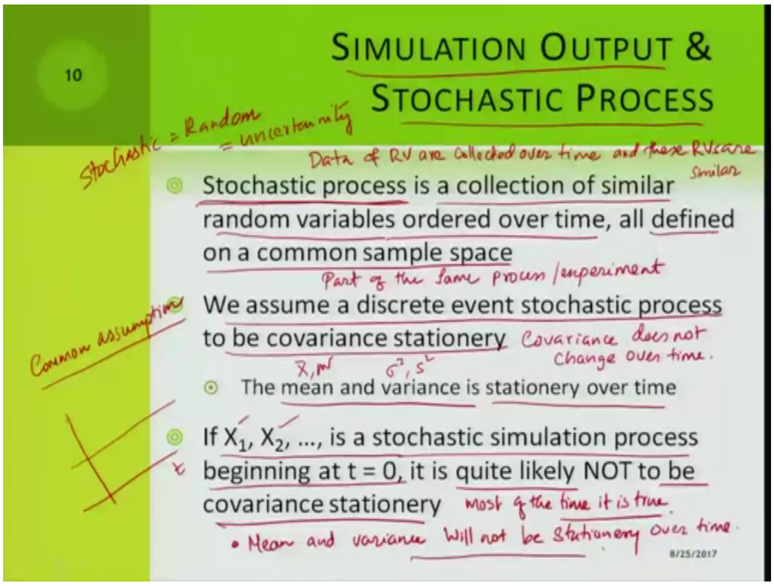
Sustainability Through Green Manufacturing Systems: An Applied Approach
Prof. Deepu Philip
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur
Dr. Amandeep Singh Oberoi
National Institute of Technology, Jalandhar

Lecture - 07
Basic Statistical Concepts for Sustainable Manufacturing Analysis Continued

Good morning, welcome to the new lecture on sustainable manufacturing. Today we are going to finish the content on statistical concepts that we started in the previous class, which are necessary to do the analysis of the simulation data. So, we will continue on, where we stop the last time. We have covered the concepts of sample space, experiments. Then we also talked about random variables. We also talked about what is a probability density function distribution functions P D F and C D F cumulative density functions or distribution functions.

We also talked about, how do we look into what is the concept of independence disjoint, all those kind of stuff. So, today we are going to talk about how do we deal with simulation output and stochastic process.

(Refer Slide Time: 01:06)



So, the simulation output and the stochastic processes is two concept; that is important for us to understand. So, when people say stochastic process, we said about stochastic is something like equal and to random. So, if I say stochastic assume that is random. Random means, you do not really know the when the next event is going to happen, there is a variability associated with it. Not variability uncertainty I means.

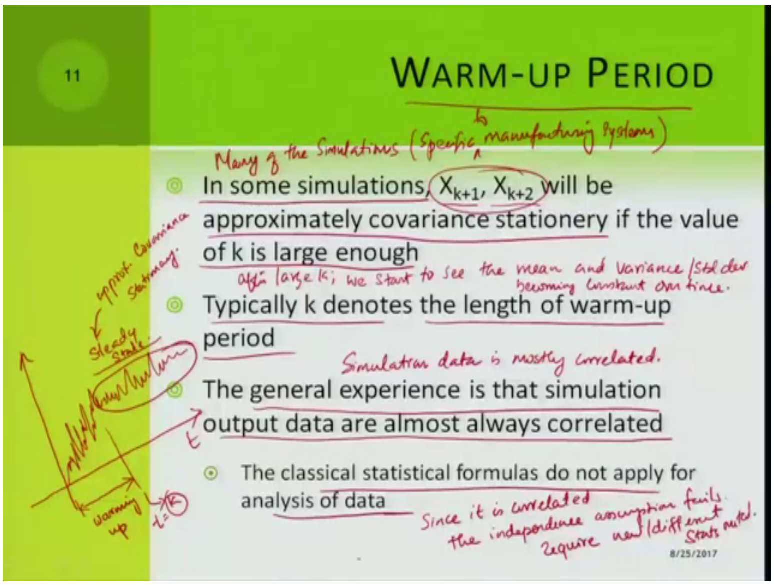
Let us say uncertainty as the word. So, a stochastic process for us, a process that is a randomness you need, is a collection of similar random variables. So, the random variables are similar, ordered over time. So, these random variables are collected over time. So, or we say as data of random variables are collected over time, and these random variables, these R V s are similar. Similar means they all part of the same process that is what we are talking about.

Well defined over a common sample space. So; that means, this is all part of the same process or experiment. So, we have basically saying that, these random variables are ordered over time, the values of the random variables are ordered over time, and they are defined over a common sample space. Now the next most important thing is that, common assumption, the common assumption in this regard, the common assumption we use, is that we assume that a discrete event stochastic process be covariance stationery, which means the covariance does not change over time; that is the idea as time progresses, the covariance of the process remains stationary. So, if you graph it you will kind of, if this is the time axis you will kind of these stationery go on change over time that is the thought process. Or another way is the mean and variance of the process, mean X bar or mu whatever and the variance, sigma square or s square; either one of them is stationary overtime. I hope that you guess I have already read the text books on statistics, and a familiar with the concept of mean variance, standard deviation, median, mode all those kind of stuff .

So, which means they do not change over time, the values remains same, but in a simulation, when you do a simulation analysis, discrete stochastic simulation analysis. If X 1 X 2 X 3 is a stochastic simulation process. So, if these are the values for random variable; that is coming from a simulation process, and the simulation is begin at time t equal to 0. It is quite likely, or most of the time it is true. What is true, that it will not be covariance stationary, or the mean and variance will not be stationary over time ok.

The idea is that they will not be stationary overtime, or the covariance of the process will change over time. So, what is, why is it important to us, why is this covariance stationary is an important thing to us, because when you want to do analysis of the simulation data. This covariance stationary is of importance, for us to identify something. What do we want to identify.

(Refer Slide Time: 05:22)



There is a concept called warm up period, every system it will take little bit time to warm up. In many other types, in some simulations, or in many of the simulations, specifically related to manufacturing systems; let us talk about manufacturing here, little bit specific to manufacturing systems. So, assume that you come to a factory at 10 in the morning, you switch on the lights, you switch on the machine, use their lot in the machines.

So, for sometime maybe 10 or 10:30 11 in the morning, it will take some time for all the entire factory to come to full flow. So, that type of a period, you know is called as a warm up period. So, in many of the situations in manufacturing, if you are doing the simulation, the X k plus 1 X k plus 2, or which means few after a large value of k. If k is large enough, after a large value of k you will see that this values of the random variables, will start to behave like approximately covariance stationery process, or you will start to see, as after large k, we start to see the mean, and variance or standard deviation whatever you want to call it.

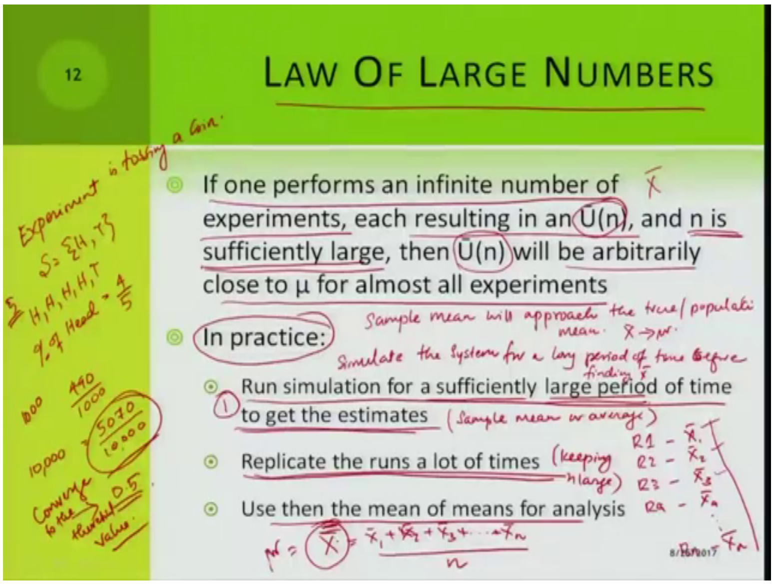
Becoming constant overtime, what we are saying here is that, in the value of k is large, the initial warm up period up period is eliminated, then you start the a covariance stationery process, and the length this k, whatever it is, it is denote or as the length of the warm up period. So, if you look at the output of a system, it will look like this. This is your time, and this is some output. We will actually start seeing it from 0, you will see all these kind of variations, we will go like this, and then of some point, but I will start seeing doing this. So, we can basically say that, up to this point is when the system was warming up, and here this is the value of k, this is the value of time period, at that time t equal to k is when the system transitions from my warm up system to a.

So, this behavior is what we call as steady state, and this is the process in which you are covariance, is almost stationary. So, this is approximately, covariance stationery, but you can think about that the mean and the variance are kind of study over time. The most important thing, that I want all of you to remember is, the typical experience from all the people who worked in the simulation field is that, the simulation output that is almost always correlated. So, simulation data, is mostly correlated, almost always correlated that is the best way for you to think about it.

Hence what is a problem with the correlation, or correlated data? The classic statistical formulas do not apply to analysis of those data. So, here we can assume that, since it is correlated, the independence, assumption fails, require new methods or different methods, different states method. So, we will see how we analyze simulation data in the input and output analysis, we will talk about that later down (Refer Time:09:52) that is somewhere related down, after we talk about the optimization and other stuff, but please remember these .These point is important for it to remember later for the entire duration of this course. Then we also talk about something called the law of large numbers. The law of large numbers is a scenario, where the concept is quite simple.

If you have a very large enough sample, then your sample mean, tends to be behaving very close to the real value of this, real mean you know expected value of the system.

(Refer Slide Time: 10:32)



So, the idea is this, if one performs an infinite number of experiments. So, if you remember we talk talked yesterday, in the experiment is, let us say tossing a coin apologize. So, you are output is, your sample space is you can get a head or a tail. So, if you do the experiment 5 times let us say, and you got the output as head head head head, and tail and if you try to estimate what is the percentage of head, then it will be 4 out of 5 is what you will get as the probability.

Which is not the real probability, but if you do this, you do this 5 times this is what it is. if you do this 1000 times, you might get somewhere close to, you know let us say you get a value of 490 out of 1000 or something like this. If you do this 10,000 times, then you would get somewhere close to maybe 507, sorry 5070 by 10,000 something like this. So, as you can see this value will slowly approach towards the theoretical probability of 0.5, that is the idea here. Most of the time the probability will pretty soon, it will converge to the theoretical value. So, the idea is that, if you do the experiment infinitely large number of times you do it large amount of time.

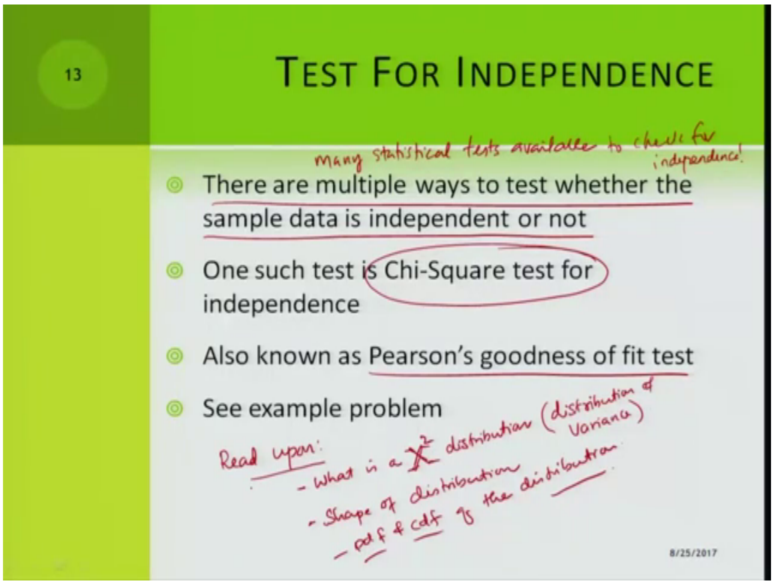
Then each experiment, each number of each time when you do this large number, you get a sample mean U bar or X bar ,whatever you want to call it, and n is sufficiently large, it is big. we are talking about 1000 10,000, those kind of numbers, then this U bar n or the X bar, whatever you want to call, it will be arbitrarily, close to mu for almost all experiments, or here you are saying that the, sample mean will approach the true, or population mean, which means X bar will tend to become mu, or mu is your population mean. What does this law of large number implies to people, who are doing simulational analysis. So, for us, when we had to do it in practice, when we are to analyze the simulation data in practice, what we do is, we run the simulation for a sufficiently large period of time.

So, we first simulate the system. So, simulate this system for a long period of time to, before finding X bar. So, you run the simulation for large period of time to get the estimate, or estimate is the, may be the sample mean, or the average. So, your large number, the end is sufficiently large. Then you repeat this many times. So, if you do this, your first the run. So, run 1 you have an X 1 bar, run 2 you have an X 2 bar, run 3, you have an X 3 bar, run 4, you have an X 4 bar like this you will continue.

So, you replicate these runs many times, then you have until nth run, you have an X n bar. Then you can use all these X bar to calculate the mean of means. So, we can calculate X bar bar, which will be equal to X 1 bar plus X 2 bar plus X 3 bar plus all the way until X n bar divided by n. So, you take the mean of these means, each mean you get from your application from there, you do the means, and this mean of means will be very close to mu or this is almost the true estimate of the population mean ,that is what we talk about the. This is how the law of large numbers is put into practice ok.

So, law of large numbers implemented into the system, is that you run the first step is you run the simulation for a large number of, sufficiently large period of time. So, that the n is sufficiently large, this concept n become sufficiently large, then using that you find the estimate, you get your X 1 bar X 2 bar X 3 bar. To do that you replicate you keep the n sufficiently large, here you are keeping n large you keep n large and replicate the whole experiment, and then use these averages to calculate the mean of means. So, use this to calculate X bar bar, here have the mean, then use that mean to do the estimates.

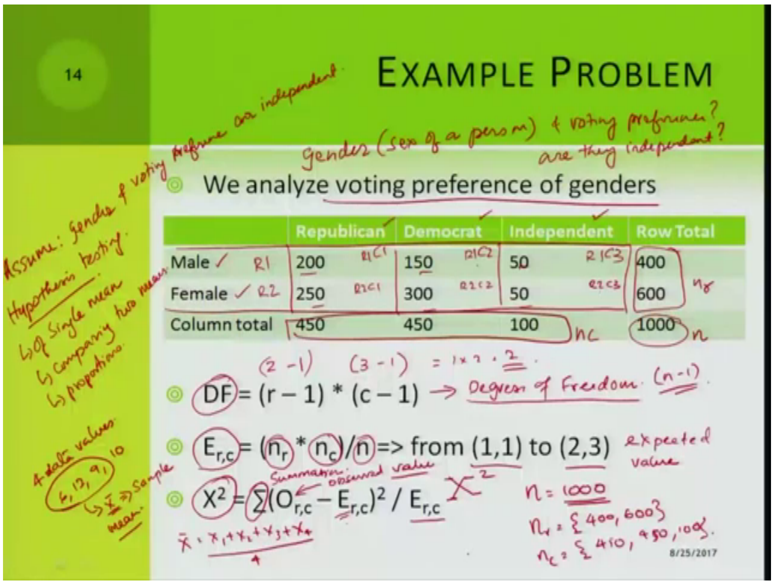
(Refer Slide Time: 16:05)



All right, one of the things that we are talked about is the importance of the concept of independence disjoint and all those kind of things. So, if you are given a bunch of data. How do you know whether the data is independent or not; is there any way to identify that. So, we will actually go through very simple example today in the class to show you how to test for independence. So, there many ways, there are multiple ways you can test, whether a sample data is independent or not. There are many statistical tests available to check for independence all right. So, check for the independence of the data using many tests, many statistical tests are available. One example tests are that we are going to cover, this is called as the chi square test for independence, or we also called as Pearson's goodness of fit test ok.

So, this test will actually help you to decide, whether the data is you know, whether we can use chi square test for do in the test of independence. I would request you guys to read upon, what is chi square; this kind of looks like this chi is a Greek variable, chi square distribution. If you read about distribution of variance, then you will get to see the chi square distribution; that is one part, then shape of distribution a P D F and C D F probability density cumulative density function of the same. So, please read these kind of things, so that you can understand what do I mean by chi square distribution. So, but anyway so, the mechanics of the text we will go through, and you will understand the concept, once you read it you augment the concept that way.

(Refer Slide Time: 18:31)



So, the example problem that we are going to analyze is that, the voting preference is it influenced by gender. So, we are trying to see, the gender, weather gender or sex of a person, and voting preference, are the independent. This is the question what we are trying to follow. So, here we begin with the assumption that, our initial thought process is that, we assume same way as in the court of law, everybody is assumed to be innocent. Here we assume that gender and voting preference are independent, and we are going to check this assumption. If we do not get evidence to verify this assumption, then we can reject this assumption and say that it is there inter related, or they are not independent they are related.

Same way as a person is assumed innocent, and biggest on the evidences that is provided in front of the court, the person is, if the guilty is proven beyond reasonable doubt, then the notion of innocence is rejected, and the person is amongst guilty. Same way this is same process. This process is typically what is called as hypothesis testing. It is available again in the statistics textbooks. I know many of you are gone through these topics, but I request you guys to read this, hypothesis testing of single mean comparing two means, all these things please do read. So, then there is also things about proportions, please do read these concepts from the statistical books, or revise your knowledge of statistics.

So, that you can follow what here it is, what we how here is that, we have two gender a male and female, and there is three parties this is the (Refer Time: 21:00) is data, the republic and democratic and independent. So, these are the stuff, and we have collected some sample about 1000 people were checked with this, and from the male that other 400 male and 600 female, or the 400 male 200 what republic in 150 vote a democrat and 50 vote in independent, of the 600 female 250 vote at republic in 300 vote at democrat and 50 vote at independent kind of a thing and. So, we are trying to find out whether the gender has an influence on the voting preference. So, the first thing that we need to learn, are the part of here is this concept is called degrees of freedom ok

So, degrees of freedom is a again a concept; that is discussed very well in this statistical text books, here the degrees of freedom is calculated as row minus 1 times column minus 1. So, we have 2 rows we are only looking at, these are, this is the data that we are looking at. So, we have 2 rows 3 columns, this is not used. So, we have 3 columns. So, this is r 1 and r 2 this is c 1 c 2 and c 3, 3 columns. So, it is since there are 2 rows as 2 minus 1, there are 3 columns; that is 3 minus 1. So, that is 2 minus 1 is 1 3 minus 1 is 2. So, it is 1 times to 2.

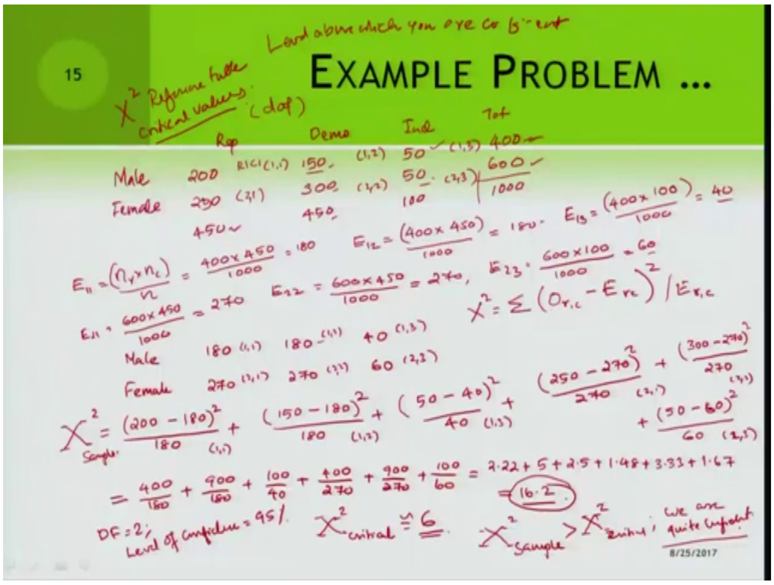
So, the degrees of freedom is 2, lot of people have this question about what is the degrees of a, degrees of freedom. Text books explain this very well, one simple as we have to look at, it is the number of. Now if you think about is a data set it is, how many of those data sets data points among the data set can be independently determined, like for example, if you have 4 data, values is given to you. Let us say four data values, let us say 6 12 9 10 something like this, and if you used to find, the next use these four values to find, then X bar which is a sample mean. So, if I give you the X bar and I give you three of these values, any of the three values, then you can find the fourth value, because you know that X bar is equal to X 1 plus X 2 plus X 3 plus X 4 divided by 4 in this case.

So, any 3 plus the X bar gives you the, you can use it to find the fourth value, or any of the three is, what is required along with the X bar. So, degrees of freedom is that, typically it is usually 1 less than the in many other people, they call it as n minus 1 or those kind of things, not necessarily true. This calculated in many different ways, but please do read the concept of degrees of freedom from the statistics books, that we are accommodated earlier, once you calculate the degrees of freedom, the next thing what you need to do. If you need to calculate the expected values, the expected value of each one of the row column E r c is the expected value calculated, which is calculated from the column 1, column 2, column 3 then. So, this is r 1 column 1, r 1 column 2, r 1 column 3.

There is r 2 column 1, r 2 column 2, and r 2 column 3, for each one of the ones. So, from 1 1 column 1 r 1 column, 1 to r 2 column 3 1 1 to 2 3, for each column you are supposed to calculate the expected value, how are you calculating the expected value number in the row multiplied by n r is number in the row multiplied by n c, number in the column divided by n, n is the total value. So, here n, little n is equal to 1000, and the n r can how values, which is 400 and 600, and n c has the values number columns has the value 450 and 100, it kind of know, what I am talking about. These are the n c values, these are the n r values, and here is your n ok.

You use that, find it, and from there you find the chi square value, this is the chi square. This is actually written like this, chi is kind of a chi square. So, the chi square value is sigma or summation. This is the summation sign, the summation of difference between O O is, this is the observed value, the observed value is this. these are the observed values minus the expected value E r c are already calculated it, and then you divide by the expected value, and then you, some those ratios and you will be able to get the answer. So, let us do one thing, let us actually solve this problem.

(Refer Slide Time: 26:28)



So, we will try to do that. So, the setup in such a way for us was that it is male, we have female, we had male data was for republican democrat, and independent and you had a total here. So, the republican data was 200, a democrat data for male was 150, and independent was 50, and the row total of 400, and for female the republican was 250, the democrat was 300, and independent was 50 with a raw total of 600. The column totals were 450 450 100, and the total sample in both cases was a 1000. So, we calculate the expected values ok.

So, if you do that. So, the expected value of 1 1 will be equal to. Remember that it is n r times n c divided by n. If you use that, the n r for the first row will be 400. This is the first row, then n c column for this one will be 450 divided by 1000, which will give you the value of approximately 180. This will be the value of 1 1 E 1 1. If you do E 1 2, which is the second 1, expected value of this that would be equal to 400 times. this is the 400, then the second 450 divided by 1000, that will give you again the value of 180, that 180 is for this particular case. then the third one is E 1 3, which is equal to, again 400 is the row value, column value is 100 divided by 1000, and you get a value of 40 in this case ok.

So, that value gets goes into this 1. So, if you follow that process, the E 2 1 will be pretty, much will be 600, that is the next row, times 450; that is the column, this column divided by 1000, that will give you a value of 270. E 2, 2 the second case, it will be 600 multiplied by 450 by 1000, and that gives you again 270, and E 2 3 expected value of 2 3. Here it will be equal to row multiplied by column divided by 1000, equal to 60, that value is again as this. So, if you write that matrix, the expected value matrix, it will look like this; male, female and you will have 180 180 40 270 270 and 60.

So, these are the expected values that gets compared one to one correspondence, then by the previous equation, if you look at it you calculate all the expected values now. Now you have to calculate the chi square. So, the chi square equation, we know that chi square is calculated by sigma summation of observed value of row column minus expected value of row column, the square of the divided by the expected value of the row column. We want to do this, then the chi square can be calculated in this regard equal to. We will take the first 1, this is your r 1 c 1 or we can call this as 1 1, this is 1 2, this is 1 3, this is 2 1, 2 2 and 2 3 that is the case. Then we have it is the first case it will be, the observed value is the for the first one 1 will be 200, and then from that the expected value for that it is 180 square of that divided by 180, that will be the first 1. Then for 1 2, it will be observed value is 150, the expected value is 180 square of that divided by, it will be divided by 180, because such a expected value E r c, this is the E r c ok.

So, this is your 1 1, 1 2, 1 3, 2 1, 2 2 and 2 3 all right plus the next one comes is, your summation, we are summing the whole thing up, the observed value for the 3 1 3 will be 50, expected value is 40 square by 40 plus, now if you do the third, the 2 1. So, this is for 1 1, 1 2, 1 3. Now we are doing is a 2 1, which will be observed value is 250, expected value is 270 square of that divided by 270, this will be 2 1 plus the 2 2 will be, the observed value is 300, the expected value is 270 square of that divided by 270, and that will be 2 2, and the last one plus will be, the observed value is 50, the expected value is 60, the square of that divided by expected value which is 60 this is your 2 3 ok.

All those values are summed up. see you will get the values equal to, this will come to 400 by 180 plus the next one will come to 900 by 180, the third one will come to 100 by 40, then fourth one will come to 400 by 270 plus then next one will come to is 900 by 270, then the third one will come to is 100 by 60, which will be equal to. If you do the math it will come to 2.22, I am already doing the two decimal places plus 5 plus 2.5 plus 1.48 plus 3.33 plus 1.67 all of this values are equal to 16.2, this is the chi square value that we get.

So, 16.2 is the chi square value of the given sample. Now how do we make a decision based on this value. So, there is a table; that is available in the statistics books, it is called the chi square reference table, which contains the critical values. So, the critical values of chi square table is calc, and you check that critical value is using degrees of freedom. So, the degrees of freedom we calculated earlier here, the d f was calculated equal to 2, and if you find that I want to prove you have to also say what is your confidence level, or the level above which, you are confident. So, that part.

So, here we are going to take the level of confidence, equal to 95 percent I am saying that, and I want to be 95 percent sure. So, if you take these values and you get this, then the critical chi square value from the table, chi square critical. This is our chi square sample that we calculated from the sample data, critical will come approximately close to 6 . So, now, we compare these values, since chi square sample is greater than chi square, chi square critical, we are quite confident. We are quite confident that our initial assumption that the gender does not influence the voting preference is not correct, and hence we reject the notion that the gender and voting preferences are independent, and we assume that gender do influence the voting reference.

So, they are not independent anymore. I assume that by this you guess are able to understand, how we are able to study the, how we can work out the concept of a independence through a simple setup data, and I would really, with this we conclude this lecture today, but I really request you guys to study this concept all by yourself little bit more. So, that we can, in the later part of the courses when we are talking about optimization and other parts, you will all be able to pick it up better.

Thank you for your time.