Total Quality Management-II
Prof. Raghunandan Sengupta
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 07
Hypothesis Testing

A very good morning, ad good afternoon, good evening my dear friends and students I am Raghunandan Sengupta from the IME department, IIT, Kanpur and welcome to this TQM two, lecture 7 which is the second class in the second week and as you do remember, we were discussing about initially about point estimation, then went into details about interval estimation and then came in to hypothesis testing and I did also mention about with a background or with a small problem of a bank loan being given by managers.

So, they would be one set of people who would default on set of people, who would not default based on that you find out alpha and beta which is type 1, type 2 errors, what are those errors? So, we did discuss about H naught null hypothesis, H A alternate hypothesis and how they are contradictory complimentary and then you try to solve, then we also did mention and I am going to mention it time and again that when you want to find out something to do with the population mean given the standard deviation of the population or the variance of the population known, we use the z distribution and z distribution as we know is symmetric with the mean value of 0 and standard deviation or variance of value of one.

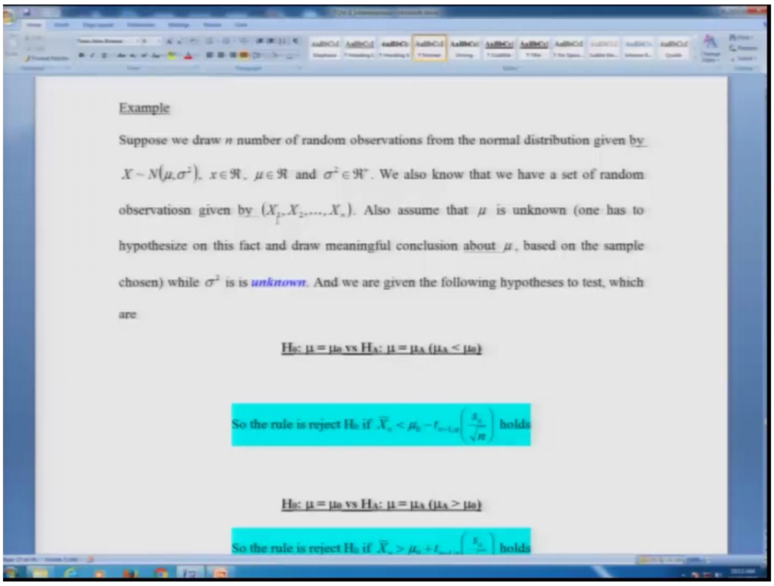
Now, in case if it is again to something to do with the population mean and in the case, the standard deviation of the variance of the population unknown, then we use the t distribution with a certain degrees of freedom being lost, why they are lost, we did discuss where we use S dash s without the dash and so on and so forth, then we did also mention that t distribution is symmetric hence when you take t to the left to the right you have t minus t alpha n minus 1 or minus k that loss of degrees of freedom whatever it is alpha or alpha by 2 depending on how the example has been framed for the hypothesis testing and on the right hand side, it will be t n minus 1 or n minus k as the case may be with alpha or alpha by 2 depending on the whether it is basically one sided, both sided and all these things.

Then we came into the second the big area about trying to find out something to do with variance the variance can be found out when you are trying to compare with the existing population one population you use the chi square with the certain degrees of freedom being lost or not being lost, it would not be lost when you know the mean value of the population, it will be lost and you will reduce it by one, if the population mean is not known because in that case you use s without the dash and in the case when population mean is known you use s with the with the dash, then you when you go to the case when you are trying to compare 2 different variability variance dispersion and all these things.

Then we use the f distribution. Now f distribution have 2 degrees of freedom m and n m being for the first population for sample size n being the second population second sample size and in that case, you are trying to use the s dashes for both the samples, in case if the population means respective means mu 1 and mu 2 this one and 2 other suffixes are not known, then you use s dimensional arrays without the dashes in both the populations for the samples and we use f m minus 1 n minus 1 accordingly.

So, so, with this, again I will continue concept will be same repeated time and again, but it is just to make you well aware and have a very clear cut picture in your mind how these frameworks are be done; obviously, there has a lot of theory here, but we would not consider it in details because it is not required for this TQM two, it will be required for a rigorous course in statistics probability and inference techniques .

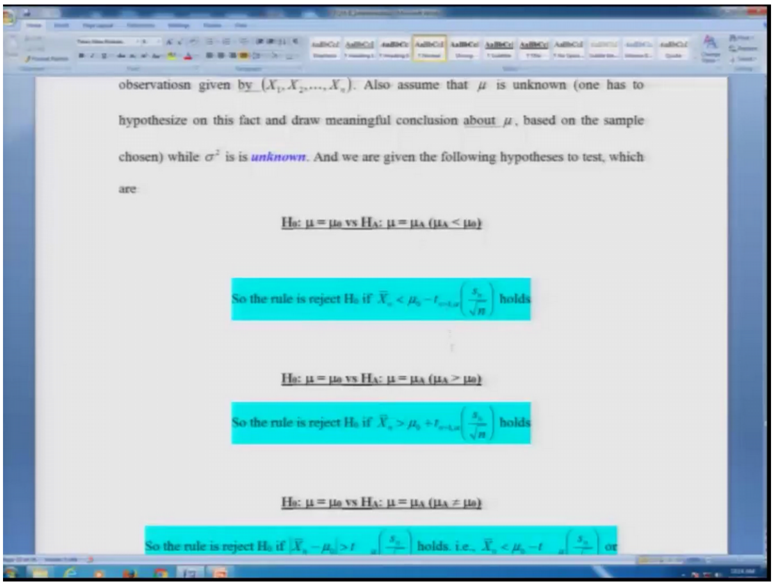
(Refer Slide Time: 04:09)



So, suppose we draw n number of random observations and r r its a normal distribution so; obviously, the observations are random variables are X capital X suffix 1 to capital A suffix n and sigma square which is the population variance is unknown then we want to compare the mean value in the H naught case, it will be the mean is equal to mu naught under H naught being true with respect to H A which is the alternative hypothesis phase mu is equal to mu a which under the alternate hypothesis and mu a less than mu not so.

Obviously, it will be the left hand side and as you saw, it is something to do with the mean do you want to first answer the question whether is z distribution or t distribution, it is not z distribution because the variance is unknown. So, you use the t distribution. So, if we are using the t distribution remember that you are using s without the dash. So, coming back to this problem it is on the left hand side. So, you will take only the left hand one tailed value. So, only remember which I am again repeating. So, if it is on the left hand side.

(Refer Slide Time: 05:12)

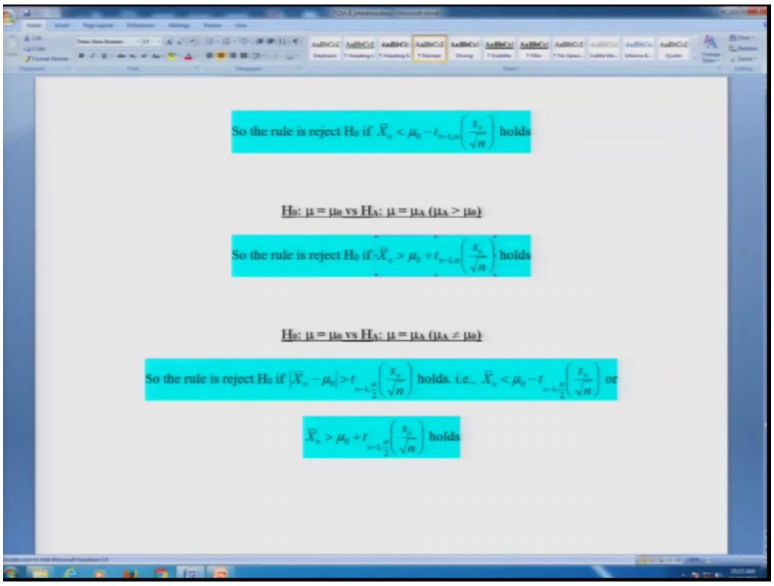


I am clicking here on the formula. So, this would be X bar n which is the sample mean for a sample size n being less than mu naught under H naught minus, but because minus sign would behave as you are looking at the left hand side, it will be minus t suffix n minus 1, as you have already lost 1 degrees of freedom number, second point importantly, it will be alpha because it is one sided and; obviously, in place of sigma you depress by S n and in case of square root of n, it remains a square root of n being divided in both the cases in the z distribution case, in the t distribution, in case it is on the if the hypothesis is basically H naught remains same, but H is greater than type.

So; obviously, you look on the right hand side, the diagrams you remember which we had already discussed it will be on the right hand side. So, in that case, the formula becomes X bar n is greater than mu naught plus because it is as it on the right hand side.

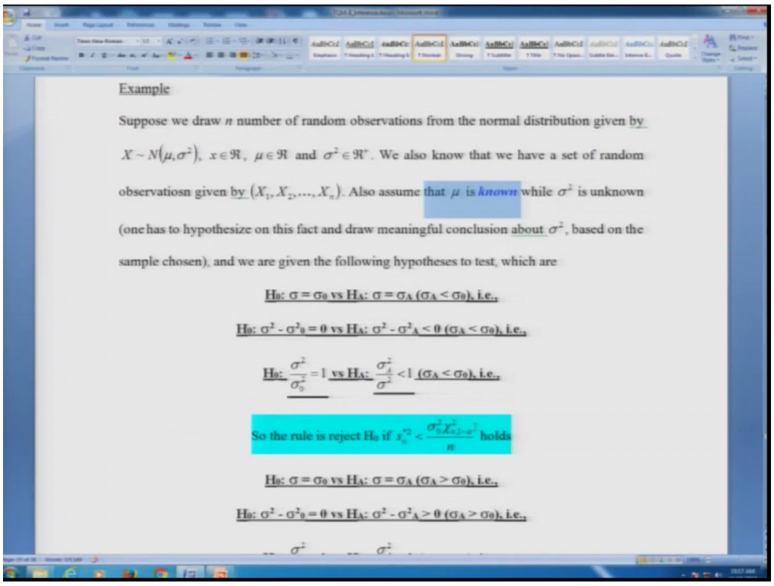
The value would bandit plus t suffix n minus 1 alpha into S n by square root of n, if it is on both sided which means H naught is mu naught and H is mu is not equal to mu naught so; obviously, it will be divided. So, the middle portion would be the whole overall area covered would be 1 minus alpha and the left hand side and right hand side, the total sum is alpha, but is equally divided and by alpha by alpha by 2 and alpha by 2. So, in this case you will reject H naught.

(Refer Slide Time: 06:36)



If this component is true that is the mod of the difference between the mean value the sample and its population value is greater than t n minus 1, n minus 1 remains same, but in case of alpha, it gets replaced by alpha by 2 and S n and square root by of n remains the same. So, if this is true; obviously, you would reject H naught. Now let us go to the case when you want to find out something to do with given the mean value is known so of the population.

(Refer Slide Time: 07:00)

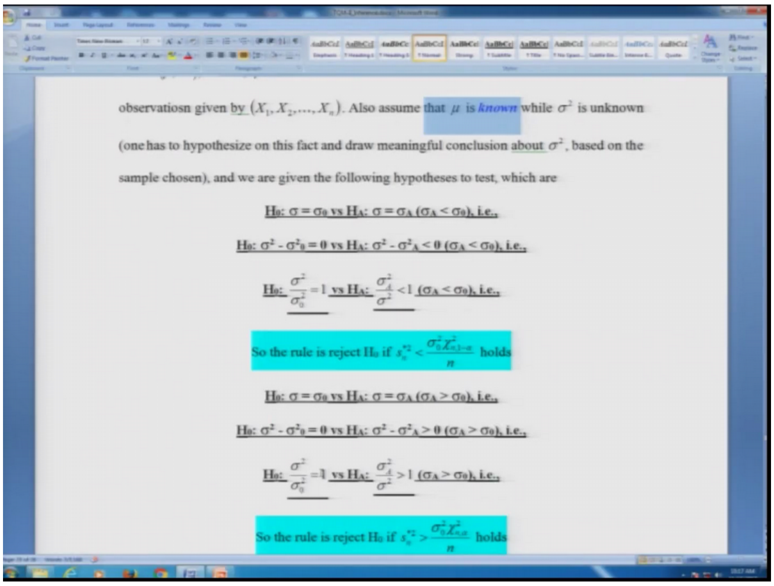


So, if it is something to do with the population variance; obviously, you will use that the chi square or the distribution, but it would as it is only one population, you will use chi square, but it would not lose any degrees of freedom because the population mean is known. So, so without going into the details because we had repeated a time and again, I will come to the rule and it will become very simple for you to understand.

So, your rule would be that you will reject H naught, if it is a left hand; left hand side hypothesis, why it is a left hand left hand sided hypothesis is that we will compare the way the ratios of the variances under H naught is equal to one and under H A, it will be less than 1 because of the less than type so; obviously, in that case S dash or s star whatever it is and this is S dash or s star because the population mean is known as mentioned here like in the blue colour where I am highlighting.

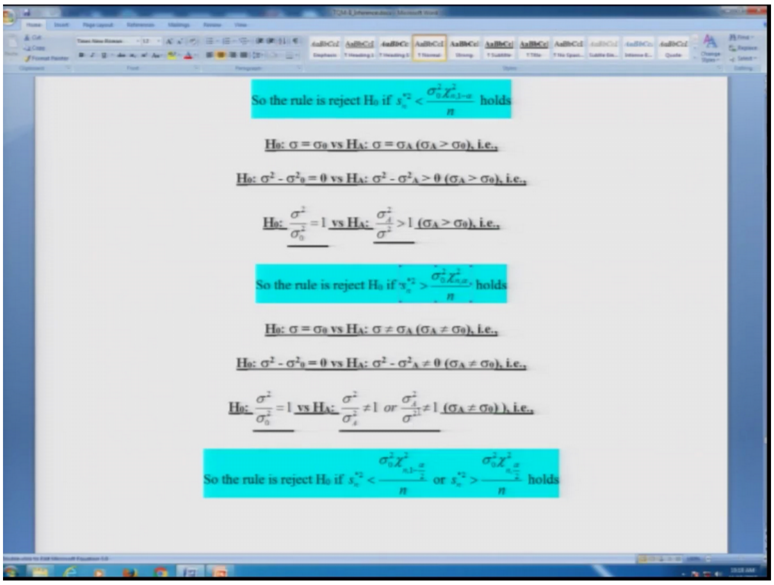
Now, so, this is less than sigma square not which is under H naught and it will be chi square n because you do not lose any degrees of freedom 1 minus alpha because on the left hand side and this is alpha naught alpha by 2 because it is one sided test divided by n in case, it is the hypothesis is of this form H naught remains the same.

(Refer Slide Time: 08:25)



H A is or the greater than type which is sigma square suffix A divided by sigma square is basically greater than 1 in that case you have on the right hand tailed test.

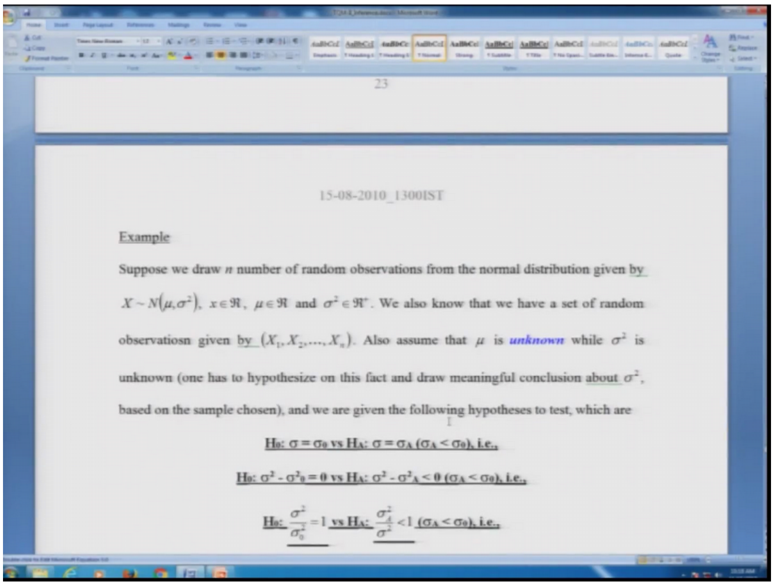
(Refer Slide Time: 08:36)



Which is S dash square is greater than sigma square naught by chi square suffix n and in this case, it is not 1 minus alpha alpha and divided by n remains as it is if it is both sided which means that you have H naught again the same, but in this case H A is when the ratios are not equal to 1 so; obviously, is 2 sided in this case, if it is 2 sided, you will reject H naught if both the left hand as well as the right hand or the right hand are true in the sense you have basically S dash being less than equal to chi square n 1 minus alpha by 2 into sigma squared suffix naught or 0 under H naught.

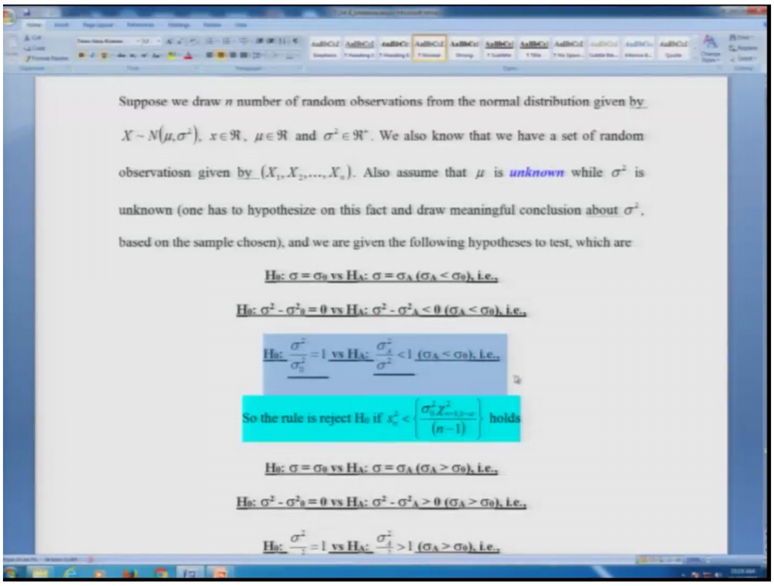
So, this is 1 minus alpha by 2 because it is 1 sided divided by n as it is and in the other case on the right hand side, it will be the everything remains the same only the greater than sign comes and in place of 1 minus alpha by 2, it becomes alpha by 2 for the suffixes of chi square, in case if you want to find out something to with the population variance given the mean value of the population being unknown.

(Refer Slide Time: 09:38)



Then the formulas which we just discussed are exactly the same with few changes and mark those changes very carefully.

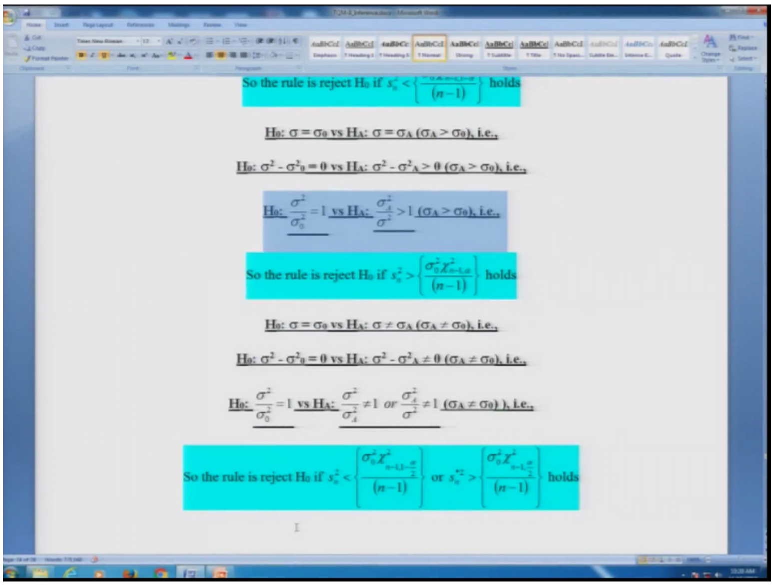
(Refer Slide Time: 09:55)



So, for the first case; hypothesis less than type so, stop pause think where are the changes S dash or s star gets replaced by s is that true let us check, yes, it is true. Now in the case, what happens to chi square, does it lose any degree of freedom, yes because you have already lost one degrees of freedom because you have used the population mean being not known being depressed by a sample mean. Hence, you have lost one degrees of freedom.

So, it will be chi square n minus 1 and also in the case of division being done by n, it is gets replaced by n minus 1. So, there are technically three changes S dash gets replaced by S chi square with a certain degree of freedom loses 1 degrees of freedom and the division on the right hand side becomes n minus 1 and not n, in case if it is a hypothesis of the greater than type again the same formula only the greater than size is replaced and as we as usual 1 minus alpha gets replaced by alpha as shown where I am basically hovering my cursor this.

(Refer Slide Time: 11:02)

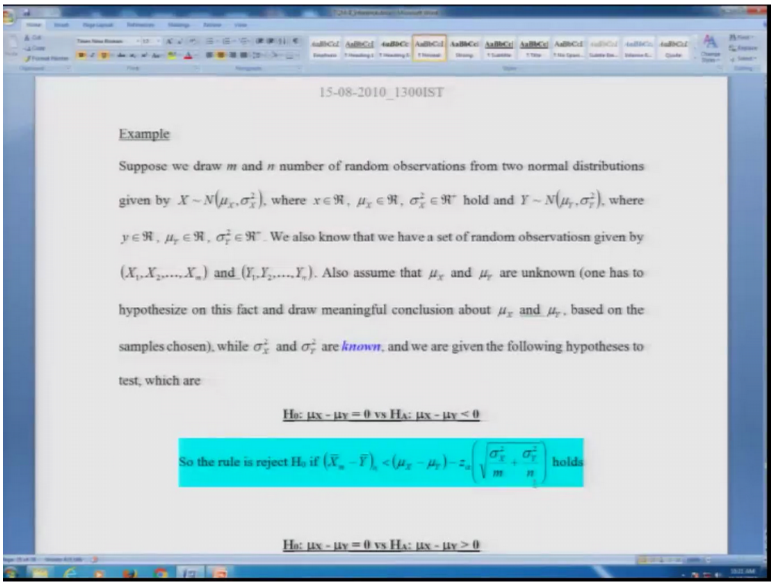


Now in case the rule says that it is of the of both tailed; that means, it is in H naught remain the same and H is basically not equal to in that case it will be s without the dash less than and as without the dash greater then, but in this case the same thing happens it in place of n, it gets replaced by n minus 1 in the denominator and the suffix is for chi square now becomes n minus 1 in both the cases and not n and; obviously, in place in the earlier case, it was 1 minus alpha now it because 1 minus alpha by 2 and the other case for the greater than type, it was alpha.

Now it gets replaced by alpha by 2. Now I want to find out something to do with the differences of the sample means considering the population means are not known and you want to compare them, consider the same example which I did discuss yesterday, the factories are producing same term of products in 2 different locations, they mean values weight or length or tensile strength or say for example, viscosity whatever the product you are trying to measure on an average, there is difference. So, if you think there is no difference.

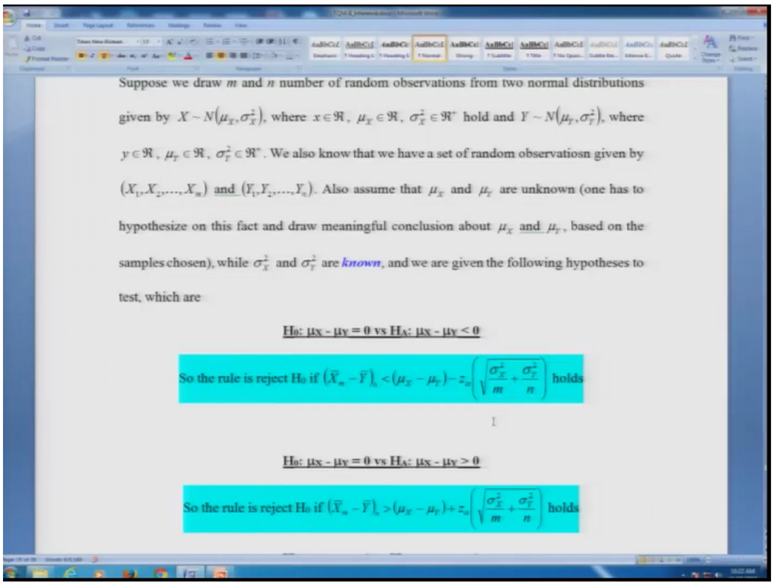
So, under the null hypothesis, they the difference between the mu 1 and mu 2 which is the mean value of population 1 and mean value of population 2 would be 0 and if you take up samples which is X bar m and X bar n m and n being the sample size which I am picking up from population 1 and population 2 when I try to compare them; obviously, it comes to the case that if the values are 0, then; obviously, it is matching with the population differences, but if it is not; obviously, you have to test the hypothesis.

(Refer Slide Time: 12:34)



So, in case, if sigma 1 and sigma 2 or sigma square X suffix X and sigma squares suffix Y for the 2 populations are known then; obviously, you will use the z distribution.

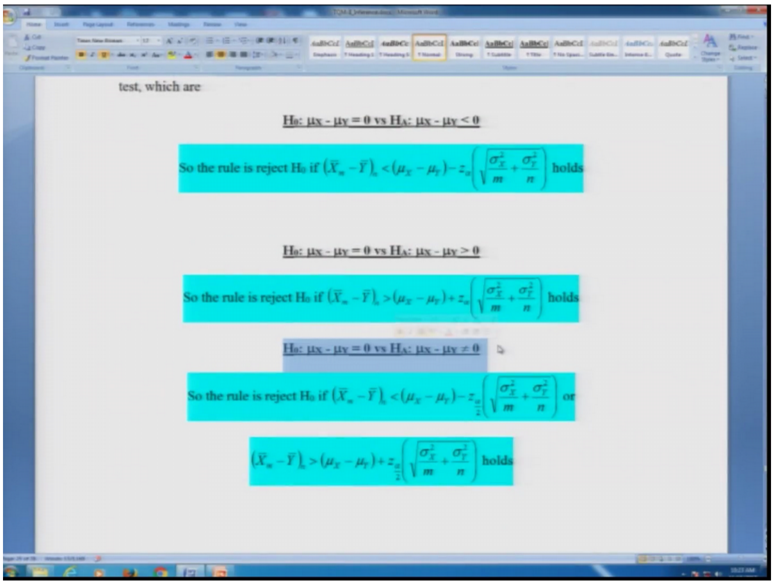
(Refer Slide Time: 12:43)



So, and the formula exactly the same the mean value of the difference of the mean value sorry this suffix n n should be below y. So, if the difference between the sample means is less than equal to the difference between the population means and; obviously, they were a minus sign because in on the left hand side z alpha comes as usual and the pooled so called variance or the variance of this 2 population becomes sigma square suffix x divided by m plus sigma square suffix n divided n sum them up and find the square root in the case, see for example, when you have the hypothesis of the greater than type.

So, what is the null hypothesis that difference between the sample means is 0 with respect to the case when the difference of the sample mean in or the population mean is sorry, population mean is less than 0 for the less than type greater than 0 for the greater than type in this formula everything remains same. So, only the less than type formula is replaced by greater than type and in front of z, it was initially minus, now it gets plus in the third example.

(Refer Slide Time: 13:46)



Where we have if we have say for example, the hypothesis null hypothesis remains the same alternate hypothesis, basically the differences of the means as not equal to 0, then ; obviously, you have in the case the values being not equal to 0 to be left hand side and right hand side which is 2 tailed. So, in this case, the sample mean is less than the difference when the population mean minus z alpha by 2 because this alpha is being equally divided on both sides of the z distribution, then the bracket part which is something to do with the pooled variance remains the same and in the other case or it is other things would be true if based on which you will reject H naught S when the difference between the sample mean is greater than the difference between the population mean plus z alpha by 2 and then multiply by the pooled sample population variance.

In case, if you want to find out something to do with now which I am just scrolling up I have not written the formula, but it will become clear to you in case, say for example, you have to do something to do with the excuse me hypothesis testing of the difference of the sample means given the population variances are not known so; obviously, you will be using in place of sigma square X and sigma square Y the corresponding C which is the standard error of sample 1 and sample 2 remembering the population mean is not known hence, you will basically lose 1 degree of freedom in population in the sample 1 1 degree of freedom in this sample.

Hence, the total degrees of freedom would be m plus n minus 2 where m is the sample size of from taken up from population 1 and n is basically the sample size taken from population 2 and you will basically have the rules accordingly law for the; then type it would be t distribution not z and for the greater than type again t distribution.

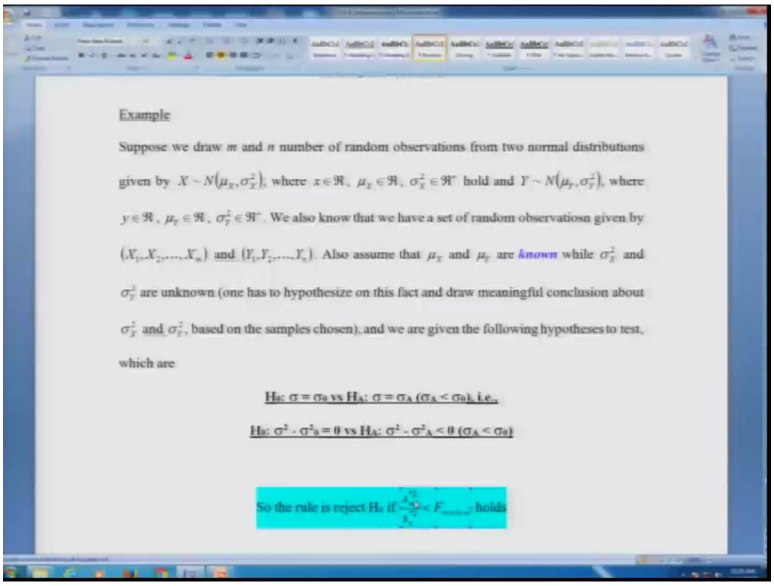
And for the not equal to type again the t distribution only changes being the distribution in all the three cases would basically lose to 2 degrees of freedom, one from the first sample, one from the second sample and in the less than type greater than type, it will basically be 1 minus alpha and alpha and in the not equal to type which is the 2 tailed, it will be 1 minus alpha by 2 and alpha by 2; so, the suffixes of t which I am talking about .

Now, I want to find out something to do with the variability of 2 products like in the same example which I gave, mean manufactured bottles are being filled up with jams and jellies carton boxes are being filled up dimensions of or different type of tubes are being measured and rather than finding all the mean values performance we want to find out the variability. So, if that is something to do variability, we immediately know that each has to something to do with the chi square or the f distribution.

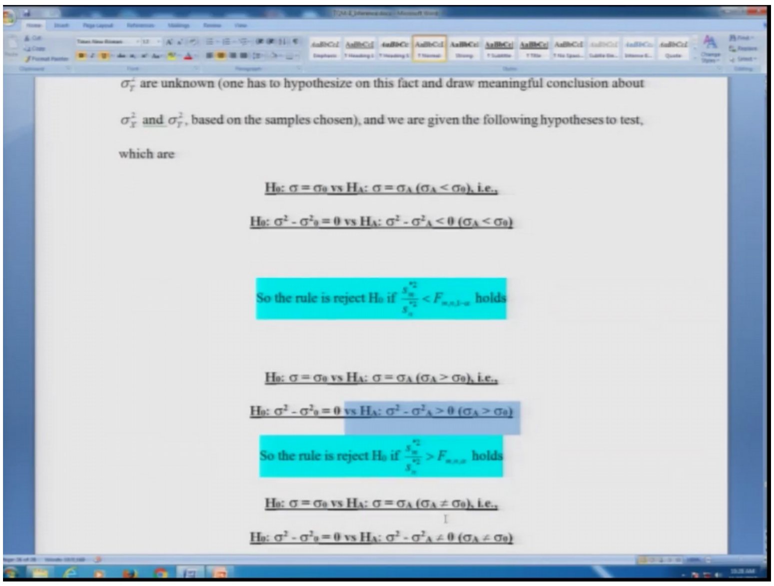
But now, we are trying to compare 2 different variabilities from 2 different populations considering their 2 different samples. So, in that case, we will use the f distribution, but now you have to pause for 1 minute, point 1, you will ask your question whether the population means unknown if it is known.

So, you know that they would not be any loss or degrees of freedom because you will be safely using s dashes in both the cases S dash suffix m or S dash suffix n depending on the sample size being m and n from population one and population 2. So, in that case, it will be f distribution suffix m comma n the m is in mango n is in Nagpur and very interestingly both the names are from synonymous to the same concept like mangoes and Nagpur's, I thought I have trying to mentioning that because that will make things very clear for you to understand the concept of m sample and n sample.

(Refer Slide Time: 17:35)



(Refer Slide Time: 17:40)

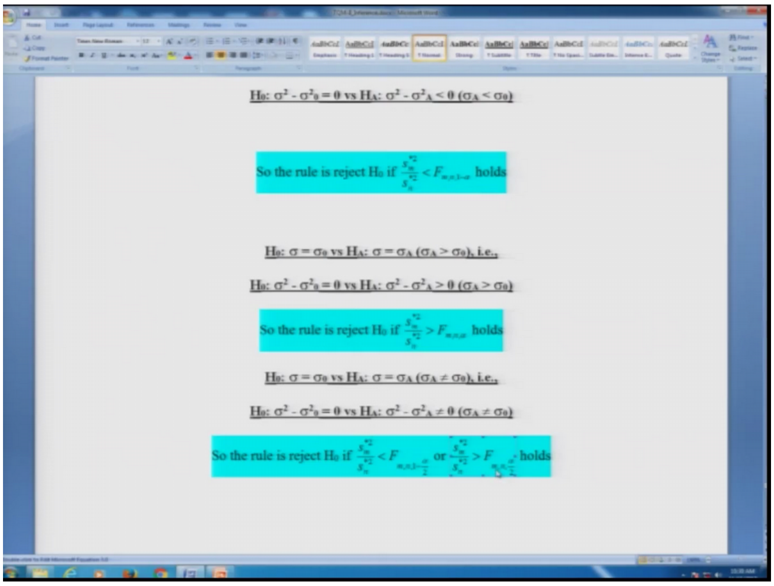


Now when we take something to do with the comparison of the sample variances; so; obviously, in the less than type so, how the less than type is being formulated the null hypothesis is that the differences between the variances or the ratios between the variances under the null hypothesis is either 0 or 1 expectedly, as the case may be and in the case when it is alternate hypothesis is the less than type the difference would be less than 0 or the ratios would be less than 1.

So, in that case, you will reject H naught if the ratio of S dash is for the first sample and the second sample is less than f m comma n minus 1 by alpha because it is the less than type for the case when you are taking the greater than type which means the differences is 0 under the null hypothesis and the, and or the ratio is one ratio means the ratios of the variances of population 1, population 2 and in the alternative hypothesis the difference is greater than 0 or the ratio of the variances for 2 population is greater than one in that case, you will reject H naught, if they if the ratio of sample variance of population 1 and population 2 is greater than f suffix m comma n comma alpha because you are on the right hand side .

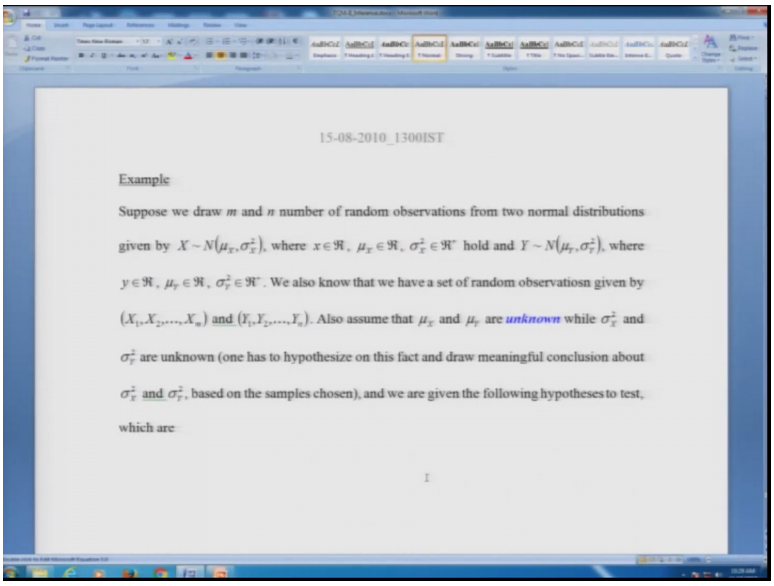
In case, if done our null hypothesis remains the same the alternative hypothesis is not equal to; that means, the differences is not equal to 0 or the ratios of the variances of population 1 and population 2 taking from trying to basically replace them with the sample variances if that ratio is not equal to 1.

(Refer Slide Time: 19:13)



So, you will reject H naught if these 2 conditions this or this are true in the in the first case, the variances ratios of the samples from some sample 1 and sample 2 corresponding to population and population 2 is less than f m comma n 1 minus alpha by 2 and in other case, it will be greater than f m comma n comma alpha by 2.

(Refer Slide Time: 09:38)



Now in the almost the last example which you are going to consider, it is again to do with them the sample variances a comparison the population variances. So, we use the sample variance, but we asked the question that whether the population mean is known it is not known. So, we can assume that we will definitely use s without the dash in both the samples and once we are assured of that and we also know that that will use the f distribution with losing one degrees of freedom, from both sample one and sample 2.

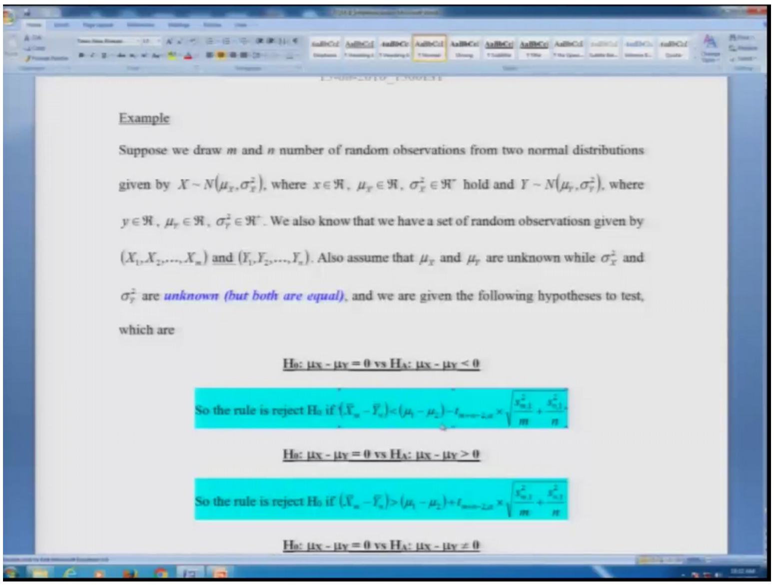
So, let us basically see the 3 cases, I am not going to repeat the hypothesis because they are exactly the same as for the example we just discussed. So, the null hypothesis remains equal to for all the three cases the alternate hypothesis consecutively R less than type greater than type not equal to. So, in this case, I will just mention them if I have seen, yes. So, I will mention them, I have not written them, but it will be easier for you to understand.

So, in this case, it will be s without the dashes the ratios of them less than f m minus 1 n minus 1 1 minus alpha by 2 in the greater than type would be s without the dashes ratios greater than f m minus 1 n minus 1 alpha by 2 and in the last case when it is not equal to it will be less than the ratios are s without the dashes ratios less than f m com[ma]- minus 1 n minus 1 comma 1 minus alpha by 2 and or it will basically be will reject H naught or the condition is s without the dashes ratios being greater than f m minus 1 n minus 1 alpha by 2.

Now, as I did mentioned; so, say for example, suppose we draw m and n observations from 2 different populations trying to basically find out the difference in the means corresponding to difference in the population means and we do not have the population means. So, you use the sample means, but only thing is that the population variances are unknown and they are unequal. So, the I will. So, first repeat the hypothesis thesis. So, the null hypothesis in all the three cases is the differences between the population means is 0 and the alternative hypothesis in the three consecutive cases are the difference between the population means is less than 0 difference between the population mean is greater than 0 and difference between the population means is not equal to 0.

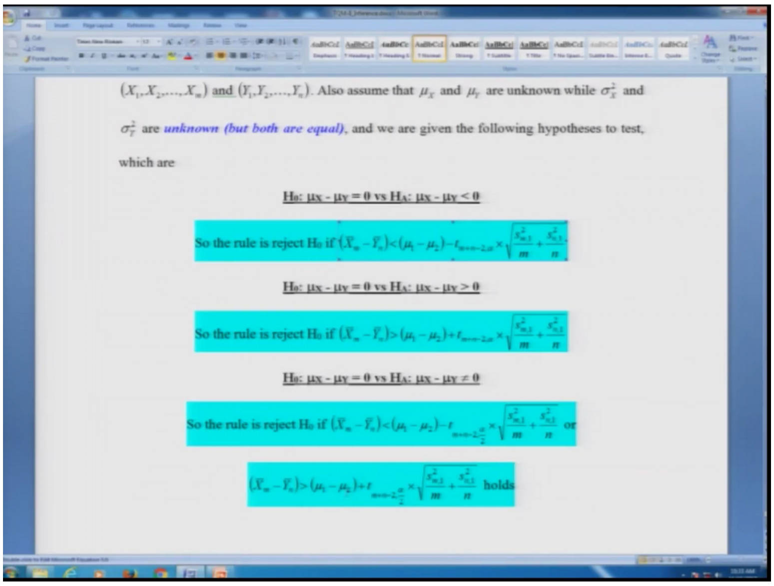
So, once you have that you will you will you will definitely first ask you a question that if as it is something to the mean should we use the z distribution or the t distribution. Now it is no more z because the creation variances are not known. So, if the partition variances not known you will use the t distribution, but when you are using the t distribution you will basically ask the question what I mean am I using am I using s without the dash or am I using s so; obviously, it is s because there you lose 1 degrees of freedom both for sample 1 and sample 2, hence, the t distribution would be with the suffix where it is m plus n minus 2 because 1 being lost as a degrees of freedom from both the population so; obviously, the rules for the less than type greater than type and not equal to type are as I mentioning consecutively like this.

(Refer Slide Time: 22:50)



The difference between the sample means being less than equal to the difference between the population means minus the t value whether t value is with suffix m plus n minus 2 comma 1 minus 1 minus alpha or alpha would not be any different for the t distribution because if you remember I did mentioned is symmetric, but only the minus sign is coming. So, it will be m plus n minus 2 and actually, it is 1 minus alpha multiplied by the pooled variance in the case of the greater than type.

(Refer Slide Time: 23:21)

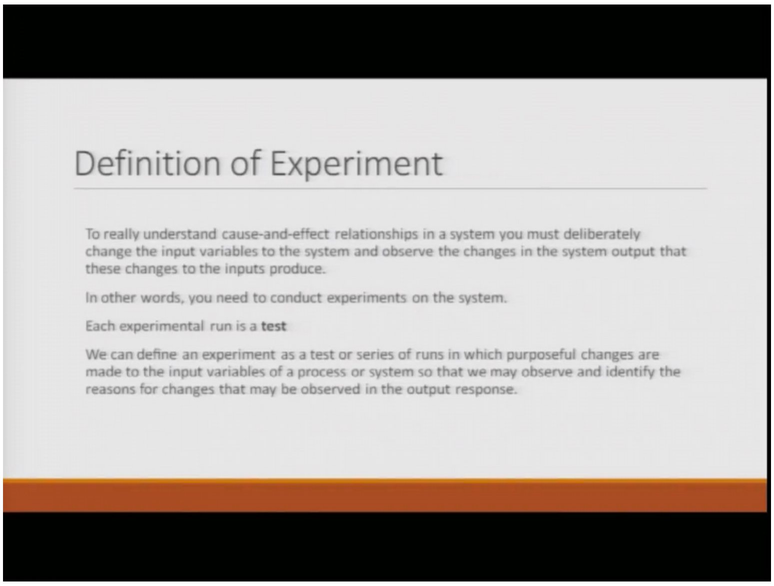


It is greater than equal to this is difference when the sample means is greater than equal to the population mean plus the same value of t, but now 1 minus alpha is replaced by alpha even though it is not written in the less than type multiplied by the pooled variance and in the case when it is not equal to type; obviously, it would be less than equal to or greater than equal to.

So, in the less than equal to H it is the difference between the sample means is less than then the difference between the population means minus t m plus n minus 2 comma alpha by 2 because is both sided again multiplied by the pooled variance and in the case it is greater than type for the not equal to type it will be greater than difference the sample mean is greater than the population mean plus t m plus n minus alpha comma alpha by 2 multiplied by the pooled variance.

Now, with this even though, I have gone a little bit fast please bear with me we will be coming back to this concept time and again and with this I will slowly start the concepts of say for example, analysis of variance would be very important and go slowly as we solve problems accordingly. So, now, we start with the analysis of variance or Anova, now first will basically describe it in detail with solved examples and slowly try to understand how the concepts which we have used like for the almost the 6 classes and almost half of the seventh one or lectures, but the concepts of hypothesis testing something to do with parametric point estimation interval estimation how they will be utilized.

(Refer Slide Time: 25:04)



So, definitions of basically what we mean my next experiment to really understand the cause and effect relationship in a system, you must deliberately change the input values to the system and observe the changes as they are happening in the system or output that these changes have been affected by the input change. So, say for example, you are trying to utilize some type of temperature control, sometimes a humidity control or some type of different type of coolant or a or a some say for example, different type of catalysts which are using for either for a mechanical process or for a chemical process for this examples which I have mentioned and you want to test that what is the change in the output is happening for the input.

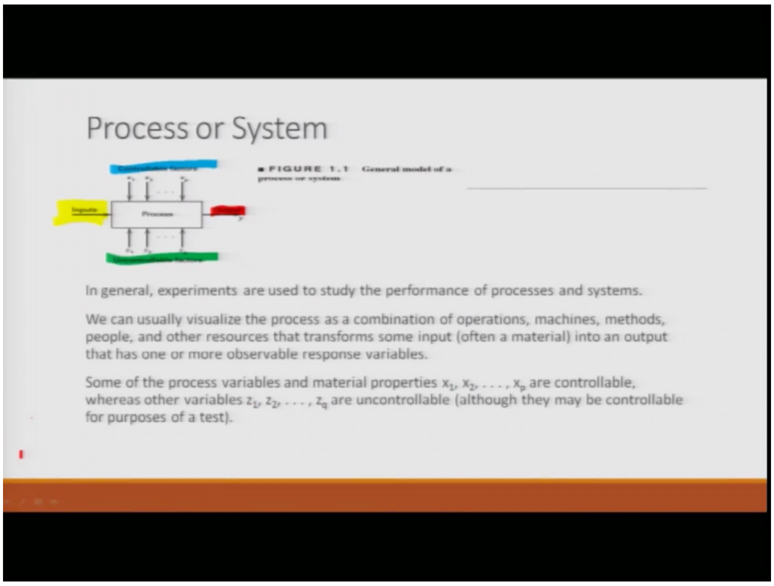
So, other examples which you consider say for example, the hypothesis hypothesis examples, we are considering and the last example which was not written, but it did discuss is that in Maharashtra the sugar cane is basically one of the biggest production of crop.

So, now you want to find out that using the change in the fertilizer or using the type different type of irrigation system what changes it will have on the output; output is basically the crop output, but acre or per kilometre square or plot hectare. So, consider this as I change the input it may be humidity it may be temperature it may be irrigation system it can be fertilizer all these things. So, or it can be some different type of mixture of soils which I am trying to utilize. So, based on that when I when I or it can be say for example, different type of initial input which I am giving may be some nutrient which I am trying to basically give to the soil or to the crop for higher yields.

So, based on those input changes I want to find out what are the output change output change is basically very simply for our example in this one which we discussed would be the increase or decrease in the per acre per acre or per hectare output. So, in other words you need to have content experiments on the system.

So, each experiment which we do to test would basically be considered to test, these type of tests would be considered not test. So, we want to conduct a test to find out what is the changes we can define an experiment as a test or a series of runs in which the purposeful changes are made to the input variables of a system or a process. So, that we may observe and identify the reasons for the changes that may be observed in the output or the responses. So, consider this.

(Refer Slide Time: 27:24)



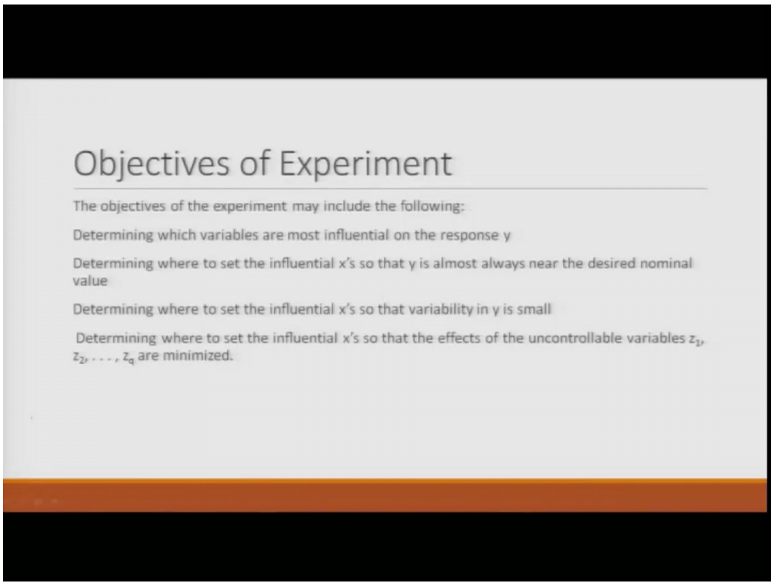
So, what we have. So, I will try to basically now highlight this. So, consider you have set of inputs which is yellow highlighted and there are basically 2 sets of variables which are there. So, I will try to highlight it again. So, one is the uncontrollable factors which you cannot control; obviously, as per the experiment and there are other sets of controllable factors and based on that you basically have an output.

So, your main task is to basically find out the comparison of the, of say for example, the input and output in order to basically understand how the whole system changed. So, with this figure let me understand and this has been taken from Montgomery book. So, all the figures with they have been given their due recognition and the IPR rights have been taken into consideration. So, I am just stating them as it is in general the experiments are used to study performance of processes and system we usually visualize the process as a combination of operations or machines methods peoples and other sources that transform some input often a material into an output that has one or more observable responses.

So, such of the process of variables on the materials are basically termed as X 1 to x p. So, which are controllable and z 1 to z q. So, this z is not nothing to do with the understand normal deviate which are uncontrollable although they may be controllable for the purpose of the experiment, but we will consider or for the test we will consider them to be not under control.

So, the objectives of the experiment would be the objective would be as following determine which variables are most influential in the response y y can be a scalar or a vector and; obviously, inputs are a vector starting from X 1 to X p or X 1 to X n depending on what terminology or nomenclature you are trying to use we will determine the where to set the influence of X s.

(Refer Slide Time: 29:27)



So, that Y is almost nearly as the desired normal value which we want determining where to set the influence of X s. So, the variability in y is also small.

So, there are 2 things we want to change the x values and we want to basically get the exact value which is something to do with the expected value, but we want to basically keep the variances as low as possible as at the same time trying to maintain the expected value would mean something to do with the consistency of the experiment if you remember we did mention that are trying to find out the unbiasedness and the consistency for the 2 important properties of finding or and will also determine.

So, these are not the main points, but only for the other absolutely very essential when you are trying to understand how the test is to be conducted and what are the important things we should remember we should also remember where to set the influences of X s. So, that the effects of the uncontrollable variables which is z one. So, z q are minimized and kept to the lowest level. So, with this I will send end the seventh lecture and continue discussing about the design of experiments slowly build up the overall atmosphere and the environment such that we will try to analyze the problems.

And also see that how later on, we will be utilizing all the concepts we just studied about the hypothesis testing and point estimation and something to do with the interval estimation and the concepts and the general framework will be utilized time and again in our design of experiments and studies and; obviously, there would be more concepts, but please bear with me we will slowly cover as and when they appear, we will discuss them in details and make it clear to all the students and all the people who are taking this course. With this, I will end the seventh lecture and let me wish you have a happy day and.

Thank very much for your attention, bye.