Floating Point Arithmetic And Errors

Floating Point Representations

- There are two formats to represent a number., one is floating point representation and the other is fixed point representation.
- The transformation of fixed point data into floating point data is known as normalization. This is done to preserve maximum number of useful information carrying digits of numbers. This transformation leads to calculation errors.

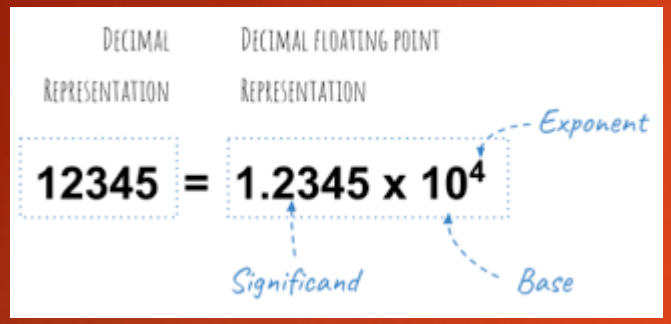
Fixed point representation : In fixed point representation, numbers are represented by fixed number of decimal places. Examples : 500.638, 4.8967 32.09

Floating point representation : In floating point representation, numbers have a fixed number of significant places.

Examples : $6.236 * 10^{3}, 1.306 * 10^{-3}$

- A floating point number has 3 parts :

1. Mantissa/significand
2. Base
3. Exponent



In scientific notation, such as 1.23 x 102 the significand is always a number greater than or equal to 1 and less than 10.

In the standard normalized floating-point numbers, the significand is greater than or equal to 0.1, and is always less than 1.

If the exponent is too small but not zero, the condition is called an underflow. If the exponent is too large and if it cannot be accommodated then, the condition is called an overflow.

Or

If the result of an arithmetic operation gives a number smaller than .1000 E^{-99} then it is called an underflow condition. Similarly, any result greater than .9999 E^{99} leads to an overflow condition.

Floating Point Arithmetic

- Floating point arithmetic is not associative. This means that in general, for floating point numbers x, y, and z:

(x+ y) + z ≠ x + (y +)z

(x. y) z ≠ x (y. z)

- Floating point arithmetic is also not distributive. This means that in general,

x . (y+z) + ≠ (x.y)+ (x .z)

So, now let us see floating point operations :

Steps for Addition and Subtraction :

- Make sure the numbers are normalized.
- Make exponents the same.
- Add mantissas together
- Normalize the result if needed.

Multiplication:

Two numbers are multiplied in the normalized floating point mode by multiplying the mantissas and adding the exponents. After the multiplication of the mantissas, the resulting mantissa is normalized as in an addition or subtraction operation, and the exponent is appropriately adjusted.

Division :

The mantissa of the numerator is divided by that of the denominator. The denominator exponent is subtracted from the numerator exponent. The quotient mantissa is normalized to make the most significant digit non-zero and the exponent is appropriately adjusted. The mantissa of the result is chopped down to 4 digits.

Significant digits

- Non-zero digits are always significant.

98 has two significant digits, and 35.9 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).

- Accuracy and precision are closely related to significant digits.
- 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} . 4.3201 has a precision of 10^{-4}.

Loss of significant digits :

- It happens when one gets too few significant digits in the subtraction of two numbers that are very close to each other.
- For eg.

1.2345678 -

1.2344444

0.0001234

Now in the above answer, there are only four significant digits(namely 1,2,3,4) . We lost 4 significant digits.

This is called 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).
- To avoid this loss of significant digits in algebraic expressions, we must rationalize these numbers.

Errors

- Error is defined as the difference between the actual value and the approximate value obtained from numerical computation.

  Suppose x is actual value and xa is approximate value,

  then Error = x-xa

Generation of errors :

- Every operation has 2 parts : operand and operator. Approximation in either of the two contributes to errors. Approximation to operands causes propagated errors and approximations to operators causes generated errors.

- Sources of error :

Different sources of error are :

- Data input errors
- Error in algorithm and
- Error during computation

- Types of error :

- Round off error
- Truncation error

Round off error : It is also known as rounding errors. It is due to the fact that floating point numbers are represented by finite precision.

- Error and accuracy are inter-related. Less the error, more the accuracy.
- Errors used for determination of accuracy are :

1. Absolute error (E_a)
2. Relative error(E_r)
3. Percentage error(E_p)

- Sources of error :

Different sources of error are :

- Data input errors
- Error in algorithm and
- Error during computation

Data input errors : 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.

Error in algorithm : Such errors occurs where infinite algorithms are used. Exact results are expected only after an infinite number of steps. As this cannot be done in practice, the algorithm has to be stopped after a finite number of steps and the results are not exact.

Error during computation: Such errors occurs when elementary operations such as multiplication and division are used the case when number of digits increases greatly so that the results cannot be held fully in computer register.

Absolute error :

- It is defined as the magnitude of the difference between the actual value (x) and the approximated value(xa).

Absolute error= | x-xa |

Relative error :

- It is defined as the ratio of absolute error and the actual value.

Relative error = | x-xa | /x

Percentage error :

- 100er=100* | x-xa | /x

Rules for rounding of numbers :

A number is rounded-off to n places after decimal by seeing (n+1) th place digit d_{n+1}, as follows:

If $d_{n+1} < 5$, then it is chopped

If $d_{n+1} > 5$, then $d_n = d_n + 1$

If $d_{n+1} = 5$, and d_n is odd then $d_n = d_n + 1$ else the number d_{n+1} is chopped.

The difference between the number x and fl(x) is the round off error.

Round-off error decreases when precision increases.

Truncation error

- It is defined as an error created by approximating a mathematical operation. It is a consequence of doing finite number of steps in a calculation that would require infinite number of steps to do exactly.
- Example is : evaluation of infinite sum
- Consider the Maclaurin Series:

$e^x=1+x+x^2/2!+x^3/3!+\ldots\ldots\ldots$

Suppose if we have to find the value of e^x when x=0.5,

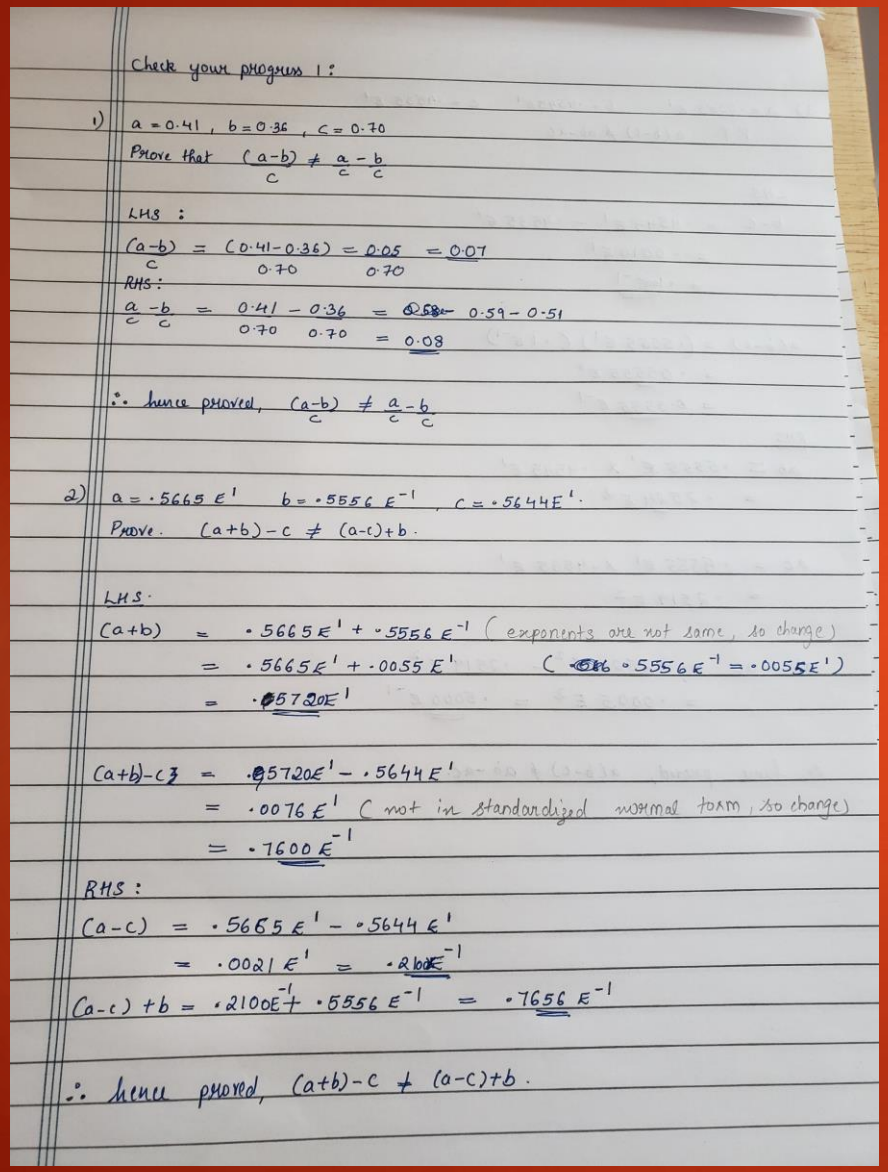
then,

$e^{0.5}=1+0.5+(0.5)^2/2!+(0.5)^3/3!+\ldots\ldots\ldots$

Suppose if we are using only first 3 terms to find the value of $e^{0.5,}$ then whatever is left over is truncation error.

- Another example for an operation that is affected by truncation error is numerical integration.

Check your progress 1



Check your progress 1:

1) $a = 0.41$, $b = 0.36$, $c = 0.70$

Prove that $\frac{(a-b)}{c} \neq \frac{a}{c} - \frac{b}{c}$

LHS :

$\frac{(a-b)}{c} = \frac{(0.41-0.36)}{0.70} = \frac{0.05}{0.70} = 0.07$

RHS :

$\frac{a}{c} - \frac{b}{c} = \frac{0.41}{0.70} - \frac{0.36}{0.70} = 0.59 - 0.51 = 0.08$

∴ hence proved, $\frac{(a-b)}{c} \neq \frac{a}{c} - \frac{b}{c}$

2) $a = .5665\,E^{1}$ $b = .5556\,E^{-1}$, $c = .5644E^{1}$.

Prove. $(a+b) - c \neq (a-c)+b$.

LHS.

$(a+b) = .5665E^{1} + .5556\,E^{-1}$ (exponents are not same, so change)

$= .5665E^{1} + .0055\,E^{1}$ ($.5556\,E^{-1} = .0055E^{1}$)

$= .5720E^{1}$

$(a+b)-c = .5720E^{1} - .5644E^{1}$

$= .0076\,E^{1}$ (not in standardized normal form, so change)

$= .7600\,E^{-1}$

RHS :

$(a-c) = .5665\,E^{1} - .5644\,E^{1}$

$= .0021\,E^{1} = .2100E^{-1}$

$(a-c)+b = .2100E^{-1} + .5556\,E^{-1} = .7656\,E^{-1}$

∴ hence proved, $(a+b)-c \neq (a-c)+b$.

3) $a = .5555\ E^1 \quad b = .4545 E^1 \quad c = .4535\ E^1$

P.T. $a(b-c) \neq ab - ac$

LHS.

$b - c = .4545\ E^1 - .4535\ E^1$

$= .0010\ E^1$

$= .1000 E^{-1}$

$a(b-c) = (.5555\ E^1)(.1\ E^{-1})$

$= .05555\ E^0$

$= 0.5555\ E^{-1}$

RHS

$ab = .5555\ E^1 \times .4545\ E^1$

$= .2524\ E^2$

$ac = .5555\ E^1 \times .4535\ E^1$

$= .2519\ E^2$

$ab - ac = .2524\ E^2 - .2519\ E^2$

$= .0005\ E^2 = .5000\ E^{-1}$

∴ hence proved, $a(b-c) \neq ab - ac$.

Check your progress 2

1) Round off the following numbers to four significant digits.

i. 450.92

We need to round off the number to four significant digits. So, $d_n=9$ and $d_{n+1}=2$

Since, $2<5$, so it is chopped.

So the answer is 450.9

ii. 48.3668

Similarly,

$d_n=6$ and $d_{n+1}=6$

Since, $6>5$, so $d_n=d_n+1$

So, the answer is 48.37

iii. 9.3265

Similarly,

d_n=6 and d_{n+1}=5

Since, d_{n+1}=5 and d_n is even, so it is chopped.

So, the answer is 9.326

iv. 8.4155

Similarly,

d_n=5 and d_{n+1}=5

Since, d_{n+1}=5 and d_n is odd, so d_n=d_n+1.

So, the answer is 8.416

v. 0.80012

d_n=1 and d_{n+1}=2

Since, 2<5 so it is chopped.

So, the answer is 0.8001

vi. 0.042514

Similarly,

d_n= 1and d_{n+1}=4

Since, 4<5 so it is chopped.

So, the answer is 0.04251

vii. 0.0049125

Similarly,

d_n=2 and d_{n+1}=5

Since, d_{n+1}=5 and d_n is even, so it is chopped.

So, the answer is 0.004912

viii. 0.00020215

d_n=1 and d_{n+1}=5

Since, d_{n+1}=5 and d_n is odd, so d_n=d_n+1.

So, the answer is 0.0002022

2) Write the following numbers in floating-point form rounded to four significant digits :

i. 100000

$.1*10^{6}$

ii. -0.002316

$-.2316*10^{-2}$

iii. -35.666

$-.3567*10^{2}$

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.

Let a= 28.483 and b=27.984

Difference= 28.483-27.984=0.499

Number of significant digits are= 3. The significant digits are : 4,9,9.

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?

$= |0.66667-(2/3)| = |(2.00001-2)/3|$

$=3.33333*10^{-6} = .0000033....<1/2(10^{-5})$

Therefore, k=5

5) Find out to how many decimal places the value 355/133 is accurate as an approximation to ?

Let p=3.14159265

Absolute error = 3.14159265 - 2.66917293

= 0.47241972

$0.47241972<1/2(10^{-1})$

Therefor k=1.

Therefore, the approximation is accurate to 1 decimal places or 2 significant digits.

Check your progress 3

1)

2) Round the number x = 2.2554 to three significant figures. Find the absolute error and the relative error.

Approximate value= 2.26

Absolute error = | true value- approximate value |

= | 2.2554-2.26 |

=.0046

Relative error = | absolute error/true value |

=.0046/2.2554

=.00204

Percentage error= .204%

3) If π = 3.14 instead of 22/7, find the relative error and percentage error.

Absolute error= | 22/7-3.14 |

= | (22-21.98)/7 |

=0.02/7=.002857

Relative error=(.002857)/(22/7)

=(.002857/22)*7

=.000909

Percentage error=.0909%

4) Determine the number of correct digits in s = 0.2217, if it has a relative error, $0.2 * 10^{-1}$.

Relative error=$0.2*10^{-1}$

True value=0.2217

Relative error=absolute error/true value

Absolute error= relative error * true value

= $0.2*10^{-1}*0.2217$

=.004434

$.004434<1/2(10^{-2})$

Therefore, number of correct digits is 2.

5) Round-off the number 4.5126 to four significant figures and find the relative percentage error.

True value= 4.5126

Approximate value=4.513

Relative error = absolute error/true value

Absolute error= | true value- approximate value |

= | 4.5126-4.513 |

=.0004

Relative error= .0004/4.5126

=.00008864

Percentage error = .008864%

Exercise 1.6

1) 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

a) 37.21829

i) Round upto 2 decimals : chop upto 2 decimals :

$.37*10^2$ $.37*10^2$

Round upto 4 decimals : chop upto 4 decimals :

$.3722*10^2$ $.3721*10^2$

(b) 0.022718

Round upto 2 decimals :

$.23*10^{-1}$

Round upto 4 decimals :

$.2272*10^{-1}$

chop upto 2 decimals :

$.22*10^{-1}$

chop upto 4 decimals :

$.2271*10^{-1}$

(c) 3000527.11059

Round upto 2 decimals :

$.30*10^{7}$

Round upto 4 decimals :

$.3000*10^{7}$

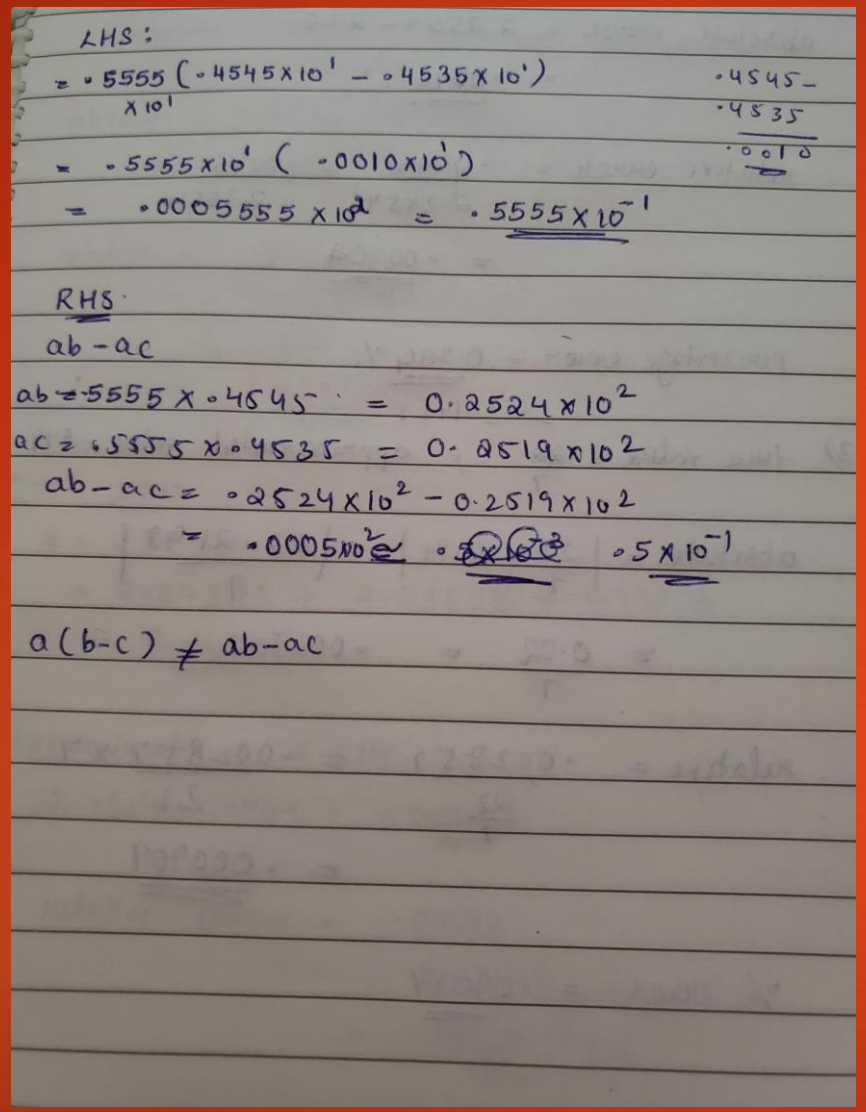
Chop upto 2 decimal :

$.30*10^7$

Chop upto 4 decimal:

$.3000*10^7$

2) Show that $a(b - c) \neq ab - ac$, where, $a = .5555 \times 10^1$, $b = .4545 \times 10^1$, $c = .4535 \times 10^1$.



3)

4) 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?

Let approximations of x and y be x* and y* respectively.

x*=0.37214

y*=0.37202

x*-y*=0.37214-0.37202

=0.00012

So, true value=x-y=0.3721448693-0.3720214371

=0.0001234322

Approximate value=x*-y*

=0.00012

Therefore, absolute error=true value-approximate value

=0.0001234322-0.00012

=0.0000034322

Relative error=absolute error/true value

=0.0000034322/0.0001234322

=0.027806=$3*10^{-2}$

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 $|x-x^*| \leq i/210^{-k}$, where k is a positive integer.

Example :

Let x* = .568 approximate to x = .5675

x – x* = –.0005

$|x-x^*| = 0.0005 = 1/2(.001) = (1/2)*10^{-3}$

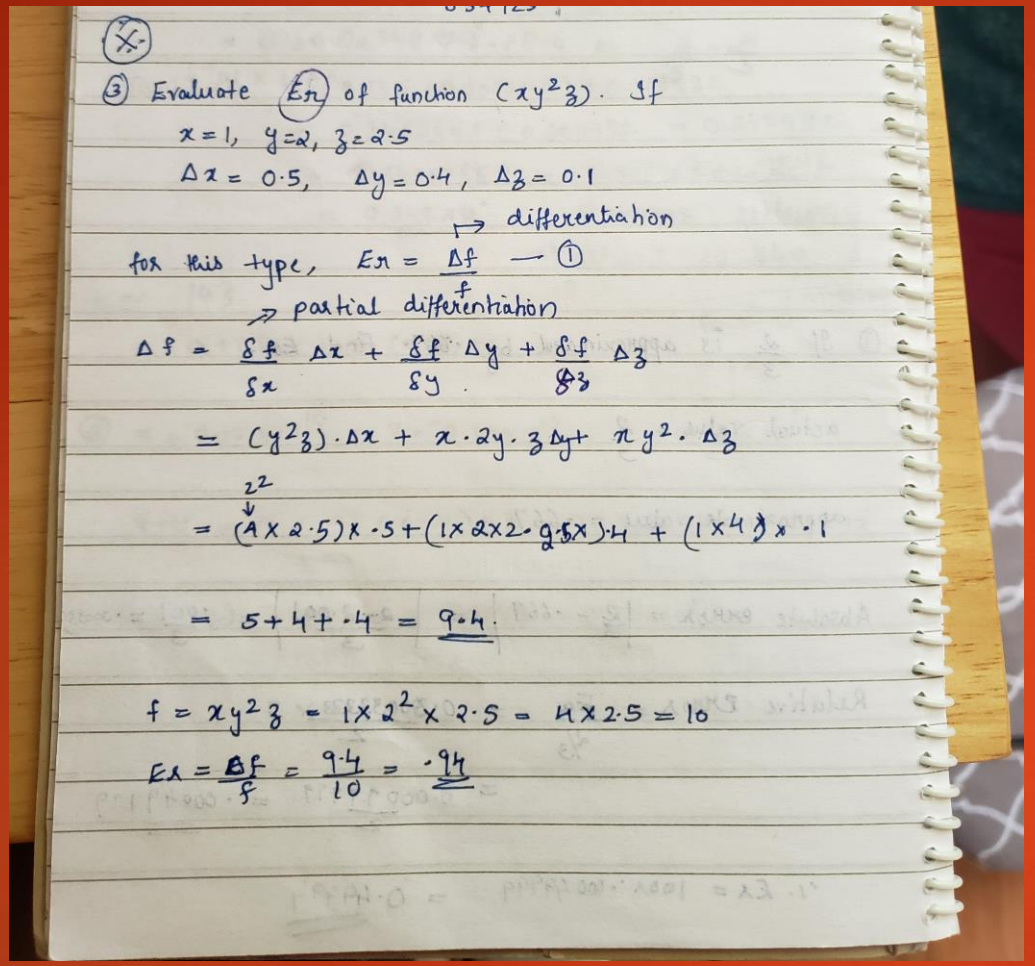
So x* approximates x correct to 3 decimal place.

Try yourself questions

1. Let $a=0.2222*10^2$,$b=0.1001*10^3$,$c=0.1002*10^3$. Prove that $a(b+c)=a*b+a*c$.
2. If 2/3 is approximated by 0.667. Find E_r.
3. Find E_r of 834.123 correct to 5 significant figures.
4. Evaluate E_r of function xy^2z, if $x=1$, $y=2$, $z=2.5$ and $\Delta x=.5$, $\Delta y=.4$, $\Delta z=.1$.

Answers

1. Try it yourself
2. 0.00049995
3. 0.000003596591 87
4. .94



(3) Evaluate Er of function (xy^2z). If

$x = 1,\ y = 2,\ z = 2.5$

$\Delta x = 0.5,\quad \Delta y = 0.4,\quad \Delta z = 0.1$

for this type, $Er = \frac{\Delta f}{f}$ — (1)

Δf → differentiation

f → partial differentiation

$$\Delta f = \frac{\delta f}{\delta x}\Delta x + \frac{\delta f}{\delta y}\Delta y + \frac{\delta f}{\delta z}\Delta z$$

$$= (y^2z)\cdot\Delta x + x\cdot 2y\cdot z\,\Delta y + xy^2\cdot\Delta z$$

$$= (2^2 \times 2.5)\times .5 + (1\times 2\times 2.5\times 2.5)\times .4 + (1\times 4)\times .1$$

$$= 5 + 4 + .4 = 9.4$$

$$f = xy^2z = 1\times 2^2\times 2.5 = 4\times 2.5 = 10$$

$$Er = \frac{\Delta f}{f} = \frac{9.4}{10} = .94$$