Lesson: Errors in Numerical Computations

Course Developer: Kapil Kumar & Chaman Singh

Department/College: Assistant Professor, Department of Mathematics, A.R.S.D. College & A.N.D. College, University of Delhi

Errors in Numerical Computations

Table of Contents

1. Learning outcomes:

After studying this chapter you should be able to understand the

- Accuracy of Numbers
- Significant Digits
- Exact Numbers
- Approximate Numbers
- Approximation by Rounded-Off Numbers
- Approximation by Chopping or Truncation
- Floating Point Representation of Numbers
- Algorithm and their Stability
- Errors in Numerical Computation
- Sources and types of Errors
- General Error Formula

2. Introduction:

The final results in a numerical computations of unknown quantities are generally approximations that's why they are not exact but involve the errors. These errors depends on the computational methods used and must be dealt with individually for each method. In this chapter, we will study the types of errors occur in the numerical computation and how to analyze these errors.

3. Accuracy of Numbers:

Accuracy is a measure of correctness of a number. By the accuracy of an approximate number, we mean the number of significant digits it has.

3.1. Significant Digits:

The digits which are used to express a number are called significant digits. Every number has a specific number of significant digits. Significant digits in a number convey the actual numerical information and are not just written down to show where the decimal point is located. Thus, in a positional notation of a number significant digits are determined as follows:

(I) All the non-zero digits are significant digits.
(II) All the zero digits which lie between two non-zero digits are significant digits.
(III) All the zero digits which lie to the right of decimal point and at the same time to the right of a non-zero digit are significant digits.

Value Addition: Note
1. Significant digits in a number are counted from left to right starting with the left most non-zero digit. 2. Zero is a significant digit except when it is used to fix the decimal point or to fill the places of unknown or discarded digits.

Example 1: Determine the number of significant digits in the following numbers.

(I) 348500 (II) 0.0032 (III) 5.67800

(IV) 27.00 (V) 0.004030 (VI) 1.0056

Solution:

(I) Number 348500 contains only four significant digits i.e., 3, 4, 8 and 5.

(II) In the number 0.0032, the first three 0's are not significant digits since they serve only to fix the position of the decimal point and indicate the place values of the other digits. Thus, the number has only two significant digits.

(III) In the number 5.67800, both the 0's are significant digits. Therefore the number 5.67800 has six significant digits.

(IV) The number 27.00 has four significant digits.

(V) In the number 0.004030, the first three 0's are not significant digits whereas the last two 0's are significant digits. Thus, the number 0.004030 has four significant digits.

(VI) The number 1.0056 has five significant digits.

3.2. Significant Digits in the Scientific Notation of a Number:

In the scientific notation, a number is written in the form of $M \text{x} 10^k$. Thus, the significant digits in a number written in scientific notation consists of all the digits explicitly significant in M.

Example 2: Determine the number of significant digits in the following numbers

(I) $6.5\text{x}10^8$ (II) $1.02\text{x}10^{-4}$ (III) $8\text{x}10^{-9}$

Solution:

(I) The number $6.5\text{x}10^8$ has two significant digits namely 6 and 5.

(II) The number $1.02\text{x}10^{-4}$ has three significant digits namely 1, 0 and 2.

(III) The number $8\text{x}10^{-9}$ has only one significant digit namely 8.

I.Q. 1

I.Q. 2

4. Numbers in Numerical Computation:

In a numerical computation two type of the numbers are used

(I) Exact Numbers
(II) Approximate Numbers

4.1. Exact Number:

A number is called exact if either it is determined by definition or it is based on counting discrete units.

For example: We know that 1 minute = 60 seconds, here 1 and 60 are exact numbers. Similarly a number of students in a class is also an exact number.

4.2. Approximate Numbers:

The numbers which represents the given numbers to a certain degree of accuracy are called approximate numbers.

There are infinitely many numbers with large number of digits which cannot be expressed by a finite number of digits.

For example: The numbers $\frac{10}{3} = 3.33333\ldots$, $\frac{50}{7} = 7.142857\ldots$ and $\pi = 3.141592\ldots$ cannot be expressed by a finite number of digits. In practice, it is desirable to limit such numbers to a manageable number of digits such as $3.33, 7.14$ and 3.1416 respectively.

This process of dropping unwanted digits is called rounding method or chopping method.

4.3. Approximation by Rounded-Off Numbers:

To round-off a number, we use the following rule. To round-off a number, we check the $(n+1)^{th}$ place digit, if

(I) The $(n+1)^{th}$ place digit is less than 5, then leave the n^{th} place digit unchanged and discard all digits to the right of n^{th} digit.

(II) The $(n+1)^{th}$ place digit is greater than 5, then add 1 to the n^{th} place digit and discard all digits to the right of n^{th} digit.

(III) The $(n+1)^{th}$ place digit is exactly 5, then add 1 to the n^{th} digit if it is odd otherwise leave it unchanged and discard all digits to the right of n^{th} digit.

Example 3: Round-off the numbers to four significant digits

(i) 0.00346512 (ii) 50.6079

(iii) 20.0358 (iv) 0.567852

Solution:

(i) 0.00346512 is rounded-off to 0.003465.

(ii) 50.6079 is rounded-off to 50.61.

(iii) 20.0358 is rounded-off to 20.04.

(iv) 0.567852 is rounded-off to 0.5678.

I.Q. 3

4.4. Approximation by Chopping or Truncation:

To approximate a number by chopping or truncation to n digits, chopped or truncate all the digits to the right of nth digit. For example, the number 63.15921 will be approximate to 63.15 by chopping to four significant digits.

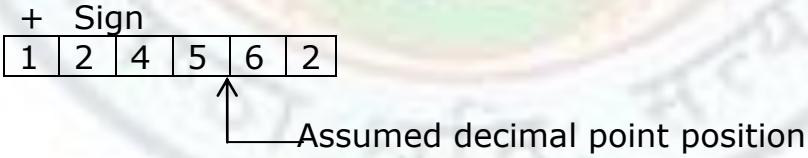
Value Addition: Note
Chopping is used in a computer with a fixed word length.

5. Floating Point Representation of Numbers:

In decimal representation of numbers, every real number is represented by a finite or infinite sequence of decimal digits. For numerical computation by the computers a number is replaced by the string of finitely many digits.

Let us assume a hypothetical computer having memory in which each location can store 6 digits and having provision to store one or more signs.

For example +1245.62 can be stored as follows



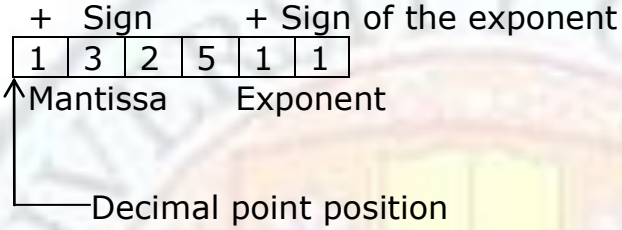
In this representation, maximum and minimum possible numbers that can be stored are 9999.99 and 0000.01 respectively in magnitude. Due to their limited range fixed-point representation are impractical in large scientific computations. Thus, this range is quite inadequate in practice..

For this a new convention is adopted which aims to preserve the maximum number of significant digits in a real number and also increase the range of values of real numbers stored. This representation is called floating point representation of numbers.

In this representation a number is expressed as a combination of a mantissa and an exponent.

The mantissa is always made greater than or equal to 0.1 and less than 1 and the exponent is the power of 10 which multiplies the mantissa. For example the number 13.25×10^{9} is represented as $.1325\times10^{11}$ and written as .1325E11 where E11 is used to represent 10^{11}. In this number mantissa is .1325 and exponent is 11.

Thus the number is stored in the memory location as



In the floating point representation, the range of the numbers that may be stored are $.1000\times10^{-99}$ *to* $.9999\times10^{99}$ in magnitude which is obviously much larger than that used earlier in fixed point notation.

Value Addition: Note
Chopping is not recommended because the corresponding error can be larger than that in rounding and is systematic. Although some computers use it because it is simpler and faster. On the other hand, some computers and calculators improve accuracy of results by doing intermediate calculations using one or more extra digits called guarding digits.

6. Algorithm:

In mathematics and computer science, Algorithm are used for calculation, data processing and automated reasoning. An algorithm is a finite sequence of rules for performing computations on a computer starting from an initial state and initial input such that at each instant the rules determine exactly what the computer has to do next. These rules describe a computation that when executed proceeds through a finite number of well-defined successive states eventually producing output and terminating at a final ending state known as "stopping rule" that makes the computer stop, thus it cannot run on indefinitely.

Features of an Algorithm:

An algorithm has five important features:

(I) **Finiteness:** An algorithm must terminate after a finite number of steps.

(II) **Definiteness:** Each step of an algorithm must be clearly defined or the action to be taken must be unambiguously specified.

(III) **Inputs:** An algorithm must specify the quantities which must be read before the algorithm can begin.

(IV) **Outputs:** An algorithm must specify the quantities which are to be outputted and their proper place.

(V) **Effectiveness:** An algorithm must be effective which means that all operations are executable.

6.1. Stability of Algorithm:

For the usefulness, an algorithm should be stable. If the small changes in the initial data produces small changes in the final results, then the algorithm is called stable algorithm. However if small changes in the initial data produces large changes in the final results then the algorithm is called unstable.

Value Addition: Underflow and Overflow
The range of exponents that a typical computer can handle is very large. If in a computation a number outside that range occurs, this is called underflow when the number is smaller and overflow when it is larger. In the case of underflow the result is usually set to zero and computation continues. Overflow causes the computer to halt.

I.Q. 4

7. Errors in Numerical Computation:

We have learnt that a computer has a finite word length so only a fixed number of digits are stored an used during computation. Therefore even if we input an exact decimal number in its converted form in the computer memory, an error is introduced in the outcome. Thus, error in numerical computation is the difference of true value of a number and the approximate value of the number.

Error = True value - Approximate value.

7.1. Sources of Errors:

In numerical computation following are the measure sources of the errors.

(I) **Algorithmic Errors:** If in a numerical computation an algorithm base on finite sequence of operations is used, then the limited steps

don't amplify the error. But if an infinite algorithm is used than the exact results are expected only after an infinite number of steps. Since this cannot be done in practice therefore algorithm has to be stopped after a finite number of steps and as a consequence the outcome results are not exact.

(II) **Input Errors:** The data used as an input in a numerical computation is rarely exact because the input data or raw data is generally collected from the experiments and the experiments give the results of only limited accuracy. Also the input data is represented in a computer for only a limited number of digits.

(III) **Computational Errors:** When the elementary operations such as multiplication and division are used in a computation, then the number of digits in the result may increase greatly. In such cases, a certain number of digits must be discarded due to the fixed number of digits storages in a computer. Therefore, the errors accumulate one after another from operation to operation.

7.2. Types of Errors:

In a numerical computation following types of errors occurs.

(I) **Inherent Errors:** Errors which are present in the input data or in the statement of a problem before its solution are called inherent errors. Such errors occurs either due to approximation of the given data or due to limitations of mathematical tables, calculators or the digital computer.

(II) **Rounding Errors:** The errors which occurs due to the rounding off the numbers during the computation are called rounding errors.

(III) **Truncation Errors:** The errors due to the using approximate results or on replacing an infinite process by a finite one are called truncation errors.

(IV) **Absolute Errors:** The numerical difference between the true value of a quantity and its approximate value is called the absolute value. The absolute error is denoted by E_a and defined as

$$\mathrm{E_a} = \left|\text{True value}\,(\mathrm{X}) - \text{Approximate value}\,(\mathrm{X'})\right| = \left|X - X'\right|$$

(V) **Relative Errors:** The ratio of the absolute error and the true value of a quantity is called the relative error. Relative error is denoted by E_r and defined as

$$\mathrm{E_r} = \left|\frac{\text{Absolute Error}}{\text{True value}}\right| = \left|\frac{\text{True value - Approximate value}}{\text{True value}}\right|$$

(VI) Percentage Error: Percentage error is denoted by E_p and defined as

$$E_p = 100.E_r\% = 100.\left|\frac{\text{True value - Approximate value}}{\text{True value}}\right|\%$$

Value Addition: Note
1. The relative and percentage errors are independent of units. **2.** If a number is correct to n decimal places, then the maximum error is $\frac{1}{2}\times 10^{-n}$. **3.** If the first significant digit of a number is k and the number is correct to n-significant digits, then the maximum relative error is $\frac{1}{k\times 10^{n-1}}$.

Example 4: If 3.33 is the approximate value of $\frac{10}{3}$, find absolute, relative and percentage errors.

Solution: Given

True value (X) = $\frac{10}{3}$

Approximate value (X') = 3.33

Error = X - X'

= $\frac{10}{3}$ - 3.33

=3.333333 - 3.33

=0.003333

$$\text{Absolute Error } (E_a) = |X - X'| = |0.003333| = 0.003333$$

$$\text{Relative Error } (E_r) = \frac{|X - X'|}{X}$$

$$= \left|\frac{0.003333}{\frac{10}{3}}\right|$$

$$= \frac{0.003333 \times 3}{10}$$

$$= 0.000999$$

$$\text{Percentage Error } (\text{E}_p) = E_r \times 100\%$$
$$= 0.000999 \times 100\%$$
$$= 0.0999\%$$

Example 5: If the true value of π is 3.1415926 and its approximate value is given by 3.1428571. Find the absolute and relative errors.

Solution: Given

True value (X) = 3.1415926

Approximate value (X') = 3.1428571

Error = X - X'

= 3.1415926-3.1428571

=-0.0012645

$$\text{Absolute Error } (\text{E}_a) = |X - X'| = |-0.0012645| = 0.0012645$$

$$\text{Relative Error } (\text{E}_r) = \frac{|X - X'|}{X}$$
$$= \left|\frac{0.0012645}{3.1415926}\right|$$
$$= 0.000402502$$

Example 6: Round-off the number 35.46735 correct to four significant digits and then calculate the absolute and relative errors.

Solution: Given

True value (X) = 35.46735

Approximate value (X') = 35.47

Error = X - X'

= 35.46735 - 35.47

=-0.00265

$$\text{Absolute Error } (\mathrm{E_a}) = |X - X'| = |-0.00265| = 0.00265$$

$$\text{Relative Error } (\mathrm{E}_r) = \frac{|X - X'|}{X}$$

$$= \frac{0.00265}{35.46735}$$
$$= 0.000074$$

Example 7: Round off the number 65468 to four significant digits and then calculate the absolute, relative and percentage errors.

Solution: Given

True value (X) = 65468

Approximate value (X') = 65470

Error = X - X'

= 65468 - 65470

=-2

$$\text{Absolute Error } (\mathrm{E_a}) = |X - X'| = |-2| = 2$$

$$\text{Relative Error } (\mathrm{E}_r) = \frac{|X - X'|}{X}$$
$$= \left|\frac{2}{65468}\right|$$
$$= 0.000030$$

$$\text{Percentage Error } (\mathrm{E}_p) = E_r \times 100\%$$
$$= 0.00003 \times 100\%$$
$$= 0.003\%$$

Example 8: Find the relative error of the number 5.6 if both of its digits are correct.

Solution: We know that if a number is correct to n decimal places then the error in the number is

Error = $\frac{1}{2} \times 10^{-n}$

Since 5.6 is correct to one decimal place, therefore

Error = $\frac{1}{2}\times 10^{-1} = \frac{1}{2}\times\frac{1}{10} = 0.05$

$$\text{Absolute Error } (E_a) = |Error| = |0.05| = 0.05$$

$$\text{Relative Error } (E_r) = \frac{|\text{Absolute Error}|}{\text{True Error}}$$

$$= \frac{0.05}{5.6}$$

$$= 0.0089$$

Example 9: If X = 2.536, find the absolute error and relative error when

(I) X is rounded-off to two decimal digits.
(II) X is truncated to two decimal digits.

Solution: Given that

X = 2.536

(I) If X is rounded-off to two decimal digits, then
Approximate value X' = 2.54
Error = X - X' = 2.536-2.54 = -0.004
Absolute Error = $|\text{Error}| = |-0.0004| = 0.0004$

$$\text{Relative Error } (E_r) = \frac{|\text{Absolute Error}|}{\text{True Error}}$$

$$= \frac{0.0004}{2.536}$$

$$= 0.000157$$

(II) If X is truncated to two decimal digits, then
Approximate value X' = 2.53
Error = X - X' = 2.536-2.53 = 0.006
Absolute Error = $|\text{Error}| = |0.006| = 0.006$

$$\text{Relative Error } (E_r) = \frac{|\text{Absolute Error}|}{\text{True Error}}$$

$$= \frac{0.006}{2.536}$$

$$= 0.00236$$

I.Q. 5

I.Q. 6

8. Error Propagation:

Theorem 1: The absolute error in the addition and subtraction of two numbers is bounded by the sum of the absolute errors in numbers.

Proof: Let X and Y are two numbers and let X' and Y' are the approximate values of X and Y respectively.

Let $\Delta X \text{ and } \Delta Y$ are errors in X and Y respectively.

(I) Errors in addition of numbers

Let Z = X + Y

and Z' = X' + Y'

then error in Z is

$$\begin{aligned}\Delta Z = Z - Z' &= \left[\left(X+Y\right)-\left(X'+Y'\right)\right] \\ &= \left[\left(X-X'\right)+\left(Y-Y'\right)\right] \\ &= \Delta X + \Delta Y\end{aligned}$$

Absolute error in Z is

$$\left|\Delta Z\right| = \left|\Delta X + \Delta Y\right| \le \left|\Delta X\right| + \left|\Delta Y\right|.$$

(II) Errors in subtraction of numbers

Let Z = X - Y

and Z' = X' - Y'

then Absolute error in Z is

$$\begin{aligned}\left|\Delta Z\right| = \left|Z - Z'\right| &= \left|\left(X-Y\right)-\left(X'-Y'\right)\right| \\ &= \left|\left(X-X'\right)-\left(Y-Y'\right)\right| \\ &= \left|\Delta X - \Delta Y\right|\end{aligned}$$

$$\left|\Delta Z\right| \le \Delta X - \Delta Y$$

Value Addition: Note
In general, Let $X = x_1 + x_2 + \ldots + x_n$ is the sum of n quantities. If ΔX is the error in X and $\Delta x_1, \Delta x_2, \ldots, \Delta x_n$ are the errors in $x_1, x_2, \ldots \text{ and } x_n$ respectively, then absolute error in X is bounded by the sum of absolute errors in its components, i.e., $\left\|\Delta X\right\| \le \left\|\Delta x_1\right\| + \left\|\Delta x_2\right\| + \ldots + \left\|\Delta x_n\right\|$.

Theorem 2: The absolute error in the multiplication and division of two numbers is bounded by the sum of the relative errors in the numbers.

Proof: Let X and Y are two numbers and let X' and Y' are the approximate values of X and Y respectively.

Let $\Delta X \text{ and } \Delta Y$ are errors in X and Y respectively.

(I) Errors in multiplication of numbers

Let Z = X.Y

and Z' = X'.Y'

then,

$$\Delta Z = \frac{\partial Z}{\partial X}\Delta X + \frac{\partial Z}{\partial Y}\Delta Y$$

$$\frac{\Delta Z}{Z} = \frac{1}{Z}\left[\frac{\partial Z}{\partial X}\Delta X + \frac{\partial Z}{\partial Y}\Delta Y\right]$$

$$= \frac{1}{XY}\left[Y\Delta X + X\Delta Y\right]$$

$$\Rightarrow \quad \frac{\Delta Z}{Z} = \left[\frac{\Delta X}{X} + \frac{\Delta Y}{Y}\right]$$

Relative error in Z is

$$\left|\frac{\Delta Z}{Z}\right| = \left|\frac{\Delta X}{X} + \frac{\Delta Y}{Y}\right| \le \left|\frac{\Delta X}{X}\right| + \left|\frac{\Delta Y}{Y}\right|$$

$$\Rightarrow \quad E_{rZ} \le E_{rX} + E_{rY}\,.$$

(II) Errors in division of numbers

Let $Z = \frac{X}{Y}$

then,

$$\Delta Z = \frac{\partial Z}{\partial X}\Delta X + \frac{\partial Z}{\partial Y}\Delta Y$$

$$\frac{\Delta Z}{Z} = \frac{1}{Z}\left[\frac{\partial Z}{\partial X}\Delta X + \frac{\partial Z}{\partial Y}\Delta Y\right]$$

$$= \frac{Y}{X}\left[\frac{1}{Y}\Delta X + \frac{-X}{Y^2}\Delta Y\right]$$

$$\Rightarrow \quad \frac{\Delta Z}{Z} = \left[\frac{\Delta X}{X} - \frac{\Delta Y}{Y}\right]$$

Relative error in Z is

$$\left|\frac{\Delta Z}{Z}\right| = \left|\frac{\Delta X}{X} - \frac{\Delta Y}{Y}\right| \le \left|\frac{\Delta X}{X}\right| + \left|\frac{\Delta Y}{Y}\right|$$

$$\Rightarrow \quad E_{rZ} \le E_{rX} + E_{rY}\,.$$

Value Addition: Note
In general, Let $X = x_1.x_2 \ldots x_n$ is the product of n quantities. If E_{rX} is the relative error in X and $E_{rx_1}, E_{rx_2}, \ldots, E_{rx_n}$ are the relative errors in $x_1,\ x_2, \ldots \ and\ x_n$ respectively, then relative error in X is bounded by the sum of relative errors in its components, i.e., $\lvert\Delta X\rvert \le \lvert\Delta x_1\rvert + \lvert\Delta x_2\rvert + \ldots + \lvert\Delta x_n\rvert$.

I.Q. 7

9. General Error Formula:

Let $z = f(x, y)$ be a function of two variables x and y. Let $\delta x \ and\ \delta y$ be the errors in x and y respectively then δz in z is given by

$$z + \delta z = f(x + \delta x,\ y + \delta y)$$

Expanding R.H.S. by Taylor's series we get

$$z + \delta z = f(x,\ y) + \left(\frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial y}\delta y\right) + \text{ terms involving higher powers of } \delta x \ and\ \delta y$$

since the errors $\delta x \ and\ \delta y$ are very small then their squares and higher powers will be very-very small therefore can be neglected, therefore

$$z + \delta z = z + \left(\frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial y}\delta y\right)$$

$$\Rightarrow \qquad \delta z = \frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial y}\delta y$$

Relative error in z is

$$\left|\frac{\delta z}{z}\right| = \left|\frac{\partial f}{\partial x}\frac{\delta x}{z} + \frac{\partial f}{\partial y}\frac{\delta y}{z}\right|$$

$$\Rightarrow \qquad \left|\frac{\delta z}{z}\right| \le \left|\frac{\partial f}{\partial x}\frac{\delta x}{z}\right| + \left|\frac{\partial f}{\partial y}\frac{\delta y}{z}\right|$$

Value Addition: Note
Let $z = f(x_1, x_2, \ldots, x_n)$ is a function of n variables and $\delta x_i \ (i = 1, 2, \ldots n)$ then relative error in z is $\left\|\frac{\delta z}{z}\right\| \le \left\|\frac{\partial f}{\partial x_1}\frac{\delta x_1}{z}\right\| + \left\|\frac{\partial f}{\partial x_2}\frac{\delta x_2}{z}\right\| + \ldots + \left\|\frac{\partial f}{\partial x_n}\frac{\delta x_n}{z}\right\|$.

Example 10: Approximate values of $\frac{1}{11}$ *and* $\frac{1}{7}$ correct to 4 decimal places are 0.0909 and 0.1429 respectively. Find the possible absolute error and relative error in the sum of 0.0909 and 0.1429. What is the smallest interval in which the exact sum of the numbers must lie ?

Solution: We know that if a number is correct to n decimal places then the error in the number is $\frac{1}{2}\times 10^{-n}$.

Since the numbers 0.0909 and 0.1429 are correct to four decimal places thus maximum error in each case is $\frac{1}{2}\times 10^{-4} = 0.00005$.

Let X = 0.0909 and Y = 0.1429 and let ΔX *and* ΔY are the error in X and Y respectively, then $\Delta X = \Delta Y = 0.00005$.

Let Z = X + Y,

Absolute error in the sum is

$$E_a = 0.00005 + 0.00005 = 0.0001.$$

Relative error in the sum is

$$E_r = \left|\frac{\Delta Z}{Z}\right| \le \left|\frac{\Delta X}{Z}\right| + \left|\frac{\Delta Y}{Z}\right| = \frac{0.00005}{0.2338} + \frac{0.00005}{0.2338} = \frac{0.0001}{0.2338} = 0.00043.$$

Now the approximate value of the sum is

$$Z' = (X + Y) \pm E_a$$

$$\Rightarrow \quad Z' = (0.0909 + 0.1429) \pm 0.0001$$

$$\Rightarrow \quad Z' = 0.2338 \pm 0.0001$$

$$\Rightarrow \quad Z' = 0.2337 \ \ or \ \ 0.2339$$

Thus, the exact sum lies in the interval (0.2337, 0.2339).

Example 11: If $\sqrt{5.5} = 2.345 \ and \ \sqrt{6.1} = 2.470$ are correct to four significant digits. Find the relative error in the difference of these numbers.

Solution: We know that if a number is correct to n decimal places then the error in the number is $\frac{1}{2} \times 10^{-n}$.

Since the numbers $\sqrt{5.5} = 2.345 \ and \ \sqrt{6.1} = 2.470$ are correct to three decimal places thus maximum error in each case is $\frac{1}{2} \times 10^{-3} = 0.0005$.

Let $X = \sqrt{5.5} = 2.345 \ and \ Y = \sqrt{6.1} = 2.470$ and let $\Delta X \ and \ \Delta Y$ are the error in X and Y respectively, then $\Delta X = \Delta Y = 0.0005$.

Let Z = Y - X,

Relative error in the difference is

$$E_r = \left|\frac{\Delta Z}{Z}\right| \le \left|\frac{\Delta X}{Z} - \frac{\Delta Y}{Z}\right| \le \left|\frac{\Delta X}{Z}\right| + \left|\frac{\Delta Y}{Z}\right| = \frac{0.0005}{0.125} + \frac{0.0005}{0.125} = \frac{0.001}{0.25} = 0.008.$$

Example 12: Find the product of the numbers 56.54 and 12.4 which are correct to given significant digits in the numbers. Find the relative error in the product.

Solution: Let X=56.54 and Y = 12.4, and let $\Delta X \ and \ \Delta Y$ are the error in X and Y respectively, then maximum error in X is $\Delta X = 0.005$ and the maximum error in Y is $\Delta Y = 0.05$.

Let Z = XY,

Relative error in the product is

$$E_r = \left|\frac{\Delta Z}{Z}\right| = \left|\frac{\Delta X}{X} + \frac{\Delta Y}{Y}\right| \le \left|\frac{\Delta X}{X}\right| + \left|\frac{\Delta Y}{Y}\right| = \frac{0.005}{56.54} + \frac{0.05}{12.4} = 0.00008 + 0.00403 = 0.00411.$$

Example 13: Find the product of the numbers 346.1 and 865.2 which are correct to four significant figures. State how many figures of the result are trustworthy?

Solution: Let X=346.1 and Y = 865.2, let $\Delta X \text{ and } \Delta Y$ are the error in X and Y respectively, then maximum error in X and Y are $\Delta X = 0.05$ and $\Delta Y = 0.05$.

Let Z = XY,

$\Rightarrow$ Z = 299446 (correct to 6 significant digits)

Maximum Relative error in the product is

$$E_r = \left|\frac{\Delta Z}{Z}\right| = \left|\frac{\Delta X}{X} + \frac{\Delta Y}{Y}\right| \le \left|\frac{\Delta X}{X}\right| + \left|\frac{\Delta Y}{Y}\right| = \frac{0.05}{346.1} + \frac{0.05}{865.2} = 0.0001444 + 0.0000578 = 0.000202.$$

Absolute error $E_a = E_r \times Z = 0.000202 \times 299446 \approx 60$.

True value of the product of the numbers lies between 299446-60=299386 and 299446+60=299506.

Mean of these values is $\frac{299386 + 299506}{2} = 299446 = 299.4 \times 10^3$ which is correct to four significant digits. There is some uncertainty about the last digit.

Example 14: Compute the percentage error in the time period $T = 2\pi\sqrt{\frac{\ell}{g}} \text{ for } \ell = 1m$ if the error in the measurement of ℓ is 0.01.

Solution: Given that

$$T = 2\pi\sqrt{\frac{\ell}{g}}$$

taking log of both sides we have

$$\log T = \log(2\pi) + \frac{1}{2}\log(\ell) - \frac{1}{2}\log(g)$$

Thus, we have

$$\frac{1}{T}\delta T = \frac{1}{2\ell}\delta\ell \qquad \text{[since } \pi \text{ and } g \text{ are constants]}$$

Relative error in T is

$$E_r = \left|\frac{1}{T}\delta T\right| = \left|\frac{1}{2\ell}\delta \ell\right|$$

$$\Rightarrow \quad E_r = \left|\frac{0.01}{2}\right|$$

$$\Rightarrow \quad E_r = 0.005$$

Percentage error in T is

$$E_p = E_r \times 100\ \%$$

$$\Rightarrow \quad E_p = 0.005 \times 100\ \% = 0.5\%.$$

Example 15: If $u = 4xy^2z^{-3}$ and errors in x, y, z be 0.001, show that maximum relative error in u at x = y = z = 1 is 0.006.

Solution: Given that

$$u = 4xy^2z^{-3}$$

$$\Rightarrow \quad \delta u = \frac{\partial u}{\partial x}\delta x + \frac{\partial u}{\partial y}\delta y + \frac{\partial u}{\partial z}\delta z$$
$$= 4y^2z^{-3}\delta x + 8xyz^{-3}\delta y - 12xy^2z^{-4}\delta z$$

Thus, the relative error in u is

$$\left|\frac{\delta u}{u}\right| = \frac{1}{u}\left|\frac{\partial u}{\partial x}\delta x + \frac{\partial u}{\partial y}\delta y + \frac{\partial u}{\partial z}\delta z\right|$$
$$= \left|\frac{\partial u}{\partial x}\frac{\delta x}{u} + \frac{\partial u}{\partial y}\frac{\delta y}{u} + \frac{\partial u}{\partial z}\frac{\delta z}{u}\right|$$
$$= \left|\frac{\delta x}{x} + 2\frac{\delta y}{y} - 3\frac{\delta z}{z}\right|$$

$$\Rightarrow \quad \left|\frac{\delta u}{u}\right| \le \left|\frac{\delta x}{x}\right| + 2\left|\frac{\delta y}{y}\right| + 3\left|\frac{\delta z}{z}\right|$$

Thus the maximum relative error in u is

$$\left|\frac{\delta u}{u}\right|_{\max} = \left|\frac{\delta x}{x}\right| + 2\left|\frac{\delta y}{y}\right| + 3\left|\frac{\delta z}{z}\right|$$

$$\Rightarrow \quad \left|\frac{\delta u}{u}\right|_{\max} = \left|\frac{0.001}{1}\right| + 2\left|\frac{0.001}{1}\right| + 3\left|\frac{0.001}{1}\right|$$

$$\Rightarrow \quad \left|\frac{\delta u}{u}\right|_{\max} = 0.001 + 0.002 + 0.003 = 0.006$$

Hence the maximum relative error in u is 0.006.

Example 16: Find the roots of the equation $x^2 - 40x + 2 = 0$ using four significant digits in the equation.

Solution: Given equation is

$$x^2 - 40x + 2 = 0$$

on comparing it with

$$ax^2 + bx + c = 0$$

we have

a = 1, b = -40 and c = 2.

Formula for the roots of the quadratic equation $ax^2 + bx + c = 0$ is

$$x_1 = \frac{1}{2a}\left(-b + \sqrt{b^2 - 4ac}\right) \quad \text{and} \quad x_2 = \frac{1}{2a}\left(-b - \sqrt{b^2 - 4ac}\right) \tag{1}$$

We also know that

$$x_1 x_2 = \frac{c}{a}$$

$$\Rightarrow \quad x_2 = \frac{c}{ax_1} \tag{2}$$

Now putting the values of a, b and c in equation (1), we have

$$x_1 = 20 + \sqrt{398} = 20.00 + 19.95 = 39.95$$

and $x_1 = 20 - \sqrt{398} = 20.00 - 19.95 = 0.05$ is poor because it involves loss of significant digits.

Now calculating the value of x_2 using equation (2) we have

$$x_2 = \frac{2}{39.95} = 0.05006$$ which is in error by less than one unit of the last digit as a computation with more digits shows.

I.Q. 8

Example 17: Find the relative error in the function $y = a_1 x_1^{m_1} a_2 x_2^{m_2} \ldots a_n x_n^{m_n}$.

Solution: Given that

$$y = a_1 x_1^{m_1} a_2 x_2^{m_2} \ldots a_n x_n^{m_n}$$

We have

$$\log y = \log a + m_1 \log x_1 + m_2 \log x_2 + \ldots + m_n \log x_n$$

We know

$$\delta y = \frac{\partial y}{\partial x_1}\delta x_1 + \frac{\partial y}{\partial x_2}\delta x_2 + \ldots + \frac{\partial y}{\partial x_n}\delta x_n$$

Thus, relative error is

$$E_r = \left|\frac{\delta y}{y}\right| = \left|\frac{\partial y}{\partial x_1}\frac{\delta x_1}{y} + \frac{\partial y}{\partial x_2}\frac{\delta x_2}{y} + \ldots + \frac{\partial y}{\partial x_n}\frac{\delta x_n}{y}\right|$$

$$\leq \left|\frac{\partial y}{\partial x_1}\frac{\delta x_1}{y}\right| + \left|\frac{\partial y}{\partial x_2}\frac{\delta x_2}{y}\right| + \ldots + \left|\frac{\partial y}{\partial x_n}\frac{\delta x_n}{y}\right| \quad (1)$$

Now we have

$$\frac{1}{y}\left(\frac{\partial y}{\partial x_1}\right) = \frac{m_1}{x_1}$$

$$\frac{1}{y}\left(\frac{\partial y}{\partial x_2}\right) = \frac{m_2}{x_2}$$

.

.

.

$$\frac{1}{y}\left(\frac{\partial y}{\partial x_n}\right) = \frac{m_n}{x_n}$$

Now putting these value in equation (1) we have

$$E_r \leq \left|\frac{m_1 \delta x_1}{x_1}\right| + \left|\frac{m_2 \delta x_2}{x_2}\right| + \ldots + \left|\frac{m_n \delta x_n}{x_n}\right|$$

$$\Rightarrow \qquad E_r \le |m_1|\left|\frac{\delta x_1}{x_1}\right| + |m_2|\left|\frac{\delta x_2}{x_2}\right| + \ldots + |m_n|\left|\frac{\delta x_n}{x_n}\right|.$$

Example 18: Prove that the absolute error in the common logarithm of a number is less than half the relative error of the given number.

Solution: Let x is any number and the common logarithm of x is denoted as

$$N = \log_{10} x = 0.43429 \log_e x$$

Thus error in N is

$$\delta N = 0.43429 \frac{\delta x}{x}$$

Absolute error in N is

$$E_a = |\delta N|$$
$$= \left|0.43429 \frac{\delta x}{x}\right|$$

$$\Rightarrow \qquad E_a < \frac{1}{2}\left|\frac{\delta x}{x}\right|$$

Thus, absolute error in the common logarithm of a number is less than half the relative error of the given number.

Example 19: What accuracy $\log_{10} 2$ should be given to obtain the roots of the equation $x^2 - 2x + \log_{10} 2 = 0$.

Solution: Given equation is

$$x^2 - 2x + \log_{10} 2 = 0$$

We know that root of the quadratic equation $ax^2 + bx + c = 0$ is given by

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Thus, roots of the equation $x^2 - 2x + \log_{10} 2 = 0$ are

$$x = \frac{2 \pm \sqrt{4 - 4\log_{10} 2}}{2} = 1 \pm \sqrt{1 - \log_{10} 2}$$

Error in roots is

$$\Delta x = \frac{1}{2}\frac{\Delta(\log_{10} 2)}{\sqrt{1-\log_{10} 2}}$$

Since roots are correct to four decimal places therefore maximum error in the roots is

$$|\Delta x| < 0.5\times 10^{-4}$$

$$\Rightarrow \quad \frac{1}{2}\frac{\Delta(\log_{10} 2)}{\sqrt{1-\log_{10} 2}} < 0.5\times 10^{-4}$$

$$\Rightarrow \quad \Delta(\log_{10} 2) < 2\times 0.5\times 10^{-4}\left(\sqrt{1-\log_{10} 2}\right) < 0.83604\times 10^{-4} \square\ 8.3604\times 10^{-5}.$$

Example 20: Find the solution of the equation $x^2 - 7x + 4 = 0$ using floating point arithmetic with 4-digit mantissa.

Solution: Given equation is

$$x^2 - 7x + 4 = 0$$

roots of the above equation are given by

$$x = \frac{7 \pm \sqrt{(-7)^2 - 4^2}}{2}$$

I.Q. 9

I.Q. 10

Exercise:

1. Round-off the following numbers correct to four significant digits:

(I) 5.612549
(II) 8.462889
(III) 0.6500025
(IV) 4589647
(V) 0.000232317
(VI) 50.00045

2. If X = 4.357, find the absolute error, relative error and percentage error when

(I) X is rounded-off to two decimal digits
(II) X is truncated to two decimal digits

3. If approximations to $\pi = 3.14159265358979\ldots$ are $\frac{22}{7}$ *and* $\frac{355}{113}$. Determine the corresponding errors and relative errors to three significant digits.

4. Let 34.65 and 71.0159 be correctly rounded to the number of digits shown. Find the smallest interval in which the exact sum of the numbers must lie.

5. If $u = \frac{4x^2y^3}{z^4}$ and errors in x, y and z be 0.001, compute the relative maximum error in u when x = y = z = 1.

6. If $u = 4x^2y^3z^{-4}$, find the maximum absolute error and maximum relative error in u when errors in x = 1, y = 2 and z = 3 respectively are equal to 0.001, 0.002 and 0.003 respectively.

7. Find the number of terms of the exponential series such that their sum gives the value of e^x correct to six decimal places at x = 1.

8. Find the smallest root of the equation $x^2 - 30x + 1 = 0$ correct to three decimal places.

9. Find the smallest root of the equation $x^2 + 100x + 2 = 0$ correct to five significant digits.

10. If $u = 5xy^2z^{-3}$, find the maximum absolute error and maximum relative error in u when errors in x = 1, y = 1 and z = 1 respectively are equal to 0.001 each.

Summary:

In this lesson we have emphasized on the followings

- Accuracy of Numbers
- Significant Digits
- Exact Numbers
- Approximate Numbers
- Approximation by Rounded-Off Numbers
- Approximation by Chopping or Truncation
- Floating Point Representation of Numbers
- Algorithm and their Stability
- Errors in Numerical Computation
- Sources and types of Errors
- General Error Formula

References:

1. Brian Bradie, A Friendly Introduction to Numerical Analysis, Peason Education, India, 2007.
2. M.K. Jain, S.R. K. Iyengar and R. K. Jain, Numerical Methods for Scientific and Engineering Computation, New Age International Publisher, India, 6th edition, 2007.