}{2}\right) f''(\xi)\right]_0^1$$

$$= h\left[f_0 + \frac{\Delta f_0}{2} + \frac{h^2}{2!} f''(\xi)\left(-\frac{1}{6}\right)\right] = h\left[f_0 + \frac{f_1 - f_0}{2} + \left(-\frac{h^2}{12}\right) f''(\xi)\right]$$

$$= h\left[\frac{f_0 + f_1}{2} - \frac{h^2}{12} f''(\xi)\right]$$

Thus, the Trapezoidal Rule is given by

$$\int_{x_0}^{x_1} f(x)dx = \frac{h}{2}(f_0 + f_1) \tag{6}$$

With error term given by

$$R_1 = -\frac{n^2}{12} f''(\xi)\, x_0 < \xi > x_n \tag{7}$$

Geometrically, $(h/2)(f_0 + f_1)$ is the area of the Trapezoid. To obtain the upper bound of the error, choose ξ in $[x_0, x_1]$ such that $f''(\xi)$ has a largest magnitude. Similarly, to find the lower bound of the error, choose ξ in $[x_0, x_1]$ such that $f''(\xi)$ has a smallest magnitude.

The error in equation (7) that is

$$\frac{h^2}{12}\min|f''(x)| \leq |R_1| \leq \frac{h^2}{12}\max|f''(x)| \tag{8}$$

is the error of only a single step, and is hence known as the local error.

2.2.1 (a) Trapezoidal Rule over $[x_0, x_n]$ — Composite Rule

It is now easy to write the Trapezoidal rule over the whole of the closed interval $[x_0, x_n]$. We write

$$\int_{x_0}^{x_n} f(x)dx = \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \ldots\ldots + \int_{x_{n-1}}^{x_n} f(x)dx \tag{5}$$

Then, using the Trapezoid Rule, we obtain

$$\int_{x_0}^{x_n} f(x)dx = \frac{h}{2}(f_0 + f_1) - \frac{h^3}{12} f''(\xi_1) + \frac{h}{2}(f_1 + f_2) - \frac{h^3}{12} f''(\xi_2) + \ldots$$

$$+ \frac{h}{2}(f_{n-1} + f_n) - \frac{h^3}{12} f''(\xi_n) \tag{10}$$

where h=(x_n-x_0)/n ; $x_{i-1} \leq \xi_i \leq x_i$ for $i = 1(1)n$, or

$$\int_{x_0}^{x_n} f(x)dx = \frac{h}{2}(f_0 + 2f_1 + 2f_2 + \ldots + 2f_{n-1} + f_n) - \frac{h^3}{12}\left[f''(\xi_1)\right.$$

$$\left. + f''(\xi_2) + \ldots\ldots\ldots + f''(\xi_n)\right]. \tag{11}$$

Equation (11) represents the Trapezoidal rule over $[x_0, x_n]$ with the error term . Hence , the composite Trapezoid Rule is given by

$$\int_{x_0}^{x_n} f(x)dx = \frac{h}{2}(f_0 + 2f_1 + 2f_2 + ... + 2f_{n-1} + f_n) \tag{12}$$

The error term is given as

$$E_n = -\frac{1}{12}h^3[f''(\xi_1) + f''(\xi_2) + ... + f''(\xi_n)], \tag{13}$$

where $x_{i-1} \le \xi_i \le x_i$ for $i = 1(1)n$. The error E_n over n steps, is the total error and is called the global error. If we now assume that $f''(x)$ is continuous on (x_0, x_n), then

$|f''(\xi i)| \le \max_{[x_0, x_n]} |f(x)| = M$. (14)

The magnitude of the global error can be written as

$$|\text{error}| \le \frac{nh^3}{12}M = \frac{(x_n - x_0)}{12}h^2M \tag{15}$$

Since $nh = (x_n - x_0)$ and M is as given in (14)

It should be noted that while the local error is $O(h^3)$, the global error is $O(h^2)$.

Remark: This rule is simple to use. It can be applied to unequally spaced argument values. Thus, it is considerably important in computing the integral of an experimentally determined function. Let (x_i, f_i), $i = 0(1)n$, be the given set of points, where $x_i - x_{i-1} = h_i$, $i = 1(1)n$, are the unequal subintervals. Then, the Trapezoidal rule over $[x_0, x_n]$ is given by

$$\int_{x_0}^{x_n} f(x)dx = \frac{1}{2}\sum_{i=1}^{n} h_i(f_{i-1} + f_i). \tag{16}$$

2.2.2 Simpson's 1/3 rd Rule over $[x_0, x_2]$

To apply this rule, we need three date values. The data values are

$(x_0, f(x_0)), (x_1, f(x_1))$ and $(x_2, f(x_2)$, Let $t = \frac{(x - x_0)}{h}$

Then equation (4) gives

$$I = \int_{x_0}^{x_2} f(x)dx = h\int_0^2 F(t)dt = h\int_0^2 [f_0 + \frac{\Delta f_0}{1!}t + \frac{\Delta^2 f_0}{2!}t(t-1) + \frac{-h^3}{3!}(t)(t-1)(t-2)f^{(3)}]dt,$$

where $x_0 \le \xi \le x_2$, Therefore,

$$I = h[f_0 t + \Delta f_0 \frac{t^2}{2} + \frac{\Delta^2 f_0}{2!}(\frac{t^3}{3} - \frac{t^2}{2}) + (\xi] + \frac{h^3}{3!}(\frac{t^4}{4} - \frac{t^3}{3} + t^2)f^{(3)}(\xi)]_0^2$$

$$= h\left[2f_0 + 2\Delta f_0 + \frac{f_2 - 2f_1 + f_0}{2}\left(\frac{8}{3} - 2\right) + (4 - 8 + 4)\frac{h^3}{3!}f^{(3)}(\xi)\right]$$

$$= \frac{h}{3}[f_0 + 4f_1 + f_2], \qquad (17)$$

where $x_0 \le \xi \le x_2$, $h = x_i - x_{i-1}$, $\quad i = 1(1)2$

We note that the coefficient of the error term vanishes. This implies that the method is of next higher order. Computing in next coefficient, we get the error term as

$$h\int_0^2 \frac{h^4}{4!}t\,|\,t-1)(t-2)(t-3)f^{(4)}(3)\,dt$$

$$= \frac{h^5}{24}\left(\frac{32}{5} - 24 + \frac{88}{3} - 12\right)f^{(4)}(n) = -\frac{h^5}{90}f^{(4)}(n), \qquad , x_0 < n < x_2$$

The error bound can be obtained by finding the maximum and minimum values of $|f^{(4)}(x)|\,on[x_0, x_2]$.

2.2.2(a) Simpson's 1/3 rule over $[x_0, x_n]$ — Composite rule

It is now easy to write the Simpson 1/3 rule over the entire closed interval $[x_0, x_n]$

Since, each sub internal required three points, we subdivide $[x_0, x_2]$ into even number of subinternals, n = 2m , so than we have odd number of points. Hence therefore, we write $\frac{x_{2m} - x_0}{(2m)}$.

$$\int_{x_0}^{x_n} f(x)dx = \int_{x_0}^{x_{2m}} f(x)dx = \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx + \ldots + \int_{x_{2m-2}}^{x_{2m}} f(x)dx$$

$$= \frac{h}{3}[(f_0 + 4f_1 + f_2) + (f_2 + 4f_3 + f_4) + \ldots + (f_{2m-2} + 4f_{2m-1} + f_{2m})]$$

$$= \frac{h}{3}[f_0 + 4f_1 + 2f_2 + 4f_3 + 2f_4 + 4f_5 + \ldots + 2f_{2m-2} + 4f_{2m-1} + f_{2m}]$$

$$\frac{n}{3}[f_0 + 4\{f_1 + f_3 + .. + f_{2m-1}\} + 2\{f_2 + .. + f_{0-m-2}\} + f_{2m}] \qquad (18)$$

The error term is given by

$$E_2 = -\frac{h^5}{90}f^{(4)}(n_1) - \frac{h^5}{90}f^{(4)}(n_2)\ldots - \frac{h^5}{90}f^{(4)}(n_m) \qquad (19)$$

$$|E_2| \le \frac{mh^5}{90}M_4, where\ M_4, = max|f^{(4)}(x)|$$

Since, $nh = \frac{(x_{2m} - x_0)}{2}$, $we\ get |\ E_2\ | \leq \frac{(x_{2m} - x_0)h^4}{180} M_4$ (20)

Since, each subinterval requires three points, we subdivide $[x_0, x_n]$ into even number of subintervals, $n = 2_m$, so that, we have odd number of points. Hence, $h = (x_{2m} - x_0)/(2_m)$.

2.2.3 Simpson's 3/8 Rule over $[x_0, x_3]$

We obtain, from Equation (4), the Simpson 3/8 rule when we set n = 3. Therefore each sub internal requires four points. We obtain the formula as

$$I = \int_{x_0}^{x_3} f(x)dx = h\int_0^3 [f_0 + \frac{\Delta f_0}{1!}t + \frac{\Delta^2 f_0}{2!}(t)(t-1) + \frac{\Delta^3 f_0}{3!}t(t-1)(t-2)]dt$$

Computing the integrals, we obtain

$$\int_{x_0}^{x_3} f(x)dx = \frac{3h}{8}[f_0 + 3f_1 + 3f_2 + f_3]. \quad (21)$$

Equation (20) is known as the Simpson 3/8 rule. The generalization over $[x_0, x_n]$, where n is an integral multiple of 3 (i.e., $n = 3m$, where m is an integer), can be done.

2.2.4 Weddle's Rule over $[x_0, x_6]$,

To apply this rule, we require 7 pairs in the interval

For $n = 6$, we obtain, from equation (4),

$$\int_{x_0}^{x_6} f(x)dx = \frac{6h}{840}[41f_0 + 216f_1 + 27f_2 + 272f_3 + 27f_4 + 216f_5 + 41f_6]. \quad (22)$$

Adding to the right hand side the term

$$\frac{6h}{840}[f_0 - 6f_1 + 15f_2 - 20f_3 + 15f_4 - 6f_5 + f_6]$$

we obtain the Weddle's rule over $[x_0, x_6]$ as

$$\int_{x_0}^{x_6} f(x)dx = \frac{3h}{10}[f_0 + 5f_1 + f_2 + 6f_3 + f_4 + 5f_5 + f_6].$$

The generalization over $[x_0, x_n]$, where n is an integral multiple of 6 (i.e., $n = 6m$, where m is an integer), can be done.

Example 1: Calculate the value of the integral

$$\int_4^{5.2} \log x dx \quad \text{by}$$

(a) Trapezoidal rule (b) Simpson's $1/3$ rule

(c) Simpson's $3/8$ rule (d) Weddle's rule.

Assume $h = 0.2$. compare the numerical solutions with the exact solution.

Solution: Taking $h = 0.2$, divide the range of integration (4, 5.2) into six equal parts.

The values of $\log x$ for each point of sub-division are given below:

X	$f(x) = \log x$
$x_0 = 4$	$f(0) = 1.3862944$
$x_0 + h = 4.2$	$f(1) = 1.4350845$
$x_0 + 2h = 4.4$	$f(2) = 1.4816045$
$x_0 + 3h = 4.6$	$f(3) = 1.5260563$
$x_0 + 4h = 4.8$	$f(4) = 1.5686159$
$x_0 + 5h = 5.0$	$f(5) = 1.6094379$
$x_0 + 6h = 5.2$	$f(6) = 1.6486586$

(a) By Trapezoidal rule, we have

$$\int_4^{5.2} \log x\,dx = \frac{h}{2}[f(0) + f(6) + 2\{f(1) + f(2) + f(3) + f(4) + f(5)\}]$$

$$= \frac{0.2}{2}[3.034953 + 2 \times 7.6207991] = 0.1(18.276551) = 1.8276551$$

(b) By Simpson's $1/3$ rule, we have

$$\int_4^{5.2} \log x\,dx = \frac{h}{3}[f(0) + f(6) + 4\{f(1) + f(3) + f(5)\} + 2\{f(2) + f(4)\}]$$

$$= \frac{0.2}{3}[3.034953 + 4(4.5705787) + 2(3.0502204)]$$

$$= \frac{0.2}{3}[3.034953 + 18.282315 + 6.1004408]$$

$$= \frac{0.2}{3} \times 27.417709 = 1.8278472.$$

(c) By Simpson's 3/8 rule, we have

$$\int_4^{5.2} \log x\,dx = \frac{3h}{8}[f(0) + f(6) + 3\{f(1) + f(2) + f(4) + f(5)\} + 2f(3)]$$

$$= \frac{3(0.2)}{8}[3.034953 + 3(6.0947428) + 2(1.5260563)]$$

$$= \frac{0.6}{8}[3.034953 + 18.284228 + 3.0521126]$$

$$= \frac{0.6}{8} \times 24.371294 = 1.827847.$$

(d) By Weddle's rule, we have

$$\int_4^{5.2} \log x\,dx = \frac{3h}{10}[f(0) + f(6) + 5\{f(1) + f(5)\} + f(2) + f(4) + 6f(3)]$$

$$= \frac{3(0.2)}{10}[3.034953 + 5(3.0445224) + 3.0502204 + 6(1.5260563)]$$

$$= \frac{0.6}{10}[3.034953 + 15.222612 + 3.0502204 + 9.1563378]$$
$$= \frac{0.6}{10} \times 30.464123 = 1.8278474$$

Exact value of the internal is $\int_4^{5.2} \log x dx = [x(\log x - 1)]_4^{5.2}$
$= [5.2(\log 5.2 - 1) - 4(\log 4 - 1)]$
$= [3.3730249 - 1.5451774] = 1.8278475$

Hence, magnitude of the errors due to different formulae are

(a) 0.0001924 (b) 0.0000003

(c) 0.0000005 (d) 0.0000001

Check Your Progress 1

1) Find the approximate value of $I = \int_0^1 \frac{dx}{1+x}$ using (i) Trapezoidal rule (two points) and (ii) Simpson's 1/3 rule (three points) .Obtain a bound for the error in the Trapezoidal Rule. The exact value of $I = In\ 2 = 0.693147$ correct to six decimal places.

.....

.....

.....

Evaluate the integral $\int_0^6 (x^2 + x + 2)dx$ using Trapezoidal rule, with h = 1.0.

.....

.....

.....

3) Evaluate the integral $\int_0^1 \frac{dx}{1+x}$ using Simpson 3/8 rule. with $h = 1/3$.

.....

.....

.....

4) Evaluate the integral $\int_1^4 x^2 dx$ using Weddle's rule, with $h = 0.5$.

.....

.....

.....

2.2.5 Remarks on Newton – Cotes Formulas

(a) These formulae are called closed quadrature formulae because they require the knowledge of the function at the end points (or at the limits of integration) as opposed to the open quadrature, formulas which do not require the knowledge of the function at the end points.

(b) Except for the Trapezoidal rule, all the formulas require the knowledge of the integrand at equispaced points. Thus, either the function should be known explicitly or at equally spaced points.

(c) The Simpson 1/3 rule is very popular. It integrates exactly a cubic function or, equivalently, gives the area under a segment of cubic curve. On the other hand, the trapezoidal rule integrates exactly a linear function.

2.3 COMPOSITE FORMULAS

We summaries all the composite formulas

Composite Trapezoidal Rule

Subdivide the interval [a, y] into M equal parts. The length of each interval is h = (b-a)/M and the points are a = $x_0, x_1 = x_0 + h, x_2 = x_0 + 2h, \ldots, x_0 + nh = x_n = b$.
Then, the composite Trapezoidal rule is given by (see Equation (12))

$$I_T = \frac{h}{2}\left[f(x_0) + 2\{(x_1) + f(x_2) + .. + f(x_{n-1})\} + f(x_n)\right]$$

$$= \frac{h}{2}\left[f_0 + 2\{f_1 + f_2 + \ldots + f_{n-1}\} + f_n\right] \qquad (23)$$

The error bound is as given in Equation (15)

Composite Simpson's rule

Subdivided the interval [a, b] into even number of subintervals giving odd number of points.

Let $h = (b-a)/2M$). The points are $x_0, x_1, .. x_2 m$.
Then, the composite Simposn's rule's given by (see Equation (18))

$$I_s = \frac{h}{3}\left[f_0 + 4\{f_1 + f_3 + .. + f_{2m-1}\} + 2\{f_2 + f_4 + .. + f_{2m-2}\} + f_{2m}\right] \qquad (24)$$

The error bound is as given in Equation (20).

Example 2: Evaluate $\int_1^6 [2 + \sin(2\sqrt{x})]dx$ using Trapezoidal rule with 11 points.

Solution: To generate 11 sample points, we use $h = [(6-1)/10] = 0.5$. The points are 1,1.5,2.0,2.5,3.0,3.5,4.0,4.5,5.0,5.5 and 6.0. Using the formula (12), we obtain

$$I = \frac{0.5}{2}\left[f(1) + f(6) + 2\{f(1.5) + f(2) + f(2.5) + f(3) + f(3.5) + f(4) + f(4.5) + f(5) + f(5.5)\}\right]$$

$= 0.25[2.90929743 + 1.01735756]$

$+0.5[2.63815764+2.30807174+1.97931647+1.68305284+1.43530410+$
$+1.24319750+1.10831775+1.02872220+1.00024140]$

$= 0.25[3.92665499]+0.5[14.42438165]$

$= 0.98166375 + 7.21219083 = 8.19385457$

Example 3: Evaluate $\int_1^6 [2+sin(2\sqrt{x})]\,dx$ using Simpsons rule with 11 points.

Solution: To generate 11 sample points we use $h=(6-1)/10=0.5$, the points are same as in the previous example. Using the Simpson's formula, we get

$$I=\frac{1}{6}[f(1)+f(6)+4\{f(1.5)+f(2.5)+f(3.5)+f(4.5)+f(5.5)\}.$$
$$+2\{f(2)+f(3)+f(4)+f(5)\}]$$
$$I=\frac{1}{6}[2.90929743+1.01735756]$$
$$+2.0\{2.30807174+1.68305284+1.24319750+1.02872220\}$$
$$+4.0\{2.63815764+1.97931647+1.43530410+1.10831775+1.00024140\}$$
$$=\frac{1}{6}[(3.92665499)+2.0(6.26304429)+4.0(8.16133735)]=8.18301550.$$

Check Your Progress 2

Evaluate the integral $I=\int_0^1 \frac{dx}{1+x}$ using (i) composite trapezoidal rule, (ii) composite Simpson's rule, with 2, 4 and 8 equal subintervals.

.....

.....

.....

Error Analysis: We have earlier devided that the errors in the composite Trapezoidal and Simpson's rules are of orders, O (h^2) and (h^4) respectively.This signifies that the error for Simpson's rule converges to zero faster than the error for the trapezoidal rule as the step size h decreases to zero. In cases where the derivatives of f(x) are known, the formulas

$$E_T(f,h)=\frac{-(b-a)f^{(2)}(c)h^2}{12} \quad \text{and} \quad E_S(f,h)=\frac{-(b-a)f^{(4)}(c)h^4}{180}$$

can be used to estimate the number of subintervals required to achieve a specified accuracy.

2.4 GAUSSIAN QUADRATURE

The numerical integration methods described so far are based on a rather simple choice of evaluation points for the function $f(x)$. They are particularly suited for regularly tabulated data, such as one might measure in a laboratory, or obtain from computer software designed to produce tables. If one has the freedom to choose the points at which to evaluate $f(x)$, a careful choice can lead to much more accuracy in evaluating the integral in question. We shall see that this method, called Gaussian or Gauss-Legendre integration, has one significant further advantage in many situations. In the evaluation of an integral on the interval α to β, it is not necessary to evaluate $f(x)$ at the endpoints, ie. at α or β, of the interval. This will prove valuable when evaluating various *improper* integrals, such as those with infinite limits.

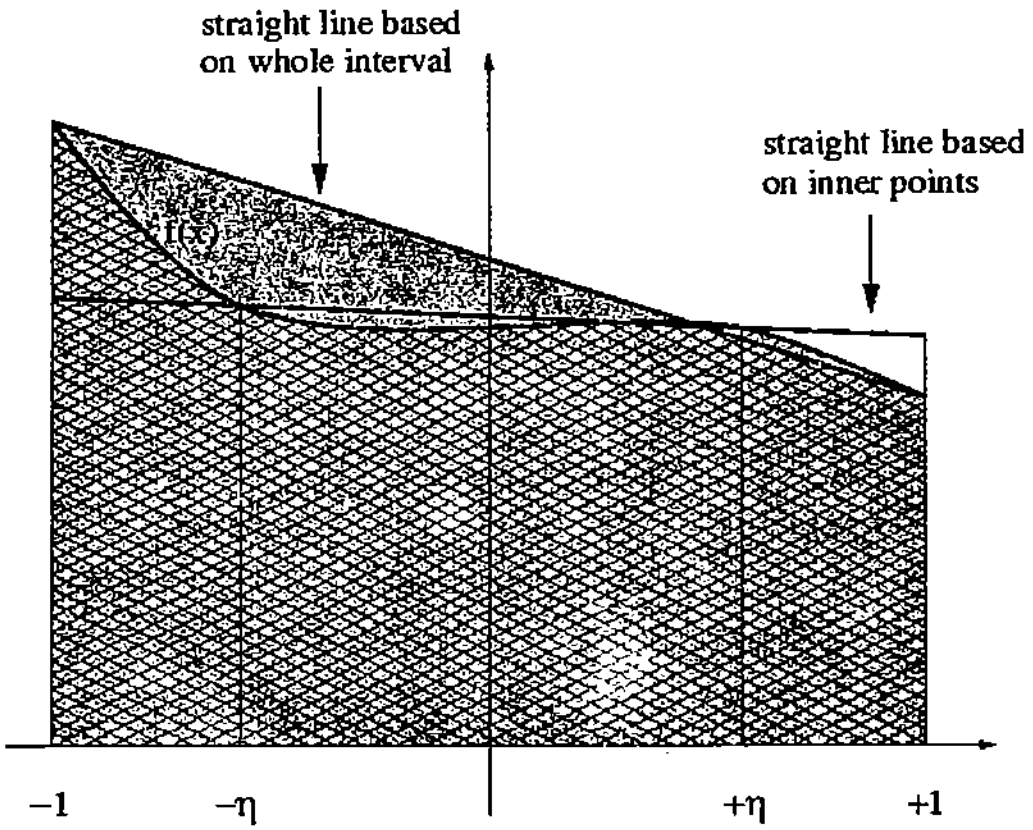


Figure 1: Comparing Trapezoidal rule of integration and Gaussian integration.

We begin with a simple example illustrated in Figure 1.

The simplest form of Gaussian integration is based on the use of an optimally chosen polynomial to approximate the integrand $f(x)$ over the interval [-1, 1]. The details of the determination of this polynomial, that is, determination of the coefficients of x in this polynomial, are beyond the scope of this presentation. The simplest form uses a uniform weighting over the interval, and the particular points at which to evaluate $f(x)$ are the roots of a particular class of orthogonal polynomials, the Legendre polynomials, over the interval[-1,1]. It can be shown that the best estimate of the

integral is given by:

$$\int_{-1} f(x)dx = \sum_{i=1}^{n} w_i f(x)$$

where x_i is a designated evaluation point, and w_i is the *weight* of that point in the sum. If the number of points at which the function *f(x)* is evaluated is *n*, then the resulting value of the integral is of the order *2n-1*. For example, a two point formula is of order 3, which is the order pf the Simpson's rule. A three point formula is of order 5. Thus the carefully designed choice of function evaluation points in the Gauss-Legendre form results in the same accuracy for about half the number of function evaluations, and thus at about half the computing effort.

The value of the abscissas and the corresponding weights in the Gauss-Legendre quadrature, for a given *n*, are known. These values are given in Table 3.

The choice of value of *n* depends on the accuracy required in a particular problem. When choosing to use *n* points, we call the method an "*n*-point Gaussian" method. The abscissas for a given *n*, are the zeros of the Legendre Polynomial of order *n*. For example , for $n = 1$, $P_1(x) = x = 0$ *gives* $x = 0$, For $n = 2$, $P_2(x) = 3x^2 - 1 =$ gives the abscissas as $x = \pm\sqrt{1/3}$, For $n = 3$, $P_3(x) = 5x^3 - 3x = 0$ gives the abscissas as $x = 0, \pm\sqrt{3/5}$ or $x = 0 \pm 0.77459667$.

Table 3: Gauss-Legendre Abscissas and Weights

n	Values of t/x	Weights	Order
2	$\pm\sqrt{1/3}$	1.0	3
3	0.0	0.88888889	
	± 0.77459667	0.55555555	5
4	± 0.33998104	0.65214515	7
	± 0.86113631	0.34785485	
5	0.0	0.56888889	9
	± 0.53846931	0.47862867	
	± 0.90617985	0.23692689	
6	± 0.23861918	0.46791393	11
	± 0.66120939	0.36076157	
	± 0.93246951	0.17132449	

For $n = 1$, the method is given by $\int_{-1} f(x)dx = f(0)$ (25)

For $n = 2$, the method is given by

$$\int_{-1}^{1} f(x)dx = f(-\sqrt{1/3}) + f(\sqrt{1/3}) \tag{26}$$

For $n = 3$, the method is given by

$$\int_{-1}^{1} f(x)dx = \frac{1}{9}[5f(-\sqrt{0.6}) + 8f(0) + 5f(\sqrt{0.6})] \tag{27}$$

The Gauss-Legendre integration formula given here computes an estimate of the integral on the interval [-1, 1]. In most cases [a, b] we may want to evaluate the integral on a more general interval, say. In this case, we use a linear transformation and reduce the interval [a, b] to [-1, 1] and then apply the Gauss-Legendre formula.

Let the transformation be $x = c + mt$. Then

for $x = a$: $a = c - m$,

for $x = b$: $b = c + m$,

The solution is c = (b + a)/2 and m = (b−2)/2. The transformation is and

$dx = mdt$. $x = \frac{1}{2}[(a+b) + (a-b)t]$

$$I = \int_{a}^{b} f(x)dx = \int_{-1}^{+1} f(c+mt)dt$$

Finally, now, we can write the Gauss-Legendre estimate of the integral as

$$I = \int_{a}^{b} f(x)dx = m\sum_{i=1}^{n} wf(c+mt_i)$$

Example 4: Evaluate the integral $I = \int_{0}^{p/2} \sin x \, dx$

Using the Gauss-Legendre Formulas. Compare with the exact solution and the values obtained by Simpson rule. The exact value is I = 1.

Solution: The required transformation is

$$x = \frac{1}{2}[(b+a) + (b-a)t] = \frac{\pi}{4}[1+t].$$

The integral becomes

$$I = \frac{\pi}{4}\int_{-1}^{1} \sin\left[= \frac{\pi}{4}(1+t)\right]$$

The two point formula gives

$$I = \frac{\pi}{4}\left[\sin\left\{\frac{\pi}{4}(1-\sqrt{\frac{1}{3}})\right\} + \sin\left\{\frac{\pi}{4}(1+\sqrt{\frac{1}{3}})\right\}\right]$$

= 0.785398163[0.325885607 + 0.945409207] = 0.998472612

The error in evaluating I is |1 −0.998472612| = 0.001527.

For N = 4 and b, we have the values of I as 0.9999999770 and 0.9999999904 respectively.

The values of I obtained by Simpson's rule for n = 3, 5 and 7 points are 1.0022798775,

1.0001345845 and 1:0000263122 respectively.

Note that the Simpson's rule is of order 3, while the Gauss-Legendre Formula are of order *2N-1*.

N	Gauss-Legendre	Simpson's 1/3
2	0.9984726135	1.0022798775
4	0.9999999770	1.0001345845
6	0.9999999904	1.0000263122

Example 5: Evaluate the integral $I = \int_0^1 \frac{dx}{1+x}$. Using Gauss-Legendre three point formula.

Solution: First we transform the interval [0, 1] to the interval [−1, 1].

Let t = $a\,x + b$. We have,

For $x = 0$: $\quad -1 = b$

For $x = 1$: $\quad 1 = a + b = a - 1$, *or* $a = 2$

Therefore, $t = 2x - 1$, *or* $x = (1 + t)/2$.

Hence, $I = \int_0^1 \frac{dx}{1+x} = \int_{-1}^1 \frac{dt}{t+3} = \int_{-1}^1 F(t)\,dt$

Guass – Lengendre three point formula gives

$$I = \frac{1}{9}\left[5F(-\sqrt{0.6}) + 8F(0) + 5F(\sqrt{0.6})\right]$$

$$\frac{1}{9}\left[5\left\{\frac{1}{3-\sqrt{0.6}} + \frac{.1}{3+0.6}\right\} + \frac{8}{3}\right] = 0.693122$$

The exact solution is $I = In2 = 0.693147$

Check Your Progress 3

1) Evaluate the integral $I = \int_1^2 \frac{2xdx}{1+x^4}$, using the Gauss-Legendre 1-point, 2-point and 3-point quadrature rules. Compare with the exact solution

.....

.....

.....

2.5 SUMMARY

In this unit, we have:

1) Derived Newton-cotes formulas.
2) Derived composite formulas.
3) Derived error in composite formulas.
4) Derived Gauss-Legendre integration formula.

2.6 SOLUTIONS/ANSWERS

Check Your Progress i

1) Using the trapezoidal rule, we have

$$I=\frac{1}{2}\left(1+\frac{1}{2}\right)=0.75 \quad Error=0.75-0.693147=0.056853$$

The error in the trapezoidal rule is given by

$$|R_1|\le\frac{(b-a)^3}{12}\max_{0\le x\le 1}|f''(x)|\le\frac{1}{12}\max_{0\le x\le 1}\left|\frac{2}{(1+x)^3}\right|\le\frac{1}{6}$$

Using the Simpson's rule, we have

$$I\square\frac{1}{6}\left(1+\frac{8}{3}+\frac{1}{2}\right)=\frac{25}{36}=0.001297.\ Error=0.75-0.693147=0.056853.$$

2) Here $x_0=0,x_1=1,x_2=2,x_3=3,x_4=4,x_5=5,x_6=6,h=1$. Therefore,

$$I=\frac{h}{2}[f(0)+2f(1)+2f(2)+2f(3)+2f(4)+2f(5)+f(6)]$$
$$=\frac{1}{2}[2+2(4)+2(8)+2(14)+2(22)+2(32)+44]=103.$$

3) Here $x_0=0,x_1=1/3,x_2=2/3,x_3=1,h=1/3$. Therefore,

$$I=\frac{3h}{8}[f(0)+3f\left(\frac{1}{3}\right)+3f\left(\frac{2}{3}\right)+f(1)]=\frac{3}{8}(\frac{1}{3})[1+\frac{9}{4}+\frac{9}{5}+\frac{1}{2}]=0.69375.$$

4) Here $x_0=1,x_1=1.5,x_2=2,x_3=2.5,x_4=3,x_5=3.5,x_6=4,h=0.5$

Therefore,

$$I=\frac{3h}{10}[f(0)+5f(1)+f(2)+6f(3)+f(4)+5f(5)+f(6)]$$
$$=\frac{3(.5)}{10}[1+5(2.25)+4+6(6.25)+9+5(12.25)+16]=21.$$

Check Your Progress 2

1) When $N = 2$, we have $h = 0.5$, Nodes are 0,0.5 and 1

We obtain

$$I_T = \frac{1}{4}\left[f(0)+2f(0.5)+f(1)\right] = \frac{1}{4}\left(1+\frac{4}{3}+\frac{1}{2}\right) = 0.708333$$

$$I_S = \frac{1}{6}\left[f(0)+4f(0.5)+f(1)\right] = \frac{1}{6}\left(1+\frac{8}{3}+\frac{1}{2}\right) = 0.694444$$

$N = 4$: $h = 0.25$, Nodes are 0,0.25, 0.5, 0.75 and 1. We obtain

$$I_T = \frac{1}{8}\left[f(0)+2\{f(0.25)+f(0.5)+f(0.75)\}+f(1)\right] = 0.697024$$

$$I_S = \frac{1}{12}\left[f(0)+4f(0.25)+2f(0.5)+4f(0.75)+f(1)\right] = 0.693254$$

$N = 8$: $h = 0.125$, nodes are 0, 0.125, 0.25, ..., 1.0

We have eight subintervals for trapezoidal rule and four subintervals for Simpson's rule. We get

$$I_T = \frac{1}{16}\left[f(0)+2\sum_{i=1}^{7} f(\frac{i}{8})+f(1)\right] = 0.694122$$

$$I_S = \frac{1}{24}\left[f(0)+4\sum_{i=1}^{4} f(\frac{2i-1}{8})+2\sum_{i=1}^{3} f(\frac{2i}{8})+f(1)\right] = 0.693155.$$

The exact value of the integral is $I = 0.693147$.

Check Your Progress 3

1) The exact solution is $I = \tan^{-1}(4) - (\pi/4)$. The required linear transformation is $x = (t + 3)/2$. The integral reduces to

$$I = \int_{-1}^{1} \frac{8(t+3)dt}{[16+(t+3)^4]} = \int_{-1}^{1} f(t)dt.$$

Using the 1-point rule, we get $I = 2f(0) = 0.4948$.

Using the 2-point rule, we get

$$I = f\left(-1/\sqrt{3}\right) + f\left(1/\sqrt{3}\right) = 0.3842 + 0.1592 = 0.5434.$$

Using the 3-point rule, we get

$$I = \frac{1}{9}\left[5f(-\sqrt{0.6})+8f(0)+5f(\sqrt{0.6})\right] = \frac{1}{9}[5(0.4393)+8(0.2474)+5(0.1379)] = 0.540($$

The exact solution is I = 0.5404.

UNIT 3 NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS

Structure

3.0 INTRODUCTION

Ordinary differential equation (ODE) occur frequently in the mathematical models encountered in science and engineering. Consequently, their numerical solution is a very large area of study.

It is well-known that the process of integration of an ODE introduces arbitrary constants. These constants are determined from the conditions imposed on the function or its derivatives. For example, a fourth order differential equation would require four conditions for the determination of the four arbitrary constants that would arise. We encounter two types of problems depending on the manner in which such conditions are specified. If all the required conditions are given at single point, we have an initial value problem (IVP). The method of solution here is direct, starting at the known point and moving step by step along the range of the integration. In other words, in an IVP, we compute the of functions $y_1(x), y_2(x), \ldots, y_n(x)$ which satisfy the given n first order ODEs.

$$y_i' = f(x, y_1, y_2, \ldots, y_n), \qquad i = 1(1)n, \tag{1}$$

Subject to n given initial conditions

$$y_i = a_i, \qquad i = 1(1)n \qquad \text{at} \quad x = x_0 \tag{2}$$

However, if the conditions are given at more than one point, then the information needed to start the computation at any single point is not sufficient and the method of solution involves either the solution of a set of simultaneous equations or the use of the estimated values are then corrected by the iteration as the calculation proceeds. This second type of problem is known as a boundary value problem (BVP). In particular, in a two point BVP, we compute the functions $y_i(x), i = 1(1)n$, which satisfy equation (1) and may take the n given conditions as

$$y_i = \alpha_i \quad \text{for certain specified values of } i \text{ at } x = x_0,$$
$$y_i = \beta_i \quad \text{for other values of } i \text{ at } x = x_1, \; i = 1(1)n \tag{3}$$

The methods of numerical solution can be derived by various means, including the finite difference formulas and the truncated Taylor series. The derivation shows that in each computation an approximation is made, and this introduces an error. The usefulness of a method depends not only on the size of the errors but also on the way in which these errors get magnified or decay, as we proceed along the range of

integration. Therefore, when using a method, we should keep in mind the concept of consistency which relates to the error introduced at a particular point and the concept of stability which relates to the growth of error, as the calculation proceeds.

3.1 OBJECTIVES

After studying this unit, you should be able to solve an ordinary differential equation

- using Euler's Method,
- using Improved Euler's Method, and
- using Runge Kutta Method.

3.2 EULER'S METHOD

This method is generally used to get the numerical solution of the differential equations, because it provides a simple procedure for computing approximations to the exact solutions.

A Simple Initial Value Problem

Let us start by looking at an initial value problem whose solution is known:

$$\frac{dy}{dt} = y, \qquad y(0) = 1$$

The solution of this IVP is $y(t) = e^t$. Using this exact solution, we will be able to determine the error in the numerical solution.

Let us suppose that we are interested in the value of the solution at $t = 1$. Since, $y(0) = 1$, we obtain,

$$\frac{dy}{dt}(0) = y(0) = 1$$

Let us now use the Taylor series first order approximation about $t = 0$ as $y(t) = y(0) + (t-0)y'(0) = 1 + t$.

Hence, the approximate value at $t = 1$, is $y(1) \approx 2$. The magnitude of error in the solution is $|e - 2.0|$.

The demerit of the above approximation is that it assumes that the derivative of the solution is a constant and equals 1 in the whole interval. We can improve the result by dividing interval into two sub-intervals. First, we will use the linear Taylor approximation based about $t = 0$ to approximate the value at $t = \frac{1}{2}$. Then, we will use a linear Taylor approximation about $t = \frac{1}{2}$ to obtain an approximate value at $t = 1$.

Using the above approximation, we get $y(\frac{1}{2}) = \frac{3}{2}$.

Hence, at $t = \frac{1}{2}$, $y'\left(\frac{1}{2}\right) = y\left(\frac{1}{2}\right) = \frac{3}{2}$. Now, Taylor series liner approximation about $t = \frac{1}{2}$, is given by $y(t) = y\left(\frac{1}{2}\right) + \left(t - \frac{1}{2}\right)y'\left(\frac{1}{2}\right) = \frac{3}{2}\left(t + \frac{1}{2}\right)$.

Hence, the approximate value at t = 1, is $y(1) \approx 2.25$. Now, |error| = |e – 2.25| is much smaller than the error obtained in the previous approximation.

You can now guess that if we use more subintervals, then the result may improve further. Suppose, that we divide [0, 1] into n subintervals. On each of the subintervals, the slope is constant (but the values of the slopes are different). The length of each subinterval is $h = \frac{1}{n}$. We will call the points that we obtain as (t_i, y_i).

Notice that $(t_0, y_0) = (0, 1)$ since this is where the initial value problem tells us to begin. To get from one step to the next, we are assuming that the solution approximately passes through (t_i, y_i). At that point, the derivative, is given by $\left(\frac{dy}{dt}\right)_{t_i} = y_i$. Hence, the linear Taylor approximation at that point is

$y(t) \approx y_i + (t - t_i)y'_i = y_i + (t - t_i)y_i$.

Therefore, we write $y(t_{i+1}) \approx y_{i+1} = (t_{i+1} - t_i + 1)y_i = (h + 1)y_i = \left(1 + \frac{1}{n}\right)y_i$.

Now, extend this concept this concept to solve a general initial value problem

$$\frac{dy}{dt} = f(t, y), \qquad y(0) = y_0 \tag{3}$$

We will form an approximate solution by taking a number of steps. We call the distance between the steps as h and the various points as (t_i, y_i). To get from one step to the next, we will form the linear Taylor approximation at $t = t_i$. The derivative at this point is given by the differential equation: $\left(\frac{dy}{dt}\right)_i = f(t, y)$. The linear Taylor approximation is then

$$\begin{aligned} y_{i+1} &= y_i + (t_{i+1} - t_i)y'(t_i) \\ &= y_i + hf(t_i, y_i) \end{aligned} \tag{4}$$

This technique is called the *Euler's Method.*

Example 1: For the IVP $y' + 2y = 2 - e^{-4t}$, $y(0) = 1$. Use the Euler's method with step size $h = 0.1$ to find approximate values of the solution at $t = 0.1$, 0.2, 0.3, 0.4, and 0.5. Compare them with the exact values of the solution at these points. Find the percentage error in the solutions.

Solution: The solution of the differential equation is $y(t) = 1 + \frac{1}{2}e^{-4t} - \frac{1}{2}e^{-2t}$.

In order to use Euler's Method we first need to rewrite the differential equation as $y' = 2 - e^{-4t} - 2y$.

Hence, $f(t, y) = 2 - e^{-4t} - 2y$. The initial values are $t_0 = 0$ and $y_0 = 1$.

Euler's method gives $y_{j+1} = y_j + hf(t_j, y_j)$. We have for $j = 0$, and $h = 0.1$.

$f_0 = f(t_0, y_0) = f(0.1) = 2 - e^{-4(0)} - 2(1) = -1$

$y_1 = y_0 + hf_0 = 1 + (0.1)(-1) = 0.9$.

Therefore, the approximation to the solution at $t_1 = 0.1$

$y(0.1) \approx y_1 = 0.9$.

For $j = 1$, we get,

$f_1 = f(0.1, 0.9) = 2 - e^{-4(0.1)} - 2(0.9) = -0.470320046$,

$y_2 = y_1 + hf_1 = 0.9 + (0.1)(-0.470320046) = 0.852967995$

Therefore, the approximation to the solution at $t_2 = 0.2$ is

$y(0.2) \approx y_2 = 0.852967995$. For j = 2, 3, 4 we obtain the following values.

$f_2 = -0.155264954$, $y_3 = 0.837441500$

$f_3 = 0.023922788$, $y_4 = 0.839833779$

$f_4 = 0.1184359245$, $y_5 = 0.851677371$

The percentage error in the numerical solutions is defined by

$$percentage\ error = \left|\frac{exact\ value - approximate}{exact value}\right| x100$$ .The results are presented in the following table

t_n	Approximation	Exact	Error
$t_0 = 0$	$y_0 = 1$	$y(0) = 1$	0 %
$t_1 = 0.1$	$y_1 = 0.9$	$y(0.1) = 0.925794646$	2.79 %
$t_2 = 0.2$	$y_2 = 0.852967995$	$y(0.2) = 0.889504459$	4.11 %
$t_3 = 0.3$	$y_3 = 0.837441500$	$y(0.3) = 0.876191288$	4.42 %
$t_4 = 0.4$	$y_4 = 0.839833779$	$y(0.4) = 0.876283777$	4.16 %
$t_5 = 0.5$	$y_5 = 0.851677371$	$y(0.5) = 0.883727921$	3.63 %

Example 2: Repeat the previous example with the approximations at $t = 1$, $t = 2$, $t = 3$, $t = 4$, and $t = 5$. Use $h = 0.1$, $h = 0.05$, $h = 0.01$, $h = 0.005$, and $h = 0.001$ for the approximations.

Solution: Below are two tables, one gives approximations to the solution and the other gives the errors for each approximation. We'll leave the computational details to you to check.

Approximations						
Time	**Exact**	**$h = 0.1$**	**$h = 0.05$**	**$h = 0.01$**	**$h = 0.005$**	**$h = 0.001$**
$t = 1$	0.9414902	0.9313244	0.9364698	0.9404994	0.9409957	0.9413914
$t = 2$	0.9910099	0.9913681	0.9911126	0.9910193	0.9910139	0.9910106
$t = 3$	0.9987637	0.9990501	0.9988982	0.9987890	0.9987763	0.9987662
$t = 4$	0.9998323	0.9998976	0.9998657	0.9998390	0.9998357	0.9998330
$t = 5$	0.9999773	0.9999890	0.9999837	0.9999786	0.9999780	0.9999774

Percentage Errors					
Time	$h = 0.1$	$h = 0.05$	$h = 0.01$	$h = 0.005$	$h = 0.001$
$t = 1$	1.08 %	0.53 %	0.105 %	0.053 %	0.0105 %
$t = 2$	0.036 %	0.010 %	0.00094 %	0.00041 %	0.0000703 %
$t = 3$	0.029 %	0.013 %	0.0025 %	0.0013 %	0.00025 %
$t = 4$	0.0065 %	0.0033 %	0.00067 %	0.00034 %	0.000067 %
$t = 5$	0.0012 %	0.00064 %	0.00013 %	0.000068 %	0.000014 %

We can see from these tables that as h decreases, the accuracy of the approximations are improved.

Example 3: Solve the initial value problem to compute approximation for $y(0.1)$, $y(0.2)$ using Euler's material with $h = 0.1$.

$\frac{dy}{dt} + 2y = 3e^{-4t}$, $y(0) = 1$. Compare with the exact solution $y(t) = \frac{(5e^{-2t} - 3e^{-4t})}{2}$.

Solution: Euler's method gives $y_{j+1} = y_j + hf(t_j, y_j)$

Where, $f(t, y) = 3e^{-4t} - 2y$.

For $j = 0$, we get, $y_1 = y_0 + hf(t_0, y_0)$.

Now, $f(t_0, y_0) = f(0,1) = 3e^0 - 2 = 1$.

We obtain, $y(0.1) \approx y_1 = 1 + (0.1)(1) = 1.1$

The exact solution is y(0.1) ~ 1.0413 and in the absolute error is 0.0587. For in next time Step $j = 1$, we get $y_2 = y_1 + hf(t_1, y_1)$.

Now, $f(t_1, y_1) = (0.1, 1.1) = 3e^{-0.4} - 2(1.1) = -0.189$

We obtain $y(0.2) \approx y_2 = 1.1 + 0.1(-0.189) = 1.0811$.

The exact solution is $y(0.2) \approx 1.0018$ and the absolute error is 0.0793.

Check Your Progress 1

1) For the IVP

$$y' - y = -\frac{1}{2}e^{t/2}\sin(5t) + 5e^{t/2}\cos(5t), \qquad y(0) = 0$$

use Euler's Method to find the approximation to the solution at $t = 1$. Use $h = 0.1$, $h = 0.05$ for the approximations and find the percentage errors.

.....

.....

.....

.....

.....

.....

3.3 RUNGE KUTTA METHOD

One member of the family of Runge–Kutta methods is so commonly used, that it is often referred to as "RK4" or as "*the* Runge–Kutta fourth order method" or as "classical Runge-Kutta fourth user method".

Let an initial value problem be specified as follows:

$y' f(t, y), \qquad y(t_0) = y_0$

Then, the RK4 method for the solution of IVP is given by the following:

$$y_{n+1} = y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \qquad (5)$$

where

$$k_1 = f(t_n, y_n)$$
$$k_2 = f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_1\right)$$
$$k_3 = f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_2\right) \quad (6)$$
$$k_4 = f\left(t_n + h, y_n + hk_3\right)$$

Thus, the next value (y_{n+1}) is determined from the present value (y_n) plus the product of the size of the interval (h) and an estimated slope. The slope is a weighted average of slopes, that is, $W_1K_1 + W_2K_2 + W_3K_3 + W_4K_4$:

- k_1 is the slope at the beginning of the interval (t_n, t_{n+1});
- k_2 is the approximate slope at the midpoint of the interval (t_n, t_{n+1}), using slope k_1 to determine the value of y at the a midpoint $t_n + (h/2)$ using Euler's method (t_n, t_{n+1});
- k_3 is again the approximate slope at the midpoint, but now using the slope k_2 to determine the y-value;
- k_4 is the approximate slope at the end of the interval (t_n, t_{n+1}), with its y-value being determined using k_3.

The weights W_1, W_2, W_3 and W_4 are determined such that the order of the method is as high as possible. We use Taylor seris expressions to find the order of the method. We obtain $W_1 = 1/6, W_2 = W_3 = 2/6,\ W_4 = 1/6$ so that the total error is of order h^4. Hence, it is called the RK4 method of fourth order.

Note we may also write the Runge-Kutta fourth order method as

$K_1 = h f(t_n, y_n)$

$K_2 = h f(t_n + h/2,\ y_n + K_1/2)$

$K_3 = h f(t_n + h/2,\ y_n + K_2/2)$

$K_4 = h f(t_n + h,\ y_n + K_3)$ (7)

$y_{n+1} = y_n + (1/6)(K_1 + 2K_2 + 2K_3 + K_4)$ (8)

Example 4: Solve the initial value problem $u' = -2tu^2$ with $u(0) = 1$ and $h = 0.2$ on the interval [0, 1]. Use the fourth order classical Runge-Kutta method.

Solution: We have $X_0 = 1$, $u_0 = 1$, $h = 0.2$, and $f(t, u) = -2tu^2$

For n=0

$$k_1 = h f(t_0, u_0) = 0$$
$$k_2 = h f\left(t_0 + \frac{h}{2}, u_0 + \frac{1}{2}k_1\right) = -0.04$$

$$k_3 = hf\left(t_0 + \frac{h}{2}, u_0 + \frac{1}{2}k_2\right) = -0.038416$$

$$k_4 = hf\left(t_0 + h, u_0 + k_3\right) = -0.0739715$$

$$u(0.2) \approx u_1 = 1 + \frac{1}{6}[0 - 0.08 - 0.076832 - 0.0739715] = 0.9615328$$

For $n = 1$

$$t_1 = 0.2, u_1 = 0.9615328$$

$$k_1 = hf\left(t_1, u_1\right) = -0.0739636$$

$$k_2 = hf\left(t_1 + \frac{h}{2}, u_1 + \frac{1}{2}k_1\right) = -0.1025754$$

$$k_3 = hf\left(t_1 + \frac{h}{2}, u_1 + \frac{1}{2}k_2\right) = -0.0994255$$

$$k_4 = hf\left(t_1 + h, u_1 + k_3\right) = -0.1189166$$

$$u(0.4) \approx u_2 = 0.9615328 + \frac{1}{6}[-0.0739636 - 0.2051508 - 0.1988510 - 0.1189166]$$

$$= 0.8620525$$

Similarly, we get, $u(0.6) \approx u_3 = 0.7352784$

$$u(0.8) \approx u_4 = 0.6097519, u(1.0) \approx u_5 = 0.5000073$$

Example 5: A ball which is at temperature 1200K is allowed to cool down in air at an ambient temperature of 300K. Assuming heat is lost only due to radiation, the differential equation for the temperature of the ball is given by

$$\frac{d\theta}{dt} = -2.2067x10^{-12}(\theta^4 - 81x10^8), \qquad \theta(0) = 1200K.$$

Find the temperature at $t = 480$ seconds using Runge-Kutta 4th order method. Assume a step size of $h = 240$ sec.

Solution: From the given differential equation, we have,

$$f(t,\theta) = -2.2067x10^{-12}(\theta^4 - 81x10^8)$$

Runge-Kutta method gives $\theta_{n+1} = \theta_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$

For $n = 0$, we have $t_0 = 0, \theta_0 = \theta(0) = 1200$

$$k_1 = f(t_0,\theta_0) = f(0,1200) = -2.2067x10^{-12}(1200^4 - 81x10^8) = -4.5579$$

$$k_2 = f\left(t_0+\frac{1}{2}h,\theta_0+\frac{1}{2}k_1h\right) = f\left(0+\frac{1}{2}(240),1200+\frac{1}{2}(-4.5579)240\right)$$

$$= f(120,653.05$$

$$= -2.2067x10^{-12}(653.05^4 - 81x10^8)$$

$$= -0.38347$$

$$k_3 = f\left(t_0+\frac{1}{2}h\theta_0+\frac{1}{2}k_2h\right) = f\left(0+\frac{1}{2}(240),1200+\frac{1}{2}(-038347)24\upsilon\right)$$

$$= f(120,1154.0)$$

$$= 2.2067 = 10^{-12}(1154.0^4 - 81x10^8) = -3.8954$$

$$k_4 = f(t_0+h,\theta_0+hk_3) = f(0+240,1200+240)(-3.8954) = f(120,265.10)$$

$$= 2.2067x10^{-12}(265.10^4 - 81x10^8) = 0.0069750$$

We obtain,

$$\theta(240) \approx \theta_1 = \theta_0 + \frac{h}{6}(k_1+2k_2+2k_3+k_4)$$

$$= 1200 + \frac{240}{6}(-4.5579 + 2(-0.38347) + 2(-3.8954) + (0.0069750)$$

$$= 1200 + 40(-13.1087) = 675.65$$

At the next step, For $n = 1$, we have, t_1=240, $y_1 = 675.65$

We obtain,

$$k_1 = f(t_1,\theta_1) = -0.44199$$

$$k_2 = f\left(t_1+\frac{1}{2}h,\theta_1+\frac{1}{2}k_1h\right) = -0.31372$$

$$k_3 = f\left(t_1+\frac{1}{2}h,\theta_1+\frac{1}{2}k_2h\right) = -0.34775.$$

$$k_4 = f(t_1+h,\theta_1+k_3h) = -0.25351$$

$$\theta(480) \approx \theta_2 = \theta_1 + \frac{h}{6}(k_1+2k_2+2k_3+k_4)$$

$$= 675.65 + \frac{240}{6}) - 0.44199 + 2(-0.31372) + 2(-0.34775) + (-0.25351)$$

$$= 675.65 + \frac{1}{6}(-2.0184)240 = 594.91K$$

Check Your Progress 2

1) Use Runge-Kutta method to solve the IVP $y' = (t-y)/2$ on $[0,0.2]$ with $y(0) = 1$. Compare the solutions with h = 0.2 and 0.1.

.....

.....

.....

2) Using Runge-Kutta method of order 4, find y(0.2) given that $y' = 3x + y/2$, $y(0) = 1$ taking $h = 0.1$.

.....

.....

.....

3) Using Runge-Kutta method of order 4, compute $y(0.2)$ and $y(0.4)$ for the IVP $10y' = x^2 + y^2$, $y(0) = 1$, taking $h = 0.1$.

.....

.....

.....

4) Apply Runge-Kutta fourth order method to find an approximate value of y when $x = 0.2$ given that $y' = x + y$ with $y(0) = 1$ and $h = 0.2$.

.....

.....

.....

.....

3.4 EXPLICIT RUNGE KUTTA METHODS

The family of explicit Runge-Kutta methods is a generalization of the RK4 method mentioned above. It is given by

$$y_{n+1} = y_n + h\sum_{i=1}^{s} b_i k_i \qquad (9)$$

where

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + c_1 h, y_n + a_{21} h k_1) \\ k_3 &= f(t_n + c_1 h, y_n + a_{31} h k_1 + a_{32} h k_2) \\ &\cdots\cdots\cdots \\ k_s &= f(t_n + c_s h, y_n + a_{s1} h k_1 + a_{s2} h k_2 + \ldots\ldots + a_{s,s-1} h k_{s-1}) \end{aligned} \qquad (10)$$

Note: We can also write the method as

$$y_{n+1} = y_n + \sum_{i=1}^{s} b_i k_i \qquad (11)$$

where

$$\begin{aligned} k_1 &= hf(t_n, y_n) \\ k_2 &= hf(t_n + c_1 h, y_n + a_{21} k_1) \\ k_3 &= hf(t_n + c_1 h, y_n + a_{31} k_1 + a_{32} k_2) \\ &\cdots\cdots\cdots \\ k_s &= hf(t_n + c_s h, y_n + a_{s1} k_1 + \ldots\ldots + a_{s,s-1} k_{s-1}) \end{aligned} \qquad (12)$$

The method is called s-stage Runge-Kutta method. To derive a method, we first fix the number of K's to be used then expand the terms by Taylor series expandsious and compare with the left hand side. The parameters are determined such that the method is of suitable order. b_i are called the wights of the method.

To specify a particular method, we need to provide the integer s (the number of stages), and the coefficients a_{ij} (for $1 \leq j < i \leq s$), b_i (for $i = 1, 2, ..., s$) and c_i (for $i = 2, 3, ..., s$). These data are usually arranged in a mnemonic device, known as a *Runge–Kutta tableau*:

$$\begin{array}{c|cccc} 0 & & & & \\ c_2 & A_{21} & & & \\ c_3 & A_{31} & a_{32} & & \\ \vdots & \vdots & & \ddots & \\ c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} \\ \hline & B_1 & b_2 & \cdots & b_{s-1} \quad b_s \end{array}$$

The Runge–Kutta method is consistent if

$$\sum_{j=1}^{n-1} a_{ij} = c_i \text{ for } i = 2, \ldots . s.$$

There are also accompanying requirements if we require the method to have a certain order p, meaning that the truncation error is $O(h^{p+1})$. These can be derived from the definition of the truncation error itself.

Examples

The RK4 method falls in this framework. Its tableau is:

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

However, the simplest Runge–Kutta method is the Euler method, given by the formula $y_{n+1} = y_n + hf(t_n, y_n)$. This is the only consistent explicit Runge–Kutta method with one stage. The corresponding tableau is:

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array}$$

An example of a second-order method with two stages is provided by the midpoint method

$$y_{n+1} = y_n + hf\left(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)\right) \tag{13}$$

The corresponding tableau is:

0	
1/2	1/2
	0 1

Another example of a two-stage, second user, explicit Runge Kutta method, is,

0	
2/3	2/3
	1/4 ¾

Or

$$y_{n+1} = y_n + hf\frac{1}{4}(k_1 + 3k_2)$$

where $k_1 = hf(t_n, y_n)$; $k_2 = hf(t_n + \frac{2}{3}h, y_n + \frac{2}{3}k_1)$ (14)

Example 6: Solve the initial-value problem $y' = (\tan y) + 1, y(1) = 1, t \in [1, 1.1]$ with step size $h = 0.025$, using the second Runge-Kutta method defined by equation (14).

Solution: The method is given by

$$k_1 = hf(t_n, y_n), k_2 = hf\left(t_n + \frac{2}{3}h, y_n + \frac{2}{3}k_1\right)$$

$$y_{n+1} = y_n + \frac{1}{4}(k_1 + 3k_2)$$

We have $t_0 = 1, y_0 = 1, h = 0.025$

For $n = 0$:

$$k_1 = hf(1,1) = 0.025(2.2557407725) = 0.063935193$$

$$k_2 = hf\left(1 + \frac{0.05}{3}, 1 + \frac{2}{3}(0.063935193)\right) = 0.067847453$$

$$y(1.025) \approx y_1 = 1 + \frac{1}{4}(k_1 + 3k_2) = 1.066869388$$

For $n = 1$:

$$k_1 = hf(1.025, 1.066869388) = 0.070338117$$

$$k_2 = hf\left(1.025 + \frac{0.05}{3}, y_1 + \frac{2}{3}k_1\right) = 0.075837684$$

$$y(1.025) \approx y_2 = y_1 + \frac{1}{4}(k_1 + 3k_2)$$

$$= 1.066869388 + \frac{1}{4}[0.070338117 + 3(0.075837684)]$$

$$= 1.141332181$$

For $n = 2$:

$$k_1 = hf(1.05, 1.141332181) = 0.079588416$$

$$k_2 = hf\left(1.05 + \frac{0.05}{3}, y_2 + \frac{2}{3}k_1\right) = 0.088251234$$

$$y(1.075) \approx y_3 = y_2 + \frac{1}{4}(k_1 + 3k_2) = 1.227417711$$

For $n = 3$:

$$k_1 = hf(1.075, 1.227417711) = 0.094921694$$

$$k_2 = hf\left(1.075 + \frac{0.05}{3}, y_3 + \frac{2}{3}k_1\right) = 0.111908193$$

$$y(1.1) \approx y_4 = y_3 + \frac{1}{4}(k_1 + 3k_2) = 1.335079279$$

3.5 SUMMARY

In this unit we have:

1) devised Euler's Method
2) devised Runge-Kutta Method
3) devised Explicit Runge-Kutta Methods

3.6 SOLUTIONS/ANSWERS

Check Your Progress 1

1) The solution to the linear first order differential equation is $y(t) = e^{t/2}\sin(5t)$. Approximations at $t = 1$ are: with $h = 0.1$, $y(1) = -0.97167$, and with $h = 0.05$, $y(1) = -1.26512$. The percentage errors are 38.54 %, 19.98 % respectively.

Check Your Progress 2

1) $h = 0.2$, $k_1 = -0.1$, $k_2 = -0.085$, $k_3 = -0.08575$, $k_4 = -0.07(425$, $u(0.2) \approx 0.9145125$, $h = 0.1$, $u(0.1) \approx 0.95035$, $u(0.2) \approx 0.911726$.

2) 1.1749

3) 1.0207, 1.0438

4) $k_1 = 0.2$, $k_2 = 0.2400$, $k_3 = 0.2440$, $k_4 = 0.2888$, $y(0.2) \approx 0.2468$

Uttar Pradesh
Rajarshi Tandon Open University

MCA-5.3

Numerical and Statistical Computing

Block

1

NUMERICAL COMPUTING-I

COURSE INTRODUCTION

Block 1 and Block 2 of the course are providing the coverage of Numerical Computing, whereas Block 3 is providing coverage of Statistical Computing. While studying these blocks, you will find the concerned importance of these topics in computers.

This course is concerned with an introduction to methods for practical solutions of problems on computers. Mathematical modelling of physical or biological problems generally gives rise to ordinary or partial differential equations, integral/integero-differential equations, or a system of algebraic equations. Once a problem has been so formulated, the next step is to solve these equations. Only a few of these problems/ equations can be solved exactly by available analytical methods and most of these cannot be solved analytically. Thus, numerical methods, together with some error analysis, must be devised for solving such problems. A method which can be used to solve a problem will be called an algorithm. An algorithm is a complete and unambiguous set of procedures leading to the solution of a mathematical problem. The selection and construction of appropriate algorithms based on numerical computation properly falls within the scope of numerical analysis. Thus, numerical analysis deals with the development and analysis of numerical methods. The awareness of the order of error in a result computed by a numerical method is of great importance, and a computable estimate of the error in the method gives an idea about the accuracy of the result obtained.

After the choice of an algorithm for solving a problem has been made, one should consider all the sources of error that may affect the results. In numerical analysis, applied to specific problems, questions concerning the quantum of the required accuracy estimates of the magnitude of the round-off error and discretitation error, about appropriate step size or the number of iterations required making allowance for corrective action etc., are considered. The course assumes the basic knowledge of calculus and certain results from linear algebra.

This course consists of three blocks and each block contains three units. The first block contains: unit 1: floating point arithmetic and errors. Here, we discuss the floating point representation of numbers, and its consequences, errors in numbers in computation, etc. In successive units of the block, we will discuss the solution of non linear equations where we put light on several iterative methods to calculate the root of a non-linear equation, and the efficiency of these methods. Finally, we will discuss the topics related to the solution of linear algebraic equations.

Again, Block 2 consists of three units: Interpolation, Numerical integration and Numerical solution to ODE. In interpolation, we consider polynomial interpolation apart from the existence and uniqueness of interpolating polynomial. Several forms of interpolating polynomials like Lagrange form and Newton's divided difference form with error terms, Numerical Differentiation, Numerical Integration and Solution of Non-linear Ordinary Differential Equations. Using interpolating polynomials, we have obtained various numerical differentiation and integration formulas together with their error analyses. Next, to solve first order ordinary differential equations, various methods like Euler's method, Taylor Series method, and Runge-Kutta method are discussed.

Finally in Block 3 we have talked about Statistical Computing: where we have discussed various topics related to the concept of probability distribution, pseudo random number generation, and regression.

This is a very elementary introduction to the subject and restricted/limited treatments of topics are presented. The third phase of numerical problem solving is

programming, wherein the suggested algorithm is transformed into a set of unambiguous step-by-step instructions to the computer. That is, after producing a flow chart, if required, the indicated procedures must be transformed into a set of machine instructions by using any scientific language like C, C++, Fortran, etc. Interested students can refer to books listed below as reference books for this.

- Numerical Methods with programs in BASIC, FORTRAN, Pascal, C++ by S. Balachandra Rao & C.K. Shantha ; University Press.
- Numerical and Statistical methods in computers by V.K. Singh; Paragon International publishers.
- Numerical Methods for Scientific and Engineering computations by M.K.Jain, S.R.K. Iyengar , and R.K. Jain; Wiley Eastern Limited.
- Probability and Statistics by Murray R. Speigel ; Shaum's Outline Series – McGraw-Hill.

BLOCK INTRODUCTION

This block contains 3 units, covering the topics related to floating point arithmetic and errors, solutions of algebraic equations, etc. In Unit 1, we discuss the floating point representation of numbers, and its consequences, errors in numbers in computation etc. In successive unit of the block, we will discuss solution of non-linear equations where we put light on several iterative methods to calculate the root of a non-linear equation, and the efficiency of these methods. Finally we will discuss the topics related to the solution of linear algebraic equations.

UNIT 1 FLOATING POINT ARITHMETIC AND ERRORS

Structure **Page Nos.**

1.0 INTRODUCTION

Numerical Analysis is the study of computational methods for solving scientific and engineering problems by using basic arithmetic operations such as addition, subtraction, multiplication and division. The results obtained by using such methods, are usually approximations to the true solutions. These approximations to the true solutions introduce errors but can be made more accurate up to some extent. There can be several reasons behind this approximation, such as the formula or method used to solve a problem may not be exact. i.e., the expression of $\sin x$ can be evaluated by expressing it as an infinite power series. This series has to be truncated to the finite number of terms. This truncation introduces an error in the computed result. As a student of computer science you should also consider the computer oriented aspect of this concept of approximation and errors, say the machine involved in the computation doesn't have the capacity to accommodate the data or result produced by calculation of a numerical problem and hence the data is to be approximated in to the limitations of the machine. When this approximated data is to be further utilized in successive calculations, then it causes the propagation of error, and if the error starts growing abnormally then some big disasters may happen. Let me cite some of the well-known disasters caused because of the approximations and errors.

Instance 1: On February 25, 1991, during the Gulf War, an American Patriot Missile battery in Dhahran, Saudi Arabia, failed to intercept an incoming Iraqi Scud Missile. The Scud struck an American Army barracks and killed 28 soldiers. A report of the General Accounting office, GAO/IMTEC-92-26, entitled *Patriot Missile Defense: Software Problem Led to System Failure at Dhahran, Saudi Arabia* reported on the cause of the failure. It turns out that the cause was an inaccurate calculation of the time since boot due to computer arithmetic errors.

Instance 2: On June 4, 1996, an unmanned Ariane 5 rocket launched by the European Space Agency exploded just forty seconds after lift-off. The rocket was on its first voyage, after a decade of development costing $7 billion. A board of inquiry investigated the causes of the explosion and in two weeks issued a report. It turned out that the cause of the failure was a software error in the inertial reference system.

Specifically, a 64-bit floating point number relating to the horizontal velocity of the rocket with respect to the platform was converted to a 16-bit signed integer. The number was larger than 32,768, the largest integer storeable in a 16-bit signed integer, and thus the conversion failed.

In this Unit, we will describe the concept of number approximation, significant digits, the way, the numbers are expressed and arithmetic operations are performed on them, types of errors and their sources, propagation of errors in successive operations etc. The *Figure 1* describes the stages of Numerical Computing.

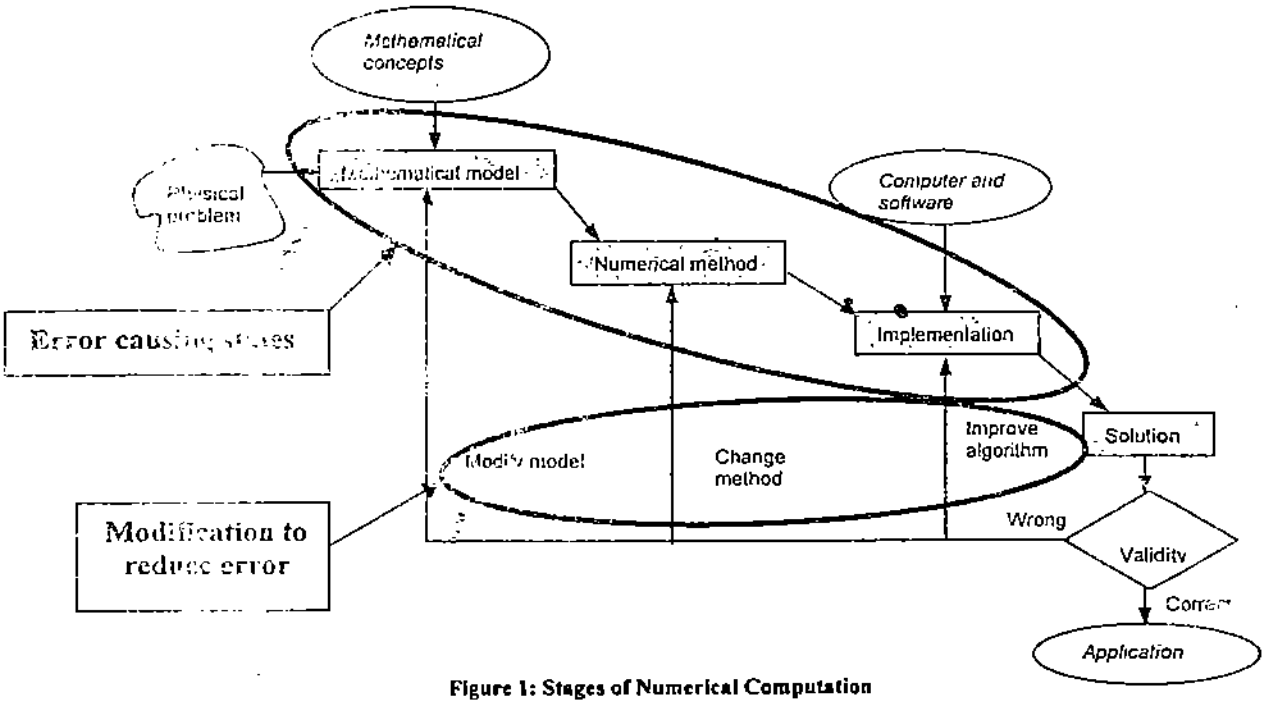


Figure 1: Stages of Numerical Computation

1.1 OBJECTIVES

After studying this unit, you should be able to:

- describe the concept of fixed point and floating point numbers representations;
- discuss rounding-off errors and the rules associated with round-off errors;
- implement floating-point arithmetic on available data;
- conceptual description of significant digits, and
- analysis of different types of errors – absolute error, relative errors, truncation error.

1.2 FLOATING POINT REPERESENTATIONS

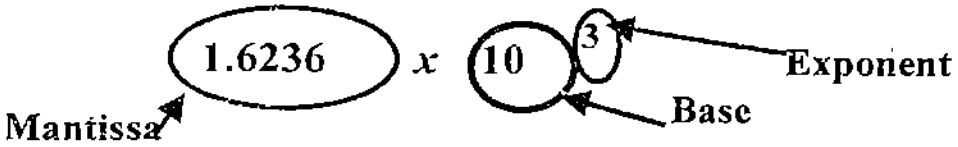
In scientific calculations, very large numbers such as velocity of light or very small numbers such as size of an electron occur frequently. These numbers cannot be satisfactorily represented in the usual manner. Therefore, scientific calculations are usually done by floating point arithmetic.

This means that we need to have two formats to represent a number, which are fixed point representation and floating point representation. We can transform data of one

format in to another and *vice versa*. The concept of transforming fixed point data into floating point data is known as normalisation, and it is done to preserve the maximum number of useful information carrying digits of numbers. This transformation ultimately leads to the calculation errors. Then, you may ask what is the benefit of doing this normalisation when it is contributing to erroneous results. The answer is simply to proceed with the calculations keeping in mind the data and calculation processing limitation of machine.

Fixed-Point numbers are represented by a fixed number of decimal places. Examples are 62.358, 1.001, 0.007 all correctly expressed up to 3rd decimal place.

Floating-Point numbers have a fixed number of significant places. Examples are 6.236×10^3 1.306×10^{-3} which are all given as four significant figures. The position of the decimal point is determined by the powers of base (in decimal number system it is 10) 1.6236×10^3.



Let us first discuss what is a floating-point number. Consider the number 123. It can be written using exponential notation as:

1.23×10^2, 12.3×10^2, 123×10^2, 0.123×10^2, 1230×10^2, etc.

Notice how the decimal point "floats" within the number as the exponent is changed. This phenomenon gives floating point numbers their name. The representations of the number 123 above are in kind of standard form. The first representation, 1.23×10^2, is in a form called "scientific notation".

In scientific computation, a real number x is usually represented in the form

$$x = \pm(d_1 d_2 \ldots\ldots d_n) x 10^m \qquad (1)$$

where $d_1, d_2, \ldots\ldots d_n$ are decimal digits and m is an integer called **exponent.** $(d_1 d_2 \ldots\ldots d_n)$ is called **significand** or **mantissa**. We denote this representation by **fl(x)**. A floating-point number is called a normalised floating-point number if $d_1 \neq 0$ or else $d_2 = d_3 = \ldots = d_n = 0$. The exponent m is usually bounded in a range

$$-M < m < M \qquad (2)$$

In scientific notation, such as 1.23×10^2 in the above example, the significand is always a number greater than or equal to 1 and less than 10. We may also write 1.23E2.

Standard computer normalisation for floating point numbers follows the fourth form namely, 0.123×10^3 in the list above.

In the standard normalized floating-point numbers, **the significand is greater than or equal to 0.1, and is always less than 1.**

In floating point notation (1), if $fl(x) \neq 0$ and $m \geq M$ (that is, the number becomes too large and it cannot be accommodated), then x is called an **over-flow number** and if

$m \leq -M$ (that is the number is too small but not zero) the number is called an **underflow number.** The number n in the floating-point notation is called its **precision.**

1.2.1 Floating Point Arithmetic

When arithmetic operations are applied on floating-point numbers, the results usually are not floating-point numbers of the same length. For example, consider an operation with 2 digit precision floating-point numbers (i.e., those numbers which are accurate up to two decimal places) and suppose the result has to be in 2 digit floating point precision. Consider the following example,

$x = 0.30x10^{1}, \quad y = 0.66x10^{-6}, \quad z = 0.10x10^{1}$

$$\begin{aligned} \text{then} \quad & x + y = 0.300000066 \times 10^{1} = 0.30 \times 10^{1} \\ & x \times y = 0.198 \times 10^{-5} \qquad = 0 \\ & z/x = 0.333\ldots \times 10^{0} \qquad = 0.33 \times 10^{0} \end{aligned} \tag{3}$$

Hence, if 0 is one of the arithmetic operations, and θ^* is corresponding floating-point operation, then we find that

$$x\ \theta^*\ y \neq x\ \theta\ y$$

However, $x\ \theta\ y = fl(x\ \theta\ y)$ (4)

1.2.2. Properties of Floating Point Arithmetic

Arithmetic using the floating-point number system has two important properties that differ from those of arithmetic using real numbers.

Floating point arithmetic is not associative. This means that in general, for floating point numbers x, y, and z:

- $(x + y) + z \neq x + (y + z)$
- $(x \,.\, y) \,.\, z \neq x \,.\, (y \,.\, z)$

Floating point arithmetic is also not distributive. This means that in general,

- $x \,.\, (y + z) \neq (x \,.\, y) + (x \,.\, z)$

Therefore, the order in which operations are carried out can change the output of a floating-point calculation. This is important in numerical analysis since two mathematically equivalent formulas may not produce the same numerical output, and one may be substantially more accurate than the other.

Example 1: Let $a = 0.345 \times 10^{0}$, $b = 0.245 \times 10^{-3}$ and $c = 0.432 \times 10^{-3}$. Using 3-digit decimal arithmetic with rounding, we have

$b + c = 0.000245 + 0.000432 = 0.000677$ (in accumulator)
$= 0.677 \times 10^{-3}$

$a + (b + c) = 0.345 + 0.000677$ (in accumulator)
$= 0.346 \times 10^{0}$ (in memory) with rounding

$a + b = 0.345 \times 10^{0} + 0.245 \times 10^{-3}$
$= 0.345 \times 10^{0}$ (in memory)

$(a + b) + c = 0.345432$ (in accumulator)
$= 0.345 \times 10^{0}$ (in memory)

Hence, we see that,
$(a + b) + c \neq a + (b + c)$.

Example 2: Suppose that in floating point notation (1) given above, n = 2 and m = 11. Consider $x = 0.10x10^{10}$, $y = -0.10x10^{10}$ and $z = 0.10x10^{1}$. Then,

$(x+y)+z = 0.1x10^{1}$ while $x+(y+z) = 0.0$.

Hence, $(x + y) + z \neq x + (y + z)$.

From the above examples, we note that in a computational process, every floating-point operation gives rise to some error, which may then get amplified or reduced in subsequent operations.

☞ Check Your Progress 1

1) Let a = 0.41, b = 0.36 and c = 0.70. Prove $\frac{(a-b)}{c} \neq \frac{a}{c} - \frac{b}{c}$.

.....

.....

.....

2) Let a = .5665E1, b = .5556E – 1, c = .5644E1. Verify the associative property for the floating point numbers i.e., prove (a + b) – c # (a – c) + b.

.....

.....

.....

Let a = .5555E1, b = .4545E1, c = .4535E1. Verify the distributive property for these floating point numbers, i.e., prove a(b – c) # ab –ac.

.....

.....

.....

1.2.3 Significant Digits

The concept of significant digits has been introduced primarily to indicate the accuracy of a numerical value. For example, if, in the number $y = 23.40657$, only the digits 23406 are correct, then we may say that y has given significant digits and is correct to only three decimal places.

The number of significant digits in an answer in a calculation depends on the number of significant digits in the given data, as discussed in the rules below.

When are Digits Significant?

Non-zero digits are always significant. Thus, 22 has two significant digits, and 22.3 has three significant digits. The following rules are applied when zeros are encountered in the numbers,

a) Zeros placed before other digits are not significant; 0.046 has two significant digits.

b) Zeros placed between other digits are always significant; 4009 kg has four significant digits.

c) Zeros placed after other digits but behind a decimal point are significant; 7.90 has three significant digits.

d) Zeros at the end of a number are significant only if they are behind a decimal point as in (c). For example, in the number 8200, it is not clear if the zeros are significant or not. The number of significant digits in 8200 is at least two, but could be three or four. To avoid uncertainty, we use scientific notation to place significant zeros behind a decimal point.

$8.200 * 10^3$ has four significant digits,

$8.20 * 10^3$ has three significant digits,

$8.2 * 10^3$ has two significant digits.

Note: Accuracy and precision are closely related to significant digits. They are related as follows:

1) Accuracy refers to the number of significant digits in a value. For example, the number 57.396 is accurate to five significant digits.

2) Precision refers to the number of decimal positions, i.e. the order of magnitude of the last digit in a value. The number 57.396 has a precision of 0.001 or 10^{-3}.

Example 1: Which of the following numbers has the greatest precision?

a) 4.3201, b) 4.32, c) 4.320106.

Solution:

a) 4.3201 has a precision of 10^{-4}
b) 4.32 has a precision of 10^{-2}
c) 4.320106 has a precision of 10^{-6}

The last number has the greatest precision.

Example 2: What is the accuracy of the following numbers?

a) 95.763, b) 0.008472, c) 0.0456000, d) 36 e) 3600.00.

Solution:

a) This has five significant digits.
b) This has four significant digits. The leading or higher order zeros are only place holders.
c) This has six significant digits.
d) This has two significant digits.
e) This has six significant digits. Note that the zeros were made significant by writing .00 after 3600.

Significant digits in Multiplication, Division, Trigonometry functions, etc.

In a calculation involving multiplication, division, trigonometric functions, etc., the number of significant digits in an answer should equal the least number of significant digits in any one of the numbers being multiplied, divided, etc.

Thus, in evaluating sin(kx), where $k = 0.097\ m^{-1}$ (two significant digits) and $x = 4.73$ m (three significant digits), the answer should have two significant digits.

Note that whole numbers have essentially an unlimited number of significant digits. As an example, if a hairdryer uses 1.2 kW of power, then 2 identical hairdryers use 2.4 kW.

1.2 kW{2 significant digit} x 2{unlimited significant digit} = 2.4 kW
{2 significant digit}

Significant digits in Addition and Subtraction

When quantities are being added or subtracted, the number of *decimal places* (not significant digits) in the answer should be the same as the least number of decimal places in any of the numbers being added or subtracted.

Keep one extra digit in Intermediate Answers

When doing multi-step calculations, *keep at least one or more significant digits in intermediate results* than needed in your final answer.

For instance, if a final answer requires two significant digits, then carry at least three significant digits in calculations. If you round-off all your intermediate answers to only two digits, you are discarding the information contained in the third digit, and as a result the *second* digit in your final answer might be incorrect. (This phenomenon is known as "round-off error.")

This truncation process is done either through rounding off or chopping, leading to round off error.

Example 3: Let x = 4.5 be approximated to $x^* = 4.49998$. Then,

$$x^* - x = -0.00002,$$

$$\frac{|x - x^*|}{x} = 0.0000044 \le 0.000005 \le \frac{1}{2}(.00001) = \frac{1}{2}10^{-5} = \frac{1}{2} \times 10^{1-6}$$

Hence, x approximates x correct to 6 significant decimal digits.

Wrong way of writing significant digits

1) Writing more digits in an answer (intermediate or final) than justified by the number of digits in the data.

2) Rounding-off, say, to two digits in an intermediate answer, and then writing three digits in the final answer.

Example 4: Expressions for significant digits and scientific notation associated with a floating point number.

Number	Number of Significant Figures	Scientific Notation	
0.00682	3	$6.82 * 10^{-3}$	Leading zeros are not significant.
1.072	4	$1.072 (* 10^0)$	Embedded zeros are always significant.
300	1	$3 * 10^2$	Trailing zeros are significant only if the decimal point is specified.
300	3	$3.00 * 10^2$	
300.0	4	$3.000 * 10^2$	

Loss of Significant Digits

One of the most common (and often avoidable) ways of increasing the importance of an error is known as loss of significant digits.

Loss of significant digits in subtraction of two nearly equal numbers:

Subtraction of two nearly equal number gives the relative error

$$r_{x-y} = r_x \frac{x}{x-y} - r_y \frac{y}{x-y}$$

which becomes very large. It has largest value when r_x and r_y are of opposite signs.

Suppose we want to calculate the number $z = x - y$ and x^* and y^* are approximations for x and y respectively, accurate to r digits and assume that x and y do not agree in the most left significant digit, then $z^* = x^* - y^*$ is as good an approximation to $x - y$ as x^* and y^* to x and y.

But, if x^* and y^* agree at left most digits (one or more), then the left most digits will cancel and there will be loss of significant digits.

The more the digits on left agrees, the more loss of significant digits. A similar loss in significant digits occurs when a number is divided by a small number (or multiplied by a very large number).

Remark 1: To avoid this loss of significant digits in algebraic expressions, we must rationalise these numbers. If no alternative formulation to avoid the loss of significant digits is possible, then we can carry more significant digits in calculation using floating-point numbers in double precision.

Example 5: Solve the quadratic equation $x^2 + 9.9x - 1 = 0$ using two decimal digit arithmetic with rounding.

Solution:

Solving the quadratic equation, we have one of the solutions as

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{-9.9 + \sqrt{(9.9)^2 - 4.1.(-1)}}{2}$$

$$= \frac{-9.9 + \sqrt{102}}{2} = \frac{-9.9 + 10}{2} = \frac{0.1}{2} = 0.05$$

while the true solutions are -10 and 0.1. Now, if we rationalize the expression, we obtain

$$= \frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{-4ac}{2a(b + \sqrt{b^2 - 4ac})}$$

$$= \frac{-2c}{b + \sqrt{b^2 - 4ac}} = \frac{2}{9.9 + \sqrt{102}} = \frac{2}{9.9 + 10} = \frac{2}{19.9} = \frac{2}{20} \equiv 0.1 \ . \ (0.1000024)$$

which is one of the true solutions.

1.3 ERROR - BASICS

What is Error?

An error is defined as the difference between the actual value and the approximate value obtained from the experimental observation or from numerical computation. Consider that x represents some quantity and xa is an approximation to x, then

Error = actual value – approximate value = x – xa

How errors are generated in computers?

Every calculation has two parts, one is operand and other is operator. Hence, any approximation in either of the two contributes to error. Approximations to operands causes propagated error and approximation to operators causes generated errors. Let us discuss how the philosophy behind these errors is related to computers.

Operand Point of View: Computers need fixed numbers to do processing, which is mostly not available. Hence, we need to transform the output of an operation to a fixed number by performing truncation of series, rounding, chopping etc. This contributes to difference between exact value and approximated value. These errors get further amplified in subsequent calculations as these values and the results produced are further utilized in subsequent calculations. Hence, this error contribution is referred to as **propagated error.**	**Operator Point of View:** Computers need some operation to be performed on the operands available. Now, the operations that occur in computers are at bit level and complex operations are simplified. There are, hence, small changes in actual operations and operations performed by computer. This difference in operations produces errors in calculations, which get further amplified in subsequent calculations. This error contribution is referred to as **generated error.**

What are the sources of error?

The sources of error can be classified as (i) data input errors, (ii) errors in algorithms and (iii) errors during computations.

Sources of Error?

Data Input	Algorithms	Computations
Input Error: The input information is rarely exact since it comes from experiments and any experiment can give results of only limited accuracy. Moreover, the quantity used can be represented in a computer for only a limited number of digits.	**Algorithmic Errors:** If direct algorithms based on finite sequence of operations are used, errors due to limited steps don't amplify the existing errors. But if algorithms with infinite steps are used, the algorithm has to be stopped after a finite number of steps	**Computational Errors:** Even when elementary operations, such as multiplication or division are used, the number of digits increases greatly so that the number cannot be held fully in register available in a given computer. In such a case, a certain number of digits must be discarded, and this again leads to error

Type of Errors?

We list below the types of errors that are encountered while carrying out numerical calculations to solve a problem.

1) Round off errors arise due to floating point representation of initial data in the machine. Subsequent errors in the solution due to this are called propagated errors.

2) Due to finite digit arithmetic operations, the computer produces generated errors or rounding errors.

3) Error due to finite representation of an inherently infinite process. For example, consider the use of a finite number of terms in the infinite series expansions of Sin x, Cos x or f(x) by Maclaurin's or Taylor Series expression. Such errors are called truncation errors.

Remark 2: Sensitivity of an algorithm for a numerical process used for computing f(x): if small changes in the initial data x lead to large errors in the value of f(x), then the algorithm is called *unstable*.

How error measures accuracy?

The two terms "error" and " accuracy" are inter-related, one measures the other, in the sense less the error is, more the accuracy is and vice versa. In general, the errors which are used for determination of accuracy are categorized as:

a) Absolute error b) Relative error c) Percentage error

Now, we define these errors.

a) **Absolute Error:** Absolute error is the magnitude of the difference between the true value x and the approximate value xa. Therefore, absolute error = $|x - xa|$.

b) **Relative Error:** Relative error is the ratio of the absolute error and actual value Therefore, relative error = $|x - xa| / x$.

c) **Percentage Error:** Percentage error is defined as,
percentage error = 100er= $100 * |x - xa|/x$.

Now, we discuss each of the errors defined above, and its propagation in detail.

1.3.1 Rounding-off Error

There are two ways of translating a given real number x into floating-point number f(x) – rounding and chopping. For example, suppose we want to represent the number 5562 in the normalized floating point representation. The representations for different values of n are as follows:

$n = 1$, $fl(5562) = .5 * 10^4$ chopped
$= .6 * 10^4$ rounded . (5)

$n = 2$, $fl(5562) = .55 * 10^4$ chopped
$= .56 * 10^4$ rounded . (6)

$n = 3$, $fl(5562) = .556 * 10^4$ chopped
$= .556 * 10^4$ rounded . (7)

Rules for rounding-off: Whenever, we want to use only a certain number of digits after the decimal point, then number is rounded-off to that many digits. A number is rounded-off to n places after decimal by seeing (n+1)th place digit d_{n+1}, as follows:

i) If $d_{n+1} < 5$, then it is chopped
ii) If $d_{n+1} > 5$, then $d_n = d_n + 1$
iii) If $d_n = 5$, and d_n is odd then $d_n = d_n + 1$ else the number d_{n+1} is chopped.

The difference between a number x and fl(x) is called the **round-off error.** It is clear that the round-off error decreases when precision increases. The round-off error also depends on the size of x and is therefore represented relative to x as

$$fl(x) = x(1 + \delta). \tag{8}$$

It is not difficult to show that

$|\delta| < .5*10^{-(n-1)}$ in rounding

while, $-10^{-(n-1)} < \delta \leq 0$ in chopping. (9)

Definition 1: Let x be a real number and x^* be a real number having non-terminal decimal expansion, then we say x^* represents x **rounded to k decimal places** if $|x - x^*| \leq \frac{1}{2}10^{-k}$, where k is a positive integer.

Example 6: If $\pi = 3.14159265$, then find out to how many decimal places the approximate value of 22/7 is accurate?

Solution: We find that

$$\left|\pi - \frac{22}{7}\right| = 0.00126449$$

Since, $0.00126449 < 0.005 = \frac{1}{2}10^{-2}$. Hence, k = 2, and we conclude that the approximation is accurate to 2 decimal places or three significant digits.

☞ Check Your Progress 2

1) Round off the following numbers to four significant digits.

(i) 450.92, (ii) 48.3668, (iii) 9.3265, (iv) 8.4155,
(v) 0.80012, (vi) 0.042514, (vii) 0.0049125, (viii) 0.00020215

.....

.....

.....

.....

2) Write the following numbers in floating-point form rounded to four significant digits.

(i) 100000, (ii) –0.0022136, (iii) –35.666

.....

.....

.....

3) The numbers 28.483 and 27.984 are both approximate and are correct up to the last digit shown. Compute their difference. Indicate how many significant digits are present in the result and comment.

.....

.....

.....

.....

4) Consider the number 2/3. Its floating point representation rounded to 5 decimal places is 0.66667. Find out to how many decimal places the approximate value of 2/3 is accurate?

.....

.....

.....

.....

5) Find out to how many decimal places the value 355/133 is accurate as an approximation to π ?

.....

.....

.....

.....

1.3.2 Absolute and Relative Errors

We shall now discuss two types of errors that are commonly encountered in numerical computations. You are already familiar with the rounding off error. These rounded-off numbers are approximations of the actual values. In any computational procedure, we make use of these approximate values instead of the true values. How do we measure the goodness of an approximation fl(x) to x ? The simplest measure which naturally comes to our mind is the difference between x and fl(x). This measure is called the error. Formally, we define error as a quantity which satisfies the identity

$$x = \text{fl}(x) + e, \tag{10}$$

If error e is considerably small, then we say that fl(x) is a good approximation of x. Error can be positive or negative. We are in general interested in the magnitude or absolute value of the error which is defined as follows

$$|e| = |x - \text{fl}(x)| \tag{11}$$

Sometimes, when the true value x is very large or very small, we prefer to study the error by comparing it with the true value. This is known as relative error and we define this error as

$$\text{relative error} = r_e = \frac{x - f(x)}{x}$$

and

$$|\text{relative error}| = \left|\frac{x - \text{fl}(x)}{x}\right| = \left|\frac{e}{x}\right| \tag{12}$$

Note that in certain computations, the true value may not be available. In that case, we replace the true value by the computed approximate value in the definition of relative error.

Theorem: If fl(x) is the n-digit floating point representation in base β of a real number x, then r_x the relative error in x, satisfies the following:

i) $|r_x| < \frac{1}{2}\beta^{1-n}$ if rounding is used.

ii) $0 \le |r_x| \le \beta^{1-n}$ if chopping is used.

For proving i), you may use the following:

Case 1. $d_{n+1} < \frac{1}{2}\beta$, *then* $fl(x) = \pm(.d_1d_2...d_n)\beta^e$

$$|x-fl(x)| = d_{n+1}.d_{n+2}...\beta^{e-n-1}$$

$$\le \frac{1}{2}\beta.\beta^{e-n-1} = \frac{1}{2}\beta^{e-n}$$

Case 2. $d_{n+1} \ge \frac{1}{2}\beta$,

$$fl(x) = \pm\{(.d_1d_2...d_n)\beta^e + \beta^{e-n}\}$$

$$|x-fl(x)| = \left|-d_{n+1}.d_{n+2}...\beta^{e-n-1} + \beta^{e-n}\right|$$

$$= \beta^{e-n-1}\left|d_{n+1}.d_{n+2} - \beta\right|$$

$$\le \beta^{e-n-1} \times \frac{1}{2}\beta = \frac{1}{2}\beta^{e-n}$$

Example 7: The true value of π is 3.14159265... In menstruation problems the value 22/7 is commonly used as an approximation to π. What is the error in this approximation?

Solution: The true value of π is π = 3.14159265.

Now, we convert 22/7 to decimal form, so that we can find the difference between the approximate value and true value. Then, the approximate value of π is $\frac{22}{7} = 3.14285714$

Therefore, absolute error = 0.00126449 and relative-error = 0.00040249966.

The round-off error of computer representation of the number π depends on how many digits are left out. Make sure that you understand each line of the following rounding off of the number π:

Number of digits	Approximation for π	absolute error	relative error
1	3.100	0.041593	0.0132%
2	3.140	0.001593	0.0507%
3	3.142	0.000407	0.0130%

Round-off errors may accumulate, propagate and even lead to catastrophic cancellations leading to loss of accuracy of numerical calculations.

☞ Check Your Progress 3

1) Let $x^* = .3454$ and $y^* = .3443$ be approximations to x and y respectively correct to 3 significant digits. Further, let $z^* = x^* - y^*$ be the approximation to $x - y$. Then show that the relative error in z^* as an approximation to $x - y$ can be as large as 100 times the relative error in x or y.

.....

.....

.....

.....

2) Round the number $x = 2.2554$ to three significant figures. Find the absolute error and the relative error.

.....

.....

.....

.....

3) If $\pi = 3.14$ instead of 22/7, find the relative error and percentage error.

.....

.....

.....

.....

4) Determine the number of correct digits in $s = 0.2217$, if it has a relative error, $\varepsilon_r = 0.2 * 10^{-1}$.

.....

.....

.....

.....

5) Round-off the number 4.5126 to four significant figures and find the relative percentage error.

.....

.....

.....

.....

1.3.3 Truncation Error

Truncation error is a consequence of doing only a finite number of steps in a calculation that would require an infinite number of steps to do exactly. A simple example of a calculation that will be affected by truncation error is the evaluation of an infinite sum. The computer uses only a finite number of terms and the terms that are left out lead to truncation error.

Numerical integration is another example of an operation that is affected by truncation error. A quadrature formula works by evaluating the integrand at a finite number of points and using smooth functions to approximate the integrand between those points. The difference between those smooth functions and the actual integrand leads to truncation error.

Taylor series represents the local behaviour of a function near a give n point. If one replaces the series by the n-th order polynomial, the truncation error is said to be **order of n**, or $O(h^n)$, where h is the distance to the given point. Consider the irrational number e

$e = 2.71828182845905\ldots$

and compare it with the Taylor series of the function $exp(x)$ near the given point $x = 0$.

$\exp(x) = 1 + x + x/2 + x^3/6 + \ldots..$

Let us check a few Taylor series approximations of the number $e = exp(1)$:

order of n	approximation for e	absolute error	relative error
3	2.500000	0.218282	8.030140%
4	2.666667	0.051615	1.898816%
5	2.708333	0.009948	0.365984%

Example 8: Find the value of e correct to three decimal places.

Solution: Recall that $e = 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \ldots\ldots\ldots$

The series is to be truncated such that the finite sum equals e to three decimal places. This means the must be less than 0.0005. Suppose that the tail starts at n = k+1. Then,

$$\sum_{n=k+1}^{\infty} \frac{1}{n!} = \frac{1}{(k+1)!} + \frac{1}{(k+2)!} \ldots\ldots + \ldots.$$

$$< \frac{1}{(k+1)!}\left[1 + \frac{1}{(k+1)} + \frac{1}{(k+1)^2} + \ldots\ldots\right.$$

$$= \frac{1}{(k+1)!}\left[\frac{(k+1)}{1-/(k+1)}\right] = \frac{1}{k!k} < 0.0005$$

For k = 6, This expression is satisfied and the truncated value of $e = 2.7181$.

1.4 SUMMARY

In this unit, we have defined the floating point numbers and their representation for usage in computers. We have defined accuracy and number of significant digits in a given number. We have also discussed the sources of errors in computation. We have defined the round-off and truncation errors and their propagation in later computations

using these values, which contains errors. Therefore, care must be taken to analyise the computations, so that we are sure that the output of computations is meaningful.

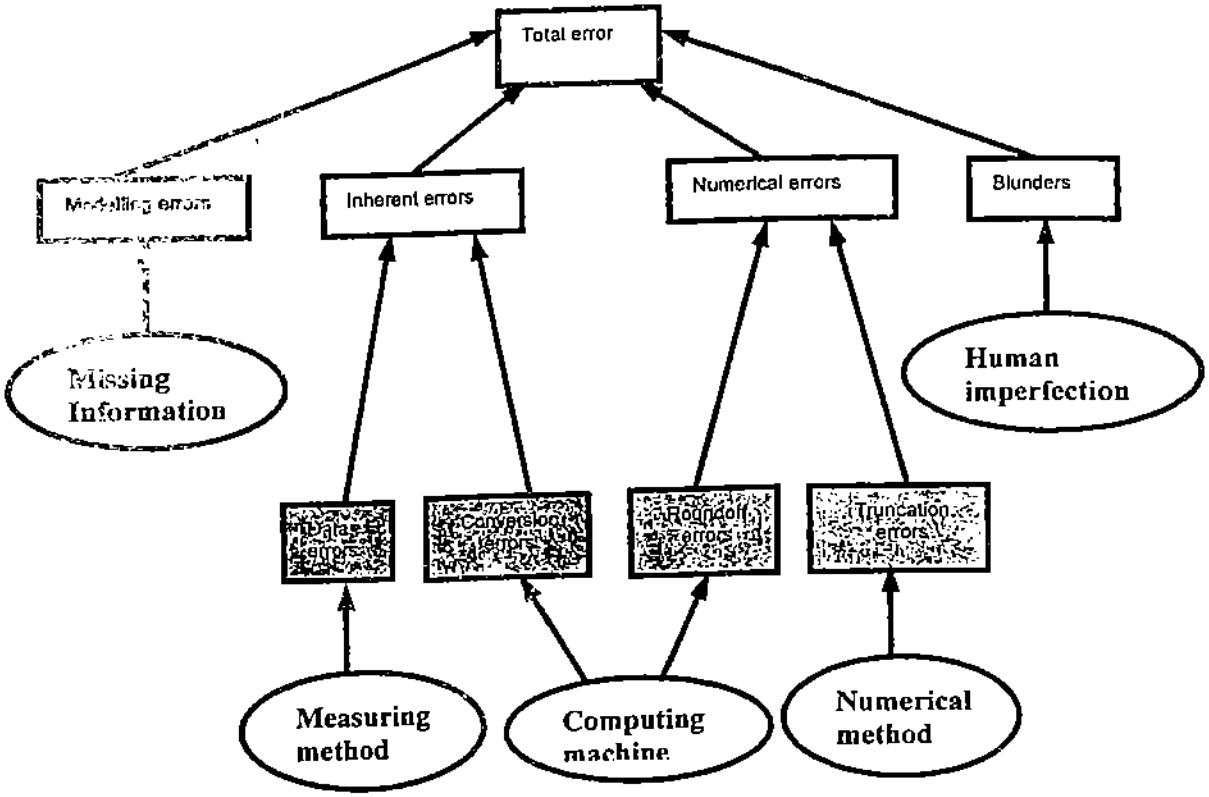


Figure 2: Types of error and their contribution to total errors

1.5 SOLUTIONS/ANSWERS

Check Your Progress 1

1) Using two decimal digit arithmetic with rounding we have,

$$\frac{(a-b)}{c} = .71 \times 10^{-1} \text{ and } \frac{a}{c} - \frac{b}{c} = .59 - .51 = .80 \times 10^{-1}$$

while true value of $\frac{(a-b)}{c} = 0.071428 \ldots$.

Therefore, $\frac{(a-b)}{c} \neq \frac{a}{c} - \frac{b}{c}$

2) Do as 1) above.
3) Do as 1) above.

Check Your Progress 2

1) (i) 50.9 (ii) 48.37 (iii) 9.326 (iv) 8.416 (v) 0.8001 (vi) 0.04251 (vii) 0.004912 (viii) 0.0002022

2) (i) $1000 * 10^2$ or $0.1000 * 10^6$ (ii) $-0.2214 * 10^{-2}$ (iii) $-0.3567 * 10^2$

3) We have 28.483 – 27.984 = 00.499. The result has only three significant digits. This is due to the loss of significant digits during subtraction of nearly equal numbers.

4) We find that $|2/3 - 0.66667| = 0.0000033... < \frac{1}{2}10^{-5}$

We find, k = 5. Therefore, the approximation is accurate to 5 decimal places.

5) Left as an exercise.

Check Your Progress 3

1) Given, $|r_x|, |r_y| \leq \frac{1}{2}10^{1-3}$

$z^* = x^* - y^* = 0.3454 - 0.3443 = 0.0011 = 0.11 \times 10^{-2}$.

This is correct to one significant digit since last digits 4 in x^* and 3 in y^* are not reliable and second significant digit of i^* is derived from the fourth digits of x^* and y^*.

Max. $|r_z| = \frac{1}{2}\ 10^{1-1} = \frac{1}{2} = (100).\left(\frac{1}{2}\right).10^{-2} \geq 100\ |r_x|, 100\ |r_y|$

2) The rounded-off number is 2.25. The absolute error is 0.0054.

The relative error is $\approx \frac{0.0054}{2.25} = 0.0024$. The percentage error is 0.24%.

3) Relative error = $\left(\frac{22}{7} - 3.14\right)\Big/\frac{22}{7} = 0.00093$. Percentage error = 0.093 %.

4) Absolute error = $0.2 * 10^{-1} * 0.2217 = 0.04493$. Hence x has only one correct digit $x \approx 0.2$.

5) The number 4.5126 round-off to four significant figures is 4.153.

Relative percentage error = $\frac{-0.0004}{4.5126} * 100 = -0.0088\%$.

1.6 EXERCISES

E1) Give the floating-point representation of the following numbers in 2 decimal digit and 4 decimal digit floating point number using (i) rounding and (ii) chopping.

(a) 37.21829
(b) 0.022718
(c) 3000527.11059

E2) Show that $a(b - c) \neq ab - ac$, where, $a = .5555 \times 10^1$, $b = .4545 \times 10^1$, $c = .4535 \times 10^1$.

E3) How many bits of significance will be lost in the following subtraction? 37.593621 – 37.584216. Assume each number is correct to seven significant digits.

E4) What is the relative error in the computation of x – y, where x = 0.3721448693 and y = 0.3720214371 with five decimal digit of accuracy?

E5) Find the smaller root in the magnitude of the quadratic equation $x^2 + 111.11x + 1.2121 = 0$, using five-decimal digit floating point chopped arithmetic.

1.7 SOLUTIONS TO EXERCISES

E1) a)

rounding	chopping
$.37 \times 10^2$	$.37 \times 10^2$
$.3722 \times 10^2$	$.3721 \times 10^2$

b)

$.23 \times 10^{-1}$	$.22 \times 10^{-1}$
$.2272 \times 10^{-1}$	$.2271 \times 10^{-1}$

c)

$.31 \times 10^2$	$.30 \times 10^2$
$.3056 \times 10^2$	$.3055 \times 10^2$

E2) Let, $a = .5555 \times 10^1$, $b = .4545 \times 10^1$, $c = .4535 \times 10^1$

$b - c = .0010 \times 10^1 = .1000 \times 10^{-1}$

$a(b - c) = (.5555 \times 10^1) \times (.1000 \times 10^{-1}) = .05555 \times 10^0 = .5550 \times 10^{-1}$

$ab = (.5555 \times 10^1)(.4545 \times 10^1) = (.2524 \times 10^2)$

$ac = (.5555 \times 10^1)(.4535 \times 10^1) = (.2519 \times 10^2)$

and $ab - ac = .2524 \times 10^2 - .2519 \times 10^2 = .0005 \times 10^2 = .5000 \times 10^{-1}$

Hence $a(b - c) \neq ab - ac$.

E3) $37.593621 - 37.584216 = (0.37593621)10^2 - (0.37584216)10^2$
$= x^* - y^* = (0.00009405)10^2$

The numbers are, correct to seven significant digits. Then, in eight digit floating-point arithmetic, the number can be written as
$z^* = x^* - y^* = (0.94050000)10^{-2}$ But as an approximation to $z = x - y$, z^* is good only to three digits, since the fourth significant digit of z^* is derived from the eighth digits of x^* and y^*, and both possibly contains errors. Here, while the error in z^* as an approximation to $z = x - y$ is at most the sum of the errors in x^* and y^*, the relative error in z^* is possibly 10,000 times the relative error in x or y^*. Loss of significant digits is, therefore, dangerous only if we wish to keep the relative error small.

Given $|r_x|, |r_y| < \frac{1}{2}10^{1-7}$, $z^* = (0.9405)10^{-2}$, is correct to three significant digits.

$$\text{Max } |r_z| = \frac{1}{2}10^{1-3} = 10000.\frac{1}{2}10^{-6} \geq (1000)|r_x| + (10000)|r_y|$$

E4) With five decimal digit accuracy $x^* = 0.37214 \times 10^0$, $y^* = 0.37202 \times 10^0$, $x^* - y^* = 0.00012$ while $x - y = 0.0001234322$.

$$\frac{\left|(x-y)-(x^*-y^*)\right|}{|x-y|} = \frac{0.0000034322}{0.0001234322} \approx 3 \times 10^{-2}.$$

The magnitude of this relative error is quite large when compared with the relative errors of x^* and y^* (which cannot exceed 5×10^{-5} and in this case it is approximately 1.3×10^{-5})

E5) Using the formula

$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$, we get, $x_1 = \frac{-111.11 + 111.09}{2} = -0.01000$

while the true solution is $x_t = -0.010910$, correct to the number of digits shown.

However, if we calculate x_1 as $x_1 = \frac{2c}{b + \sqrt{b^2 - 4ac}}$, we get

$$x_1 = \frac{-2 \times 1.2121}{111.11 + 111.09} = \frac{-2.4242}{222.20}$$

$$= -\frac{24242}{2222000} = -0.0109099 = -.0109099$$

which is accurate to five digits.

UNIT 2 SOLUTION OF ALGEBRAIC AND TRANSCENDENTAL EQUATIONS

2.0 INTRODUCTION

We often come across equations of the forms $x^4 - 3x^3 + x^2 + 6x - 5 = 0$ or $e^x + x - 2 = 0$ etc. Finding one or more values of x which satisfy these equations is one of the important problems in Mathematics.

An equation of the type $f(x) = 0$ is algebraic if it contains power of x, that is, f(x) is a polynomial. The equation is called transcendental, if it contains powers of x, exponential functions, logarithm functions etc.

Example of algebraic equatoins:

$$2x = 5, \quad x^2 + x = 1, \quad x^7 = x(1 + 2x).$$

Example of transcendental equations

$$x + \sin x = 0, \quad e^{\sqrt{x}} = x, \quad \tan x = x.$$

As we know, direct methods can be used to solve the polynomial equations of fourth or lower orders. We do not have any direct methods for finding the solution of higher order polynomial equations or transcendental equation. In these cases, we use numerical methods to solve them.

In this unit , we shall discuss some numerical methods which give approximate solutions of an equation f(x) = 0. These methods are iterative in nature. An iterative method gives an approximate solution by repeated application of a numerical process. In an iterative method, we start with an initial solution and the method improves this solution until it is improved to acceptable accuracy.

Properties of polynomial equations:

i) The total number of roots of an algebraic equation is the same as its degree.

ii) An algebraic equation can have at most as many positive roots as the number of changes of sign in the coefficients of f(x).

iii) An algebraic equation can have at most as many negative roots as the number of changes of sign in the coefficient of f(– x).

iv) If $f(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \ldots + a_{n-1}x + a_n$ have roots $\alpha_1,\alpha_2,\ldots,\alpha_n$, then the following hold good:

$$\sum_i \alpha_i = \frac{-a_1}{a_0}, \quad \sum_{i<j} \alpha_i\alpha_j = \frac{a_2}{a_0}, \quad \prod_i \alpha_i = (-1)^n \frac{a_n}{a_0}.$$

2.1 OBJECTIVES

After studying this unit, you should be able to :

- find an initial guess of a root;
- use bisection method;
- use Regula-falsi method;
- use Newton's Method;
- use Secant Method, and
- use successive iterative method.

2.2 INITIAL APPROXIMATION TO A ROOT

All the numerical methods have in common the requirement that we need to make an initial guess for the root. Graphically, we can plot the equation and make a rough estimate of the solution. However, graphic method is not possible to use in most cases. We with to determine analytically an approximation to the root.

Intermediate Value Theorem

This theorem states that if f is a continous function on [a, b] and the sign of f(a) is different from the sign of f(b), that is f(a)f(b) < 0, then there exists a point c, in the interval (a, b) such that f(c) = 0. Hence, any value c ε (a, b) can be taken as an initial approximation tc the root.

Example 1: Find an initial guess to find a root of the equation, $2x - \log_{10} x = 7$.

Solution: Let $f(x) = 2x - \log_{10} x - 7$. The values of function f are as given in *Table 1*.

Table 1

X	1	2	3	4
f(x)	– 5	– 3.301	– 1.477	0.397

We find f(3)f(4) < 0. Hence, any values in (3, 4) can be taken as an initial guess.

Example 2: Estimate an initial guess to find a root of the equation,
$2x - 3 \sin x - 5 = 0$.

Solution: Let $f(x) = 2x - 3 \sin x - 5$. Note that $f(-x) = -2x + 3 \sin x - 5$ which is always negative. Therefore, the function f(x) has no negative real roots. We tabulate the values of the function for positive x, in Table 2.

Table 2

x	0	1	2	3
f(x)	– 5	– 5.5224	– 3.7278	0.5766

Since f(2) and f(3) are of opposite signs, a root lies between 2 and 3. The initial guess can be taken as any value in (2, 3).

2.3 BISECTION METHOD

This is one of the simplest methods and is based on the repeated application of the intermediate value theorem.

The bisection method is defined as follows:

i) Find an interval (a, b) in which in root lies, using intermediate value theorem.
ii) Direction the interval (a, b). Let c = (a + b)/2. If f(c) = 0, then x = c is the root and the root is determined. Otherwise, use the intermediate value theorem to decide whether the root lies in (a, c) or (c, b).
iii) Repeat step using the interval (a, c).
iv) The procedure is repeated while an length the last interval is less than the desired accuracy. The mid point of this last interval is taken as the root.

Example 3: Use bisection method to find a positive root of the equation

$$f(x) = 0.5\, e^x - 5x + 2$$

Solution: We find that f(0) = 2.5 and f(1) = −1.6408. Therefore, there is a root between 0 and 1. We apply the bisection method with a=0 and b= 1. The mid point is c = 0.5 and f(0.5) = 0.32436. The root now lies in (0.5, 1.0).

The tabulated values are shown in *Table 3*.

Table 3

a	b	midpoint (c)	f(a)	f(b)	f(c)
0	1	0.5	2.5	-1.6408591	0.32436064
0.5	1	0.75	0.32436064	-1.6408591	-0.6915
0.5	0.75	0.625	0.32436064	-0.6915	-0.190877
0.5	0.625	0.5625	0.32436064	-0.190877	0.06502733
0.5625	0.625	0.59375	0.06502733	-0.190877	-0.063367
0.5625	0.59375	0.578125	0.06502733	-0.063367	0.00072137
0.578125	0.59375	0.5859375	0.00072137	-0.063367	-0.0313502
0.578125	0.5859375	0.58203125	0.00072137	-0.0313502	-0.0153212
0.578125	0.58203125	0.58007813	0.00072137	-0.0153212	-0.0073016
0.578125	0.580078125	0.57910156	0.00072137	-0.0073016	-0.0032906
0.578125	0.579101563	0.57861328	0.00072137	-0.0032906	-0.0012847
0.578125	0.578613281	0.57836914	0.00072137	-0.0012847	-0.0002817
0.578125	0.578369141	0.57824707	0.00072137	-0.0002817	0.00021984
0.57824707	0.578369141	0.57830811	0.00021984	-0.0002817	-3.093E-05
0.57824707	0.578308105	0.57827759	0.00021984	-3.093E-05	9.4453E-05
0.578277588	0.578308105	0.57829285	9.4453E-05	-3.093E-05	3.1762E-05
0.578292847	0.578308105	0.57830048	3.1762E-05	-3.093E-05	4.1644E-07
0.578300476	0.578308105	0.57830429	4.1644E-07	-3.093E-05	-1.526E-05
0.578300476	0.578304291	0.57830238	4.1644E-07	-1.526E-05	-7.42E-06
0.578300476	0.578302383	0.57830143	4.1644E-07	-7.42E-06	-3.502E-06
0.578300476	0.57830143	0.57830095	4.1644E-07	-3.502E-06	-1.543E-06
0.578300476	0.578300953	0.57830071	4.1644E-07	-1.545E-06	-5.631E-07
0.578300476	0.578300714	0.5783006	4.1644E-07	-5.631E-07	-7.333E-08
0.578300476	0.578300595	0.57830054	4.1644E-07	-7.333E-08	1.7156E-07

After 24 iterations we see that the smaller root can be found in the interval [.578300476, .578300595]. Therefore, we can estimate one root to be 0 .5783005 One of the first things to be noticed about this method is that it takes a lot of iterations to

get a high degree of precision. In the following error analysis, we shall see method as to why the method is taking so many directions.

2.3.1 Error Analysis

The maximum error after the i^{th} iteration using this process is given by

$$\varepsilon_i = \frac{|b-a|}{2^i}$$

Taking logarithms on both sides and simplfying, we get

$$i \geq \frac{[\log(b-a) - \log \varepsilon_i]}{\log 2} \qquad (1)$$

As the interval at each iteration is halved, we have $(\varepsilon_{i+1} / \varepsilon_i) = (1/2)$. Thus, this method converges linearly .

Example 4 : Obtain the smallest positive root of $x^3 - 2x - 5 = 0$, correct upto 2 decimal places.

Solution : We have $f(x) = x^3 - 2x - 5$, $f(2) = -1$ and $f(3) = 16$. The smallest positive root lies in (2, 3). Therefore, $a = 2$, $b = 3$, $b - a = 1$, we need solution correct to two decimal places, that is,

$\varepsilon \leq 0.5(10^{-2})$, from (1), we get

$$i \geq \frac{\log 1 - \log[0.5(10^{-2})]}{\log 2} = \frac{-\log(0.005)}{\log 2} \approx 8.$$

This shows that 8 iterations are required to obtain the required accuracy. Bisection method gives the iteratived values as $x_1 = 2.5$, $x_2 = 2.25$, ..., $x_8 = 2.09$. Then $x \approx 2.09$ is the approximate root.

2.4 REGULA FALSI METHOD

Let the root lie in the interval (a, b). Then, P(a, f(a)), Q(b, f(b)) are points as the curve. Join the points P and Q. The point of intersection of this, with the X-axis, c, line is taken as the next approximation to the root. We determine by the intermediate value theorem, whether the root now lies in (a, c) or (c, b) we repeat the procedure. If x_0, x_1, x_2,..... are the sequence of approximations, then we stop the iteration when $|x_{k+1} - x_k| <$ given error tolerance.

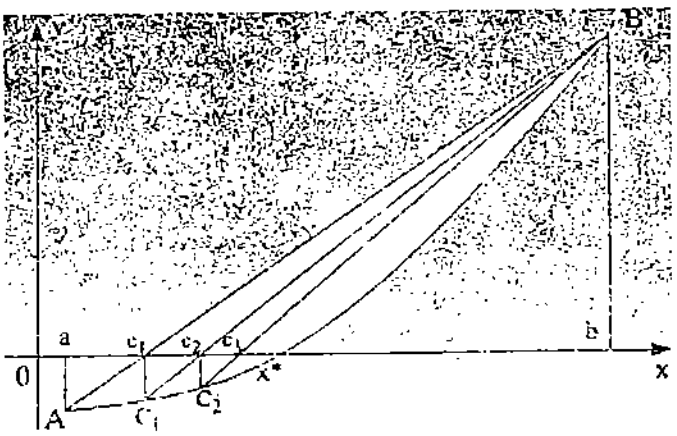


Figure 2.0: Regula falsi method

The equation of line chord joining (a, f(a)), (b, f(b)) is

$$y - f(a) = \frac{f(b) - f(a)}{b - a}(x - a).$$

Setting y = 0, we set the point of intersection with X-axis as givens

$$x = c = a - \frac{b - a}{f(b) - f(a)} f(a)$$
$$= \frac{af(b) - bf(a)}{f(b) - f(a)}$$

If we denote $x_0 = a$, $x_1 = b$, then the iteration formula can be written as

$$x_{n+1} = \frac{x_{n+1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}, \quad n = 1,2,..... \tag{2}$$

The rate of convergence is still linear but faster than that of the bisection method. Both these methods will fail if f has a double root.

Example 5: Obtain the positive root of the equation $x^2 - 1 = 0$ by Regua Falsi method.

Solution: Let $f(x)=x^2 - 1$. Since $f(0) = -1$, $f(2) = 3$, Let us take that the root lies in (0, 2). We have $x_0 = 0$, $x_1 = 2$.

Then, using (2), we get

$$x_2 = \frac{x_0 f(2) - x_1 f(0)}{f(2) - f(0)} = \frac{0 - 2(-1)}{3 + 1} = 0.5, \quad f(0.5) = -0.75$$

The root lies in (0.5, 2.0), we get

$$x_3 = \frac{0.5f(2) - 2.0f(0.5)}{f(2) - f(0.5)} = \frac{0.5(3) - 2.0(-0.75}{3 + 0.75} = 0.8$$

f(0.8) = -0.36. The root lies in (0.8, 2). The next approximation

$$x_4 = \frac{0.8(3) - 2.0(-0.36)}{3 + 0.36} = 0.9286, \quad f(0.9286) = -0.1377.$$

We obtain the next approximations as $x_5 = 0.9756$, $x_6 = 0.9918$, $x_7 = 0.9973$, $x_8 = 0.9990$. Since, $|x_8 - x_7| = 0.0017 < 0.005$, the approximation $x_8 = 0.9990$ is correct to decimal places.

Note that in this problem, the lower end of the interval tends to the root, and the minimum error tends to zero, but the upper limit and maximum error remain fixed. In other problems, the opposite may happen. This is the property to the regula falsi method.

2.5 NEWTON'S METHOD

This method is also called Newton-Raphsan method. We assume that f is a differentiable function in some interval $[a, b]$ containing the root.

We first look at a "pictorial" view of how Newton's method works. The graph of $y = f(x)$ is plotted in *Figure 3.1*. The point of intersection x = r, is the required root.

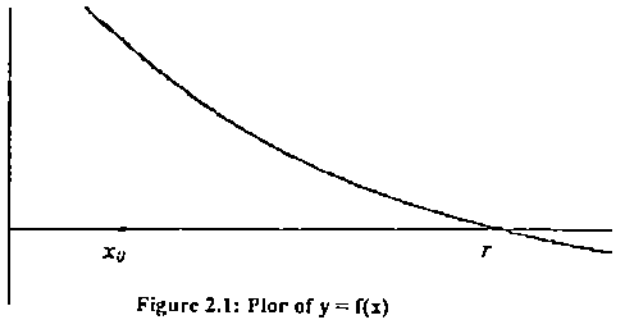


Figure 2.1: Plor of y = f(x)

Let x_0 be an initial approximation of r. Then, (x0, f(x0)) is a point as the curve (*Figure 3.2*).

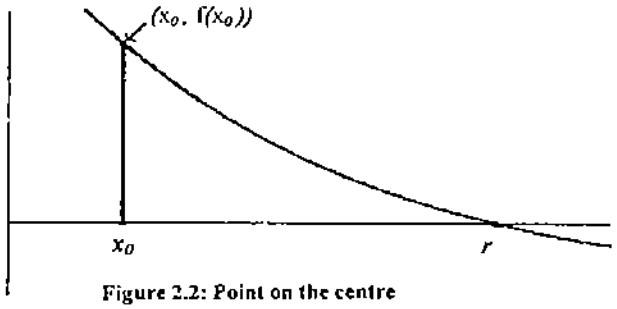


Figure 2.2: Point on the centre

Draw the tangent line to the curve at the point $(x_0 , f(x_0))$. This line intersects the x-axis at a new point, say x_1 (*Figure 3.3*).

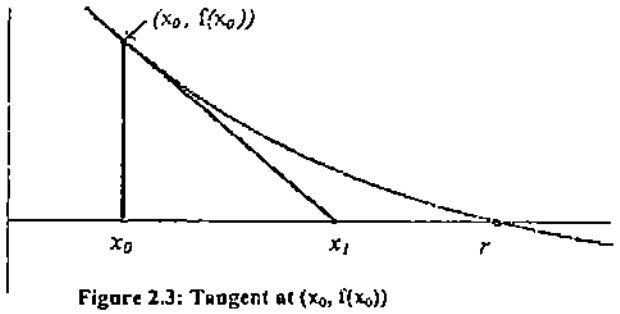


Figure 2.3: Tangent at $(x_0, f(x_0))$

Now, x_1 is a better approximation to r, than x_0 . We new repeat this process, yielding new points $x_2 , x_3 ,$ ... until we are "close enough" to r. *Figure 3.4* shows one more iteration of this process, determining x_2 .

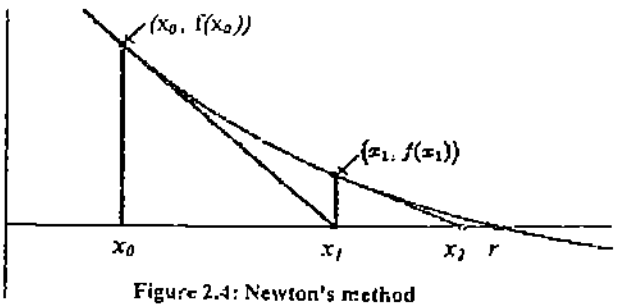


Figure 2.4: Newton's method

Now, we derive this method algebraically. The equation of the tangent at $(x_0, f(x_0))$ is given by

$y - f(x_0) = f'(x_0)(x - x_0)$

Where $f'(x_0)$ is the slope of the curve at $(x_0, f(x_0))$. Setting $y = 0$, we get the point of intersection of the tangent with x-axis as

$$y - f(x_0) = f'(x_0)(x - x_0), \text{ or } x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

But, this is our next approximation, that is

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Iterating this process, we get the Newton-Rephson as

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad \text{for } n = 0, 1, 2,..... \tag{3}$$

Example 6: Find the smallest positive root of $x^7 + 9x^5 - 13x - 17 = 0$.

Solution : Let $f(x) = x^7 + 9x^5 - 13x - 17$, we have $f(0) < 0$, $f(1) < 0$ and $f(1)f(2) < 0$. Hence, the smallest positive root lies in (1, 2). We can take any value in (1, 2) or one of the end points as the initial approximation. Let $x0 = 1$, we have,
$f'(x) = 7x^6 + 45x^4 - 13$. The Newton-Raphson method becomes

$$x_{n+1} = x_n - \frac{x_n^7 + 9x_n^5 - 13x_n - 17}{7x_n^6 + 45x_n^4 - 13}, \qquad n = 0,1,2.....$$

Starting with $x_0 = 1$, we obtain the values given in Table 4.

Table 4

n	x_n	$f(x_n)$	$f'(x_n)$	x_{n+1}
0	1	–20	39	1.512820513
1	1.512820513	52.78287188	306.6130739	1.340672368
2	1.340672368	12.33751268	173.0270062	1.269368397
3	1.269368397	1.46911353	133.1159618	1.258332053
4	1.258332053	0.03053547	127.6107243	1.258092767
5	1.258092767	0.00001407	127.4932403	1.258092657

After 6 iterations of Newton's method, we have

$$|x_6 - x_5| = |1.258092657 - 1.258092767| = 0.000000110.$$

Therefore, the root correct to 6 decimal places is $r = 1.258092657$.

Possible drawbacks:

Newton's method may not work in the following cases:

i) The x-values may run away as in *Figure 2.5(a)*. This might occur when the x-axis is an asymptote.

ii) We might choose an x-value that when evaluated, the derivative gives us 0 as in *Figure 2.5(b)*. The problem here is that we want the tangent line to intersect the x-axis so that we can approximate the root. If x has a horizontal tangent line, then we can't do this.

iii) We might choose an x_0 that is the beginning of a cycle as in *Figure 2.5(c)*. Again it is hoped that the picture will clarify this.

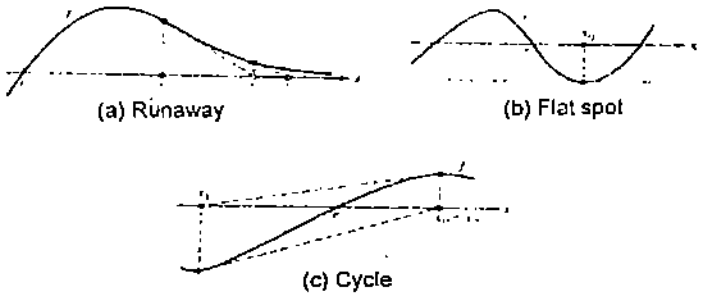


Figure 2.5: Divergence of Newton's method

However, the dificulties posed have one artifical. Normally, we do not encounter such problems in practice. Newton-Raphson methods is one of the powerful methods available for obtaning a simple root of f(x) = 0.

2.5.1 Error Analysis

Let the error at the n^{th} step be defined as

$$e_n = x_n - x$$

Then the error at the next step is

$$x + e_{n+1} = x + e_n - \frac{f(x+e_n)}{f'(x+e_n)}.$$

Explaing in Taylor Series, we obtain

$$e_{n+1} = e_n - \frac{f(x) + e_n f'(x) + \frac{1}{2}e_n^2 f''(x) + ...}{f'(x) + e_n f''(x) + ...} \qquad (4)$$

Since, x is a root, we have f(x) = 0. Then,

$$e_{n+1} = e_n - \frac{e_n f'(x)[1 + \frac{1}{2}e_n \frac{f''(x)}{f'(x)} + ...]}{f'(x)[1 + e_n \frac{f''(x)}{f'(x)} + ...]}$$

$$= e_n - e_n[1 + \frac{1}{2}e_n \frac{f''(x)}{f'(x)} + ...][1 + e_n \frac{f''(x)}{f'(x)} + ...]^{-1}$$

$$= e_n - e_n[1 + \frac{1}{2}e_n \frac{f''(x)}{f'(x)} + ...][1 - e_n \frac{f''(x)}{f'(x)} + ...]$$

$$= e_n - e_n[1 - \frac{1}{2}e_n \frac{f''(x)}{f'(x)} + ...]$$

$$= \frac{1}{2}e_n \frac{f''(x)}{f'(x)} e_n^2 + ...] \qquad (5)$$

We can neglect the centric and higher powers of e_n, as they are much smallest than e_n^2, (e_n is itself a small number).

Notice that the error is squared at each step. This means that the number of correct decimal places *doubles* with each step, much faster than linear convergence. We call it quadratic convergence.

This sequence will converge if

$$\left|\frac{f''(x)}{f'(x)}e_n^2\right| < |e_n|, \qquad |e_n| < 2\left|\frac{f''(x)}{f'(x)}\right| \tag{6}$$

If f' is not zero at the root (simple root), then there will always be a range round the root where this method converges.

If f' is zero at the root (double root), then the convergence becomes linear.

Example 7: Compute the square root of a, using Newton's method. How does the error behave?

Solution: Let $x = \sqrt{a}$, or $x^2 = a$. Define $f(x) = x^2 - a$. Here, we know the root exactly, so that we can see how well the method converges.

We have the Newton's method for finding a root of f(x) = 0 as

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - a}{2x_n} = \frac{x_n^2 + a}{2x_n} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) \tag{7}$$

Starting with any suitable initial approximation to $\sqrt{a}$, we find $x_1, x_2, \ldots$, which convege to the required value.

Error at the n^{th} step is $e_n = x_n - \sqrt{a}$. Substituting, we get

$$\begin{aligned} e_{n+1} &= \frac{(e_n + \sqrt{a})^2}{2(\sqrt{a} + e_n)} - \sqrt{a} \\ &= \frac{2a + 2e_n\sqrt{a} + e_n^2}{2(\sqrt{a} + e_n)} - \sqrt{a} \\ &= \frac{2(\sqrt{a} + e_n)\sqrt{a} + e_n^2}{2(\sqrt{a} + e_n)} - \sqrt{a} \\ &= \frac{e_n^2}{2(\sqrt{a} + e_n)} \end{aligned} \tag{8}$$

If $a = 0$, this simplifies to $e_n/2$, as expected. Here, we are finding the root of $x^2 = 0$, which gives a double root x = 90.

Since $a>0$, e_{n+1} will be positive, provided e_n is greater than $-\sqrt{a}$, i.e provided x_n is positive. Thus, starting from any positive number, all the errors, except perhaps the first, will be positive.

The method converges when,

$$|e_{n+1}| = \left|\frac{e_n^2}{2(\sqrt{a} + e_n)}\right| < |e_n| \tag{9}$$

or $\quad e_n < 2(e_n + \sqrt{a})$

which is always true. Thus, the method converges to the square root, *starting from any positive number*, and it does so quadratically.

We now discuss another method, which does not require the knowledge of the derivative of a function.

2.6 SECANT METHOD

Let x_0, x_1 be two initial approximations to the root. We do not require that the root lie in (x_0, x_1) as in Regula Falsi method. Therefore, the approximations x_0, x_1 may lie on the same side of the root. Further, we obtain the sequence of approximations as $x_2, x_3, \ldots\ldots$ At any stage, we do not require or check that the not lies in the interval (x_k, x_{k-1}). The derivation of the method is same as in the Regula Falsi method. (*Figure 2.6*)

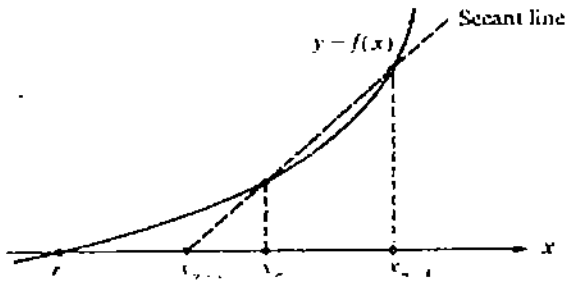


Figure 2.6: Secant method

The method is given by (see Equation (21))

$$x_{n+1} = \frac{x_{n-1} f_n - x_n f_{n-1}}{f_n - f_{n-1}} \qquad (10)$$

We compute x_2 using x_0, x_1; x_3 using x_1, x_2; and so on.

The rate of convergence of the method is super linear (1.618), that is, it works better than the Regula Falsi method.

Example 8: Use secant method to find the roots of the equation $f(x) = 0.5e^x - 5x + 2$.

Solution: We have $f(-x) = 0.5e^{-x} + 5x + 2 > 0$ for all x. Hence, there is no negative root.

We obtain,
$f(0) = 2.5$, $f(1) = -1.6408$, $f(2) = -4.3055$, $f(3) = -2.9572$,
$f(4) = 902990$, $f(x) > 0$ for $x > 4$.

Therefore, the given function has two roots, one root in (0, 1) and the second root in (3, 4).

For finding the first not, we take $x_0 = 0$, $x_1 = 1$, and compute the sequence of approximations $x_2, x_3, \ldots$

For finding the second root, we take $x_0 = 3$, $x_1 = 4$ and compute the sequence of approximations $x_2, x_3, \ldots\ldots$

The results are given in *Table 5*.

Table 5

x_{n-1}	x_n	x_{n-1}	x_n	x_{n+1}
0	1	2.5	–1.640859086	0.60373945
1	0.603739453	–1.6408591	–0.104224624	0.57686246
0.603739453	0.576862465	–0.1042246	0.002909403	0.57830459
0.576862465	0.578304589	0.0029094	–1.68E-06	0.57830058
0.578304589	0.578300578	–1.65E-05	–2.57E-09	0.57830058
0.578300578	0.578300577	–2.57E-09	1.11E-15	0.57830058
0.578300577	0.578300577	1.11E-15	0	0.57830058
3	4	–2.9572315	9.299075017	3.24128244
4	3.241282439	9.29907502	–1.423168098	3.34198736
3.241282439	3.341987358	–1.4231681	–0.572304798	3.40972316
3.341987358	3.409723161	–0.5723048	0.079817605	3.40173252
3.409723161	3.401432525	0.07981761	–0.003635061	3.40179365
3.401432525	3.401793651	–0.0036351	–2.15E-09	3.4017958
3.401793651	3.401795804	–2.15E-05	5.87E-09	3.4017958
3.401795804	3.401795804	5.87E-09	–7.11E-15	3.4017958

The two roots are 0.57830058, 3.4017958 correct to all decimal places given.

2.7 METHOD OF SUCCESSIVE ITERATION

The first step in this method is to write the equation $f(x) = 0$ in the form

$$x = g(x). \tag{11}$$

For example, consider the equation $x^2 - 4x + 2 = 0$. We can write it as

$$x = \sqrt{4x - 2}, \tag{12}$$

$$\text{or as } x = (x^2 + 2)/4, \tag{13}$$

$$\text{or as } x = \frac{2}{4 - x} \tag{14}$$

Thus, we can choose from (11) in several ways. Since, $f(x) = 0$ is the same as $x = g(x)$, finding a root of $f(x) = 0$ is the same as finding a root of $x = g(x)$, i.e. finding a fixed point α of $g(x)$ such that $\alpha = g(\alpha)$. The function $g(x)$ is called an ***iteration function*** for solving $f(x) = 0$.

If an initial approximation x_0 to a root α is provided, a sequence $x_1, x_2, \ldots$ may be defined by the iteration scheme

$$x_{n+1} = g(x_n) \tag{15}$$

with the hope that the sequence will converge to α. The successive iterations for solving $x = e^{-x}/3$, by the method $x_{n+1} = e^{-x_n}/3$ is given in *Figure 2.7*.

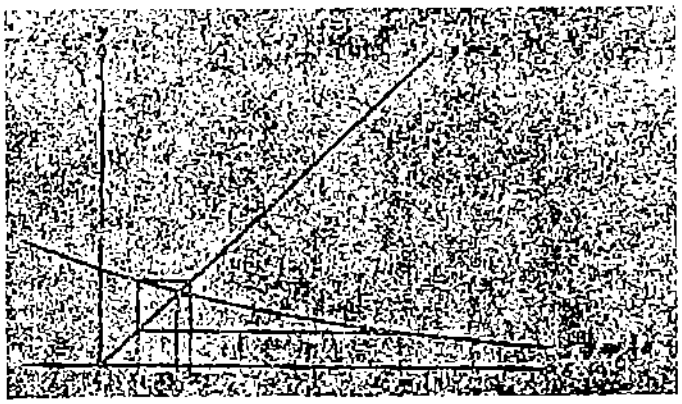


Figure 2.7: Successive iteration method

The method converges if , for some constant M such that 0< M <1, the inequality

$$| g(x) - g(\alpha) | \leq M | x - \alpha | \tag{16}$$

holds true whenever $|x - \alpha| \leq |x_0 - \alpha|$. For, if (16) holds, we find that

$$| x_{n+1} - \alpha | = | g(x_n) - \alpha | = | g(x_n) - g(\alpha) | \leq M | x_n - \alpha | \tag{17}$$

Using this relation recursively, we get

$$| x_{n+1} - \alpha | \leq M | x_n - \alpha | \leq M^2 | x_{n-1} - \alpha | \leq M^n | x_0 - \alpha | \tag{18}$$

Since, $0 < M < 1$, $\lim M^n = 0$ and thus $\lim x_n = \alpha$.

Condition (16) is satisfied if the function g(x) possesses a derivative $g'(x)$ such that $| g'(x) | < 1$ for $| x - \alpha | < | x_0 - \alpha |$ If x_0 is close to α, then we have

$$| x_{n+1} - \alpha | = | g(x_n) - g(\alpha) | \leq g'(\xi) | x_n - \alpha | \tag{19}$$

for some ξ between x_0 and α. Therefore, condition for convergence is

$$| g'(\xi) | < 1, \text{ or } | g'(x) | < 1. \tag{20}$$

Example 9 : Let us consider the equation $f(x) = x^3 + x - 2$. It has only one real root at $x = 1$. There are several ways in which $f(x)=0$ can be written in the desired form, $x = g(x)$.

For example, we may write $x = x + f(x) = g(x)$ and write the method as

$$x_{n+1} = x_n + f(x_n) = x_n^3 + 2x_n - 2$$

In this case, $g'(x) = 3x^2 + 2$, and the convergence condition is

$$|g'(x)| = 3x^2 + 2 < 1, \qquad or \qquad 3x^2 < -1.$$

Since, this is never true, this arrangement doesn't converge to the root.

An alternate rearrangement is

$$x_{n+1} = 2 - x_n^3$$

This method converges when

$$|g'(x)| = |-3x^2| < 1, \quad or \quad x^2 < \frac{1}{3}, \quad or \quad |x| < \frac{1}{\sqrt{3}}.$$

Since this range $[-1/\sqrt{3}, 1/\sqrt{3}]$ does not include the root x = 1, this method will not converge either.

Another rearrangement is

$$x_{n+1} = \sqrt[3]{2 - x_n}$$

In this case, the convergence condition becomes

$$\frac{1}{3}|(2-x_n)|^{-2/3} < 1, \quad or \quad (2-x_n)^{-2} < 3^3, \quad or \quad |x_n - 2| > \sqrt{27}.$$

Again, this range does not conain the root.

Another rearrangement is

$$x_{n+1} = \frac{2}{x_n^{\,2} + 1} \tag{21}$$

In this case, the convergence condition becomes

$$\frac{4|x|}{(1+x^2)^2} < 1, \qquad 4|x| < (1+x^2)^2$$

This inequality is satisfied when x > 1. Hence, if we start with such an x, the method will converge to the root.

Let $x_0 = 1.2$, Then, from (21), we obtain the sequence of approximations as
$x_1 = 0.8197$, $x_2 = 1.1963$, $x_3 = 0.8227$, $x_4 = 1.1927$, $x_5 = 0.8255$, $x_6 = 1.1894$, ...

The approximations oscillate about x = 1 and converge very slowly.

2.8 SUMMARY

In this unit, we have discussed the acutal meaning of root. In general, root determination of an equation is a tedious exercise. So, to handle such tasks, we have discussed some basic, simple, but still powerful methods of root determination. The methods discussed in this unit are Bisection method, Regular falsi method, Newton's method, Secan method and Successive iteration method.

2.9 EXERCISES

E1) In the following problems, find the intervals of length 1 unit, in which the roots lie
(a) $12x^3 - 76x^2 + 131x - 42 = 0$; (b) $4x^2 + 8x - 21 = 0$
(c) $x - e^{-x} = 0$ (d) $x = 2 \cos x$

E2) Find all the roots in Problems 1(a), (b), (c) by ergular falth method, secant method and Newton-Raghson method.

E3) Find the smaller roots in Problems 1(b) and the root in 1(c), by successive iteration method.

E4) Show that the equation $x^3 - 6x - 1 = 0$, has a root in the internal (–1, 0). Obtain this root using the successive iteration method.

2.10 SOLUTIONS TO EXERCISES

E1) (a) (0, 1), (1, 2), (3, 4) (b) (– 4, –3), (1, 2)
(c) (0, 1) (d) (0.5, 1.5)

E2) (a) 0.5, 1.2, 3.5 (b) – 3.5, 1.5
(c) 0.567143

E3) (a) Use $x_{n+1} = x_n - 0.05(4x_n^2 + 8x_n - 21)$ with $x_0 = 1.4$
(b) Write $x_{n+1} = e_n^{-xn}$

E4) Write $x_{n+1} = (x_n^3 - 1) / 6$; $x_0 = -0.5, -0.167449$

UNIT 3 SOLUTION OF LINEAR ALGEBRAIC EQUATIONS

Structure **Page Nos.**

3.0 INTRODUCTION

Systems of linear equations arise in many areas of study, both directly in modelling physical situations and indirectly in the numerical solution of other mathematical models. Linear algebraic equations occur in the linear optimization theory, least square fitting of data, numerical solution of ordinary and partial differential equations, statistical interference etc. Therefore, finding the numerical solution of a system of linear equations is an important area of study.

From study of algebra, you must be familiar with the following two common methods of solving a system of linear equations :

1) By the elimination of the variables by elementary row operations.
2) By the use of determinants, a method better known as *Cramer's rule*.

When smaller number of equations are involved, Cramer's rule appears to be better than elimination method. However, Cramer's rule is completely impractical when a large number of equations are to be solved because here n+1 determinants are to be computed for n unknowns.

Numerical methods for solving linear algebraic systems can be divided into two methods, **direct** and **iterative.** Direct methods are those which, in the absense of round-off or other errors, yield exact solution in a finite number of arithmetic operations. Iterative methods, on the other hand, start with an initial guess and by applying a suitable procedure, give successively better approximations.

To understand, the numerical methods for solving linear systems of equations, it is necessary to have some knowledge of properties of the matrices. You might have studied matrices, determinants and their properties in your linear algebra course.

In this unit, we shall discuss two direct methods, namely, **Gauss elimination method** and **LU decomposition method**, and two iterative methods, *viz.*; **Jacobi method, Gauss – Seidel method** and **Successive over relaxation method**.These methods are frequently used to solve systems of linear equations.

3.1 OBJECTIVES

After studying this unit, you should be able to:

- state the difference between direct and iterative methods for solving a system of linear equations;
- learn how to solve a system of linear equations by Gauss elimination method;
- understand the effect of round off errors on the solution obtained by Gauss elimination method;
- learn how to modify Gauss elimination method to Gaussian elimination with partial pivoting to avoid pitfalls of the former method;
- learn LU decomposition method to solve a system of linear equations;
- learn how to find inverse of a square matrix numerically;
- learn how to obtain the solution of a sytem of linear equations by using an iterative method, and
- state whether an iterative method will converge or not.

3.2 GAUSS ELIMINATION METHOD

One of the most popular techniques for solving simultaneous linear equations is the Gaussian elimination method. Karl Friedrich Gauss, a great 19th century mathematician, suggested this elimination method as a part of his proof of a particular theorem. Computational scientists use this "proof" as a direct computational method. The approach is designed to solve a general set of n equations and n unknowns

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n &= b_2 \\ \cdot \quad\quad \cdot \quad\quad \cdot \quad\quad \cdot & \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n &= b_n \end{aligned} \qquad (1)$$

In matrix form, we write Ax = b, where

$$A = \begin{bmatrix} a_{11} & a_{12} & .. & a_{1n} \\ a_{21} & a_{22} & .. & a_{2n} \\ .. & .. & .. & .. \\ a_{n1} & a_{n2} & .. & a_{nn} \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \\ . \\ x_n \end{bmatrix}, b = \begin{bmatrix} b_1 \\ b_2 \\ . \\ b_n \end{bmatrix}$$

Gaussian elimination consists of two steps:

1) *Forward Elimination:* In this step, the elementary row operations are applied on the augmented matrix [A|b] to transform the coefficient matrix A into upper triangular form.

2) *Back Substitution:* In this step, starting from the last equation, each of the unknowns is found by back substitution.

Forward Elimination of Unknowns: In this first step the first unknown, x_1 is eliminated from all rows below the first row. The first equation is selected as the pivot equation to eliminate x_1. So, to eliminate x_1 in the second equation, one divides

the first equation by a_{11} (hence called the pivot element) and then multiply it by a_{21}. That is, same as multiplying the first equation by a_{21}/a_{11} to give

$$a_{21}x_1 + \frac{a_{21}}{a_{11}}a_{12}x_2 + \ldots + \frac{a_{21}}{a_{11}}a_{1n}x_n = \frac{a_{21}}{a_{11}}b_1$$

Now, this equation is subtracted from the second equation to give

$$\left(a_{22} - \frac{a_{21}}{a_{11}}a_{12}\right)x_2 + \ldots + \left(a_{2n} - \frac{a_{21}}{a_{11}}a_{1n}\right)x_n = b_2 - \frac{a_{21}}{a_{11}}b_1$$

or $\quad a_{22}x_2 + \ldots + a_{2n}x_n = b_2$

where $a_{22} = a_{22} - \frac{a_{21}}{a_{11}}a_{12}, \ldots,$

$a_{2n} = a_{2n} - \frac{a_{21}}{a_{11}}a_{1n}$, $b_2 = b_2 - \frac{a_{21}}{a_{11}}b_1$.

This procedure of eliminating x_1, is now repeated for the third equation to the n^{th} equation to reduce the set of equations as

$$\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 \\
&a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2 \\
&a_{32}x_2 + a_{33}x_3 + \ldots + a_{3n}x_n = b_3 \\
&\vdots \\
&a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n = b_n.
\end{aligned} \qquad (2)$$

This completes the first step of forward elimination. Now, for the second step of forward elimination, we start with the second equation as the pivot equation and a'_{22} as the pivot element. So, to eliminate x_2 in the third equation, one divides the second equation by a'_{22} (the pivot element) and then multiply it by a'_{32}. That is, same as multiplying the second equation by a'_{32}/a'_{22} and subtracting from the third equation. This makes the coefficient of x_2 zero in the third equation. The same procedure is now repeated for the fourth equation till the nth equation to give

$$\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 \\
&a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2 \\
&a_{33}x_3 + \ldots + a_{3n}x_n = b_3 \\
&\vdots \\
&a_{n3}x_3 + \ldots + a_{nn}x_n = b_n
\end{aligned} \qquad (3)$$

The next steps of forward elimination are done by using the third equation as a pivot equation and so on. That is, there will be a total of (n–1) steps of forward elimination. At the end of (n–1) steps of forward elimination, we get the set of equations

$$\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 \\
&a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2
\end{aligned}$$

$$a_{33}^{-}x_3 + \ldots + a_n^{-}x_n = b_3^{-} \qquad (4)$$

$$a_{nn}^{(n-1)}x_n = b_n^{(n-1)}$$

Back Substitution: Now, the equations are solved starting from the last equation as it has only one unknown. We obtain

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}$$

Now, we solve the (n−1)th equation to give

$$x_{n-1} = \frac{1}{a_{n-1,n-2}^{(n-2)}}\left[b_{n-1}^{(n-2)} - a_{n-1,n}^{(n-2)}x_n\right]$$

since x_n is determined.

We repeat the procedure until x_1 is determined. The solution is given by

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}$$

and

$$x_i = \frac{b_i^{(i-1)} - \sum_{j=i+1}^{n} a_{ij}^{(i-1)}x_j}{a_{ii}^{(i-1)}}, \text{ for } i = n-1, n-2, \ldots, 1 \qquad (5)$$

Example 1: Solve the following linear system of equations
$x_1 + x_2 + x_3 = 3$,
$4x_1 + 3x_2 + 4x_3 = 8$,
$9x_1 + 3x_2 + 4x_3 = 7$

using the Gauss elimination method.

Solution: In augmented form, we write the system as

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 4 & 3 & 4 & 8 \\ 9 & 3 & 4 & 7 \end{array}\right]$$

Subtracting 4 times the first row from the second row gives

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & -1 & 0 & -4 \\ 9 & 3 & 4 & 7 \end{array}\right]$$

Subtracting 9 times the first row from the third row, we get

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & -1 & 0 & -4 \\ 0 & -6 & -5 & -20 \end{array}\right]$$

Subtracting 6 times the second row from the third row gives

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & -1 & 0 & -4 \\ 0 & 0 & -5 & 4 \end{array}\right]$$

Restoring the transformed matrix equation gives

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & -5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ -4 \\ 4 \end{bmatrix},$$

Solving the last equation, we get $x_3 = \frac{-4}{5}$. Solving the second equation, we get $x_2 = 4$ and the first equation gives $x_1 = 3 - x_2 - x_3 = 3 - 4 + \frac{4}{5} = \frac{-1}{5}$.

Example 2: Use Gauss Elimination to solve

$10x_1 - 7x_2 = 7$

$-3x_1 + 2.099x_2 + 6x_3 = 3.901$

$5x_1 - x_2 + 5x_3 = 6$

correct to six places of significant digits.

Solution : In matrix form , we write

$$\begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.099 & 6 \\ 5 & -1 & 5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 3.901 \\ 6 \end{bmatrix}$$

Multiply the first row by 3/10 and add to the second equation, we get

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.001 & 6 \\ 5 & -1 & 5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 6.001 \\ 6 \end{bmatrix}$$

Multiply the first row by 5/10 and subtract from the third equation, we get

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.001 & 6 \\ 0 & 2.5 & 5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 6.001 \\ 2.5 \end{bmatrix}$$

This completes the first step of forward elimination.

Multiply the second equation by 2.5/(–0.005)= –2500 and subtract from the third equation, we obtain

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.001 & 6 \\ 0 & 0 & 15005 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 6.001 \\ 15005 \end{bmatrix}$$

We can now solve the above equations by back substitution. From the third equation, we get

$15005x_3 = 15005$, or $x_3 = 1$.

Substituting the value of x_3 in the second equation, we get

$-0.001x_2 + 6x_3 = 6.001$, or $-0.001x_2 = 6.001 - 6 = 0.001$, or $x_2 = -1$

Substituting the values of x_3 and x_2 in the first equation, we get

$10x_1 - 7x_2 = 7$, or $10x_1 = 7 + 7x_2 = 0$, or $x_1 = 0$.

Hence, the solution is $[0 \quad -1 \quad 1]^T$.

3.3 PITFALLS OF GAUSS ELIMINATION METHOD

There are two pitfalls in the Gauss elimination method.

Division by zero: It is possible that division by zero may occur during forward elimination steps. For example, for the set of equations

$10x_2 - 7x_3 = 7$
$6x_1 + 2.099x_2 - 3x_3 = 3.901$
$5x_1 - x_2 + 5x_3 = 6$

during the first forward elimination step, the coefficient of x_1 is zero and hence normalisation would require division by zero.

Round-off error: Gauss elimination method is prone to round-off errors. This is true, when there are large numbers of equations as errors propagate. Also, if there is subtraction of almost equal numbers, it may create large errors. We illustrate through the following examples.

Example 3: Solve the following linear equations

$$10^{-5}x + y = 1.0,$$
$$x + y = 2.0 \qquad (6)$$

correct to 4 places of accuracy.

Solution: For 4 places of accuracy the solution is , $x \approx y \approx 1.0$.
Applying the Gauss elimination method, we get (by dividing with the pivotal element)

$$x + 10^5 y = 10^5$$
$$(1 - 10^5)y = 2.0 - 10^5.$$

Now, $10^5 - 1$ when rounded to four places of accuracy, become is 10^5. Similarly, $10^5 - 2$ when rounded to four places of accuracy becomes 10^5.

Hence, from the second equation we get, $10^5 y = 10^5$, or $y = 1.0$.

Substituting in the first equation, we get x = 0.0, which is not the solution.
Such errors can also arise when we perform computations with less number of digits.
To avoid these computational disasters, we apply partial pivoting to gauss elimination.

3.4 GAUSS ELIMINATION METHOD WITH PARTIAL PIVOTING

We perform the following modification to the Gauss elimination method. At the beginning of each step of forward elimination, a row interchange is done, if necessary, based on the following criterion. If there are n equations, then there are $(n-1)$ forward elimination steps. At the beginning of the k^{th} step of forward elimination, we find the maximum of

$$|a_{kk}|, |a_{k+1,k}|, \ldots\ldots\ldots\ldots, |a_{nk}|$$

That is, maximum in magnitude of this elements on or below the diagonal element.

Then, if the maximum of these values is $|a_{pk}|$ in the p^{th} row, $k \le p \le n$, then interchange rows p and k.. The other steps of forward elimination are the same as in Gauss elimination method. The back substitution steps remain exactly the same as in Gauss elimination method.

Example 4: Consider Example 3. We now apply partial pivoting on system (6).

Solution: We obtain the new system as

Since, $a_{11} < a_{21}$, we interchange the first and second rows (equations).

$$x + y = 2.0$$
$$10^{-5}x + y = 1.0$$

On elimination , we get second equation as y = 1.0 correct to 4 places. Substituting in the first equation, we get x = 1.0, which is the correct solution.

3.5 LU DECOMPOSITION METHOD

The Gauss elimination method has the disadvantage that the right-hand sides are modified (repeatedly) during the steps of elimination). The LU decomposition method has the property that the matrix modification (or decomposition) step can be performed independent of the right hand side vector. This feature is quite useful in practice. Therefore, the LU decomposition method is usually chosen for computations.

In this method, the coefficient matrix into a product of two matrices is written as

$$\mathbf{A = L\,U} \qquad (7)$$

where L is a lower triangular matrix and U is an upper triangular matrix.

Now, the original system of equations, $\mathbf{A\,x = b}$ becomes

$$\mathbf{LU\,x = b} \qquad (8)$$

Now, set $\mathbf{U\,x = y}$, then, (8) becomes

$$\mathbf{Ly = b} \qquad (9)$$

The rationale behind this approach is that the two systems given in (9) are both easy to solve. Since, L is a lower triangular matrix, the equations, $\mathbf{Ly = b}$, can be solved for y using the forward substitution step. Since U is an upper triangular matrix, $\mathbf{U\,x = y}$ can be solved for x using the back substitution algorithm.

We define writing A as LU as the Decomposition Step. We discuss the following three approaches of Decomposition using 4×4 matrices.

Doolittle Decomposition

We choose $l_{ii} = 1$, i=1, 2, 3, and write

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 \\ l_{41} & l_{42} & l_{43} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \tag{10}$$

Because of the specific structure of the matrices, we can derive a systematic set of formulae for the components of **L** and **U** .

Crout Decomposition:

We choose $u_{ii} = 1$, i = 1, 2, 3, 4 and write

$$\begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} & u_{14} \\ 0 & 1 & u_{23} & u_{24} \\ 0 & 0 & 1 & u_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \tag{11}$$

The evaluation of the components of **L** and **U** is done in a similar fashion as above.

Cholesky Factorization:

If A is a symmetric and positive definite matrix, then we can write the decomposition as

Where L is the lower triangular matrix

$$L = \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \tag{12}$$

We now describe the rationale behind the choice of $l_{ii} = 1$ in (10) or $u_{ii} = 1$ in (11). Consider the decomposition of a 3×3 matrix as follows.

$$\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$\begin{bmatrix} l_{11}u_{11} & l_{11}u_{12} & l_{11}u_{13} \\ l_{21}u_{11} & l_{21}u_{12} + l_{22}u_{22} & l_{21}u_{13} + l_{22}u_{23} \\ l_{31} & l_{31}u_{12} + l_{32}u_{22} & l_{31}u_{13} + l_{32}u_{23} + l_{33}u_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{13}$$

We note that L has 1+2+3=6 unknowns and U has 3+2+1=6 unknowns, that is, a total of 12 unknowns. Comparing the elements of the matrices on the left and right hand sides of (13), we get 9 equations to determine 12 unknowns . Hence, have 3 arbitrary parameters, the choice of which can be done in advance. Therefore, to make computations easy we choose $l_{ii} = 1$ in Doolittle method and $u_{ii} = 1$ in Crout's method.

In the general case of decomposition of an N ×N matrix, L has 1+2+3+....+N= $[N(N+1)/2]$ And U also has N(N+1)/2 unknowns, that is a total of $N^2 + N$ unknowns comparing the elements of A and the product LU, we obtain N^2 equations. Hence, we have N arbitrary parameters. Therefore, we choose either $l_{ii} = 1$ or $u_{ii} = 1$,i = 1,2,....n,

Now , let us give the solution for the Doolittle and Crout decomposition.

Doolittle Method: Here $l_{ii} = 1$, i = 1 to N. In this case, generalisation of (13) gives

$u_{1j} = a_{1j}$, j = 1 to N
$l_{i1} = a_{i1} / a_{11}$, i = 2 to N
$u_{2j} = a_{2j} - l_{21} . u_{1j}$, j=2 to N
$l_{i2} = (a_{i2} - l_{i1} u_{12}) / u_{22}$, i = 3 to N, and so on

Crout's Method: Here $u_{ii} = 1$, i = 1 to N . In this case , we get

$l_{i1} = a_{i1}$, i=1 to N
$u_{1j} = a_{1j} / a_{11}$, j= 2 to N
$l_{i2} = a_{i2} - l_{i1} u_{12}$, i = 2 to N
$u_{2j} = (a_{2j} - l_{21} u_{1j}) / l_{22}$, j= 3 to N, and so on

Example 5: Given the following system of linear equations, determine the value of each of the variables using the LU decomposition method.

$$6x_1 - 2x_2 = 14$$
$$9x_1 - x_2 + x_3 = 21$$
$$3x_1 - 7x_2 + 5x_3 = 9$$

Solution: We write A = LU, with $u_{ii} = 1$ as

$$\begin{bmatrix} 6 & -2 & 0 \\ 9 & -1 & 1 \\ 3 & +7 & 5 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} l_{11} & l_{11}u_{12} & l_{11}u_{13} \\ l_{21} & l_{21}u_{12} + l_{22} & l_{21}u_{13} + l_{22}u_{23} \\ l_{31} & l_{31}u_{12} + l_{32} & l_{31}u_{13} + l_{32}u_{23} + l_{33} \end{bmatrix}$$

We obtain $l_{11} = 6$, $l_{21} = 9$, $l_{31} = 3$, $l_{11}u_{12} = -2$, $u_{12} = -1/3$;

$l_{11} u_{13} = 0$, $u_{13} = 0$

$l_{21}u_{12} + l_{22} = -2$, $l_{22} = -1 + 3 = 2$; $l_{21}u_{13} + l_{22} u_{23} = 1$

$u_{23} = 1/2$, $l_{31}u_{12} + l_{32} = +7$, $l_{32} = +7 + 1 = 8$;

$l_{31}u_{13} + l_{32}u_{23} + l_{33} = 5$, $l_{33} = 5 - 4 = +1$.

Hence, $L = \begin{bmatrix} 6 & 0 & 0 \\ 9 & 2 & 0 \\ 3 & 8 & +1 \end{bmatrix}, U = \begin{bmatrix} 1 & -1/3 & 0 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}$

Solving $L_y = b$, $\begin{bmatrix} 6 & 0 & 0 \\ 9 & 2 & 0 \\ 3 & 8 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 14 \\ 21 \\ 9 \end{bmatrix}$

We get, $y_1 = \frac{14}{6} = \frac{7}{3}$; $9\, y_1 + 2\, y_2 = 21$, $y_2 = \frac{1}{2}(21-21) = 0$

$3\, y_1 + 8\, y_2 + y_3 = 9$, $y_3 = 9 - 7 = 2$.

Solving $U_x = y$, $\begin{bmatrix} 1 & -1/3 & 0 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7/3 \\ 0 \\ 2 \end{bmatrix}$

We get, $x_3 = 2$, $x_2 + \frac{1}{2}\, x_3 = 0$, $x_2 = -1$; $x_1 - \frac{1}{3}\, x_2 = \frac{7}{3}$, $x_1 = \frac{7}{3} - \frac{1}{3} = 2$.

The solution vector is [2 −1 2].

Solving the system of Equations

After decomposing A as A = LU, the next step is to computate the solution. We have,

LUX = b, set UX = y

Solve first Ly = b, by forward substitution. Then, solve Ux = y, by backward substitution to get the solution vector x.

3.6 ITERATIVE METHODS

Iterate means repeat. Hence, an iterative method repeats its process over and over, each time using the current approximation to produce a better approximation for the true solution, until the current approximation is sufficiently close to the true solution -- or until you realize that the sequence of approximations resulting from these iterations is not converging to the true solution.

Given an initial guess or approximation $x^{(0)}$ for the true solution x, we use $x^{(0)}$ to find a new approximation $x^{(1)}$, then we use $x^{(1)}$ to find the better approximation $x^{(2)}$, and so on. We expect that $x^{(k)} \rightarrow x$ as $k \rightarrow \infty$; that is, our approximations should become closer to the true solution as we take more iterations of this process.

Since, we do not actually have the true solution x, we cannot check to see how close our current approximation $x^{(k)}$ is to x. One common way to check the closeness of $x^{(k)}$ to x is instead by checking how close $Ax^{(k)}$ is to Ax, that is, how close $Ax^{(k)}$ is to b.

Another way to check the accuracy of our current approximation is by looking at the magnitude of the difference in successive approximations, $|x^{(k)} - x^{(k-1)}|$. We expect $x^{(k)}$ to be close to x if $|x^{(k)} - x^{(k-1)}|$ is small .

The Jacobi Method

This method is also called Gauss – Jacobi method. In Jacobi method, the first equation is used to solve for x_1, second equation is used to solve x_2 etc. That is,

$$x_1 = \frac{1}{a_{11}}\left[b_1 - (a_{12}x_2 + \ldots\ldots + a_{1n}x_n)\right]$$

$$x_2 = \frac{1}{a_{22}}\left[b_2 - (a_{21}x_1 + a_{23}x_3 \ldots\ldots + a_{2n}x_n)\right]$$

If in the ith equation

$$\sum_{j=1}^{n} a_{i,j}x_j = b_i \qquad (15)$$

we solve for the value of x_i, we obtain ,

$$x_i = (b_i - \sum_{j=1}^{n} a_{i,j}x_j)/a_{i,j} \qquad (16)$$

This suggests an iterative method defined by

$$x_i^{(k)} = (b_i - \sum_{j=1} a_{i,j}x_j^{(k-1)}/a_{i,j} \qquad (17)$$

is the Jacobi method. Note that the order in which the equations are solved is ·vant, since the Jacobi method treats them independently. For this reason, the ·obi method is also known as the *method of simultaneous displacements*, since the ·lates could in principle be done simultaneously.

·obi method can be written in matrix notation.

Let A be written as A = L + D + U, where L is strictly lower triangular part, D the diagonal part and U is strictly upper triangular part.

$$L = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ a_{21} & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{bmatrix}, U = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ 0 & 0 & \ddots & a_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, D = \begin{bmatrix} a_{11} & & & 0 \\ & a_{22} & & \\ & & & \\ 0 & & & a_{nn} \end{bmatrix}$$

Therefore, we have

$(L + D + U)\,X = b$, or $DX = -(L + U)\,x + b$

Since, $a_{ii} \neq 0$, D^{-1} exists and is equal to

$D^{-1} = \text{diag}\,(1/a_{11}, 1/a_{12}, \ldots\ldots 1/a_{nn})$.

Inverting D, we write the intersection as

$$X^{(k+1)} = -D^{-1}(L + U)\,X^{(k)} + D^{-1}\,b \qquad (18)$$

$$= M_J\,x^{(k)} + C \qquad (19)$$

where $M_J = -D^{-1}(L + U)$ and $C = D^{-1}\,b$.

The matrix M_J is called the iteration matrix. Covergence of the method depends on the properties of the matrix M_J.

Diagonally dominant: A matrix A is said to be diagonally dominant if

$$| a_{ii} | \geq \sum_{j=1, i \neq j}^{n} | a_{ij} | \qquad (20)$$

with inequality satisfied for atleast one row.

Convergence: (i) The Jacobi method converges when the matrix **A** is diagonally dominant. However, this is a sufficient condition and necessary condition.

(ii) The Jacobi method converges if Spectral radius (M_J) < 1. Where Spectiral radius of a matrix = max $=| \lambda |$ and x_i are eigenralues of M_J.This is a necessary and sufficient condition. If no intial approximation is known, we may assume $X^{(0)} = 0$.

Exercise 1: Are the following matrices diagonally dominant?

$$\mathbf{A} = \begin{bmatrix} 2 & -5.81 & 34 \\ 45 & 43 & 1 \\ 123 & 16 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 124 & 34 & 56 \\ 23 & 53 & 5 \\ 96 & 34 & 129 \end{bmatrix}$$

Solution: In A, all the three rows violate the condition (20). Hence, A is not diagonally dominant.

In B, in the row $| -129 | = 129 < 96 + 34 = 130$. Therefore, B is also not diagonally dominant.

Example 2: Solve the following system, of equations
$x + y - z = 0$, $-x + 3y = 2$, $x - 2z = -3$ by Jacobi Method, both directly and in matrix form. Assume the initial solution vector as $[0.8 \quad 0.8 \quad 2.1]^T$.

Solution : We write the Jacobi method as

$$x^{(k+1)} = -y^{(k)} + z^{(k)}, \; y^{(k+1)} = \frac{1}{3}(2 + x^{(k)}), \; z^{(k+1)} = \frac{1}{2}(3 + x^{(k)})$$

with $x^{(0)} = 0.8$, $y^{(0)} = 0.8$, $z^{(0)} = 2.1$, we get the following approximations.

$x^{(1)} = 1.3$, $y^{(1)} = 0.9333$, $z^{(1)} = 1.9$;
$x^{(2)} = 0.9667$, $y^{(2)} = 1.1$, $z^{(2)} = 2.5$;
$x^{(3)} = 1.0500$, $y^{(3)} = 0.9889$, $z^{(3)} = 1.98335$;
$x^{(4)} = 0.99445$, $y^{(4)} = 1.01667$, $z^{(4)} = 2.025$;
$x^{(5)} = 1.00833$, $y^{(5)} = 0.99815$, $z^{(5)} = 1.997225$;
$x^{(6)} = 0.988895$, $y^{(6)} = 1.00278$, $z^{(6)} = 2.004165$,
$x^{(7)} = 1.001385$, $y^{(7)} = 0.99630$, $z^{(7)} = 1.99445$;
$x^{(8)} = 0.99815$, $y^{(8)} = 1.00046$, $z^{(8)} = 2.00069$;
$x^{(9)} = 1.00023$, $y^{(9)} = 0.99938$, $z^{(9)} = 1.999075$.

At this stage, we have,

$| x^{(9)} - x^{(8)} | = 0.002$, $| y^{(9)} - y^{(8)} | = 0.0019$, $| z^{(9)} - z^{(8)} | = 0.0016$.

Therefore, the 9th interaction is correct to two decimal places.

Let us represent the matrix A in the form

$$A = L + D + U = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -2 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

We have,

$$Mj = -D^{-1}(L+U) = -\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & -1/2 \end{bmatrix}\begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 & 1 \\ 1/3 & 0 & 0 \\ 1/2 & 0 & 0 \end{bmatrix}$$

$$c = D^{-1}b = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & -1/2 \end{bmatrix}\begin{bmatrix} 0 \\ 2/3 \\ 3/2 \end{bmatrix}$$

Therefore, Jacobi method gives,

$$X^{(k+1)} = \begin{bmatrix} 0 & -1 & 1 \\ 1/3 & 0 & 0 \\ 1/2 & 0 & 0 \end{bmatrix} X^{(k)} + \begin{bmatrix} 0 \\ 2/3 \\ 3/2 \end{bmatrix}$$

The intial approximation is given as $X^{(0)} = [0.8 \quad 0.8 \quad 2.1]^T$

Then, we have

$$X^{(1)} = \begin{bmatrix} 0 & -1 & 1 \\ 1/3 & 0 & 0 \\ 1/2 & 0 & 0 \end{bmatrix}\begin{bmatrix} 0.8 \\ 0.8 \\ 2.1 \end{bmatrix} + \begin{bmatrix} 0 \\ 2/3 \\ 3/2 \end{bmatrix} = \begin{bmatrix} 1.3 \\ 0.9333 \\ 1.9 \end{bmatrix}$$

which is same as $X^{(1)}$ obtained earlier.

Since, the two procedures (direct and in matrix form) are identical, we get the same approximations $x^{(2)}$, ... $x^{(9)}$. The exact solution is $x = [1 \quad 1 \quad 2]^T$.

Note that the coefficient matrix A is not diagonal dominant. But, we have obtained the solution correct to two decimal places in 9 interactions. This shows that the requirement of A being diagonal dominant is a sufficient condition.

Example 3: Solve by Jacobi's method the following system of linear equations.

$$2x_1 - x_2 + x_3 = -1$$
$$x_1 + 2x_2 - x_3 = 6$$
$$x_1 - x_2 + 2x_3 = -3.$$

Solution: This system can be written as

$$x_1 = 0.5\,x_2 - 0.5\,x_3 - 0.5$$
$$x_2 = -0.5\,x_1 + 0.5\,x_3 + 3.0$$
$$x_3 = -0.5\,x_1 + 0.5\,x_2 - 1.5$$

So the Jacobi iteration is

$$\begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{bmatrix} = \begin{bmatrix} 0.0 & 0.5 & -0.5 \\ -0.5 & 0.0 & 0.5 \\ -0.5 & 0.5 & 0.0 \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \end{bmatrix} + \begin{bmatrix} -0.5 \\ 3.0 \\ -1.5 \end{bmatrix}$$

Since, no intial approximation is given, we start with $x^{(0)} = (0, 0, 0)^T$. We get the following approximations.

$X^{(1)} = [-0.5000 \quad 3.0000 \quad -1.5000]^T$
$X^{(2)} = [1.7500 \quad 2.5000 \quad 0.2500]^T$
$X^{(3)} = [0.6250 \quad 2.2500 \quad -1.1250]^T$
$X^{(4)} = [1.1875 \quad 2.1250 \quad -0.6875]^T$
$X^{(5)} = [0.9063 \quad 2.0625 \quad -1.0313]^T$
$X^{(6)} = [1.0469 \quad 2.0313 \quad -0.9219]^T$
$X^{(7)} = [0.9766 \quad 2.0156 \quad -1.0078]^T$
$X^{(8)} = [1.0117 \quad 2.0078 \quad -0.9805]^T$
$X^{(9)} = [0.9941 \quad 2.0039 \quad -1.0020]^T$
$X^{(10)} = [1.0029 \quad 2.0020 \quad -0.9951]^T$
$X^{(11)} = [0.9985 \quad 2.0010 \quad -1.0005]^T$
$X^{(12)} = [1.0007 \quad 2.0005 \quad -0.9988]^T$
$X^{(13)} = [0.9996 \quad 2.0002 \quad -1.0001]^T$
$X^{(14)} = [1.0002 \quad 2.0001 \quad -0.9997]^T$

After 14 iteractions, the errors in the solutions are

$$|x_1^{(14)} - x_1^{(13)}| = 0.0006, \; |x_2^{(14)} - x_2^{(13)}| = 0.0001, \; |x_3^{(14)} - x_3^{(13)}| = 0.0004.$$

The solutions $x^{(14)}$ are therefore almost correct to 3 decimal places.

The Gauss-Seidel Method

We observe from Examples 2 and 3 that even for a 3 x 3 system, the number of iterations taken by the Jacobi method (to achieve 2 or 3 decimal accuracy) is large. For large systems, the number of iterations required may run into thousands. Hence, the Jacobi method is slow. We also observe that when the variable x^i is being iterated in say the k-th iteration, the variables, $x_1, \ldots, x_{i-1}$ have already been updated in the k-th iteration. However, these values are not being used to compute $x_i^{(k)}$. This is the disadvantage of the Jacobi method. If we use all the current available values, we call it the Gauss-Seidel method.

Therefore, Gauss – seidel method is defined by

$$x_i^{(k)} = (bi - \sum_{j<i}^{n} a_{i,j} x^{(k)} - \sum_{j>i}^{n} a_{i,j} x_i^{(k-1)}) / a_{i,i} \qquad (21)$$

Two important facts about the Gauss-Seidel method should be noted. First, the computations in (21) are serial. Since, each component of the new iterate depends upon all previously computed components, the updates cannot be done simultaneously as in the Jacobi method. Second, the new iterate depends upon the order in which the equations are being used. The Gauss-Seidel method is sometimes called the *method of successive displacements* to indicate the dependence of the iterates on the ordering. If this ordering is changed, the *components* of the new iterate (and not just their order) will also change.

To derive the matrix formulation, we write,

$AX = (L + D + U)\,X = b$ or $(L + D)\,X = -UX + b.$

The Gauss-Seidel method can be expressed as

$$X^{(k+1)} = -(L + D)^{-1} U X^{(k)} + (L + D)^{-1} b$$
$$= M_G X^{(k)} + C \qquad (22)$$

where $M_G = -(L + D)^{-1} U$ is in iteration matrix and $C = (L + D)^{-1} b$.

Again, convergence depends on the properties of M_G if Spectral radius$(M_G) < 1$, the iteraction converges always for any intial solution vector. Further, it is known that Gauss-Seidel method converges atleast two times faster than for Jacobi method.

Example 4: Solve the system in Example 2, by the Gauss-Seidel method. Write its matrix form.

Solution: Gauss-Seidel method for solving the system in Example 2 is given by

$$x^{(k+1)} = -y^{(k)} + z^{(k)},\ y^{(k+1)} = \frac{1}{3}(2 + x^{(k)}),\ z^{(k+1)} = \frac{1}{2}(3 + x^{(k)})$$

with $x^{(0)} = 0.8$, $y^{(0)} = 0.8$, $z^{(0)} = 2.1$, we obtain the following results.

$x^{(1)} = 1.3,\ y^{(1)} = 1.1,\ z^{(1)} = 2.15;$
$x^{(2)} = 1.05,\ y^{(2)} = 1.01667,\ z^{(2)} = 2.025;$
$x^{(3)} = 1.00833,\ y^{(3)} = 1.00278,\ z^{(3)} = 2.004165;$
$x^{(4)} = 1.001385,\ y^{(4)} = 1.00046,\ z^{(4)} = 2.00069;$
$x^{(5)} = 1.00023,\ y^{(5)} = 1.000077,\ z^{(5)} = 2.000115;$

The errors after the 5^{th} iteractions are

$$|x^{(5)} - x^{(4)}| = 0.0012,\ |y^{(5)} - y^{(4)}| = 0.00038,\ |z^{(5)} - z^{(4)}| = 0.00057.$$

In 5 iteractions, we have for two place accuracy, while 9 iteractions we required in the Jacobi method.

The matrix function can be written as

$$X^{(k+1)} = -\begin{bmatrix} 1 & 0 & 0 \\ -1 & 3 & 0 \\ 1 & 0 & -2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} X^{(k)} + \begin{bmatrix} 1 & 0 & 0 \\ -1 & 3 & 0 \\ 1 & 0 & -2 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \\ -3 \end{bmatrix}$$

$$= +\frac{1}{6}\begin{bmatrix} -6 & 0 & 0 \\ -2 & -2 & 0 \\ 3 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} X^{(k)} = \frac{1}{6}\begin{bmatrix} -6 & 0 & 0 \\ -2 & -2 & 0 \\ -3 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \\ -3 \end{bmatrix}$$

$$= +\frac{1}{6}\begin{bmatrix} 0 & -6 & 6 \\ 0 & -2 & 2 \\ 0 & -3 & 3 \end{bmatrix} X^{(k)} - \frac{1}{6}\begin{bmatrix} 0 \\ -4 \\ -9 \end{bmatrix}$$

Starting with $X^{(0)} = [\,0.8 \quad 0.8 \quad 2.1]^T$, we get the same iterated values as above.

Example 5: Solve the system given in Example 3 by Gauss-Seidel method.

Solution: For Gauss- Seidel iterations, the system in Example 3 can be written as

$$x_1^{(k+1)} = 0.5\,x_2^{(k)} - 0.5\,x_3^{(k)} - 0.5$$
$$x_2^{(k+1)} = -0.5\,x_1^{(k+1)} + 0.5\,x_3^{(k)} + 3.0$$
$$x_3^{(k+1)} = -0.5\,x_1^{(k+1)} + 0.5\,x_2^{(k+1)} + 1.5$$

Start with (0, 0, 0), we get the following values

$X^{(1)} = [-0.5000 \quad 3.2500 \quad 0.3750]^T$
$X^{(2)} = [0.9375 \quad 2.7188 \quad -0.6094]^T$
$X^{(3)} = [1.1641 \quad 2.1133 \quad -1.0254]^T$
$X^{(4)} = [1.0693 \quad 1.9526 \quad -1.0583]^T$
$X^{(5)} = [1.0055 \quad 1.9681 \quad -1.0187]^T$
$X^{(6)} = [0.9934 \quad 1.9939 \quad -0.9997]^T$
$X^{(7)} = [0.9968 \quad 2.0017 \quad -0.9976]^T$
$X^{(8)} = [0.9996 \quad 2.0014 \quad -0.9991]^T$
$X^{(9)} = [1.0003 \quad 2.0003 \quad -1.0000]^T$
$X^{(10)} = [1.0001 \quad 1.9999 \quad -1.0001]^T$

After 10 iterations, the errors in solutions are
$|x_1^{(10)} - x_1^{(9)}| = 0.0002, |x_2^{(10)} - x_1^{(9)}| = 0.0.0004, |x_3^{(10)} - x_3^{(9)}| = 0.0001.$

The solutions are correct to 3 decimal places.

3.7 SUMMARY

In this unit, we have discussed direct and interative methods for solving a system of linear equations. Under these categories of methods, used to solve a system of linear equations, we have discussed Gauss elimination method and LU decomposition method. We have also discussed the method of finding inverse of a square matrix. Further, under the category of iterative methods for root determination we have discussed Jacobi and Gauss Seidel method.

3.8 EXERCISES

E1. Solve the following systems using the Gauss elimination method.

(a) $3x_1 + 2x_2 + 3x_3 = 5,$
$x_1 + 4x_2 + 2x_3 = 4,$
$2x_1 + 4x_2 + 8x_3 = 8,$

(b) $3x_1 + x_2 + x_3 = 1.8,$
$2x_1 + 4x_2 + x_3 = 2.7,$
$x_1 + 3x_2 + 5x_3 = 4.0.$

(c) $x_1 - x_2 + x_3 = 0,$
$2x_1 + 3x_2 + x_3 - 2x_4 = -7$
$3x_1 + x_2 - x_3 + 4x_4 = 12,$
$3x_2 - 5x_3 + x_4 = 9$

(d) $3x_1 + x_2 = 5,$
$x_1 + 3x_2 + 6x_3 = 6$
$4x_2 + x_3 + 3x_4 = 7$
$x_3 + 5x_4 = 8,$

E2. Solve the following systems using the LU decomposition method.

(a) $3x + y + z = 3,$
$x + 4y + 2z = 0,$
$2x + y + 5z = 4,$

(b) $2x + y + z = 5,$
$x + 3y + 2z = 4,$
$-y + y + 6z = 4,$

(c) $4x + y + 2z = 3.6,$
$x + 3y + z = 2.5,$
$2x + y + 2z = 4.0.$

(d) $3x + y = -2,$
$x + 3y - z = 0,$
$-y + 7z = 13.$

E3. For problems in 1(a), (b); 2(a), (b), (c), (d), obtain the solution to 3 decimals using the Jacobi and Gauss Seidel methods. Write the matrix formulations also. Assume the initial solution vectors respectively as

(i) $[0.8, 0.6, 0.5]^T$,
(ii) $[0.3, 0.3, 0.6]^T$,
(iii) $[0.9, -0.6, 0.6]^T$,
(iv) $[1.9, 0.2, 0.9]^T$,
(v) $[0.2, 0.5, 1.1]^T$,
(vi) $[-1.1, 0.9, 2.1]^T$.

3.9 SOLUTIONS TO ANSWERS

1. (a) 1, ½, ⅓ . (b) 0.3, 0.4, 0.5.
   (c) 1, –1, –2, 2. (d) 3/2, ½, ½, 3/2.

2. (a) 1, –1/2, 1/2. (b) 2, 0, 1.
   (c) 0.3, 0.4, 1. (d) –1, 1, 2, (You can also try LL^T decomposition)

3. Refer to Page 49 and 52.