Data Analysis and Decision Making – I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 19
Hypothesis Testing: A

Very good morning, good afternoon, good evening, my dear friends welcome to this data analysis and decision making 1 course, under the NPTEL MOOC series and this is as you see this is the 19th lecture and this course is for 12 weeks 30 hours. Each week being 5 lectures, each lecture being for half an hour and the we are in the fourth week and I am Raghu Nandan Sengupta from the IME Department, IIT Kanpur.

So, if you remember in the end of the eighteenth lecture, we are discussing the table, where you had 2 statement, which is the H 0 and H A, which is alternative hypothesis being for H 0 H A and this H 0 being basically the hypothesis which you want to disprove, which is the actual statement based on the fact that what is the information, which is being given by the by the provider; that means, whose state when you are going to disprove, which is the null hypothesis h 0.

Now, as I also said that with respect to the mother nature and the information being provided by the person, who whom you are going to disprove and your plan of action also they can be such 4 different outcomes, in one cases can be where H, I will just use the simple words in order to make you understand consider H 0 being true actually and you are proving that H 0 is true. So, there is no error.

Another case is that H A is true and you are proving H A is true, there is no error, I am using the word H 0 and H A in the simple sense, the other 2 cases 2 cases are H 0 being true, but H A is being proved true. So, there is an error and the fourth one is H A being true and H 0 being proved true, there is another a, which is basically the alpha and beta. And what is alpha and beta? I will try to basically, explain that with a simple example.

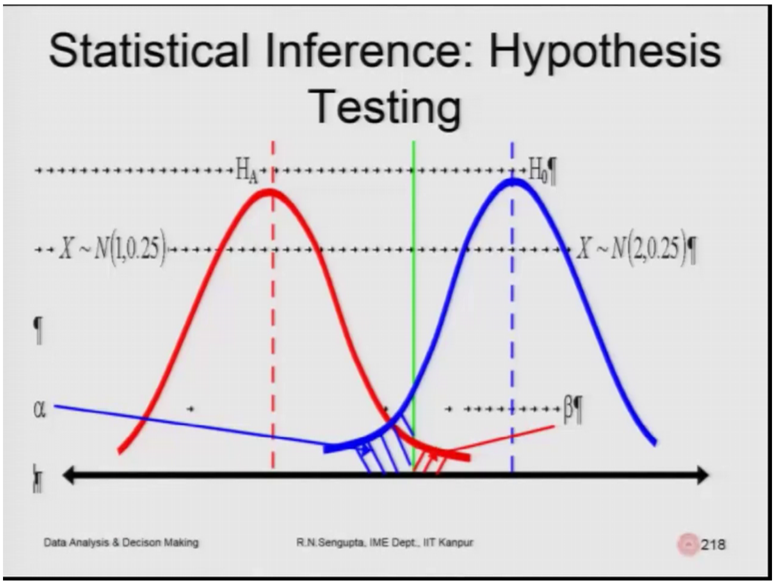
Now, in this hypothesis testing remember, whenever you are making the hypothesis testing, the fact remains that these outcomes on the rules, which we will consider are all coming from, the case of interval estimation and point estimation. So, there is no new concepts, which are being going to be discussed. It only the way, how you tackle the problem would be basically more relevant.

So, consider this case an, I will and basically, build up a story here. Consider you are a bank manager in a public sector bank on a private bank, whatever and you are the company on the organization on the bank. They have said that, you should disburse loans and the loans are such that, you would basically agree to the fact that giving the loans would mean, that you want to basically, have a set of information being provided by the customer so that, you can make your loan decision accordingly.

Consider, there was different type of questioners, when a question has being, what is the person's age profile, where is the person working, is it a government organization, is it a private organization, does he or she have a business, What is his or her recitation address, does he or she own a house, does he or she owned a car or does he or she owned a scooter, is the spouse of the person working, what is the, that does the person have a bank account in that bank or does the person has the person ever defaulted. And all these things and are being provided or asked from the person, who has asked for a loan and for each of these there are some points.

So, the, my bank manager, which is you. You provide some points for each of these questions and basically, collate the points add those points and if the number of points is basically, say for example, 60 and above or 70 and above. Consider whatever it is 60 and above and then the loan is sanctioned.

(Refer Slide Time: 04:27)



Statistical Inference: Hypothesis Testing

Data Analysis & Decison Making

R.N.Sengupta, IME Dept., IIT Kanpur

218

So, consider for the time being. The green line which is there in front of you is the line, which basically de-markets the point 60. So, any person on to the right, who score is there would be given that loan any person on to the left of the green line; that means, as you see this, this slide on to the left of the line would not be provided any loan.

Now, then you will be asking, what are these red and blue curves? Now each person, who is being who whose score is basically more than 60; so, technically that person would have different scores. So, if that there are different scores, we will mean that the distribution of the scores. For our case, we are considering a simplistic example, the person would basically or the set of persons, who the scores would basically be normally distributed.

So, the blue line which you see is the normal distributed score points for the persons, who would definitely be given a loan, now on the same context the set of persons, who would be denied the loan would also have a distribution. We will also consider the distribution to be normal and that normal distribution is again given by the red line.

Now, a red curve now remember this, blue dotted line vertical and the red dotted and vertical are the average scores, which would be there corresponding to the set of people, who are given the loan and the set and the red line, but dot line is basically, set is the average score for those set of peoples, who would not be given the loan. Now see the green line passes in such a way that, they are 2 areas and what are these 2 areas? I do not want to basically, mark anything because, it has been colored quite nicely. So, it is easy for you to understand.

So, consider this red hashed lines, this red hashed lines actually means these are the people to whom erroneously, we give the loan, but the probability of getting back the loan is 0, such that it would basically be a bad debt for you as the manager bank manager. On the other hand, the blue hashed lines which is on the left hand side, where I am hovering the pen are those though those set of peoples, who should have been given the loan, but they are denied the loan, which means that in the long run their opportunity cost lost, which is business loss.

Now, if I consider that these are both errors; that means, the red hashed line are those set of peoples who should not have been be given the loan, but they are given the loan and the blue hash lines are those set of peoples, who should have been given the loan, but they are denied the loan. So, one is basically bad debt and one in the second one is basically, opportunity cost lost. Now these are errors as I mentioned these are the alpha and beta errors, which you have just discussed in the table beforehand.

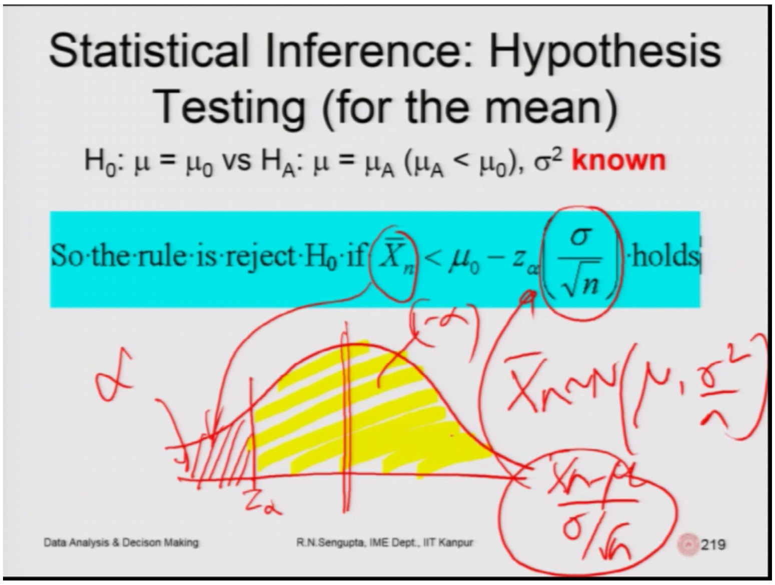
Now, see the peculiarity of this in this interesting example so; obviously, you will be tempted to reduce beta and both basically, reduce alpha also so; obviously, if both are 0, then your life is done because, you have never given out to a person, who would basically bad debt. Bad debt, the person will not return the loan and the blue one; obviously, although set of peoples, you will definitely like to give the loan because, that would or else it would be opportunity cost lost for you.

Now, consider the green line is shifted vertically either on to the left or the right. So, it is and I am basically talking from my own perspective. So, it is basically moved to the left, it would mean that beta increases and alpha decreases on the other hand you try to basically shift at the green vertical line onto the right. So, in that case beta decreases and alpha increases.

Now which means that, trying to basically reduce both of them at the same time is not possible. So, what you will try to do from the statistical perspective is basically, try to minimize the sum of alpha and beta and then basically, try to solve the problems accordingly. So, when you are trying to basically do the happiness testing, you will keep the values of beta fixed at a certain level and then basically try to understand, what is the value of alpha and if you remember that alpha is basically related to 1 minus alpha, which is the level of significance.

So, in this problem, it basically states and the table also which we saw earlier is that there would be 2 type of errors alpha and beta and we will try to basically, see that how it can be reduced in how the problem can be stated in such a way, that given the set of information form beta, we will try to basically find that what is alpha and based on the alpha, we will try to do the experiment accordingly.

(Refer Slide Time: 09:12)



Statistical Inference: Hypothesis Testing (for the mean)

H_0: $\mu = \mu_0$ vs H_A: $\mu = \mu_A$ ($\mu_A < \mu_0$), σ^2 **known**

So the rule is reject H_0 if $\bar{X}_n < \mu_0 - z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)$ holds

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 219

α
$(1-\alpha)$
Z_α
$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$

So let us basically, consider the now the rules and remember the rules are exactly what we have done in the interval estimation. So, there is no change and even if they change, I will basically make all things clear to you and once the rules are basically covered, I will come to the problem solving later on. So, consider the first rule in this case, you are trying to basically test the hypothesis a null hypothesis that mu is equal to mu 0 versus the case that in under H A, which is the alternative hypothesis, the mu A value is less than mu 0, which means that mu A value would be on to the left of H mu 0.

So, now I will basically draw the diagram in order to make you understand. Now when you and I will use the red color here, so, what you want is certain portion on to the left and you want to basically prove or disprove this and consider this distribution is normal. So, the mean value for the normal distribution under H 0 would be basically mu 0.

So, what do you want to find out is that that mean value, which is given you will try to basically, find out that this is you ask a question. Yes this is a test something to do with the mean. So, if that is the case; obviously, you know that it will be either the Z distribution or the t distribution then, you ask the question is the standard deviation of the variance of the distribution known? The answer is yes, if the answer is yes; obviously, use the Z table.

So, now if it is Z table, what you will try to do is that, you will try to basically find out the LCL, in that case such that any X bar mean, which is the mean value of that sample, which you are taking up is on to the left hand side. You will basically reject H 0 based on the overall problem formulation, which has been done. So, what you want to do is that that, you will try to basically find out the mean value for the sample. If the mean value of the sample is lying on the left hand side; so this is what. This is H 0, which is here and this value sigma by square root of n is basically based the fact, that that the X bar n, which is the mean value of the sample is distributed normal with a mu value of as the mean of the expected value and the variance been sigma square by n.

Hence basically, when you convert into a Z normal value, it will be X bar n minus mu by sigma by square root of n. So hence this is going here. So, if mu x minus and now where does that Z alpha combine? And remember that I am harping on the fact that is Z alpha. So, it will change depending on the example, 2 things will change number 1, whether Z is the case or it should be T because, you remember that T would be the case, when the standard deviation on the variance of the population is not known point one and; obviously, t would have is own corresponding degrees of freedom, which is not there for Z.

The second thing is remember that that alpha value, which you are putting in now many of the cases it, would be 1 minus alpha or alpha or alpha by 2, depending on the example. So, these words may not be making any sense to you immediately, but as you see the diagram, it will become clear to you these diagrams, I will basically draw as you proceed with this or problem solving or explaining the concept.

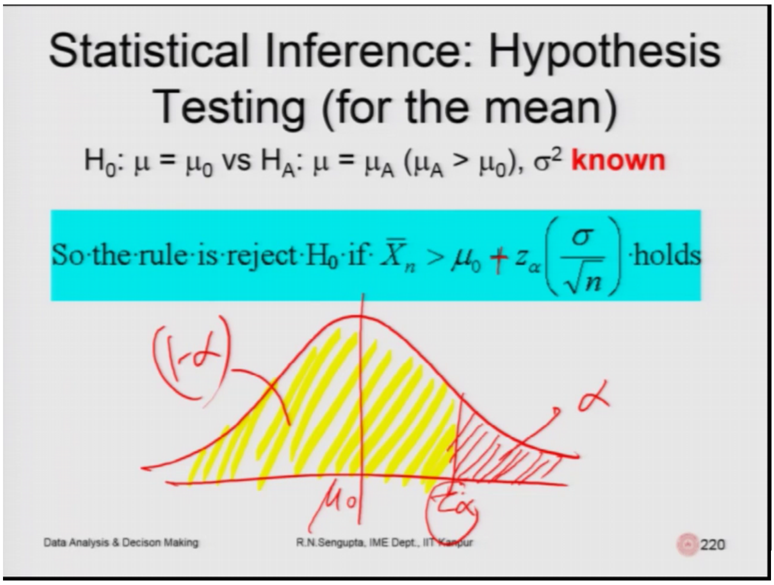
Now, in this case X bar n, who should be less than less than on to the left hand side of mu 0 minus this value; that means, you are finding out some Z alpha here and this Z alpha is minus so; obviously, this would be how far the standard deviation is from the is from the mean value and the value of Z alpha basically, gives you what is the level of confidence which you have.

Now, the level of confidence is very important let me consider. So, if you are talking about the level of contra-variance, which is 1 minus alpha. So, if this whole area is alpha, then in this case, the area on to the right would basically be 1 minus alpha. So, the whole area which you come covering under the distribution, one part is alpha which is left and the right hand side is basically, 1 minus alpha.

So hence, you will basically, the rule is that you will reject H 0, if X bar n is less than mu 0 minus Z alpha sigma by square root of n. Now before you proceed to the next rule and basically build up the story here, now consider that you want to find out that, the null hypothesis is mu is equal to mu 0 versus alternate hypothesis is mu is equal to mu A, but the fact is that mu A is on to the right hand side, which means that is greater than H 0. So, what do we do?

So, in this case; obviously, you have to consider the right hand side, which is where I am hovering the pen and the rule would be like this.

(Refer Slide Time: 14:33)



Statistical Inference: Hypothesis Testing (for the mean)

H_0: $\mu = \mu_0$ vs H_A: $\mu = \mu_A$ ($\mu_A > \mu_0$), σ^2 **known**

So the rule is reject H_0 if $\bar{X}_n > \mu_0 + z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)$ holds

Data Analysis & Decison Making

R.N.Sengupta, IME Dept., IIT Kanpur

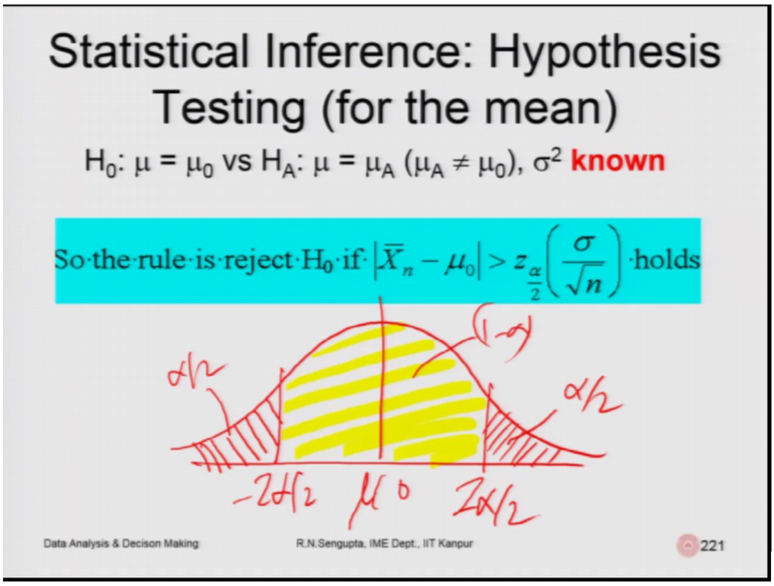
220

So, the; this is mu 0, this is the value, this is Z alpha and this will be 1 minus alpha, which is the level of significance. This is mu 0, this values alpha, this value is 1 minus alpha.

So in this case, you will try to basically find out and reject H 0, if X bar n which is the sample mean is on to the right hand side of the upper control limit, which is mu 0 plus Z alpha into sigma by square root of n, this value of sigma by square root of n is constant, this Z alpha is now here, this plus my plus sign is coming due to the fact, this is on to the right hand side in the other case, it came due to it was on the left hand side.

Now, the third problem: so, if the, what you are seeing are exactly the lower control limit and the upper control limit for the less than type and the greater than type. Now the third question for the similar type of problem, for finding out something to do with the mean were provided. The mu value as shown here, where I am basically pointing my finger is that, that sigma square from the population is known what happens to the problem, if now we frame the problem as H 0 mu is equal to mu 0 under H 0 and under H A mu is equal to mu A, but mu A is not equal to mu 0; that means, it can be either on the left hand side or the right hand side.

(Refer Slide Time: 16:24)



Statistical Inference: Hypothesis Testing (for the mean)

H_0: $\mu = \mu_0$ vs H_A: $\mu = \mu_A$ ($\mu_A \neq \mu_0$), σ^2 **known**

So the rule is reject H_0 if $|\bar{X}_n - \mu_0| > z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$ holds

Data Analysis & Decison Making

R.N.Sengupta, IME Dept., IIT Kanpur

221

So, how do we frame the problem? Now let me draw the diagram.

This is mu 0; this is minus Z alpha by 2. Remember this is by 2 by 2, this I am harping time and again and I will come to that again, this is Z alpha by 2. So, this area is alpha by 2, alpha by 2, this area is now 1 minus alpha.

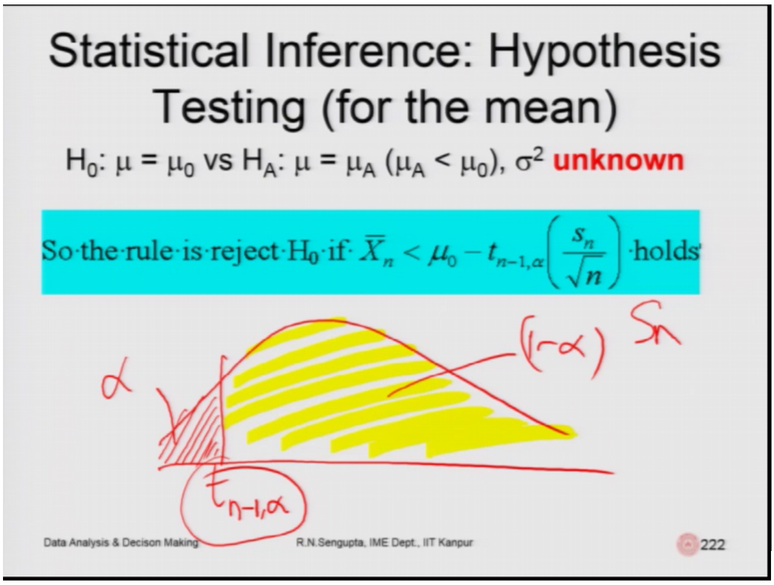
Now, see the problem here, problem is not the problem that we are facing, but how we are basically able to frame it. Now in this case as I said mu is equal to mu 0 and under and under H 0 and under H mu is equal to mu A, but mu is not equal to mu 0 and sigma square, which is the population variance or the standard deviation square for the population is known.

So, in this case you will reject H 0, if the sample mean is to the is basically, given the fact that it is either on the left hand side or the right hand side or it can you be the complementary set of it is not in between. Now the values, which you have put for Z alpha by 2 are being put in such a way, that their right hand side of the area, on to the right hand side of plus Z alpha by 2 is alpha by 2 and you are basically, considering a symmetric distribution and the left hand side of the area, on to the left of minus Z alpha by 2 is also alpha by 2.

So whole the sum is basically, alpha by 2 by n alpha by 2 is alpha and; obviously, the middle area, which is the level of confidence, which you are already formulated depending on the problem formulation, which you are already seen for the interval estimation would be 1 minus alpha. So in this case, what we do in these 3 examples 3 not the examples, I would not say 3 different formulations of the same problem is that in one case, it is less than time. The second case, it is greater than time and the fourth third case is basically, not equal to and you basically utilize the same for problem formulation for the interval estimation. But remember that for the less than time, it will be minus Z alpha for the greater than time, it will be plus Z alpha and for the not equal to it will be minus Z alpha by 2 plus Z alpha by 2, such that the interval which you have is basically, 1 minus alpha, which is the level of significance and that can be either on to the left hand side or on the right hand side or it can be in the middle depending on how you have basically, been able to formulate the problem.

Now, the question is that we are going to answer now if these things cleared you will basically ask that what happens and how do we frame the problem for the hypothesis testing? If the case is that we want to find out something to do with the mean of the population provided the standard deviation of the population on the variance of the population is unknown, then do we use the same distribution Z? The answer is no, what distribution will we use? We will use the t distribution and; obviously, there would be corresponding degrees of freedom, which are basically, now explain.

(Refer Slide Time: 19:53)



Statistical Inference: Hypothesis Testing (for the mean)

H_0: $\mu = \mu_0$ vs H_A: $\mu = \mu_A$ ($\mu_A < \mu_0$), σ^2 **unknown**

So·the·rule·is·reject·H_0·if· $\overline{X}_n < \mu_0 - t_{n-1,\alpha}\left(\frac{s_n}{\sqrt{n}}\right)$ ·holds

α
$(1-\alpha)$
S_n
$t_{n-1,\alpha}$

Data Analysis & Decison Making

R.N.Sengupta, IME Dept., IIT Kanpur

222

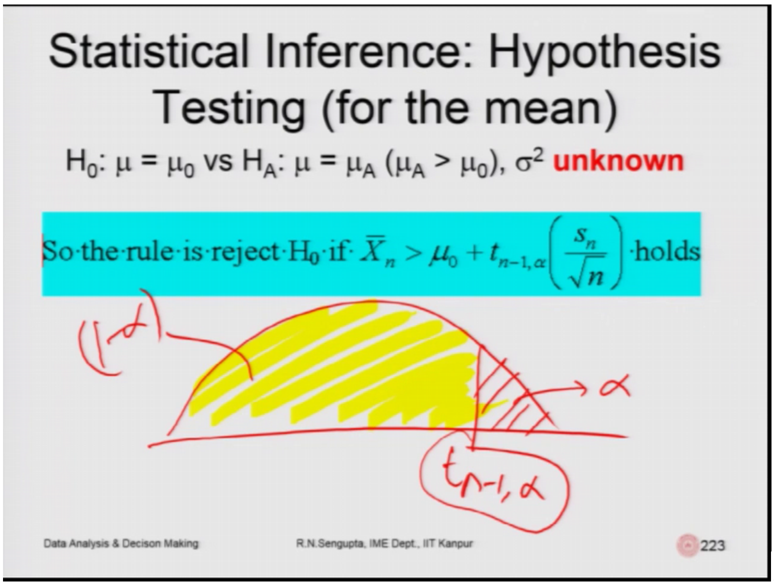
Now, consider the hypothesis testing as H 0 mu is equal to mu 0 versus H A which is mu is equal to mu A, but mu is less than mu 0 and in another case, that sigma square is unknown; that means, the population variance is unknown. So, in this case again, you have the same type of distribution, I am using the figures are not important, what is important? Is basically where the values lie and which side?

So, consider this is the t distribution. This now would basically, do with the t value. So, in this case it if X bar n is on to the left hand side let me. So, this area would basically be alpha, this is a level of significance which you have and this would be given by n minus 1, remember that that is important and alpha because, alpha is basically on to the left hand side, why it is n minus 1 alpha? Because you have long you have you utilized 1 degrees of freedom hence, as your lost 1 degrees of freedom trying to find out the first moment, which is the mean hence t distribution would basically, would have lost 1 degrees of freedom, which is basically now become n minus 1.

So in this case, the sample mean should be less than equal to the population mean, under H 0 is mu 0 minus. The value of t n minus 1 into alpha, this is this value, which I have I am just giving the concept without the sign for t is t n suffix n minus 1 alpha and S n, why S n? Because you have already not you have do not have any information of the population mean. So, you will use the sample mean as the best estimate for the population mean and as you have used the set of observations want to end, for the first time to find out the best estimate for the population mean using the sample mean, hence the degrees of freedom is reduced by 1 and; obviously, this would be square divide by square root of 1 as the formula it is.

Now, the question would arise that, what is the hypothesis for the case, when it is greater than type. In that case H 0 under H 0 you have mu is equal to mu 0 under H mu is equal to mu A, but is greater than H 0 and sigma square, which is the population variance is unknown and the simply in this case, it is on the right hand side.

(Refer Slide Time: 22:24)



Statistical Inference: Hypothesis Testing (for the mean)

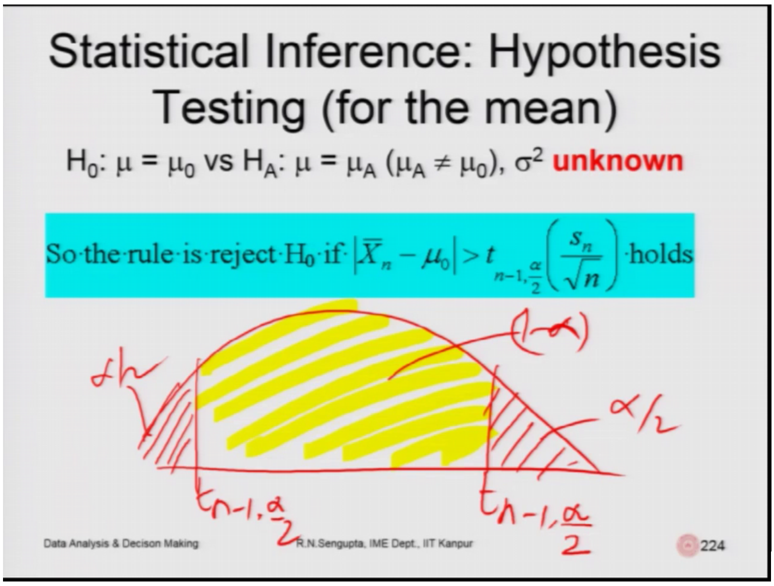
H_0: $\mu = \mu_0$ vs H_A: $\mu = \mu_A$ ($\mu_A > \mu_0$), σ^2 **unknown**

So the rule is reject H_0 if $\bar{X}_n > \mu_0 + t_{n-1,\alpha}\left(\frac{s_n}{\sqrt{n}}\right)$ holds

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 223

So, this would basically be t n minus 1 alpha, this area is alpha, this area is 1 minus alpha. So basically, utilized in such a way that the left hand side is 1 minus alpha, which is the degrees of freedom. In this case, other it was degrees of freedom, in degrees of freedom is basically n minus 1, the level of signal is A 1 minus alpha and here t n minus 1, would basically be the statistic based on which, you will basically do the test. So, this is a greater than type and this sample mean should be on the right hand side, that is mu 0 under H 0 plus that, the value as you moving on to the right hand side, depending on the level of significance, that value is t n minus 1 alpha into S n by square root of n.

(Refer Slide Time: 23:27)



Statistical Inference: Hypothesis Testing (for the mean)

H_0: $\mu = \mu_0$ vs H_A: $\mu = \mu_A$ ($\mu_A \neq \mu_0$), σ^2 **unknown**

So the rule is reject H_0 if $|\overline{X}_n - \mu_0| > t_{n-1,\frac{\alpha}{2}}\left(\frac{s_n}{\sqrt{n}}\right)$ holds

Data Analysis & Decison Making

R.N.Sengupta, IME Dept., IIT Kanpur

224

Third similar type of problem under this concept is, what is the rejection rule of the on the rule. Based on the fact that under H 0 mu is equal to mu 0 under H A, mu is equal to mu A, but mu A is not equal to H 0 and sigma square, which is the population variance is unknown. So again, same t distribution, I am trying to basically draw the diagrams in order to make you understand.

So, these are the t values. So, this is t n minus 1, t n minus 1. Now look at this important thing, this whole area would be alpha by 2, this whole area on to the right is alpha by 2. So hence, that total addition is alpha by 2 by alpha by 2 and hence, the degree the level of significance for this is alpha by 2 and this is alpha by 2, such that the overall area, which is in between is 1 minus alpha which is level of significance.

So here, basically trying to redefine the lower control limit and the upper control limit, in the lower control limit, it is mu 0 minus and what does the minus value is two n minus 1 comma alpha by 2 into S n. S n remains the same because, you are utilizing the S without the dash provided, you have lost 1 degrees of freedom because, the population mean is not known, you utilize observations from the sample to in order to estimate the population parameter hence, you have lost 1 degrees of freedom divided and here the S n would be divided by square root of n.

In the other case, when you go to the right hand side, everything remains the same, it is underneath the upper control limit will be mu 0 plus t n minus 1 comma alpha by 2 into S n divided by square root of n. So, we have already seen that the rules that how you will basically, formulate the hypothesis testing rules provided, the sample mean was being utilized in order to basically, find out some hypothesis statements related to the population mean and in the first 3 instances, because the rules are 3 for the case, when the variance was known. So, you use the Z distribution once, on to the left hand side they use Z alpha, I am just harping on the statistic, Z alpha with a minus sign; obviously, because on to the left.

The second case is that, and the first case; obviously, in that case, it was the H 0 was less than type. In the second rule, under the case when you want to find out something to do with the hypothesis testing for the population mean provided the variance was known and it was H 0 was of the greater than time, then the rule was basically, something to do with on to the right hand side, where it was Z alpha and; obviously, minus and plus would come, if they are less then time and greater than time, when the rule was basically to try to find out something to do with the null hypothesis and the alternative hypothesis was H A, where H A basically, states that mu is not equal to mu not. So, in that case the Z values becomes minus Z alpha by 2 and plus Z alpha by 2, because the left hand side area and the right hand side area are equally dispersed. So hence, the sum is basically alpha and the level of confidence, which you have in between, is 1 minus alpha.

Then we basically, went to the fact that we want to find out something to do with the hypothesis testing corresponding to the sample of the population mean using the sample mean, but the variance of the population is unknown hence, we will basically utilize the case that, we are used using the t distribution, but losing 1 degrees of freedom because, the population being not known hence, you will try to utilize the sample mean as the best estimate and there we basically, find found out that, if it is less than type for the case for H 0, where mu is less than mu 0 provided the population variance is unknown, we use the t distribution minus t n minus 1 comma alpha and; obviously, it will be S n divided by square root of n.

When the case, the second rule was basically H 0 where H mu A is greater than type of greater than mu 0; so, you will basically, utilize the plus value and they shall be t n minus 1 comma alpha into S, S divided by square root and n, I am only harping on the right hand time. The other values, where mu 0 remains the same and in the third rule basically, we saw that if it was not equal to rule; that means, under H 0 H A, if mu A was not equal to H mu 0. So, in that rule you basically, have the left hand side and the right hand side. Left hand side basically would be t n minus 1 alpha by 2 into S n divided by square root of n and in the other hand, it will be for the right hand side, it will be plus t n minus 1 alpha by 2 into S n divides divided by square root of n.

So; obviously, we will see that, how we can formulate the problem for the standard deviation, how we can form where the problem provided the mean value being known on are not known, we will continue discussing that in the twentieth lecture, which will basically be the last lecture for the fourth week have a nice day and.

Thank you very much.