Environmental Remediation of Contaminated Sites
Prof. Bhanu Prakash Vellanki
Department of Civil Engineering
Indian Institute of Technology – Roorkee

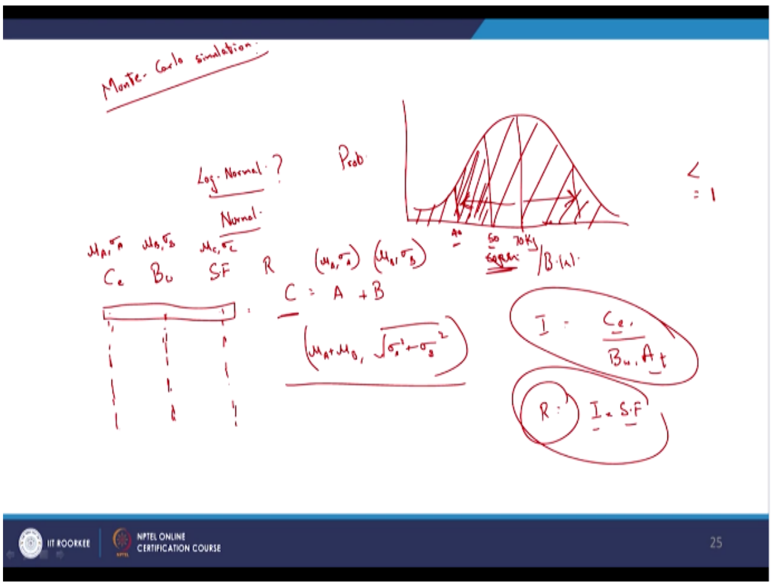
Lecture - 08
Risk Assessment
Stochastic Approach

Hello everyone, again welcome back to the latest lecture session. So again a quick recap of what we have been up to right. We have looked at or we have been looking at anyway a risk assessment right in the context of trying to understand the extent to which you need to remediate a particular contaminated site right and in that context, we looked at two approaches or you know one in greater detail anyway the deterministic approach.

In that case, let us say we estimated various variables with a point estimate. For example, body weight, I think we took it to be 70 kgs right. Though, we do know let us say the exposed concentration, exposed population pardon me will not have or everyone in that particular exposed population will not have only 70 kgs right. So obviously the risks associated will vary from one person to the other.

But you know that particular variation in risk is not going to be captured when you look at the deterministic approach right. So in that context, we talked about treating the variables as variables right and we looked at or try to understand briefly anyway the probability distribution functions right.

(Refer Slide Time: 01:29)



And right we looked at probability distribution functions or how is the probability distributed and typically let us say you know for that particular example let us say or let us say this can be concentration or typically let us say body weight let us say right. So we have the mean to be 70 kgs and then the standard deviation let us say right that would capture I think mu +or- sigma will capture I believe 67% of the data right and so on.

So here let us say if I want to get the concentration of particular or you know what is the probability let us say the body weight will be within a particular range as in 30 to 40 or 40 to 50 how do I get? So if this is the case for 40 and for 50, so what are the chances that the body weight would be between 40 and 50 now? I need to get the area with under this graph between these two limits I guess right.

Again, what would be the probability or the total probability or what would be the total area here let us see under this graph now. It should be <or=1 because the probability can never be >1 obviously right. That is just one aspect that we need to keep in mind right. So in this context, we also talked about log normal distribution let us say right log normal and in that case, what did we, how did we understand this particular distribution log normal.

We call the particular variable to be log normally distributed let us say right. When the natural logarithm of that particular variable let us say is normally distributed, we consider that particular variable to be log normally distributed right and typically in our case or you know in our case of risk assessment, we come across normal distributions and log normal distributions right.

So you know again taking this forth let us say if I have C=A+B let us say and mu A and sigma A and mu B and sigma B and I want to calculate you know or estimate C let us say. So what is one particular way obviously, the simplest case would be let us say mu A+mu B standard deviation would be the square root of sigma A square+sigma B square right. So obviously this is the simple case right which obviously we are typically not faced with.

For example, if I look at the intake let us say, intake has I believe exposure concentration, body weight in the denominator, averaging time here let us say right and so on and risk would obviously have intake and the slope factor. So there are multiple other variables right and different kinds of what you say variables too right and how do I take these into account and the relevant complex calculations and try to come up with my particular calculation for risk here right.

So in this context, typically what you want to do right, how do you want to go about it, so you are going to let us say create, not create I guess, generate a particular set of data that would conform to that particular distribution of that particular variable. For example, let us say if we assume that this exposure concentration that we have here is normally distributed let us see.

So I am going to generate a set of data let us say which are going to have the particular mean and standard deviation that would be that would that we have for this exposure concentration, similarly for body weight and so on. So once I have that as in I say generate a set of data for exposure concentration, body weight and slope factor right such that the relevant variables are not variables I guess.

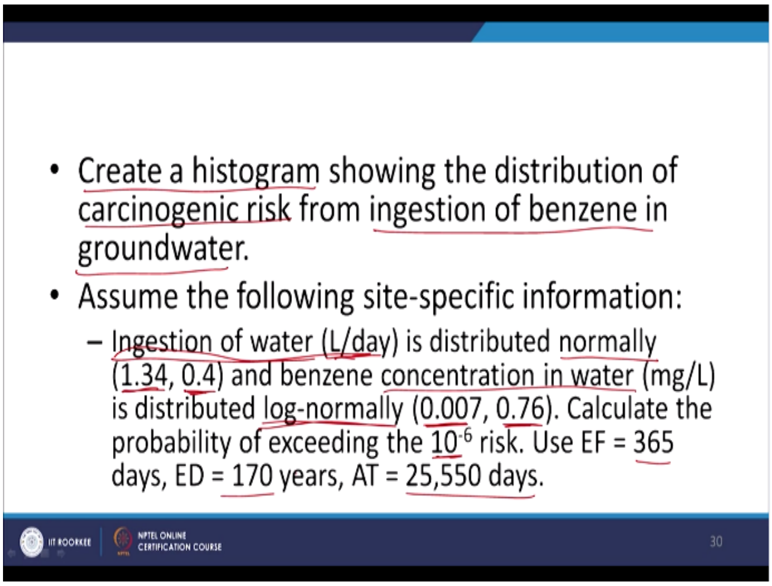
They follow their particular distributions so mu A sigma A, mu B sigma B, mu C sigma C right. I am going to calculate let us say not calculate, come up with different what do we say sets of data that would conform to that particular distribution function let us say right and how can I then calculate the risk. So I am going to calculate the risk as in I will consider this particular set let us say.

And then calculate the risk according to my relevant formulae here right. So here obviously, the case needs to be that they need to be randomly generated right. Obviously, I am trying to look at all the permutations or combinations that are feasible, let us say person with low body weight exposed to very high concentrations let us say or person with high body weight exposed to very low concentrations.

So you know different cases out there right, all such permutations and combinations obviously need to be taken into account in this particular case. So obviously where does this come from? That comes from Monte Carlo simulation what is called I believe Monte Carlo simulation. So in that case, again what is that about? You generate random numbers let us say that conform to that particular distribution function.

And again we are going to have an example soon, so we are going to generate random numbers and use those particular data that we have or random numbers that we have generated to be to then calculate your variable or in this context risks that you are going to look at let us see right. So let us see what we have out here and I have an example here.

(Refer Slide Time: 06:46)



So here we are going to have to create a histogram right showing the distribution of carcinogenic risk again from only one pathway the ingestion of benzene in groundwater right and we have the relevant conditions. So here looks like instead of the point estimate for ingestion of water let us say, we say that the exposure concentration or you know the concentration that they are ingesting has a mean of 1.34 standard deviation of 0.4 and is distributed normally.

And again benzene concentration I guess let us say okay ingestion rate I guess is liter per day is normally distributed. As in, I might drink more water let us say, maybe I am more physically active and another person maybe not or maybe his body does not need as much amount of water. So obviously the rate of water intake is going to be little different and that is what we have here.

The amount of water ingested per day is distributed normally with its mean and standard deviation here and the benzene concentration in that water again it seems to be log normally distributed right. Again, what this log normal distribution about? It means that the natural logarithm of this concentration in water is normally distributed and that particular distribution has 0.007 as its mean and 0.76 as the standard deviation.

So then we are asked to calculate the probability of risk exceeding 10 power -6 which is obviously the typical value right and expose your frequency and duration, averaging time. Exposure duration again this is from (()) (08:13) the relevant standard book or reference that we are following for this course and you know the relevant example from there and maybe there is a minor typo here.

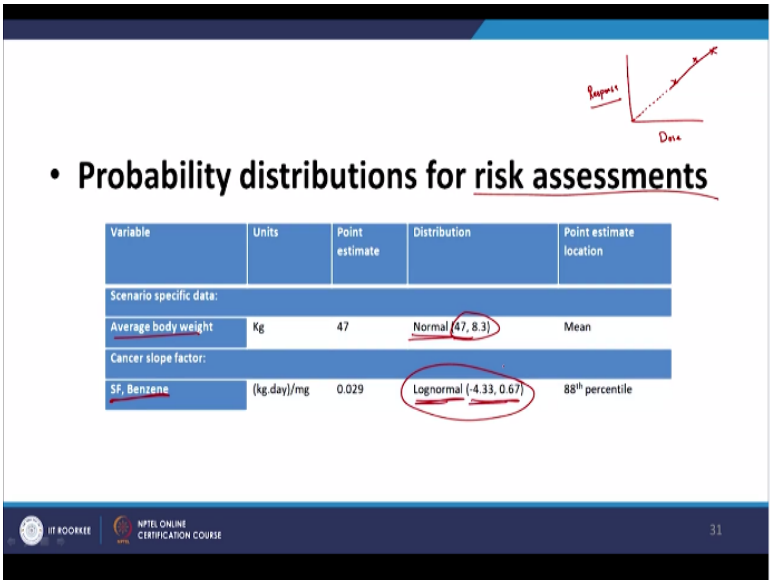
Because the exposed duration obviously seems unreasonable, so we are going to look at that and obviously maybe we can just change it to 17 years which seems a more reasonable value and look at that particular case too. So again how do I go about this? As we talked about it, we are going to look at Monte Carlo simulation right and what are we going to do in that particular case, we are going to let us say come up with or generate random numbers that would conform to let us say for the case of ingestion of water.

We are going to come up with the set of data or random or generate random numbers that would conform to let us say this particular distribution which is normally distributed with a mean of 1.34 and standard deviation of 0.4. Similarly, we are going to come up with the values for the different other variables, not come up with the values maybe, come up with generating random numbers for the different variables.

And then use those particular random numbers for each of these variables to be able to calculate your risk I guess right. Again, I think Monte Carlo simulation I think the root lies in the relevant person who came up with it. I think he was analyzing I think gambling and a particular case of what do we say flipping the coin I guess in a particular gambling house or I think Casino.

Anyway, I think you can look up the relevant history in that context and they were trying to predict let us say the relevant outcome and such and in that case you know, he came up with this particular I guess time-tested procedure which is logically sound too I guess right.

(Refer Slide Time: 09:49)



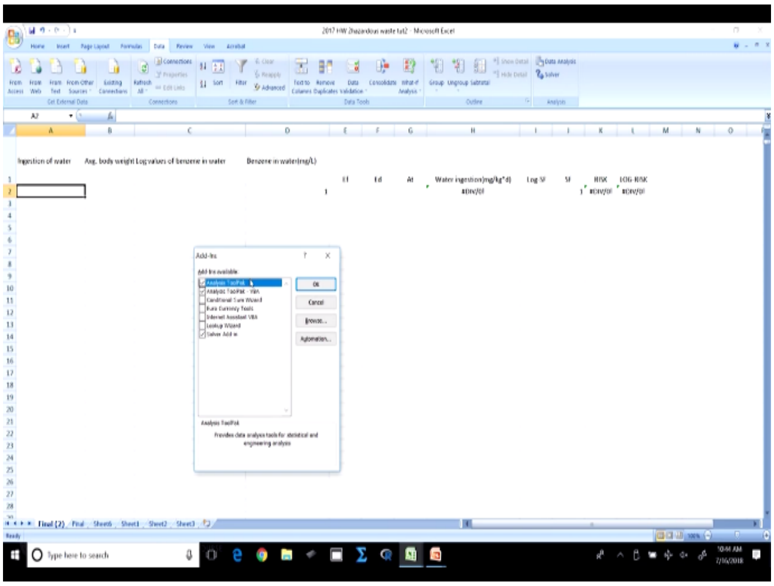
Again, moving on, so again what are some of the other what we say variables for which we use in the risk assessment. So obviously body weight and here we are going to have distribution to be normal and with mean and standard deviation and slope factor benzene rather than having a point estimate though, we are going to have a log normal distribution that these relative mean and standard deviation.

So these are for the log normal distribution as in the mean after you take the natural logarithm of the slope factor right. So again slope factor, how do we get that, we get that from the toxicity studies and such if you remember, again we have this dose here and response here which is a carcinogenic tumor or such and we have high doses that we are going to look at let us see and I fit some particular model.

And we have linearized a multistage model right, we are going to assume that it is linear even at lower concentrations and again this is from accelerated trials as in over a shorter period of time at high concentrations again from animals to humans. So there are considerable uncertainties involved and thus obviously in this stochastic approach rather than using a particular point estimate, we try to understand the particular variation by considering that.

It is a log normal distribution with the relevant mean and standard deviation. So let us go forth and look at how to get this done right.

(Refer Slide Time: 11:09)



So here I have the relevant excel sheet here. So ingestion of water right and I am going to look at that what is the ingestion of water, so I am trying to generate a data set that would be what we say representative of that particular true population. So in this context, let us say if I use lesser or what we say generated lesser random numbers let us say 100 or 10 and so on let us say.

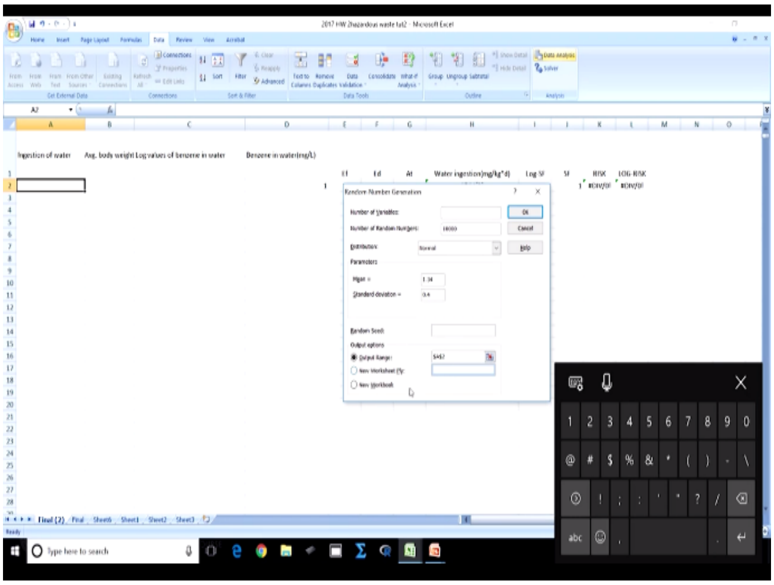
It might not capture the true picture or you might not get the smooth distribution that you are looking for. Obviously, to get a smooth distribution what do you need to do, you need to obviously increase your particular number of random numbers that you are going to generate. So because I looked at this earlier let us say looks like at around 10,000. Let us say I am going to get the relevant what do you say smooth distribution function.

So I am going to look at that but I would ask you to try it out with maybe 1000 or 100 random numbers and look at how the relevant solution changes I guess. So in this case, we are looking at ingestion of water and then I am going to go to data and then data analysis right and then I am going to go to random number generation here, random number generation.

So in case you do not have this in your particular excel sheet let us say or excel file, you need to go to options excel options and then go to add-ins let us say and then manage add-ins and then add it from this particular toolpak here let us say right analysis toolpak I guess right and solver add-in if you want to, certainly analysis toolpak. So again I come back to where I was earlier data. So data analysis, so I am trying to create 10,000 random numbers that would conform to that particular distribution.

And I think I have it out here what are the values here? It is normally distributed with 1.34 and 0.4 being its mean and standard deviation.

(Refer Slide Time: 13:08)



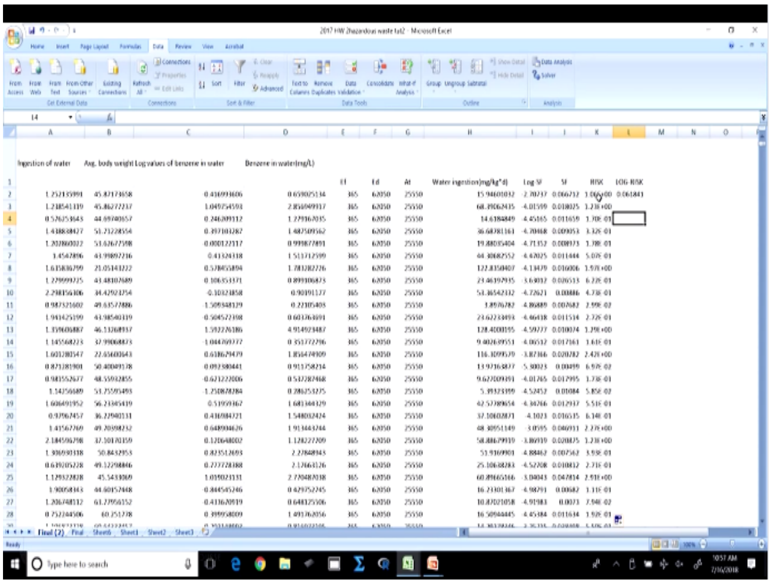
So I am going to go back here, 1.34 right so it is going to be equal to 1.34 and I think the standard deviation was 0.4 right and output range let us say should be here and obviously here I say that the number of random numbers that I want is 10,000. Again, the bigger be data said that I am looking at or sample size obviously. The truer your prediction will be to the actual what we say solution or the actual case right say.

So here obviously I am looking at 10,000 and it is normally distributed. There are other distributions here but obviously I am going to look at normal and so on and so forth and then I am going to ask you to come up with this particular data right. So obviously it is going to take some time. Again to understand what we are up to, what are we doing, we have different variables rather than looking at one-point estimate to be able to capture the variation inherent within a particular scenario let us say right.

You are going to have different particular combinations as in one particular example I just mentioned was a person with you know greater body weight exposed to lower concentrations, a person with maybe lower body weight exposed to higher concentrations and maybe drinking more water and such let us say right. Maybe this particular guy with the lower body weight let us say is working out there in that particular field a landfill.

And maybe the guy with the greater body weight exposed to lower concentration typically sits only in the office let us say right. So to look at these aspects as in one person will be associated with higher risk, one person with lower risk and so on. We look at or come up with these distribution functions that would try to you know give us a true picture or truer picture let us say or better picture of what the scenario is for that particular variable.

(Refer Slide Time: 14:52)



Once we do that, we generate random numbers let us say for each of these variables and try to look at these particular or try to calculate the risk from these particular random generators. The key here is random right, so we are going to assume that the greater, as we keep increasing the sample size obviously not assume I guess, you can see that out there. You see that the risk let us say are going to be what do we say or you are going to get or approach the truer picture as you increase the sample size let us say.

As in, I would think in layman's terms let us say as if you are capturing all the inherent variations or permutations and combinations within a site. So here I have the ingestion of water, I have that. So the next tab I have I guess average body weight, let us see what I have there and body weight I think I have it out here, average body weight again normally distributed with 47 being the average and standard deviation being 8.3.

So I am going to have to get that again, so again data analysis, so random number generation. So 10,000, the mean now is going to be equal to so I am going to set it up in the relevant cell here and again still normal distributed and 10,000. Let me just make sure I have the relevant values 47 and 8.3. So I am going to have the mean to be 47 here and standard deviation to be 8.3 and then everything seems fine okay right.

So again what would happen let us say, if you rather than taking 10,000 what do we say or generating 10,000 random numbers, you consider only let us say 100 or such let us say. Let us say then again as I guess is logical to you know as I talked about or use the layman's terms, you are looking at different combinations or you know permutations out there let us say right. So all those particular combinations might not be captured right.

And thus let us say your particular final solution when you are calculating the risk let us say, well you know the distribution let us say will not be what can I say I guess uniform, not uniform pardon me, it would not be smooth right. The distribution would not be smooth, so if you want to get to the truer picture obviously, you need to have a greater set of data and thus that is why we are going with 10,000 I guess.

So log values of benzene and benzene in water, so I believe I have log values of benzene in water here because we say that the benzene in water is distributed log normal right. So the logarithm or natural logarithm of this benzene concentration has a mean of 0.007 and standard deviation of 0.76. So again I am going to generate that particular data here let us see. So again go back to data analysis, random number generator.

It is still normal as in I do not have the case here for log normal distribution as you see. So normal again 10,000 and here I think I have 0.007 and here 0.76. Let me double- check that, 0.007 and 0.76 that is how the benzene concentration is distributed. As you see, let us say the mean is low right 0.007 I guess, again I think it is milligram per liter but as you see compared to this particular mean, the standard deviation is very high but again that is for log normal.

Let us look at our distribution again out here. So 0.007, output range 2 I guess right and this particular cell right and I have the relevant case, the log natural logarithm of this benzene concentration in water has a mean of 0.007 and standard deviation of 0.76 let us say right. So

I am going to generate that and obviously here in the second or the fourth column, I have benzene in water.

So how do I calculate that and now I am generating the natural logarithm of benzene concentration in water. So obviously to get the benzene concentration, actual benzene concentration in water, I am going to come up with the exponential or use the exponential function I guess right. So let us see where we are. So obviously because I am generating a lot of data or in this case 10,000 data sets I guess.

This particular system is taking time I know but that is usual. So here we have the benzene concentration in water. How do I get that from the log normal distribution? Here I have the log or natural logarithm values of benzene in water. So thus to get the benzene concentration in water, I am going to have to calculate it the exponential of this particular cell right. So obviously we have that and I have it for all the data out there.

And exposure frequency, let us see what we have out here, so exposure frequency is 365 days and exposure duration is 170 years right and let me go back to my particular case here. Exposure frequency is 365 days let us say and exposure duration is I believe 170 years as I mentioned you know probably a typo but you know let us just calculate that and you can calculate for a more suitable value of 17 years maybe.

Averaging time, again because you are looking at carcinogenic risk is going to be equal to the lifetime which is 70 years let us say right, 70 years and that is 25,550 days. So here that is going to be equal to 25, 550 days let us say and here I have the calculation for intake I guess. Ingestion of body water let us say benzene concentration, exposure, frequency, duration/body weight and averaging time.

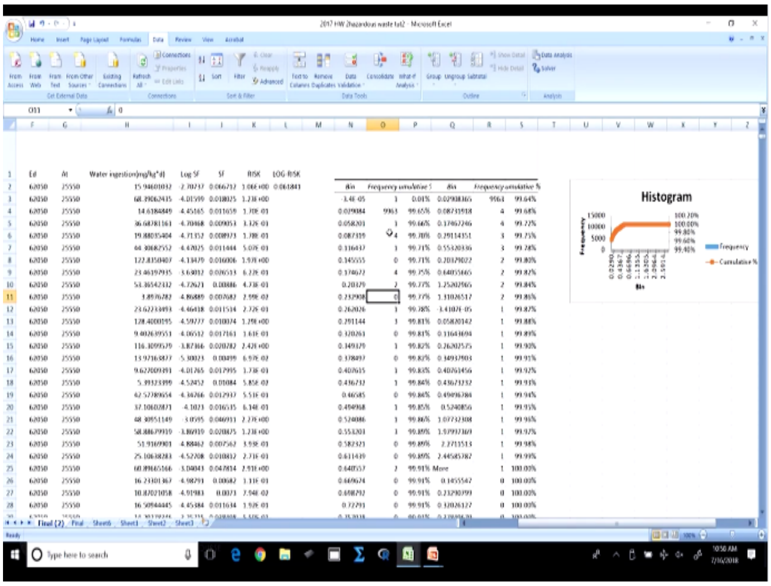
So have that out here, so again similarly I am going to get it for all the particular cells yes and again here I have the logarithm of the slope factor and thus again need to generate the relevant data. So let us go back to that particular cell here right, so I think the slope factor has the log normal distribution as we can see out here right and the particular distribution has a mean of -4.33 and standard deviation of 0.67.

So let me try to plug that in. Again, how do I do that? Data analysis, random number generation right, again 10,000 data sets and here I need to plug the relevant values in -4.33 let us say right and the output cell has to be out here and what is the standard deviation please, let us go back to what we have, it is 0.67 pardon me, it is 0.67 and let us say I am going to say okay. So here we have the data for the log normal distribution.

Again, thus I have it labeled as logarithm of slope factor, again keep in mind that it is natural logarithm. So I have generated 10,000 random numbers and now to get the slope factor obviously, it has to be the exponential of this particular cell let us say. So this is the actual slope factor, so I calculated that and risk obviously is intake*slope factor in this context of carcinogenic risk right.

That is what we have water ingestion*slope factors. So again what do we have, we have 10,000 cases of ingestion of water or types of ingestion of rate of intake of water, 10,000 body weights let us say and 10,000 concentrations of benzene in water but same case or we are assuming that it is uniformly distributed for exposure, frequency, duration and averaging time and so on and so forth and we are calculating the risk now let us say and here I have the risk.

(Refer Slide Time: 22:51)

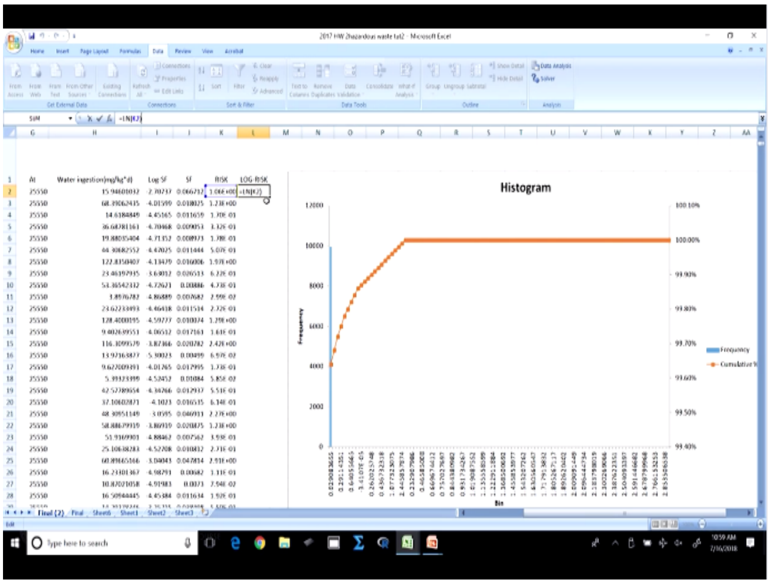


And let us say now I want to capture how or understand how this particular risk is distributed let us say. How do I do that? I can look at data analysis, go to histogram out here okay and what is my input range here, obviously it is K here and bin range I am going to ask it to calculate that automatically but if you want to, you can give a particular bin range. So output range let us say I am going to want it have it somewhere here, that seems fine.

Sorted histogram, cumulative percentages and chart output, this will help me understand the data in a better manner but here I believe I cannot give it in this way because I am going to have this particular text here. So it starts from K to and I believe since we asked for 10,000 data sets, it landed 10,001 I guess right, so 10,001 okay. So let us see what we have here, so I guess excel came up with its own bin and frequency right.

And obviously what do I have here, this graph the way it is as you see here, let us say it skewed to the left.

(Refer Slide Time: 23:59)

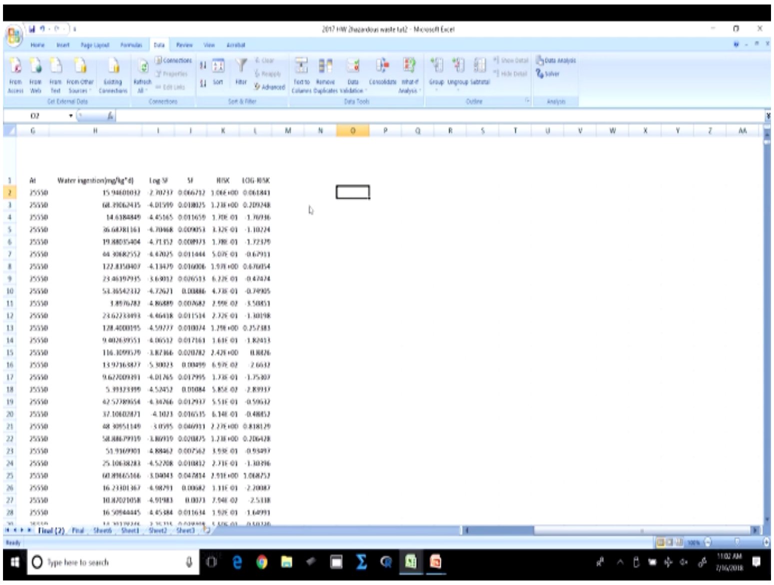


And then here you see almost everything out here let us say right and again why is that the case because let us say we asked excel to come up with its own bin size let us say right or with its automatic bin size and obviously here as you see within a particular bin let us say, the frequency let us say is 9963 as in almost all the data is within a particular bin size let us say or again as you can understand from here what does this mean though?

If it is skewed in this way that it is not normally distributed, it has a different kind of distribution. So probably what would this mean, it probably means that it has a log normal distribution right and that is not unreasonable because we looked at normal distribution, log normal distribution and uniform distribution in our particular calculation of this risk right. That is what we have here.

And thus let us say you know it is not unreasonable or you know it is maybe logical to understand that we are again going to have a log normal distribution. So to understand that or capture that true risk let us say what am I having here, I am going to take the natural logarithm of that particular risk let us say and I am going to calculate that, yes I have that.

(Refer Slide Time: 25:07)



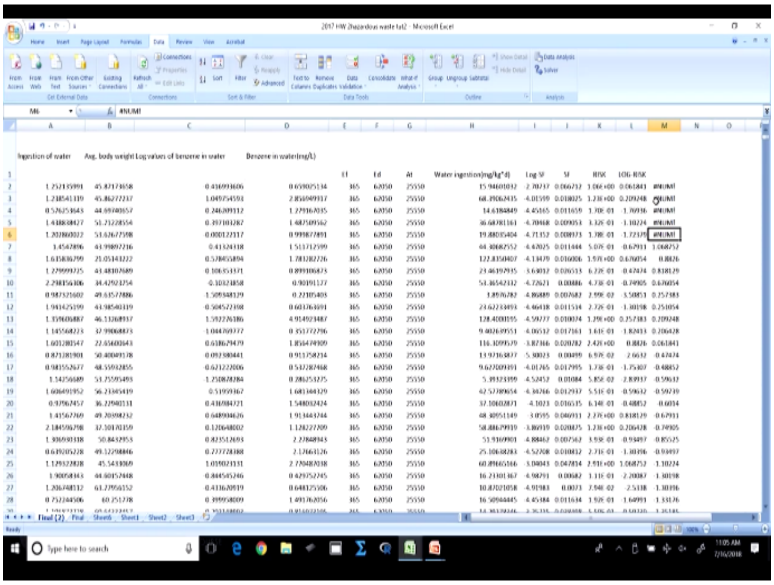
So I am going to delete this particular risk from here let us say right and here I am going to now conduct the data analysis on this particular set. Histogram again and here I am going to now go with K and now it is L here right, the next column. So again I want to understand the true nature or you know try to visualize it in a better manner let us say right. So thus I am again going for the log normally distributed or assuming that it is going to be log normally distributed I took the logarithm of the risk and that is what I have here.

And again I am asking excel to come up with the particular, so here the issue seems to be that you know the input range contains non-numeric data. As in, there are some cases maybe because I came up with 10,000 cases of risk. There are maybe you know cases where it has non-numeric data. So what does this mean now?

As in, I took 10,000 cases right, so within all these different permutations and combinations let us say, you know there might be some cases let us say where you know in the denominator or such it does not make sense and then we might get non-numeric data I guess right. So obviously histogram you know it cannot take that non-numeric data into the context. So let us say let me try something out here right.

So typically though what can you do, you need to go through this particular 10,000 list let us say. You need to go through that particular 10,000 list, look at where the what we say non-numeric data is and then you know obviously get that out of the picture or delete that let us say right. So obviously you know manually doing that is going to be relatively different. So I am obviously going to try to use excel here let us say.

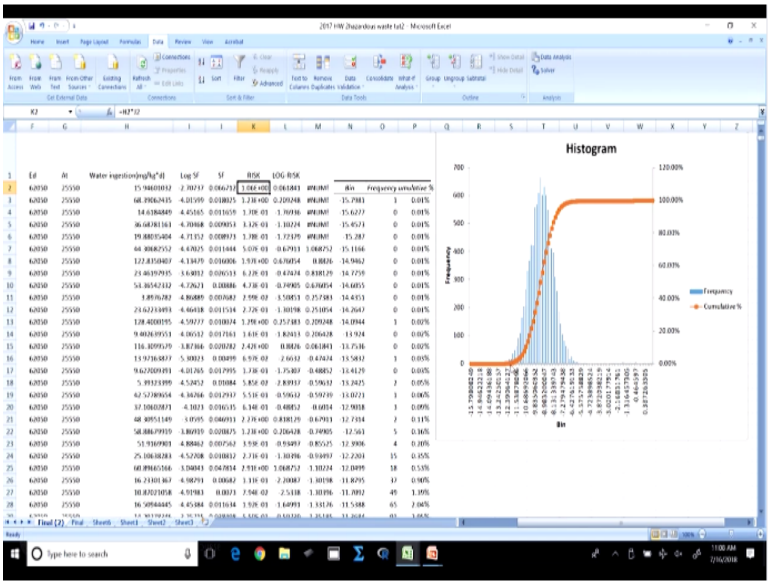
(Refer Slide Time: 26:56)



And what am I going to do? I am going to try to sort it here let us say. So sort it, continue with current selection and sort log values from smallest to largest okay and so I am trying to look at this particular value here. So I copy pasted the values in this particular column and then I sorted from smallest to largest and now see that there are what we say 4 or 5 cells rather that came up with what we say non-numeric values.

Again, within 10,000 set of data and relevant permutation and combinations so that is not unexpected. So thus we could not obviously come up with the histogram. So I am going to not consider this data and ask the you know data analysis or excel to come up with the particular histogram here. So let us say here I am going to look at obviously column M here let us say right.

(Refer Slide Time: 27:48)



And I am going to say it is going to be M here right and I need to start from cell number 7 right, so cell number 7 here, so this probably needs to be a capital M okay. So and M2 seems fine, so let us look at what data we have out here and now you see now that this particular graph is remarkably smooth or you know if not remarkably pretty smooth and again what do we have here.

Excel came up with its own bin size out here and column N and in O the frequency let us say and then again the probability here, I guess this is the cumulative but again you can calculate the probability obviously based on this frequency by 10,000 or this 9995 I guess right. Again, how did we mention that we can get the probability as in the number of events by that particular bin size let us say.

And how do we get the probability distribution function let us say, that particular probability by the bin size I guess right. I guess I messed it up there. So how do I get the probability? The number of events in that bin/the total number of events, so in that case it would be let us say for this particular bin it will be 1/9995 because I took 10,000-995 right. So in that case, again I can come up with relevant probability and the probability distribution function.

So having plotted that I see that this particular logarithm of the risk let us say has a normal distribution. As in, what do I understand from that, as in the risk has a log normal or follows a log normal distribution right. So again this is what we have here, so earlier if you looked at the previous graph though we see that you know we could not make much sense out of it. Why is that, because it was log normally distributed right say right?

And that is why it was skewed to the left, it as in just the risk was skewed to the left but when I took the logarithm of these values, the natural logarithm I see that now it follows a normal distribution and again here we have we can understand the data in a better manner. Obviously, though we are yet to calculate what is it now the probability that the risk will exceed 10 power -6, so how do I do that?

I can sort them out as in I did sort them out here let us say or I can sort out the risk here let us say right and look at those values that would exceed 10 power -6 and obviously calculate the probability which we are going to do or I can look at or calculate the natural logarithm of 10 power -6 plotted out here let us say wherever that is and look at that cumulative risk and see you know what are the chances of this probability exceeding 10 power -6 I guess right or 10 power -6 risk let us see.

Because it is going to be the area obviously right, so again we are going to you know move on to this or you know discuss this in great detail again in the next session but for today what have we done, we looked at stochastic approach, Monte Carlo simulation, as in we considered random number generation let us say for different variables depending on their type of distribution.

Then, calculated the risk for each of those particular sets of data we came up with and then we saw that the risk was actually log normally distributed and that is what we have here and we are still obviously yet to calculate the what we say actual aspect we are supposed to, what is that now, the case that or probability that the risk would be >10 power -6 let us say right. So again, we can look at the area under this particular graph or obviously count the number of cases where the risk exceeded 10 power -6.

So we are going to look at both those factors or aspects in the next session I guess right and I guess with that I am going to be done for today and thank you.