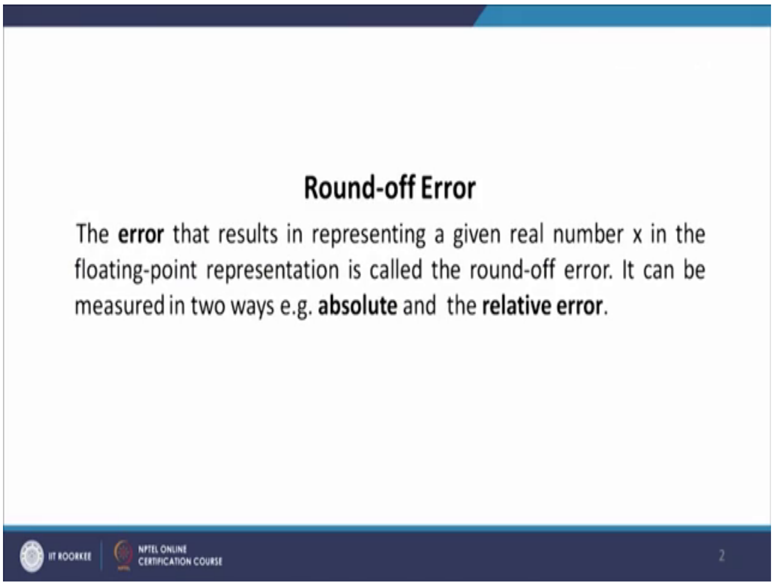
Numerical Linear Algebra
Dr. P. N. Agrawal
Department of Mathematics
Indian Institute of Technology, Roorkee

Lecture - 22
Round-off error

Hello friends, welcome to my lecture on round off errors. Now the error that results in representing a given real number x in the floating-point representation is called round off error.

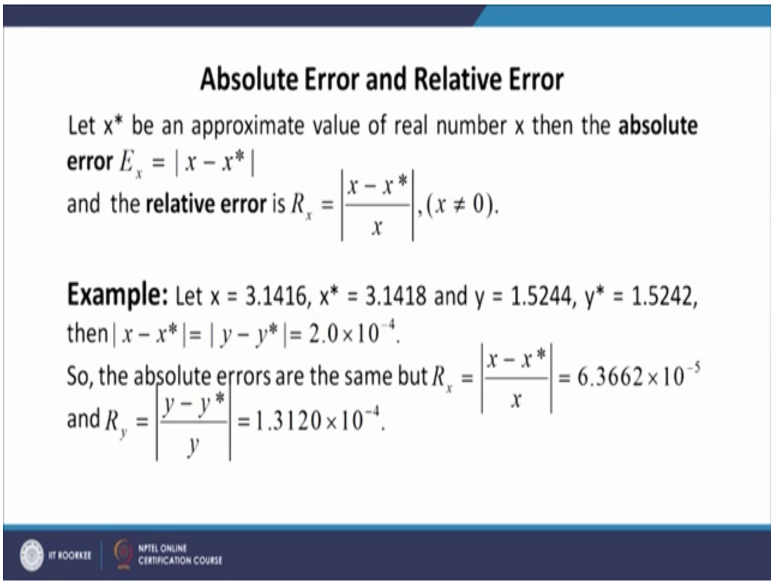
(Refer Slide Time: 00:28)



We have seen that when you represent floating point representation you in a machine, there are 2 types of base to do that. One is that chopping and the other one is rounding. So, both of them caused error which is called as the round off errors ok.

So, it now this error can be measured in 2 ways one is that absolute error, and the other one is relative error.

(Refer Slide Time: 00:59)

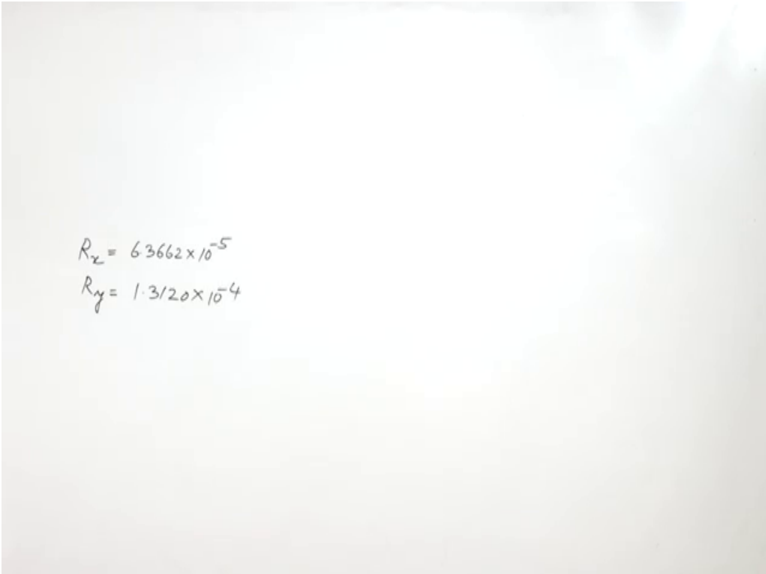


In the case let us now define what do we mean by absolute error. So, let x star be an approximate value of a real number x, then the absolute error x is equal to mod of x minus x star. And the relative error is defined as Rx equal to mod of x minus x star over x assuming that of course, x is not equal to 0.

For example, let us consider, x is equal to 3.1416 and x star is equal to 3.1418, y equal to 1.5244, and y star equal to 1.5242. Then mod of x minus x star and mod of y minus y star both are same, and they are equal to 2.0 into 10 to the power minus 4. So, the absolute error in both the cases is same it is 2.0 into 10 to the power minus 4.

Now, but if you calculate the relative error Rx, in the case of x, then x minus x star over x mod of that is come comes out to be 6.3662 into 10 to the power minus 5. While the relative error in y r y comes out to be mod of y minus y star over y is equal to 1.3120 into 10 to the power minus 4.

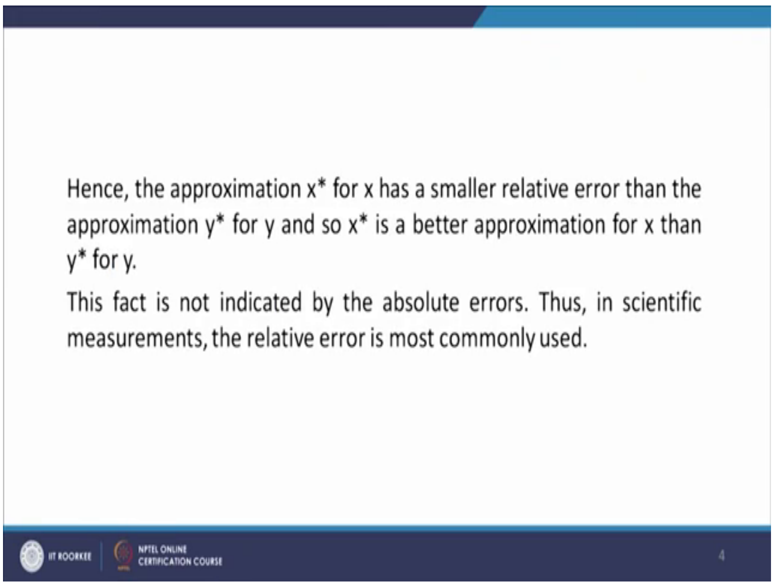
(Refer Slide Time: 02:23)



So, Rx is equal to Rx is equal to 6.366 3, 6.3662 into 10 to the power minus 5. And Ry is equal to 1.3120 into 10 to the power minus 4.

Now, you can see that this is 6.36 6 2 into 10 to the power minus 5, and Ry is 1.3120 into 10 to the power minus 4. So, Rx is smaller than Ry ok. And therefore, we can say that the approximation x star for x has a smaller relative error, then the approximation y star for y.

(Refer Slide Time: 03:16)

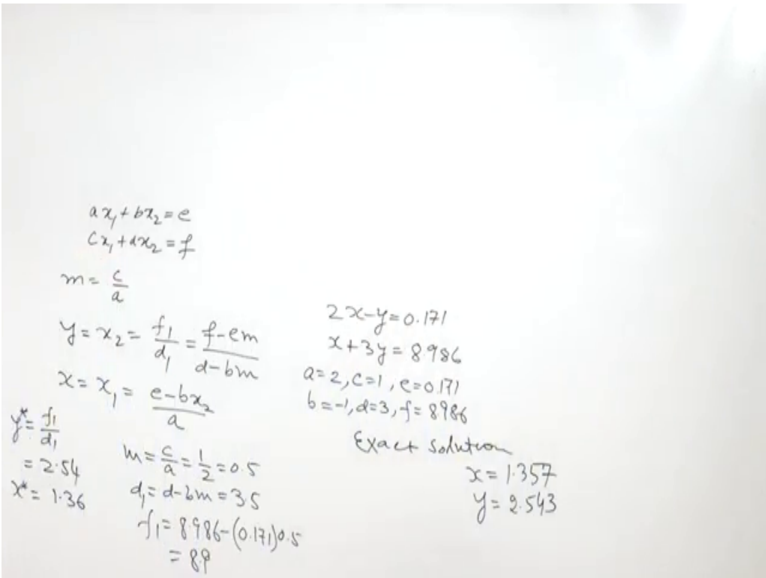


And so, we can say x star gives a better approximation to x the approximation y star y.

Now, this fact is not indicated by the absolute error, you can see because absolute error in both the cases is same mod of x minus x star is equal to mod y minus y star is same in so, the this fact is not indicated by the absolute error, but and therefore, the relative error is most commonly used in scientific measurements.

Now, let us take up the case of system of linear equations, and solve it by Gaussian elimination. We have discussed in our previous lecture, the Gaussian elimination is scheme we consider the equations as ax 1, plus b x 2 equal to e cx 1 plus d x 2 equal to f.

(Refer Slide Time: 04:12)



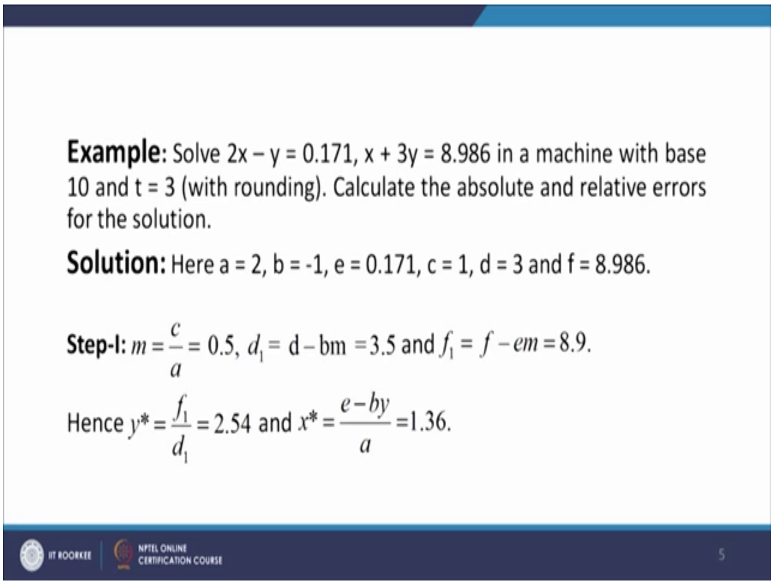
And we had m equal to c by a there. What we did was we multiplied the first equation by c by a assuming a to be not equal to 0 and then subtracting the first equation from the second equation. So, when you multiply the first equation by c by a, the coefficient of x one will become c. So, cx 1 will cancel cx 1 when you subtract after multiplication by c by a from the second equation. And we got x 2 equal to f 1 upon d 1 f 1 by d equal to f minus e m, and d 1 was d minus b m.

So, having found x 2, then one can find x one from the equation 1, by e minus bx 2 divided by a. So, here in this case we have in this example 2 x minus y equal to 0.171, and we have x plus 3 y equal to 8.986.

Now so, when we can calculate m equal to c y so, a is equal to 2 here, c is equal to 1, b is equal to minus 1, and d equal to 3, e equal to 0.171, and f equal to 8.986 ok. So, m equal to c by a when you calculate m equal to c by a here, c by a means 1 by 2. So, m comes out to be 0.5. And d 1 equal to d minus bm, d 1 equal to d minus bm ok. So, put the value of d d is equal to d is equal to 3 b is equal to minus 1, and m equal to point 5, d 1 comes out be 3.5, and we have done f 1 ok.

So, f 1 is equal to f, f is equal to 8.986 minus e is equal to 0.171 into m, m is equal to 0.5. So, this comes out to be 8.9, 8.9 here we are taking base 10 and t equal to 3 t equal to 3 with rounding.

(Refer Slide Time: 07:38)

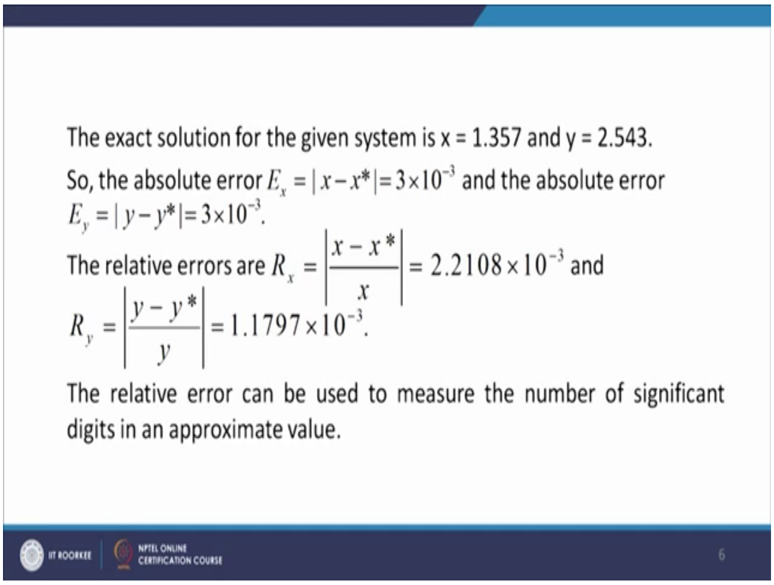


So, I am writing the values with t equal to 3, when you calculate d 1 in the digital computer. And use t equal to 3 with rounding, then you will get d 1 is equal to 3.5 and f 1 equal to 8.9. And so, the value of y star, y here x 1 x 2 in this example are actually x and y. So, this is y and this is x. So, the value of y star y star is the value of obtain from f 1 over d 1. So, I call that as y star.

So, y star equal to f 1 by d 1, and this comes out to be equal to 2.54, 2.54, and then the value of x star we are putting star because their approximate values there we are taking t equal to 3. So, x star comes out to be 1.36 6. So, when you do these computations in the digital computer, with t equal to 3 and rounding. Then using the rounding rule, we will get these values of x star and y star.

Now, the exact solution for the given system is x equal to exact solution is; exact solution is x equal to 1.357, and y equal to 2.543. So, the absolute error will be equal to if you calculate the absolute error ex equal to mod of x minus x star it is 3 into 10 to the power minus 3 and the absolute error in y mod of y minus y star is 3 into 10 the power minus 3.

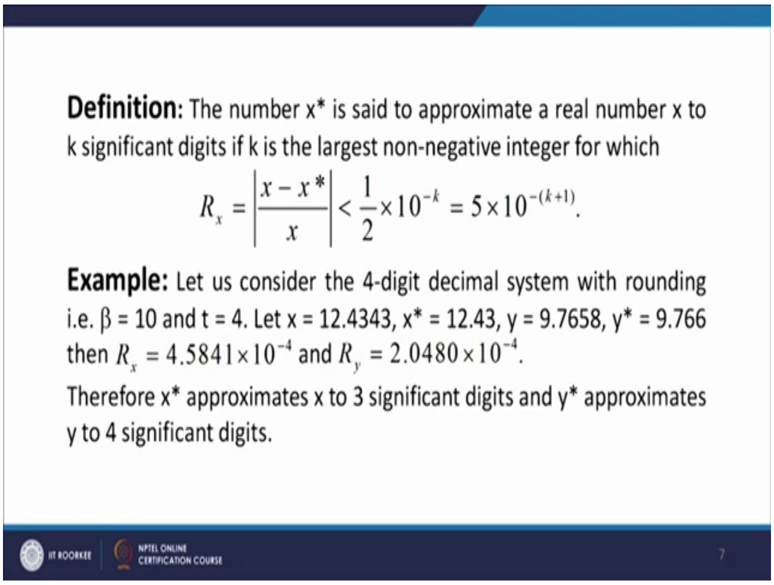
(Refer Slide Time: 09:45)



Again, with t equal to 3, the relative errors come out come out to be relative errors are Rx equal to mod of x minus x star over x which is 2.2108 into 10 to the power minus 3 and Ry equal to mod of y minus y star upon y which is 1.1797 into 10 to the power minus 3.

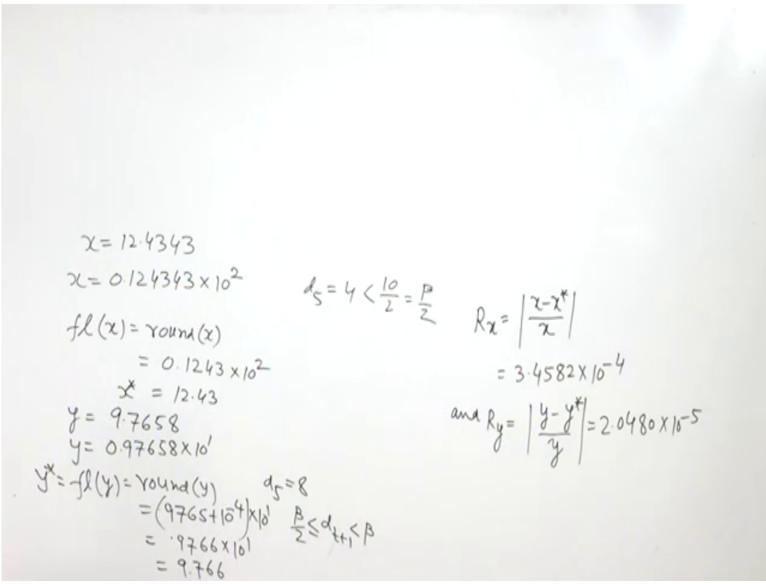
And so, we can say that the relative error, now the relative error can be used to measure the number of significant digits in an approximate value, we are going to see that how a relative error can be used to measure the number of significant digits in an approximate value.

(Refer Slide Time: 10:36)



Let us now consider the 4-digit decimal system with rounding ok, we are taking beta equal to 10, t equal to 4, if you take x equal to 21.4343 ok.

(Refer Slide Time: 10:47)



Then in the floating-point representation x will be written as 0.124343 into 10 to the power 2 ok.

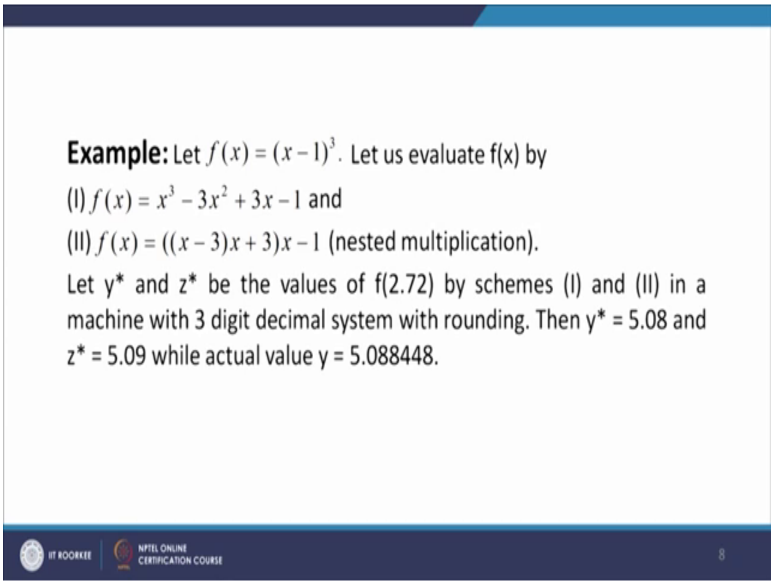
So, here you can see we are taking t equal to 4; that means, we have to written 4 significant digits. So, this is d 1 d 2 d 3 d 4. Now d 5, d 5 is 4 and 4 is less than 10 by 2 that is beta by 2 ok. So, what we will do? We will write it as 0.1243; we will leave it we will write it as 0 0.d1d2 d 3 d 4 ok. And remaining digits we discard into 10 to the power 2. So, this will be 12.43.

Now, if you take y equal to 9.7658, then in the floating-point representation y is equal to ah, 0.97658 into 10 to the power 1 ok. So, f l y equal to round y. Now d 1 is 9, d 2 is 7, d 3 is 6, d 4 is 5, d 5 is 8 ok. D 5 is equal to 8 so, this d 5 lies in the range beta by 2 less than or equal to dt plus 1 less than beta, beta is equal to 10 ok. T is equal to 4 ok. So, d 5 is 8 so, the; so, we will what we will do? We will add 2.9765, we add 10 to the power minus d equal to 4 ok, t is equal to 4. So, we add 10 to the power minus 4. So, this becomes into 10 to the power 1. So, this becomes 0.9766 into 10 to the power 1. So, this is 9.766. So, this is y star, and this is you are x star. So, x star is 12.43 y star is equal to 9.766.

Let us find the relative errors Rx and Ry. So, Rx is equal to mod of x minus x star divide by x, then this is 3.4582 into 10 to the power minus 4. And Ry Ry is equal to 2.0480 into 10 to the power minus 5.

Now, let us look at the definition, from the definition it is clear that in the case of Rx where k is equal 3 while in the case of Ry k is equal to 4. So, x star approximates 2 x to 3 significant digits, and y star approximates to y to 4 significant digits.

(Refer Slide Time: 14:48)

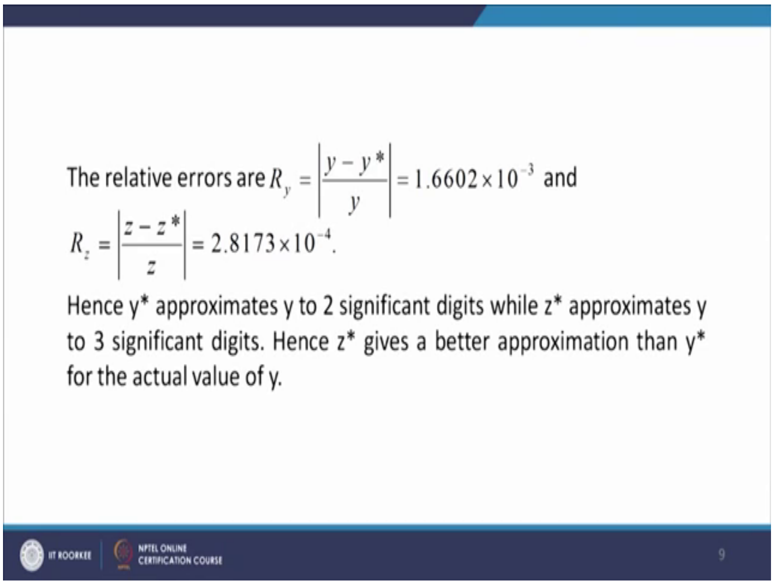


Now let us take the example function evolution, let us say fx equal to x minus 1 whole cube we can evaluate fx by 2 by direct method fx equal to x cube minus 3 x square plus 3 x minus 1, by expending x minus 1 whole cube, and another one is by nested multiplication which we discussed in the previous lecture.

So, when we do the nested multiplication, the x cube minus 3 x square plus 3 x minus 1 will be written as x minus 3 into x plus 3 and then multiplied by x minus 1. Now let us say y star and z star be the values of f 2 points 7 2. So, we are taking x equal to 2.72 here. And evaluate the value of f x, when x is equal to 2.72 by schemes 1 and 2.

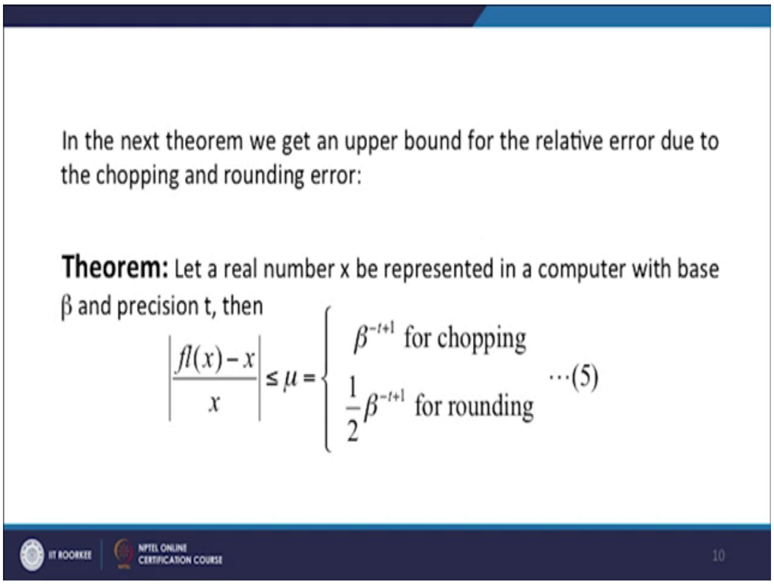
So, when we apply the scheme 1; that is, fx equal to x cube minus 3 x square plus 3 x minus 1, and evaluate f equal to f at 2.72, the value of y star comes out to be 5.08 with 3-digit decimal system with rounding. And the z star comes out to be 5.09 by the nested multiplication. So, actual value is 5.088448, and you can see that z star is closer to the actual value then y star. So, z star actually it is calculated by nested multiplication. So, we see that nested multiplication is a better numerical scheme.

(Refer Slide Time: 16:33)



Now, the let us calculate the relative errors, the relative error in y that is Ry is equal to mod of y minus y star over y, which is 1.6602 into 10 to the power minus 3, and the relative error Rz is mod of y minus z star divided by y, which is equal to 2.8173 into 10 to the power minus 4. So, y star approximates y to 2 significant digits, while z star approximates y to 3 significant digits. And therefore, we can say that z star gives better approximation to y better approximation to y then y star. So, the nested multiplication scheme is better than the direct method.

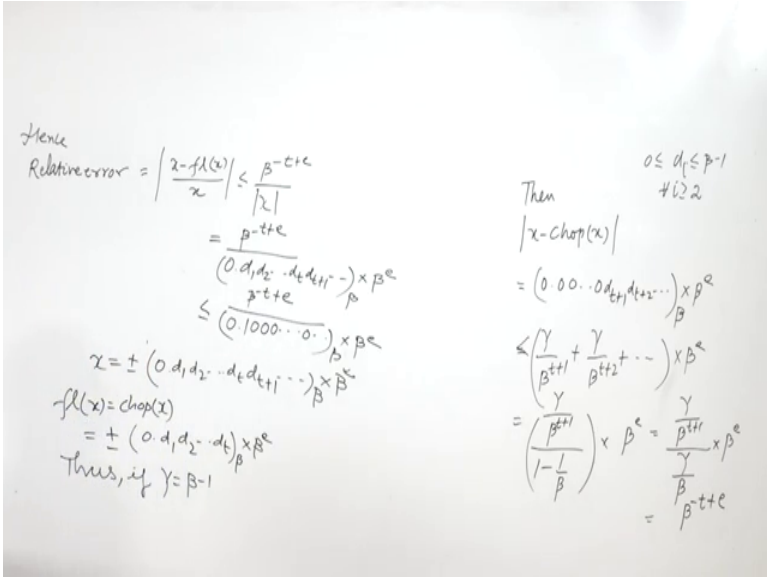
(Refer Slide Time: 17:18)



Now, in the next round we get an upper bound for the relative error due to the chopping and rounding error. We have seen that when the floating-point representation of real number in the detail machine is used. There are 2 kinds of errors, there are 2 kinds of way in which error occurs chopping and rounding.

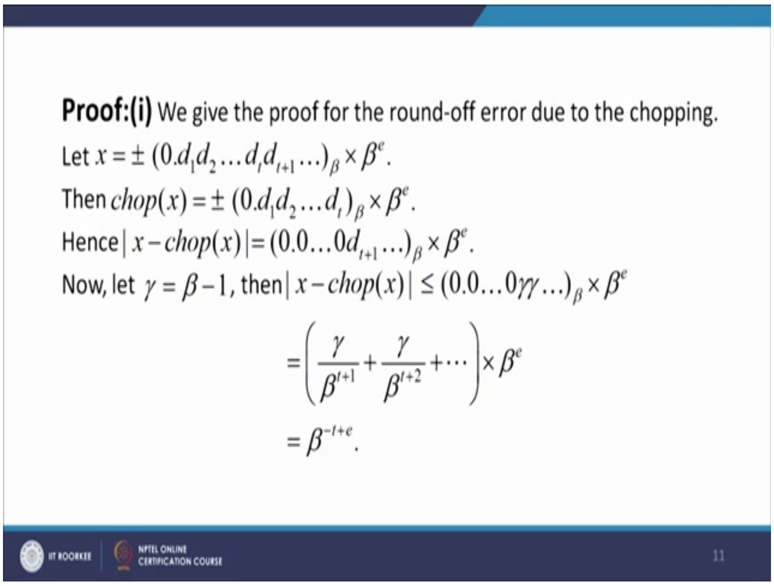
So, let us see what is the upper bound for the relative error, due to the chopping and rounding let us assume that a real number x be represented in a computer with base beta and precision t, then so, precision t means t number of significant digits ok. So, relative error we know, relative error is x minus fl x divided by x.

(Refer Slide Time: 18:09)



Now, this is less than this is defined as less than or equal to beta mu, mu is the upper bound this is in the case of chopping, it is beta to the power minus t plus 1 for chopping. And in the case of rounding, it is half of this value. Now let us move this theorem. So, we give the proof first we give the proof for the round off error due to chopping. So, x is equal to in the floating-point representation x is written like this.

(Refer Slide Time: 19:15)



Now, we are trying to find the upper bound this, upper bound for the chopping case ok. So, flx equal to chop x to we are taking we are taking precision t ok. So, precision t means we are taking t significant digits. So, this means after dt, we discovered all the remaining digits. So, we have plus minus 0.d1d2 and so on dt beta to the power e ok thus, if you define gamma equal to beta minus 1.

If you define gamma equal to beta minus 1, then mod of x minus chop x minus chop x chop x means x minus flx this will be equal to from x we are subtracting chop x. So, and we are taking modulus. So, 0 point now d 1 d 2 dt they are exactly same. So, we will have 0 0 and so on 0, and then we will have dt plus 1 dt plus 2 and so on ok.

This means that first t places are 0s, after that we have dt plus 1 dt plus 2 and so on, the now let us see that in the case of normalised floating-point representation. d 1 is lying between one and beta minus 1, but d 2 d 3 and so on, all dis for I greater than or equal to they are mod than equal to 0, but less than or equal to beta minus 1 ok.

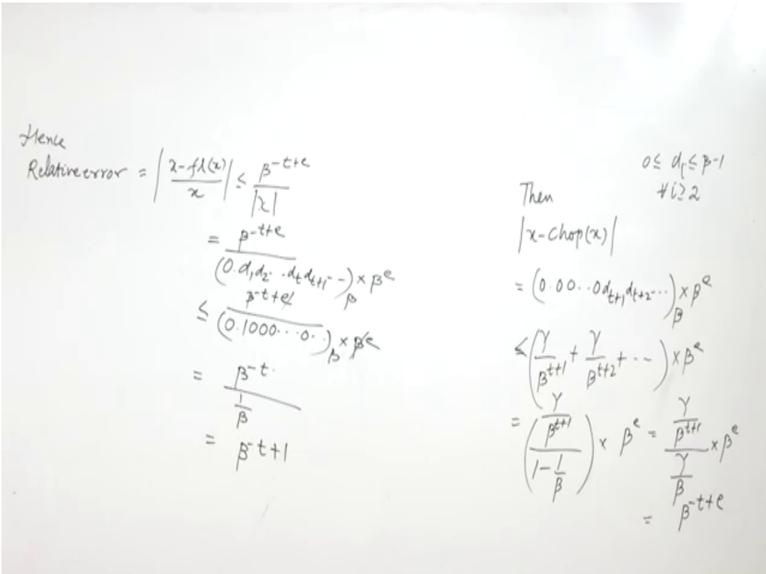
So, d di is lying between 0, and beta minus 1 for all I greater than or equal to 2 ok. So, this means that we shall have this is less than or equal to dt plus 1, dt plus 1 will be less than or equal to beta minus 1 and beta minus 1 we are taking as gamma. So, gamma over beta to the power t plus 1 ok beta is the base and this is t plus 1 position, then gamma over beta to the power t plus 2 and so on.

All this digits dt plus 1 dt plus 2 they are less than or equal to beta minus 1 that is they are less than or equal to gamma into ok. So, this is this beta we do not write; now in to beta to the power e this is infinite series the geometric series, where 1 by beta is the common ratio. So, gamma beta to the power t plus 1 divided by 1 minus 1 by beta we have into beta to the power e and this is equal to. So, this is beta gamma upon beta to the power t plus 1 divided by beta minus 1 upon beta, beta minus 1 is gamma. So, gamma upon beta into beta to the power e ok and this is nothing but beta to the power minus t plus e.

Now, let us find the relative error ok. So, hence relative error equal to mod of x minus flx divided by x ok. Mod of x minus flx is chop x. So, this is less than or equal to beta to the power minus t plus e divided by x ok, mod of x. Mod of x is this quantity. So, this is equal to beta to the power minus t plus e divided by now this is 0.d1d2 dt dt plus 1 and so on, beta into beta to the power e ok.

Now, we have to make it less than or equal to, this means that we put the minimum values of d d 1 d 2 dt dt plus 1 and so on. So, that this quantity is maximised. So, beta to the power minus t plus e divided by now minimum value of d 1 is one because d 1 lies from 1 2, I mean between one and beta minus 1. So, 0.1 and then d 2 d t d 3 d t dt plus 1 can be taken as 0s and so on into beta to the power e. And this gives us this cancels with this beta e e e and what we get beta to the power minus t divided by now this is 1 over beta ok 1 over beta, then 0 over beta is square, then 0 over beta cube and so on.

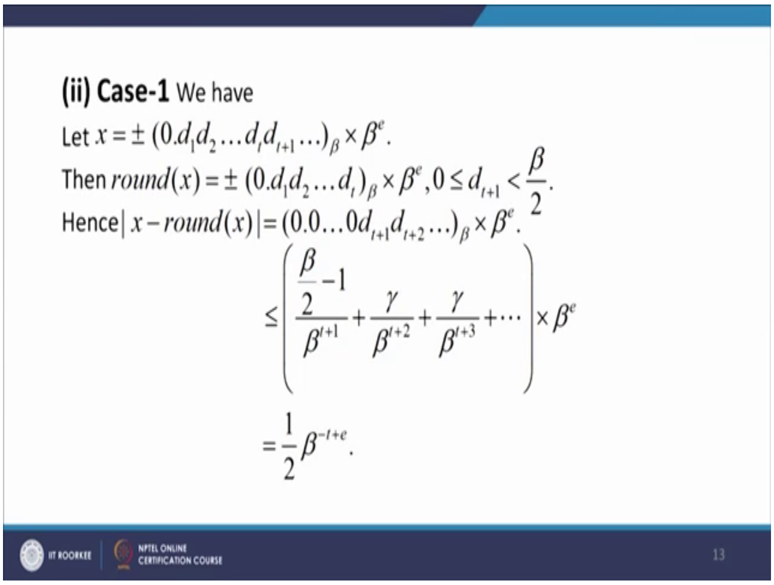
(Refer Slide Time: 25:20)



So, this is equal to beta to the power minus t plus 1. So, this is the upper bound in the case of chopping.

Now, we go to the next case, where we have rounding. So, in the case of rounding, again let us say x is equal to plus minus 0.d1d2 dt dt plus 1, base is beta into beta to the power a, and beta to the power e now first we consider the case, when 0 is less than or equal to dt plus 1 less than beta by 2.

(Refer Slide Time: 26:09)

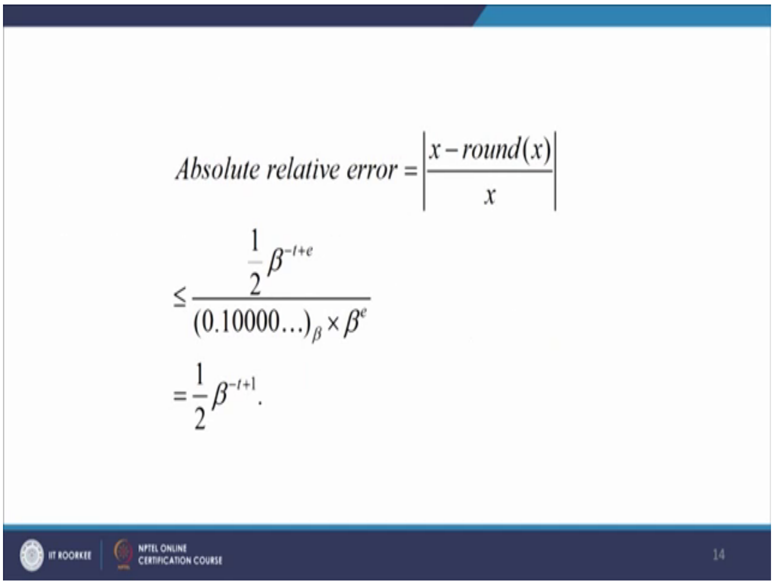


Then the rounding of x gives plus minus 0.d1d2 and so on dt base beta into beta to the power e. And therefore, mod of x minus flx, or we can say round x is equal to 0 point, now first d first t places are same in x and round of x. So, we have 0 0 0 and then we have dt plus 1 dt plus 2 and so on with base beta into beta to the power e.

Now, when we look at the place dt plus 1 dt plus 1 is less than beta by 2. Therefore, dt plus 1 is less than or equal to beta by 2 minus 1. So, dt plus 1 is less than or equal to beta by 2 minus 1 will give you, beta y y 2 minus 1 upon beta to the power t plus 1. And then dt plus 2 dt plus 3 and so on, they are all less than or equal to beta minus 1 the condition is only on dt plus 1.

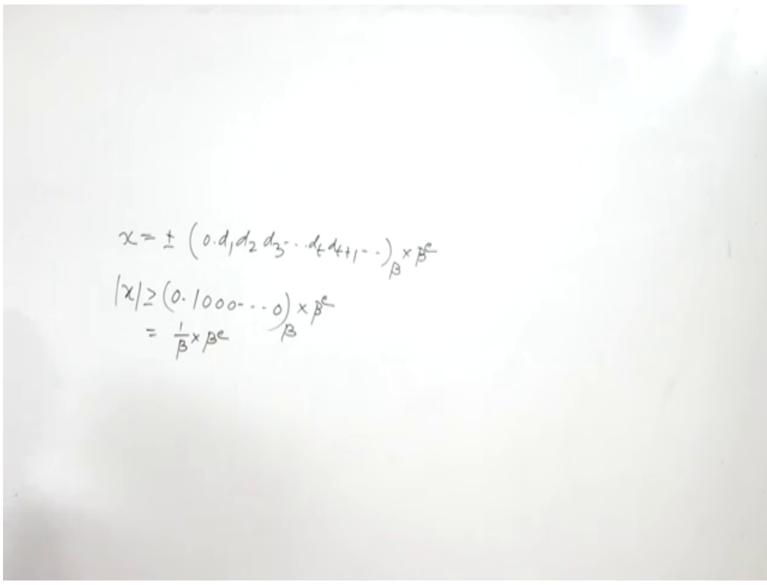
So, we have written them as gamma, because of the each one of them is less than or equal to beta minus 1. So, we are and beta minus 1 we are assuming as gamma. So, gamma over beta to the power t plus 2 plus gamma over beta to the power t plus 3 and so on into beta to the power e, and when you evaluate this infinite series hence where gamma over beta plus 2 is the first term. And 1 by beta is the geometric ratio, then and then you simplify this it turns out that it is 1 by 2 beta to the power minus t plus e.

(Refer Slide Time: 28:58)



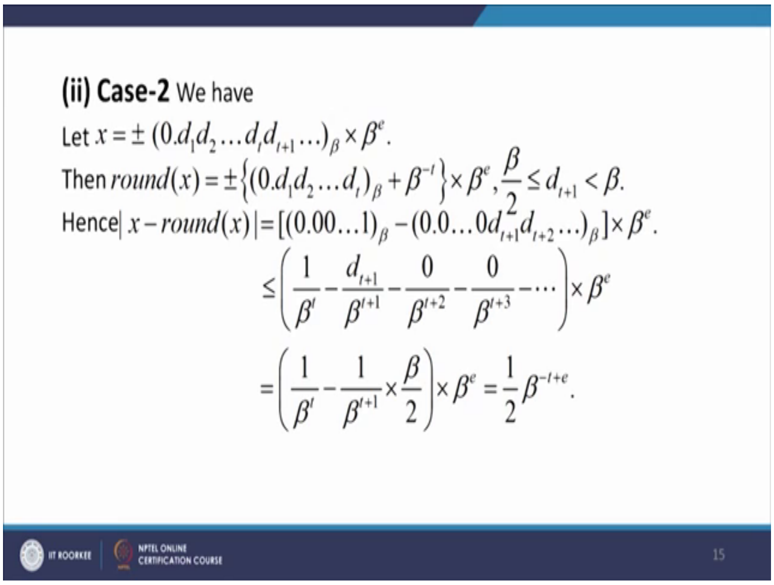
So, absolute relative error in this case will be or you can say relative error in this case will be mod of x minus round of x divided by x. Now so, we have x minus round of x less than or equal to 1 by 2 beta to the power minus t, and in the denominator, we are writing x is equal to plus minus 0-point d 1 d 2 d 3 dt dt plus 1 and so on.

(Refer Slide Time: 28:20)



Beta into beta to the power e beta to the power now, we want make it this quantity less than or equal to less than or equal to means we put the minimum values of d 1 d 2 d 3 and so on. So, mod of x is less than or equal to greater than or equal to in the denominator we put minimum values. So, this is greater than or equal to 0.1 minimum value of d 1 is one and then d 2 d 3 they are all 0s beta into beta to the power e. So, we have this and this is equal to 1 by beta in to beta to the power e. So, beta 2 power e will cancel and we will get 1 by 2 beta to the power minus t plus 1.

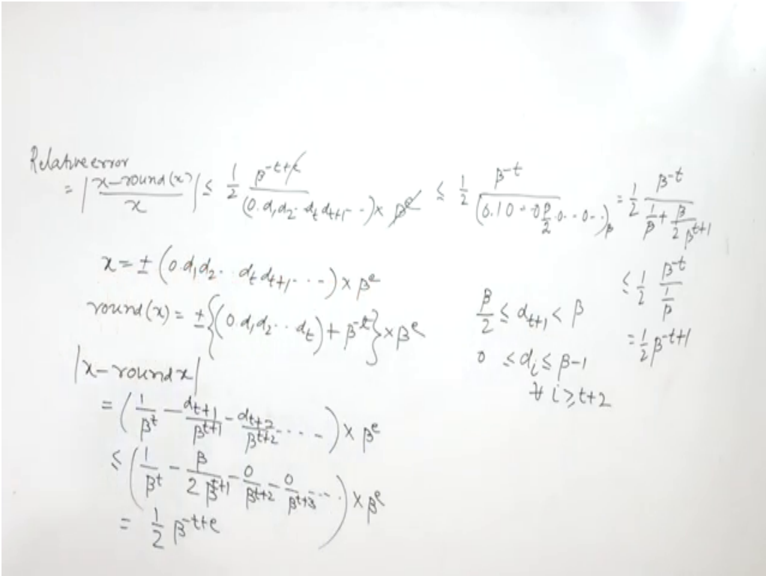
(Refer Slide Time: 29:26)



Now, let us go to the next for case here in the case of rounding ok. So, again x is equal to plus minus 0.d1d2 and so on dt dt plus 1, with base beta into beta to the power e, and when we do the rounding in the second case, in the second case dt plus 1 is greater than or equal to beta by 2, but less than beta. So, then we add beta to the power minus t to the 2 dt, we add 1 by beta to the power t.

So, so, x mod of x minus round x in this case will be what x is equal to plus minus 0.d1d2 dt dt plus 1 and so on into beta to the power e.

(Refer Slide Time: 30:15)



Round x is equal to plus minus 0-point d 1 s 2 dt, plus beta to the power minus t into 10 to beta to the power e. So, 2 dt we are adding 1 by beta to the power t. So, this will become greater so, we will have mod of x minus round x. Numerically it will be more than this quantity numerically will be more than this, because to the tth decimal we are adding 1 by beta to the power t ok.

So, this is equal to now what will happen? d 1 d 2 dt dt dt will cancel, but dt plus 1 by beta to the power t we have. So, we will have this less than or equal to 1 by beta to the power t, and then we have dt plus 1 upon beta to the power t plus 1 and so on. We are we are subtracting them so, this is equal this is this is equal to 1 by beta to the power t minus ok, this is less than this is equal to this, and this is how much? this is less than or equal to 1 by beta to the power t. So, we put now we are these are subtracting. So, this means we have to put in order to make less than or equal to we have to put minimum values of dt plus 1 dt plus 2 and so on ok. And this will give you what? Dt plus 1 yeah dt plus 1 is beta by 2 is less than or equal to dt plus 1 and less than beta ok.

So, we want to maximise this. So, we put minimum values ok, this is for the dt plus and dt di 0 less than or equal di less than or equal to beta minus 1 for all i bigger than or equal to t plus 2 ok. So, dt plus 2 and so on we will put as 0s and dt plus 1 we will put as beta y 2 ok. So, 1 by beta t minus beta by 2 into bt plus 1, minus 0 upon beta t to the power t plus 2 and so on into beta to the power e ok. And this will give you so, this will beta will cancel, and we will get 1 over beta to the power t minus 1 over 2 beta to the power t so, 1 by 2 beta to the power minus t plus e.

Now, we go to the relative error. So, relative error will be mod of x minus round x divided by x ok. And this is less than or equal to 1 by 2 beta to the power minus t plus e, x is equal to mod of x is equal to 0.d1d2 dt dt plus 1 and so on, into beta to the power beta to the power e ok. And what we will get?

Now, we have to put minimum values here of d 1 d 2 d t dt plus 1. So, that we get upper bound. So, 1 by 2 and this beta to the power e will cancel with this, we have beta to the power minus t, in the denominator minimum value of d 1 is 1. So, 0.1 d 2 dt they are all 0's minimum values are 0's dt plus 1 has minimum value beta by 2 ok. Minimum values are beta by 2. So, we have and dt plus 1 they can all be taken as 0. So, we will have ok, this is now this is how much? 1 by 2 beta to the power minus t divided by 1 by beta, plus and this is the t plus 1 th position, beta by 2 into beta to the power t plus 1 ok.

Now, this is further less than or equal to this is a positive quantity, we can drop this and 1 by 2 beta to the power minus t divided by 1 by beta. And we get it as 1 by 2 beta to the power minus t plus 1. So, in the case of dt plus 1 more than or equal to beta by 2 plus less than beta again we get half of beta to the power minus t plus 1 ok.

(Refer Slide Time: 26:09)

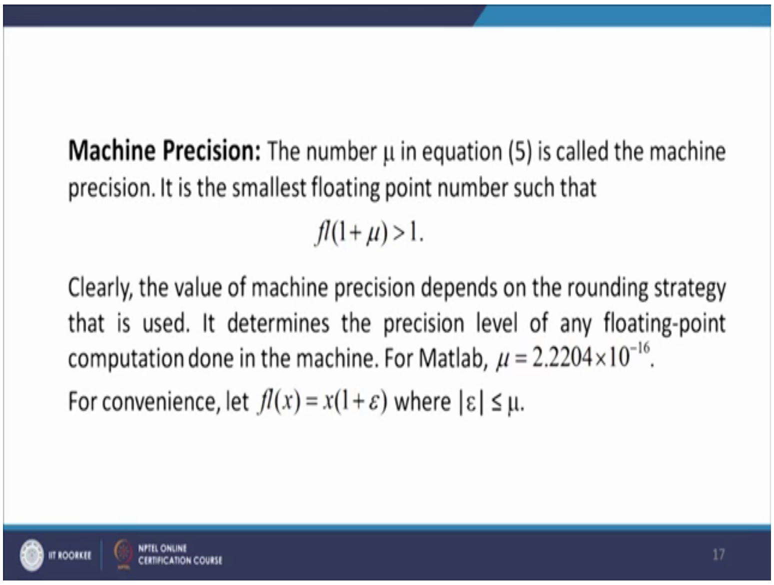
$$Relative\ error = \left|\frac{x - round(x)}{x}\right|$$

$$\leq \frac{\frac{1}{2}\beta^{-t+e}}{(0.100\ldots0d_{t+1}00\ldots)_\beta \times \beta^e}$$

$$= \frac{\frac{1}{2\beta^t}}{\frac{1}{\beta} + \frac{\beta}{2} \times \frac{1}{\beta^{t+1}}} \leq \frac{1}{2}\beta^{-t+1}.$$

Now, we go to a machine precision. The number mu in the equation number 5, this equation number 5 the number mu is called the machine precision.

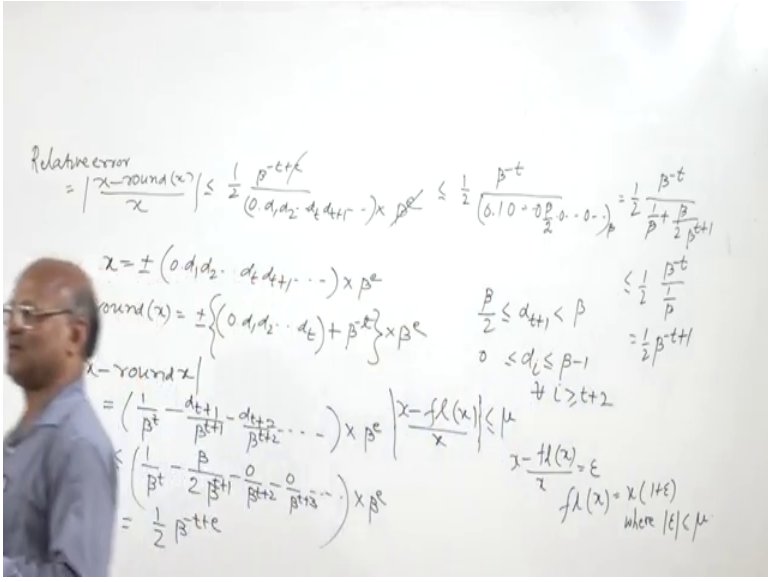
(Refer Slide Time: 36:18)



It is this minus floating-point number such that fl of 1 plus mu is greater than 1. The value of machine precision depends on the rounding strategy that we use. You can see that in the case of chopping it is beta to the power minus t plus 1 while in the case of rounding it is half of that in the case of chopping. So, it determined the precision level of any floating-point computation done in the machine.

In the case of MATLAB this mu is 2.2204 into 10 to the power minus 16. Now when we do the of addition multiplication division in floating point representation we shall be taking flx equal to x into 1 plus epsilon ok.

(Refer Slide Time: 37:26)



See we have we have flx equal to x minus flx mod of x minus flx divided by x this is equal to this is relative error this is less than or equal to mu we have taken ok.

So, what we do is let us take x minus flx divided by x equal to mu x is equal to epsilon, then we shall write flx equal to x times 1 plus epsilon where epsilon is such that mod of epsilon is less than or equal to mu. With that I would like to end my lecture.

Thank you very much.