Exact and Approximate Numbers:

An approximate number is a number slightly differ from the exact number X and it is used in place of the letter in calculations.

For eg.

Exact number	Approximate values
$\sqrt{2}$	1.41 or 1.414 etc
π	3.14 or 3.141 or 3.14159 etc
4/3	1.33333333……

Significant Digits

The digits that are used to express a number are called significant digit or significant figure. A significant digit of an approximate number is any non-zero digit in its decimal representation or any zero lying between significant digits or used as place holder to indicate a retained place. The digit 1, 2, 3, 4, 5, 6, 7, 8, 9 are significant digit '0' is also a significant figure except when it is used to fix the decimal point or to fill the place of unknown or discarded digit.

For eg.

(i) The numbers 7845, 3.589, 0.4758 contain four significant digits

(ii) The numbers 0.00386, 0.000587, 0.0000296 contain three significant digits since zeros only help to fix the position of the decimal point.

(iii) The number 45000, 7300 also have two significant digits.

Two notational conventions:-

The significant figure of a number in positional notation consist of

(a) all non-zero digits.

(b) zero digits which (i) lie between significant digits.

(ii) lie to the right of decimal point and at the same time to the right of non-zero digit.

(iii) are specifically indicate to be significant –The significant in a number written in scientific notation $\left(M\times10^{n}\right)$ consists of all the significant digits explicitly in M. significant counted from left to right starting with left most non-zero digits.

Numbers	**Significant Figures**	**No. of significant digits**
3.1416	3,1,4,1,6	5
0.66667	6,6,6,6,7	5
4.0687	4,0,6,8,7	5
5090	5,0,9	3
7.00	7,0,0	3
0.000620	6,2,0	3
5.2×10^{4}	5,2	2
3.506×10	3,5,0,6	4
8×10^{-3}	8	1
4.6300×10^{4}	4,6,3,0,0	5

Errors (Types of errors)

Sources of errors may be identified as:

1) Inherent errors
2) Round off errors
3) Truncation errors
4) Generated errors
5) Propagation errors
6) Absolute errors
7) Relative errors
8) Percentage errors

1) **Inherent errors:** Error which are already present in the statement of a problems before its solution are called inherent errors. Such errors arise due to the given data before approximate or due to the limitations of mathematical tables, calculate or the digital computer. Inherent error can be minimised by taking better data.

2) **Round off errors:** These errors are arise due to cutting off extra digits and retaining as many as desired. Number are rounded off according to following rules:
   To round off a number of 'n' significant digits, discard all the digits to the right of the nth digit and if this discarded number is
   a) less than half a unit in the nth place leave the nth digit unaltered.
   b) greater than half a unit in the nth place, increase the nth digit by unity.
   c) exactly half a unit in the nth place, increase the nth digit by unity if it is odd otherwise leave it unchanged. The number thus rounded off is said to be correct to 'n' significant figures.
3) **Propagation error:** The error resulting due to an arithmetic operation on two numbers which are rounded off to the desired significant figure is called propagation error.
4) **Truncation errors:** while evaluating the function in series expansion form, the term after a finite number of terms are omitted, resulting the function called truncation error or chopping off error.

   for example:- if $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} \ldots\ldots\ldots. = \mathrm{X}$ is replaced by

   $$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} = \mathrm{X}', \text{ then truncation error is } \mathrm{X} - \mathrm{X}'.$$

5) **Generated errors:** Computer after performing a arithmetic operation on any two numbers stores in the result in memory with finite digits resulting in error called generated error.
6) **Absolute error:** Absolute is the numerical difference between the exact value of a quantity and its approximate value. If X is the exact value of a quantity and X' is its approximate value, then $\mathrm{E_a} = |\mathrm{X} - X'|$ is called absolute error.
7) **Relative error:** The relative error is the ratio of the absolute error of the number to the absolute value of the corresponding exact number X,

   i.e. $$\mathrm{E_r} = \frac{\mathrm{E_a}}{|\mathrm{X}|} \quad \text{or} \quad \mathrm{E_r} = \frac{|\mathrm{X} - X'|}{|\mathrm{X}|} \quad \text{or} \quad \mathrm{E_r} = \left| \frac{\mathrm{X} - X'}{X} \right|$$

8) **Percentage error:** If X is the exact value of a quantity and x is its approximate value, then $E_p = \left|\frac{X - X'}{X}\right| \times 100$ or $E_p = E_r \times 100$ is called percentage error.

Important rules –

Rule1. If the number X is correct to n decimal places, then error = $\frac{1}{2} \times 10^{-n}$

Rule2. If the number is correct to n significant digit, then maximum relative error $\le \frac{1}{2} \times 10^{-n}$

Rule3. If X' is the approximate value of X, the absolute error after truncating to k digits from n is

$$|X - X'| < 10^{n-k}$$

Rule4. If X' is the approximate value of X, the absolute error after rounding off k digits from n

$$|X - X'| < \frac{1}{2} 10^{n-k}$$

Rule5. If X' is the approximate value of X, then relative error after truncating to k digits

$$\left|\frac{X - X'}{X}\right| < 10^{-k+1}$$

Rule6. If X' is the approximate value of X, then relative error after rounding off to k digit then

$$\left|\frac{X - X'}{X}\right| < \frac{1}{2} \times 10^{-k+1}$$

Rule7. If the number is correct to d decimal places, then absolute error $\le \frac{1}{2} \times 10^{-d}$

Rule 8 If a number is correct to 'n' significant figure and the first significant digit of the number is k then the relative error

$$E_r < \frac{1}{k \times 10^{n-1}}$$

Example 1: Round off the following number to three significant figures 7.893, 12.865, 6.4356, 3.567, 84767, 5.8250.

Sol: - The numbers given below are rounded off to three significant figures:
7.893 to 7.89, 12.865 to 12.9, 6.4356 to 6.44, 3.567 to 3.57, 84767 to 84800, 5.8254 to 5.82.

Example 2: Round off the following number to four significant figures 6.284359, 9.864651, 12.464762, 1.6583, 30.0567, 0.859378, 3.14159.

Sol: - The numbers given below are rounded off to four significant figures:
6.284359 to 6.284, 9.864651 to 9.865, 12.464762 to 12.46, 1.6583 to 1.658, 30.0567 to 30.06, 0.859378 to 0.8594, 3.14159 to 3.142.

Example 3: Round off the number 865250 to four significant figures and compute absolute error E_a, relative error E_r, Percentage error E_p.

Sol: - Let exact number X = 865250

and round off the number 865250 to four significant figures = 865200

Let approximate number $X' = 865200$

Then absolute error = $E_a = |X - X'| = |865250 - 865200| = |50| = 50$

∴ Absolute error= $E_a = 50$

Relative error= $E_r = \left|\frac{X-X'}{X}\right| = \frac{|X-X'|}{|X|} = \frac{50}{865250} = 0.0000671$

$= 6.71 \times 10^{-5}$

∴ Relative error= $E_r = 6.71 \times 10^{-5}$

Percentage error= $E_p = \left|\frac{X-X'}{X}\right| \times 100 = 6.71 \times 10^{-5} \times 100 = 6.71 \times 10^{-3}$

∴ Percentage error= $E_p = 6.71 \times 10^{-3}$

Example 4: Round off the number 37.46235 to four significant figures and compute absolute error E_a, relative error E_r, Percentage error E_p.

Sol: - Let exact number X = 37.46235

and round off the number 37.46235 to four significant figures = 37.46000

Let approximate number $X' = 37.46000$

Then absolute error = $E_a = |X - X'| = |37.46235 - 37.46000| = 0.00235$

∴ Absolute error= $E_a = 0.00235$

Relative error= $E_r = \left|\frac{X-X'}{X}\right| = \frac{|X-X'|}{|X|} = \frac{0.00235}{37.46235} = 0.0000627$

$= 6.27 \times 10^{-5}$

∴ Relative error= $E_r = 6.27 \times 10^{-5}$

Percentage error= $E_p = \left|\frac{X-X'}{X}\right| \times 100 = 6.27 \times 10^{-5} \times 100 = 6.27 \times 10^{-3}$

∴ Percentage error= $E_p = 6.27 \times 10^{-3}$

Example 5: Find the absolute error if the number X = 0.00545828 is

(i) Truncated to three decimal digits

(ii) Rounded off to three decimal digits

Sol: - Given the exact number X = 0.00545828 = 0.545828×10^{-2}

and after truncating the 0.00545828 to three decimal places = 0.545×10^{-2}

Therefore approximate number $X' = 0.545 \times 10^{-2}$

Then absolute error = $E_a = |X - X'| = |0.545828 \times 10^{-2} - 0.545 \times 10^{-2}|$

$= |(0.545828 - 0.545) \times 10^{-2}|$

$= |0.00828 \times 10^{-2}|$

$= 0.00828 \times 10^{-2}$

$= 0.828 \times 10^{-5} < 10^{-2-3} = 10^{-5}$

∴ Absolute error= $E_a = 0.828 \times 10^{-5}$

Given the exact number X = $0.00545828 = 0.545828 \times 10^{-2}$

and after rounding off the 0.545828×10^{-2} to three decimal places = 0.546×10^{-2}

Therefore approximate number $X' = 0.546 \times 10^{-2}$

Then absolute error = $E_a = |X - X'| = |0.545828 \times 10^{-2} - 0.546 \times 10^{-2}|$

$=|(0.545828 - 0.546) \times 10^{-2}|$

$= |-0.000172 \times 10^{-2}|$

$= 0.000172 \times 10^{-2}$

$= 0.172 \times 10^{-5} < 0.5 \times 10^{-5}$

∴ Absolute error= $E_a = 0.172 \times 10^{-5}$

Example 6: Find the relative error if the number X = 0.004997 is

(i) Truncated to three decimal digits

(ii) Rounded off to three decimal digits

Sol: - Given the exact number X = $0.004997 = 0.4997 \times 10^{-2}$

and after truncating the 0.4997×10^{-2} to three decimal places = 0.499×10^{-2}

Therefore approximate number $X' = 0.499 \times 10^{-2}$

Relative error = $E_r = \left|\frac{X-X'}{X}\right| = \left|\frac{0.4997\times10^{-2}-0.499\times10^{-2}}{0.4997\times10^{-2}}\right|$

$= \left|\frac{(0.4997-0.499)\times10^{-2}}{0.4997\times10^{-2}}\right|$

$= \left|\frac{0.0007}{0.4997}\right|$

$=|0.0014|$

$= 0.14 \times 10^{-2}$

$= 0.140 \times 10^{-2} < 10^{1-3} = 10^{-2}$

∴ Relative error = $E_r = 0.140 \times 10^{-2}$

Given the exact number X = $0.004997 = 0.4997 \times 10^{-2}$

and after rounding off the 0.4997×10^{-2} to three decimal places = 0.5×10^{-2}

Therefore approximate number $X' = 0.499 \times 10^{-2}$

Relative error = $E_r = \left|\frac{X-X'}{X}\right| = \left|\frac{0.4997\times10^{-2}-0.5000\times10^{-2}}{0.4997\times10^{-2}}\right|$

$= \left|\frac{(-0.0003)\times10^{-2}}{0.4997\times10^{-2}}\right|$

$= \left|-\frac{0.0003}{0.4997}\right|$

$= |-0.0006|$

$= 0.0006$

$= 0.06 \times 10^{-4}$

$= 0.06 \times 10^{-3+1} < 0.5 \times 10^{-3+1}$

$\therefore$ Relative error $= E_r = 0.06 \times 10^{-4}$

Limiting absolute error:

The limiting error of an approximate number denoted by Δx is any number not less than the absolute error of that number.

From the definition, we have $E_a = |X - X'| \leq \Delta x$

Therefore X lies within the range $X' - \Delta x \leq X \leq X' + \Delta x$

Thus we can write $X = X' \pm \Delta x$

Then ΔX is an upper limit on the magnitude of the absolute error and is said to be measure absolute accuracy.

Similarly the quantity $\frac{\Delta x}{|X|} \approx \frac{\Delta x}{|X'|}$ measures the relative accuracy.

Limiting Relative error:

The limiting relative error δx of a given approximate number X' is any number not less than the relative error of that number.

From the definition, we have $E_r \leq \delta x$

i.e. $\frac{E_r}{|X|} \leq \delta x$

$\Rightarrow$ $E_r \leq |X| \, \delta x$

In practical situations $X \approx X'$. Therefore we may use $\Delta x \leq |X'| \, \delta x$

If Δx denotes the limiting absolute error of X', then $E_r = \frac{E_a}{X} \leq \frac{\Delta x}{X' - \Delta x}$

Thus we can write $\delta x = \frac{\Delta x}{X' - \Delta x}$, for the limiting error of the number X'.

Propagation Error:

Propagation error in an operation arithmetic operation occurs due to approximate values of the numbers taken by computer with only finite digits

(i) **Propagation of error in addition operation:** Let X_A and Y_A be the approximate values of two numbers where exact values are X and Y. Let e_X and e_Y be the errors in X and Y respectively.

$\therefore$ $e_X = X - X_A$ & $e_Y = Y - Y_A$

Let Z denote the sum of X and Y . i.e. $Z = X + Y$.

Then Z_A is the approximate value of Z is given by $Z_A = X_A + Y_A$.

The error in Z is $e_Z = Z - Z_A$

Now $Z = X + Y$

$\Rightarrow$ $Z_A + e_Z = (X_A + e_X) + (Y_A + e_Y)$

$\Rightarrow$ $Z_A + e_Z = (X_A + Y_A) + (e_X + e_Y)$

$$\Rightarrow \quad (X_A + Y_A) + e_Z = (X_A + Y_A) + (e_X + e_Y)$$

$$\Rightarrow \quad e_Z = e_X + e_Y$$

Therefore the error in the sum of two numbers is equal to the sum of their errors.

Note: If e_X and e_Y of opposite signs, then the resultant error e_Z is reduced.

(ii) **Propagation of error in subtraction operation:** Let X_A and Y_A be the approximate values of two numbers where exact values are X and Y. Let e_X and e_Y be the errors in X and Y respectively.

$$\therefore \quad e_X = X - X_A \quad \& \quad e_Y = Y - Y_A$$

Let Z denote the difference of X and Y . i.e. $Z = X - Y$.

Then Z_A is the approximate value of Z is given by $Z_A = X_A - Y_A$.

The error in Z is $\quad e_Z = Z - Z_A$

Now $\quad Z = X - Y$

$$\Rightarrow \quad Z_A + e_Z = (X_A + e_X) - (Y_A + e_Y)$$

$$\Rightarrow \quad Z_A + e_Z = (X_A - Y_A) + (e_X - e_Y)$$

$$\Rightarrow \quad (X_A - Y_A) + e_Z = (X_A - Y_A) + (e_X - e_Y)$$

$$\Rightarrow \quad e_Z = e_X - e_Y$$

$$\Rightarrow \quad |e_Z| = |e_X - e_Y|$$

$$\Rightarrow \quad |e_Z| \le |e_X| + |e_Y|$$

The absolute error of a difference of two numbers is less than or equal to the sum of their absolute errors.

(iii) **Propagation of error in multiplication operation:** Let X_A and Y_A be the approximate values of two numbers where exact values are X and Y. Let e_X and e_Y be the errors in X and Y respectively.

$$\therefore \quad e_X = X - X_A \quad \& \quad e_Y = Y - Y_A$$

Let Z denote the product of X and Y . i.e. $Z = X.Y$.

Then Z_A is the approximate value of Z is given by $Z_A = X_A . Y_A$.

The error in Z is $e_Z = Z - Z_A$

Now $\quad Z_A = X_A . Y_A$

$$\Rightarrow \quad Z - e_Z = (X - e_X).(Y - e_Y)$$

$$\Rightarrow \quad Z - e_Z = XY - (Y e_X + X e_Y) + e_X e_Y$$

Now neglecting the product $e_X e_Y$ which is very small being the product of very small errors e_X and e_Y, we have

$$e_Z \approx Y e_X + X e_Y$$

Dividing both sides by $Z\ (= X.Y)$, we get

$$\frac{e_Z}{Z} \approx \frac{Y e_X}{Z} + \frac{X e_Y}{Z}$$

$$\Rightarrow \quad \frac{e_Z}{Z} \approx \frac{Y e_X}{X.Y} + \frac{X e_Y}{X.Y}$$

$$\Rightarrow \quad \frac{e_Z}{Z} \approx \frac{e_X}{X} + \frac{e_Y}{Y}$$

Where $\frac{e_Z}{Z}, \frac{e_X}{X}, \frac{e_Y}{Y}$ are the relative errors in Z, X and Y respectively?

Therefore the relative error in the product of two numbers is approximately equal to the sum of the relative error of the factors of the product.

(iv) **Propagation of error in division operation:** Let X_A and Y_A be the approximate values of two numbers where exact values are X and Y. Let e_X and e_Y be the errors in X and Y respectively.

$$\therefore \quad e_X = X - X_A \quad \& \quad e_Y = Y - Y_A$$

Let Z denote the quotient of X and Y . i.e. $Z = \frac{X}{Y}$.

Then Z_A is the approximate value of Z is given by $Z_A = \frac{X_A}{Y_A}$.

The error in Z is $\quad e_Z = Z - Z_A$

$$\Rightarrow \quad e_Z = \frac{X}{Y} - \frac{X_A}{Y_A}$$

$$\Rightarrow \quad e_Z = \frac{X}{Y} - \frac{X - e_X}{Y - e_Y}$$

$$\Rightarrow \quad e_Z = \frac{X}{Y} - \frac{X\left(1 - \frac{e_X}{X}\right)}{Y\left(1 - \frac{e_Y}{Y}\right)}$$

$$\Rightarrow \quad e_Z = \frac{X}{Y} - \frac{X}{Y}\left(1 - \frac{e_X}{X}\right)\left(1 - \frac{e_Y}{Y}\right)^{-1}$$

$$\Rightarrow \quad e_Z \approx \frac{X}{Y} - \frac{X}{Y}\left(1 - \frac{e_X}{X}\right)\left(1 + \frac{e_Y}{Y}\right)$$

Since $\left(1-\frac{e_Y}{Y}\right)^{-1} \approx 1+\frac{e_Y}{Y}$ by the binomial series and neglecting the higher powers of small relative error term $\frac{e_Y}{Y}$.

$$\therefore \qquad e_Z \approx \frac{e_X}{Y} - \frac{Xe_Y}{Y^2} + \frac{e_X e_Y}{Y^2}$$

Neglecting the last term $\frac{e_X e_Y}{Y^2}$ which is very small, being the product of very small errors e_X and e_Y, we have

$$e_Z \approx \frac{e_X}{Y} - \frac{Xe_Y}{Y^2}$$

Dividing both sides by $Z\left(=\frac{X}{Y}\right)$, we get

$$\frac{e_Z}{Z} \approx \frac{e_X}{ZY} - \frac{X e_Y}{ZY^2}$$

$$\frac{e_Z}{Z} \approx \frac{e_X}{\frac{X}{Y}Y} - \frac{X e_Y}{\frac{X}{Y}Y^2}$$

$$\frac{e_Z}{Z} \approx \frac{e_X}{X} - \frac{e_Y}{Y}$$

Hence the relative error in the quotient of two numbers is the difference of their relative errors.

(v) **Propagation of error in power operation:** Let $Z = X^p$. where p is rational number

Let X_A and Z_A be the approximate values of two numbers where exact values are X and Z. Let e_X and e_Z be the errors in X and Z respectively.

$$\therefore \qquad e_X = X - X_A \qquad \& \qquad e_Z = Z - Z_A$$

Now $\qquad Z_A = X_A^p$.

$$\Rightarrow \qquad Z - e_Z = (X - e_X)^p$$

$$\Rightarrow \qquad Z - e_Z = \left[X\left(1-\frac{e_X}{X}\right)\right]^p$$

$$\Rightarrow \qquad Z - e_Z = X^p\left[1 - p\left(\frac{e_X}{X}\right) + \ldots\ldots\right]$$

$$\Rightarrow \qquad Z - e_Z = X^p - pX^{p-1}e_X + \ldots\ldots$$

Neglecting the terms involving the second and higher powers of the small relative error $\left(\frac{e_X}{X}\right)$, we get $Z_A \approx X^p - pX^{p-1} e_X$

or $Z_A \approx Z - pX^{p-1} e_X \qquad (\because Z = X^p)$

or $Z - Z_A \approx pX^{p-1} e_X$

or $e_Z \approx pX^{p-1} e_X$

Dividing both sides by $Z(= X^p)$, we get

$$\frac{e_Z}{Z} \approx \frac{pX^{p-1} e_X}{Z}$$

or $$\frac{e_Z}{Z} \approx \frac{pX^{p-1} e_X}{X^p}$$

or $$\frac{e_Z}{Z} \approx pX^{-1} e_X$$

or $$\frac{e_Z}{Z} \approx p\left(\frac{e_X}{X}\right)$$

Thus the relative error in the p^{th} power of a number is equal to p- times the relative error in the number.

Example 1.4. *Find the absolute error and relative error in $\sqrt{6} + \sqrt{7} + \sqrt{8}$ correct to 4 significant digits.*

Sol. We have $\sqrt{6} = 2.449$, $\sqrt{7} = 2.646$, $\sqrt{8} = 2.828$

$\therefore \qquad S = \sqrt{6} + \sqrt{7} + \sqrt{8} = 7.923.$

Then the absolute error E_a in S, is

$$E_a = 0.0005 + 0.0007 + 0.0004 = 0.0016$$

This shows that S is correct to 3 significant digits only. Therefore, we take $S = 7.92$

Then the relative error E_r is

$$E_r = \frac{0.0016}{7.92} = 0.0002.$$

Example 1.5. *The area of cross-section of a rod is desired upto 0.2% error. How accurately should the diameter be measured?* *(Pune, B. Tech., 2003)*

Sol. If A is the area and D is the diameter of the rod, then $A = \pi\left(\frac{D}{2}\right)^2 = \frac{\pi}{4} D \,.\, D$.

Now error in area A is 0.2% *i.e.*, 0.002 which is due to the error in the product $D \times D$.

We know that if E_a is the absolute error in the product of two numbers X and Y, then

$$E_a = X_{aY}E + YE_{aX}$$

Here $X = Y = D$ and $E_{aX} = E_{aY} = E_D$, therefore

$$E_a = DE_D + DE_D \quad \text{or} \quad 0.002 = 2DE_D$$

Thus $E_D = 0.001/D$ *i.e.*, the error in the diameter should not exceed $0.001\ D^{-1}$.

■ **Example 1.6.** *Find the product of the numbers 3.7 and 52.378 both of which are correct to given significant digits.*

Sol. Since the absolute error is greatest in 3.7, therefore we round off the other number to 3 significant figures *i.e.* 52.4.

$\therefore$ Their product $P = 3.7 \times 52.4 = 193.88 = 1.9388 \times 10^2$

Since the first number contains only two significant figures, therefore retaining only two significant figures in the product, we get

$$P = 1.9 \times 10^2$$

Generated error – If errors occurs after an arithmetic operation due to finite digit arithmetic then that is called generated error for the next operation such error will result in propagated error

Machine Epsilon—Let the true number be expressed in the general form as

$$\text{True } x = \left(f_x + g_x \times 10^{-d}\right) \times 10^{E}$$

$$= f_x \times 10^{E} + g_x \times 10^{E-d}$$

$$= \text{approximate } x + \text{error}$$

Therefore chopping error = $g_x \times 10^{E-d}$; $0 \le g_x < 1$. where g_x represent the truncation part of the number in normalized form, d is the number of digits permitted in mantissa and E is the exponent. The absolute error due to chopping is

$$e_r = \left|\frac{g_x \times 10^{E-d}}{f_x \times 10^{E}}\right|$$

Obviously, relative error is maximum when g_x is the maximum and f_x is minimum. Maximum value of g_x is less than 1 and minimum value of f_x is 0.1. Therefore the absolute value of relative error will satisfy following relation

$$e_r \le \left|\frac{g_x \times 10^{E-d}}{f_x \times 10^{E}}\right| = 10^{-d+1}$$

$$e_r \le 10^{-d+1}$$

The maximum relative error given above is called Machine Epsilon. The name 'machine' indicate that this value is machine dependent, because of length of mantissa 'd' is machine dependent.

Therefore for a decimal machine, this uses chopping machine epsilon $\in = 10^{-d+1}$.

Similarly for a machine which uses round off $e_r \le \left|\frac{0.5 \times 10^{E-d}}{0.1 \times 10^{E}}\right| = 0.5 \times 10^{-d+1}$

Therefore machine epsilon $\in = 0.5 \times 10^{-d+1}$; for rounding off

Note 1:- Machine epsilon represents upper bound off error due to floating point representation. It also suggest that data can be represented in machine with ‘d’ significant decimal digits and relative error does not depend upon the size of number.