Numerical Methods.

* **Floating point Arithmetic** :- Floating point arithmetic also known as real point arithmetic, uses the numbers with fractional parts as operands and it is used in most of the computations.
  - fixed point
  - floating point.

Fixed point is used to represent integers and the floating point is used to represent real numbers.
Floating point number system $f(\beta, k, m, M)$ can be defined as a subset of the real numbers system which is characterized by the parameters:

β: The Base.
k: The number of digits in the base β
m: minimum exponent
M: Maximum exponent.

Elements of $f(\beta, k, m, M)$ can also be expressed as

$$x = \pm(0.d_1 d_2 \cdots d_n)_\beta$$

where $d_1, d_2 \cdots, d_n$ are all digits in the base β and all d's lies between 0 and β.
The fractional parts can also be written as

$$(0.d_1 d_2 \cdots d_n)_\beta = d_1 \times \beta^{-1} + d_2 \times \beta^{-2} + \cdots + d_n \times \beta^{-n}.$$

which is called mantissa.
eg: 79.54056, 8.100432 etc.
mantissa → 54056 mantissa 100432.
Shifting the mantissa to the left till the non zero digit is called normalization.
eg Normalize 42.53×10^6 ?

PTO.

Scanned with CamScanner

After normalization

$.4253 \times 10^8$

Now mantissa is 4253

Exponent is 8 (ie 10^8.)

Significant figures:- The digits 1, 2, 3, ... 8, 9 which are used to express a number are known as significant digits or figures. Zero also plays a role of significant digit ~~and~~ except when it is used to fix the decimal point to fill the places of unknown digits. To find significant digits, the following kept in mind.

① If the number is in positional notation, then the significant figures in the number consists of

(i) All non zero digits

(ii) The zero digit which lies between significant digits and lies to the right of decimal point and at the same time, to the right of a non-zero digit.

eg. find significant digit of 0.000789

The significant digits are 7, 8, 9.

② If it is in scientific notations (ie $k \times 10^n$), then the significant digits are all digits explicitly in k.

eg. 6×10^{-4} The significant digit is 6.

Note:- Significant digits are counted from left to Right starting with the left most non zero digits.

Scanned with CamScanner

Errors in Computation:-

1. Machine Error:- Computers have a finite word length - usually a 32 bit or a 64 bit word length. Since, most of the numbers can not be represented exactly in say, 32 bits, error are automatically introduced when these numbers are used in computation.
   The accuracy of a computer is called machine epsilon.

2. Inherent Error:- These are errors which are present in problems itself. Such errors arise due to wrong formulation of the problem or unsuitable soln procedure or inaccuracies in the data etc.

3. Round-off Errors:- Real numbers contain an infinite number of digits. In scientific and engineering computation, we express a real number x in the floating point form as

$$x = \pm 0 \cdot d_1 d_2 \cdots d_n \cdots \times 10^m$$

where $d_1 d_2 \cdots d_n \cdots$ are natural numbers b/w 0 and 9, m is a +ve or -ve integer and is called the exponent. Each digit $d_1 d_2 \cdots d_n \cdots$ except the leading zeros, is called a significant digit. Since we can not retain infinite number of digits in a number, we round-off the number to, say n significant digits.

To round-off a number to n significant digits, we discard all the digits to the right of the n^{th} digit if $(n+1)^{th}$ digit is less than 5.
- If the $(n+1)^{th}$ digit is more than 5, then we increase the n^{th} digit by 1.

Scanned with CamScanner

- If the $(n+1)^{th}$ digit is equal to 5, then we increase the n^{th} digit by 1 if it is odd. and leave the n^{th} digit unchanged if it is even.

eg:

$1.6583 = 0.16583 \times 10^1$

$= 0.1658 \times 10^1$ (Rounded to 4 significant digits)

$= 0.166 \times 10^1$ (3)

$= 0.17 \times 10^1$ (2)

$= 0.2 \times 10^1$ (1)

* NOTE: Let x be an exact number, and x^* be its approximation.

If $|x - x^*| \leq 0.5 \times 10^{-k}$ or

$|x - x^*| \leq 5 \times 10^{-(k+1)}$

Then x^* represents x accurate to k significant digits.

Now, we define:

Absolute Error $= |x - x^*|$.

Relative Error $= \frac{|x - x^*|}{|x|}$ or $\frac{|x - x^*|}{|x^*|}$.

Percentage error $= \frac{|x - x^*|}{|x|} \times 100$

eg. $x = \pi = 3.14159265$ is app. by $\pi = \frac{22}{7} = x^*$

Then

$|x - x^*| = |\pi - \frac{22}{7}| = 0.00126 < 5 \times 10^{-(2+1)}$

$\therefore$ Approximation is accurate to 2 decimal places (digits).

Scanned with CamScanner

4. Truncation Error:- Truncation error arises due to the use of approximate formulas which are generally obtained by truncating an infinite series, i.e. is taking only a finite number of terms in an infinite series.

We consider the Taylor series with a remainder to study the truncation error. We-

$$f(x) = f(x_0) + (x-x_0)f'(x_0) + \cdots + \frac{(x-x_0)^{m+1}}{(m+1)!} f^{m+1}(x_0) + \frac{(x-x_0)^m}{m!} f^m(x_0) + \cdots \quad \text{---(1)}$$

Taylor series expansion of $f(x)$ about $x = x_0$.

$x_0 \in [a,b]$.

Now if we retain first m terms, we get-

$$f(x) = P_{m+1}(x) = f(x_0) + (x-x_0)f'(x_0) + \cdots + \frac{(x-x_0)^{m+1}}{(m+1)!} f^{m+1}(x_0) +$$

The Neglected Part of the series

Principal Part $\boxed{\frac{(x-x_0)^m}{m!} f^m(x_0)} + \frac{(x-x_0)^{m+1}}{(m+1)!} f^{m+1}(x_0) + \cdots$

also forms an infinite series. The first term in this series is called the Principal Part of the truncation error or simply the truncation error (T.E.).

$$\therefore \text{T.E.} = \frac{(x-x_0)^m}{m!} f^m(\xi) \qquad x_0 < \xi < x.$$

Since ξ is unknown function of x, we obtain a bound on T.E. as

$$\boxed{|\text{T.E.}| \leq \frac{1}{m!} \max_{[a,b]} |x-x_0|^m \cdot M_m}$$

$$\boxed{M_m = \max_{[a,b]} |f^m(x)|}$$

Scanned with CamScanner

Example 1. Find the significant digits in the following:

(i) 9353 (iii) 53.07 (iii) 0.0460

Solution: (i) 9353 The significant digits are 9, 3, 5, 3.

(ii) 53.07 The significant digits are 5, 3, 0, 7.

(iii) 0.0460 The significant digits are 4, 6, 0.

Example 2. Round off the following numbers to three significant digits.

(i) 8.894 (ii) 23.865 (iii) 9.4356 (iv) 5.8254

Solution: (i) 8.894 becomes 8.89

(ii) 23.865 becomes 23.9

(iii) 9.4356 becomes 9.44

(iv) 5.8254 becomes 5.82

Example 3. Round off the numbers 665250 and 27.46235 to four significant dig and compute Ea, Er, Ep in each case.

Solution: (i) 665250 is rounded off to four significant digits = 665200

Here $Y = 665250$ and $Y' = 665200$

$$Ea = |Y - Y'| = |665250 - 665200| = 50$$

$$Er = \left|\frac{Y - Y'}{Y}\right| = \frac{50}{665250} = 7.52 \times 10^{-5}$$

Scanned with CamScanner

and $Ep = 100\,Er = 100 \times 7.52 \times 10^{-5}$
$= 7.52 \times 10^{-3}$

(ii) 27.46235 is rounded off to four significant digits = 27.46.

In this case $Y = 27.46235$ and $Y' = 27.46$.

Therefore, $Ea = |Y - Y'| = |27.46235 - 27.46|$
$= 0.00235$

$$Er = \left|\frac{Y - Y'}{Y}\right| = \frac{0.00235}{27.46235} = 8.56 \times 10^{-5}$$

$Ep = 100\,Er = 100 \times 8.56 \times 10^{-5} = 8.56 \times 10^{-3}$

Example 4. If 0.333 is the approximate value of $\frac{1}{3}$, find absolute, relative and percentage error.

Solution: Here $Y = \frac{1}{3} = 0.333333$ and $Y' = 0.333$

$Ea = |Y - Y'| = |0.333333 - 0.333| = 0.000333.$

$$Er = \left|\frac{Y - Y'}{Y}\right| = \frac{0.000333}{0.333333} = 0.000999.$$

$Ep = 100\,Er = 100 \times 0.000999 = 0.0999\%.$

Example 5. Evaluate the sum $S = \sqrt{5} + \sqrt{7} + \sqrt{8}$ to 4 significant digits and find its absolute and relative error.

Solution: As $\sqrt{5} = 2.236$ (After rounding off)

$\sqrt{7} = 2.646$

$\sqrt{8} = 2.828$

$S = \sqrt{5} + \sqrt{7} + \sqrt{8} = 2.236 + 2.646 + 2.828$
$= 7.710$

But the true value of $\sqrt{5} + \sqrt{7} + \sqrt{8} = 7.71024641331057$

Therefore, $Ea = |7.71024641331057 - 7.710| = 0.00024641331057$

And $$Er = \frac{Ea}{S} = \frac{0.00024641331057}{7.710} = 0.000031960$$
$= 3.196 \times 10^{-5}$

Example 6. Suppose 1.732 is used as an approximation to $\sqrt{3}$ then find the absolute and relative error.

Solution: Here Y (True value) = 1.732050807 and $Y' = 1.732$

Therefore, $Ea = |Y - Y'| = |1.732050807 - 1.732| = 0.000050807$

and $$Er = \left|\frac{Y - Y'}{Y}\right| = \frac{0.000050807}{1.732050807} = 0.000029333$$
$= 2.933 \times 10^{-5}$

Scanned with CamScanner