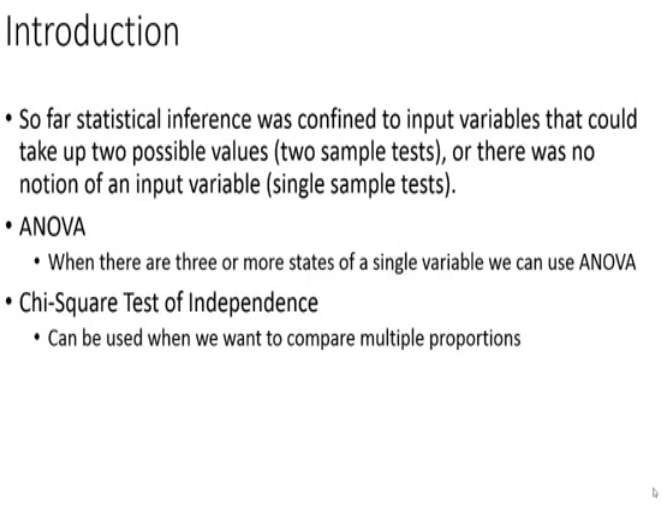
Introduction of Data Analytics
Prof. Nandan Sudarsanam and
Prof. B Ravindran
Department of Management Studies and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 03
Lecture – 14
Inferential Statistics – ANOVA and TOI

Hello and welcome to our lecture on ANOVA or also known as analysis of variance; and, that chi square test of independence – the TOI is test of independence – the chi square test of independence. We present this as the last lecture in the series on inferential statistics. But, it is only fair to say that, while some set of techniques are presented is a part of inferential statistics. Most of the techniques in statistics use statistical inference in some way or the other. So, for instance, after this lecture, we will be talking about regression and so on. And, there is a significant component there, which is associated with statistical inference. But, in any case, this lecture, we are going to talk about the analysis of variance and the test of independence.

(Refer Slide Time: 01:01)



So far, statistical inference was confined to input variables that could take up two possible values; and, that was the two sample cases – the two sample tests, where you would compare two different samples. And, the variable of interest would be for instance, either male or female. So, we came up with an example, where we said the

heights of tenth standard boys in public schools versus the heights of tenth standard girls in public schools. So, if you were interested in seeing if the average height of a boy in tenth standard in public schools in India is equal to the average height of a girl in tenth standard in public schools in India; here the output – essentially you can think of there are two variables that are involved and you can think of the output variable as one of height and your input variable you can think of as gender. So, you have two different sets of data and you are comparing them.

Now, there are also these cases, where it does not make sense to think of these as input and output variables; that is just confusing. So, and, these are the single sample test cases, where you just have one variable height and you have defined all the other parameters around it. And so, it could be something like I have already defined that I am interested in studying the average height of boys in tenth standard in public schools in India; and, I am interested in seeing that average height is less than a 120 centimeters or something – some such number. There is no concept of it; there is just one variable, which is height and that is it and you are comparing it to some number that you had in your head. So, that is essentially summarizing the two sample case and the single sample case.

For the first time with something like the ANOVA, we take on a case, where this variable, which is essentially like this input variable, can take on two or more states. So, essentially, you will see it taken three or more states. And, that is the real differentiator about ANOVA. So, just to kind of give you a little bit more color on this, a standard application of ANOVA would be something like the following, where you are still dealing with an output variable that is typically quantitative and continuous like for instance, like height. But, your input variable need not have just two states, it can have three or more states; and, you are still probably interested in comparing their average. So, extending the example that we saw in the two sample t-test case; if you wanted to ask a question such as – is the average height of tenth standard boys in Tamil Nadu equal to the average height of tenth standard boys in public schools in Maharashtra; is it equal to the average height in Karnataka.

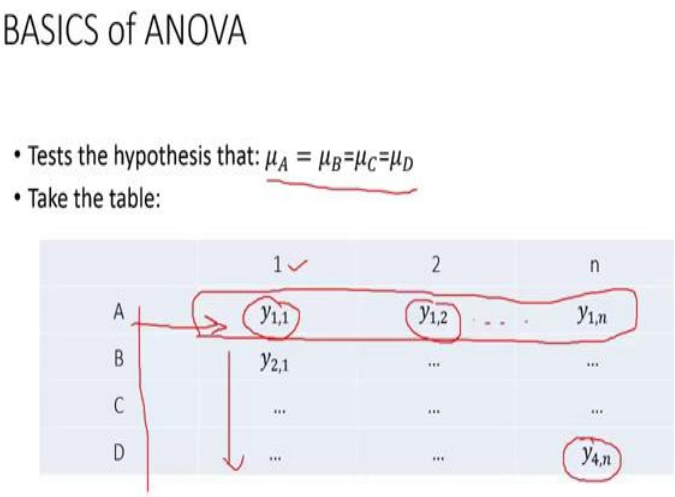
So, you are not just interested in comparing two sets of samples; but, you had multiple sets of samples. And typically, what you are comparing, the output essentially, variable is still like a continuous quantitative variable like height. It can be anything else; but, essentially, it is a quantitative variable. And, you have multiple discrete settings of your

input variables. So, you are not just interested in comparing boys versus girls or method A versus method B; but, you are now going from… In the t-test, you are comparing method A versus method B; you can now say is method A equal to method B equal to method C equal to method D or are they different, or a some subset of them different from the others. So, that is essentially an application of ANOVA.

Now, the chi-square test of independence is one that can be used when you want to compare again multiple proportions. When you are interested in two variables; and, both of these variables are categorical variables. So, it could be the same thing like Tamil Nadu, Karnataka. So, the different states of India could be one of the variables. And, the other variable could be the number of people at different age groups. So, people between – number of people between 0 to 20 years – between 20 to 40 years and 40 to 60 years. So, there is one variable, which is the state, which is a categorical variable. And, here is another variable age, which I have in some sense made categorical. But, essentially, I am just interested in these two variables, which are categorical and I am just interested in looking to see there is any relationship between them.

Another way to think of chi-square test of independence is see it as some form of an extension of your proportion z-test. In a proportion z-test, you would typically handle problems of the following nature. You will say if you took a question like our men more likely to purchase a certain product than women. That would be a proportion z-test, because you have two categories: men and women. It would be – has to be precise; it would be a two sample proportion z-test, because you would have two categories; men would be one category; women would be the other category. And, their likelihood of purchasing a certain product would be represented by a binary. So, upstream of 1's and 0's. So, 30% women purchase this product out of a 100 samples; and, 20% men purchase this product out of 25 samples. But, you take that same problem, where you had just two categories, which is purchased or not purchased, men and women, and you extend that to multiple categories on both dimensions. So, that is what we did when we said state could be multiple different states; and, we said age could be multiple different ages. So, if you are able to take these two variables and extend them to multiple categories, you can use a test of independence in sets.

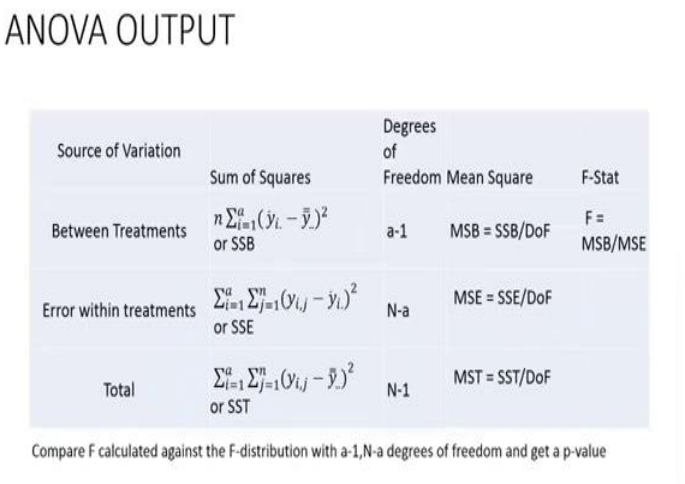
(Refer Slide Time: 07:25)



So, let us first jump into the ANOVA and try and explain the core concept and the math behind it. The idea behind the ANOVA again would be to test a hypothesis of this nature, which is $\mu_A = \mu_B = \mu_C = \mu_D$. A typical t-test for instance would have just looked at something like is $\mu_A = \mu_B$. So, I think a very natural question that quite often comes up is what so special about 2 versus 3? Why is it that for 2, you can use the t-test? 3 and more… It said the ANOVA is built for this kind of multiple comparisons; but, it works just as well even if you just want to compare two samples. So, you can essentially replace your t-test with a two sample t-test with an ANOVA and you would be doing something mathematically identical.

So, before we go any further, we have identified the hypothesis. Now, let us get some nomenclature ready. So, these are the different samples. So, this is the sample corresponding to A; and, A could be anything; it could be fertilizer A versus fertilizer B versus fertilizer C. A could represent the state Karnataka; B could represent Tamil Nadu; C could represent Maharashtra, whatever. So, it depends on the question. Let us take fairly consistent examples. So, let us take the example of a height of boys in public schools in tenth standard for four different states. So, these are the four different states. And, this is the dataset. So, this is data point 1. So, this is data point 1 for state 1. So, that is represented as $y_{1,1}$. Now, data point 2 for state 1 is represented as $y_{1,2}$ and it goes on till $y_{1,n}$. And similarly, it goes in this direction; $y_{2,1}$; and ultimately, you have $y_{4,n}$ in the bottom corner. Now, having defined this, I am going to throw you in some sense in the deep end with the formulas of how an ANOVA is actually computed.

(Refer Slide Time: 09:38)



So, this is the idea. And so, let us just go through this step by step. The idea here is to compute an F-statistic. And, what if we learnt about… If there is one thing we have learnt about hypothesis testing, it is that you come up with a hypothesis, which we did; and a hypothesis was said $\mu_A = \mu_B = \mu_C = \mu_D$. And, the alternate hypothesis by the way out there is that, not all μs are equal, because there are many ways in which mu A – that you can violate the statement $\mu_A = \mu_B = \mu_C = \mu_D$. You can violate it by saying $\mu_A = \mu_B \neq \mu_C = \mu_D$ or you can put that not equal to anywhere in that equation. So, quite simply, the alternate hypothesis is that, not all the μs are equal. So, we took care of the first step, which is to create a null and alternate hypothesis. The second step is to do some kind of computation to come up with the test statistic. We are going to talk a little bit about how… First, we are going to talk about the overall structure, which is the math behind this table leads to an F-statistic with a - 1 degrees of freedom and n - 1. If you remember, we discussed about how the F-statistic has a numerator degrees of freedom and a denominator degrees of freedom. And, you will calculate their F-statistic and the rest of the hypothesis testing is the same. You will then calculate a p value and then choose to reject or not reject.

Now, let us take one step back. We say the F-statistic that we have computed here is nothing but something called MSB divided by MSE. We are going to talk a little bit further about these terms. But, for now, you can you can take this; again go one step back to these two values. And, you technically do not need to know the mean square total or any of these terms. So, this entire part you do not need for actual calculation of the test

statistics. But, it is there to give you a bigger picture. So, again we say this MSB is nothing but SSB divided by something called the degrees of freedom. And, that is fairly obvious in terms of what math needs to happen there; and, again the same concept with MSE. So, at least we know how things flow from one side to the next.

Now, let us get to the core of the formula, which is calculating the sum of squares between and the sum of squares error. The sum of squares between is essentially $\bar{y}_{i.} - \bar{\bar{y}}..$; what do we mean by that? So, let me first… We will talk about the terms here and then I will come to n and a and so on. Essentially, what $\bar{y}_{i.}$ means is $\bar{y}_{A.}$, is nothing but the average of all these terms. $\bar{\bar{y}}..$ is nothing but the average of all the numbers here. So, if I want a row-wise addition, I say $\bar{y}_{A.}$. And, similarly, I would say $\bar{y}_{B.}$ for this row. So, this would be $\bar{y}_{A.}$; this could be $\bar{y}_{B.}$ and so on. And, the idea is that, if you are going to change A, B and so on; and, A, B and C and D can be coded as 1, 2, 3, 4. Then, I might as well just say $\bar{y}_{i.}$ and put that through a loop essentially. So, that is what I am doing here. I am saying $\bar{y}_{i.}$ to say I am going from i = 1 to a; where, a is the total number of treatments, which in my particular case is 4. So, there are four different treatments or there are four different sets of samples that I am comparing. So, essentially, I am taking i from 1 to 4 and I am taking each of those means that I calculate. So, I am essentially taking this average, this average, this average, this average separately. I am taking each of those and I am subtracting that from the grand mean – the overall mean. The overall mean is nothing but every single output variable. This entire table in a sense; this entire table average is $\bar{y} - \bar{\bar{y}}..$

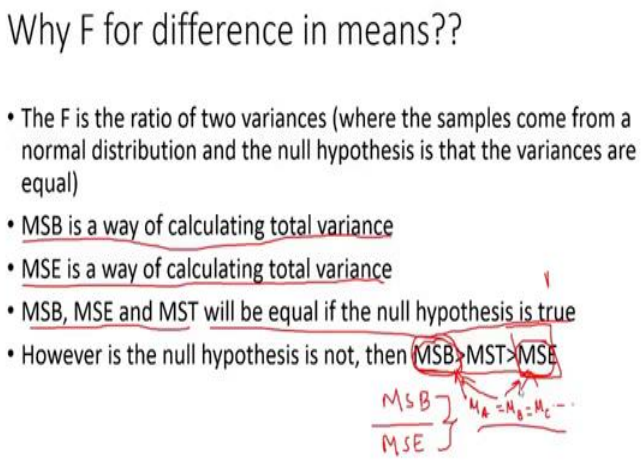
So, essentially, what I am doing out here is, If you look at this formula very similar to standard deviation. But, it is essentially the standard deviation of the means. So, there is an overall mean and you have individual treatment means. So, I am essentially looking at out here this term looks… It is not the standard deviation because I have not done the square root and I am not divided by the number – n - 1. In this case, it would be a - 1. So, I have not divided that. So, that is why it is still; I am going to divide it. Once I divide it, it is kind of like variance; but, I have not divided it yet. So, out here it is essentially like a sum of squares. Once it gets divided by that a - 1, then it becomes that MSB. But, in

concept, this is… And, we are going to talk about the multiplication by n in a second. But, in concept, this is a lot like the standard – the variance of the means. So, it is like the variance. So, if you were to compute a set of means for each row, a set of like $\bar{y}_{i.}$ for each row, the standard deviation of these values or rather the variance of these values is essentially what you are computing with MSB.

Now, what is the concept behind here? Here what you are doing is you are looking at the standard deviation of each data point. So, i , j just means that i goes from 1 to a; which means I am going through each row and I am going through each column out here; j goes from 1 to n. So, j is going from 1 to n; i is going from 1 to a. So, it is literally like I am going through each data point and all through this matrix; and, I am looking at the deviation of each data point from its row average. We saw these two are the same terms that we are using here. The concept here is that, we are looking at essentially the standard deviation within each row. I am not looking at the deviation of y_{11} from the grand mean y bar bar; but, I am looking at the deviation of y_{11} from this rows – essentially, this rows average. And, I am averaging that across every row. So, it is almost like I am taking a row-wise average of every data point and I am averaging that. So, that is called the sum of squares error.

And, again you are dividing by the total number of points that you see; it is the set of N – capital N by the way out here. There is a difference between the small n and the capital N; capital N is essentially nothing but the total number of data points. So, small a is the total number of treatments; in this case, it is equal to 4, n; we have not actually defined out here; but, essentially, it is the total number of columns. So, that is the number of data points that each combination has. And, capital N is the total number of data points. So, essentially, it is nothing but a times n. So, a times n is capital N. So, using that nomenclature, again calculating MSE also should be obvious. So, the mechanics of how you do the ANOVA with the F-test and calculate a p value should at this point be fairly clear to you.

(Refer Slide Time: 18:57)

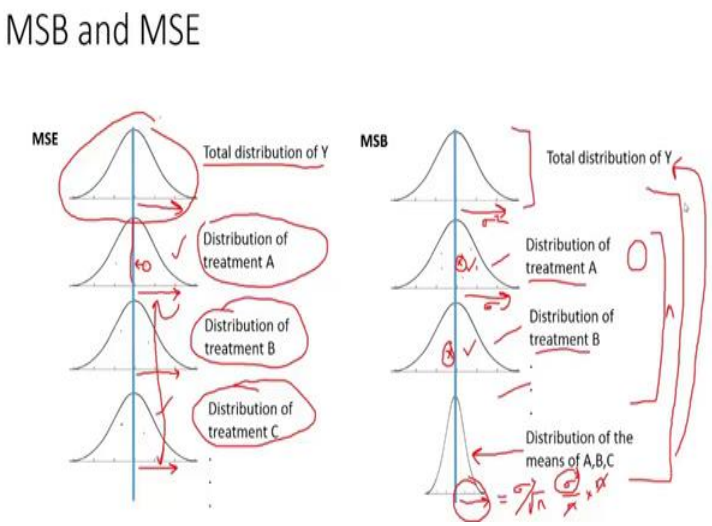


But, let us spend a few minutes and try and get a little bit more intuition on this; which is, why are we using an F-test? Which we know from earlier experience is used to see the difference between two variances. We use the two sample F-test for detecting the difference between two variances. Now, why are we using that to calculate the difference between means? And, the core idea is the following that, the mean square between is one way of calculating the total variance. Now, mean square error is another way of calculating total variance. Now, these three, which is mean square between mean square error and mean square total, which is what is described here are all going to be equal in a sense. Again they are going to be statistically equal, not actually equal.

If the null hypothesis is true; so, if μ– if we go back to the null hypothesis, which is that $\mu_A = \mu_B = \mu_C = \mu_D$; then, MSB, MSC and MSD would be statistically equivalent. However, if it is not true, then you will find that MSB > MSD > MSE. So, you have got two different ways of computing variance. And, what you are doing is you are doing an F-test of these two different variance calculations. So, you are doing an F-test on these two different variance calculations and you will not reject the null hypothesis, if the null hypothesis is true, which is $\mu_A = \mu_B = \mu_C = \mu_D$. Now, if those means are not equal, then these two computations of variance do not really represent the total variance. And, this computation of variance becomes an over estimate and this computation of variance becomes an under estimate of the overall variance. And therefore, a test of whether these two are different variances will show up to be true. Essentially you will wind up rejecting the null hypothesis that these two variances are equal; and thereby, commenting on the

fact that the null hypothesis of means was probably not true. So, let us get a little bit more intuition on why this becomes if μ_A … if the null hypothesis is not true. And, here the null hypothesis is that, $\mu_A = \mu_B = \mu_C$ …. If this null hypothesis is not true, let us take a look at why this computation becomes an overestimate and this becomes an underestimate.

(Refer Slide Time: 21:47)



So, here is the first case where the null hypothesis is true. So, the null hypothesis is true; that means, $\mu_A = \mu_B = \mu_C = \mu_D$; which means the mean of μ_A . So, this is the distribution of A; this is the distribution of B; and, this is the… So, this is just a sample case. So, there is A, B, C. And, here are the distribution. And, hey, look at the means of this the same. And, here I have gone further and even made the distributions look identical. So, if you were to take samples from A, samples from B, samples from C and you are just going to put them all in one bucket called total distribution of Y, this is how it would look, because essentially, since these distributions are identical and their means are identical, it looks this is also identical.

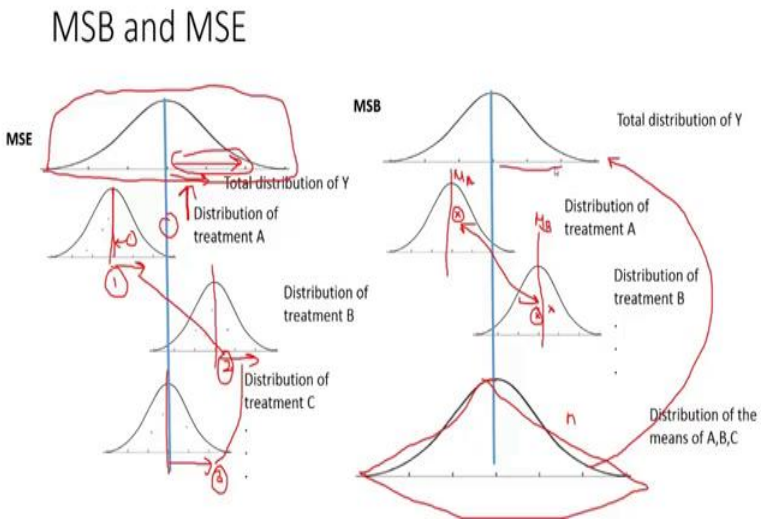
So, how do you compute mean-squared error? The way you compute mean-squared error is – essentially, look at the variance of each distribution; you basically take each data point and look at the standard deviation or variance of the data point with respect to its mean. So, you are essentially computing the variance of each of these distributions and you are just averaging them. And, because the variance of these distributions are going to be the same, that is going to match the variance of the total distribution of y; and, we just explained why.

Now, let us step to mean square between. Mean square between is… So, here is the total distribution of Y and you are interested in capturing the variance of the total distribution of Y. And, the way you are going to do that is you are going to take the mean of distribution A; you are going to take a sample mean of distribution of A. And, as you have known from the earlier part of this course, if you take 5 data points or 6 data points on A and you compute a sample mean, that sample mean need not exactly fall on the population mean. So, you might get another value here for distribution B. Now, as long as the null hypothesis is true, which is the distribution A equals distribution B equals distribution C equals [FL] then you are going to get some sampling distribution if you take the means of each of these distributions. And, if you take the means of each of these distributions; then if all of these are equal, then you are going to get a new distribution called the distribution of the means of A, B, C.

And, what is it that we have discussed about the variance of this distribution? We have discussed that as long as you are sampling from essentially the same distribution – you are sampling from the same distribution, because A now looks identical to B looks identical to C. They all have the same means. As long as you are doing that, we know that this standard deviation is equal to the overall standard deviation, that is, the standard deviation of the total distribution or you can think of it as the standard deviation of any of these distributions, because they are all identical – divided by square root of n. Or, you can think of it – you can think of it in variance terms and say $\sigma 2/_n$. So, as long as I can compute this variance and then move this and multiply it by n, I could cancel this out and I can get an estimate of this variance. So, that is what I am trying to do.

What I am trying to do out here is I am calculating means from each of these distributions and I am taking the standard deviation of those means; I am taking the variance of those means. I am taking the variance of those means and multiplying it by n with the belief that, if the null hypothesis is true, that should be a great estimate of the variance of the total distribution; which is fine. That is going to work great as long as the null hypothesis is true. So, as long as the null hypothesis is true, this approach to calculating the variance of this distribution – variance of this distribution; and, this approach to calculating the variance of this distribution. So, the MSE approach to calculating variance of the total distribution and MSB approach to calculating the variance of the total distribution – both should work perfectly fine.

(Refer Slide Time: 26:22)



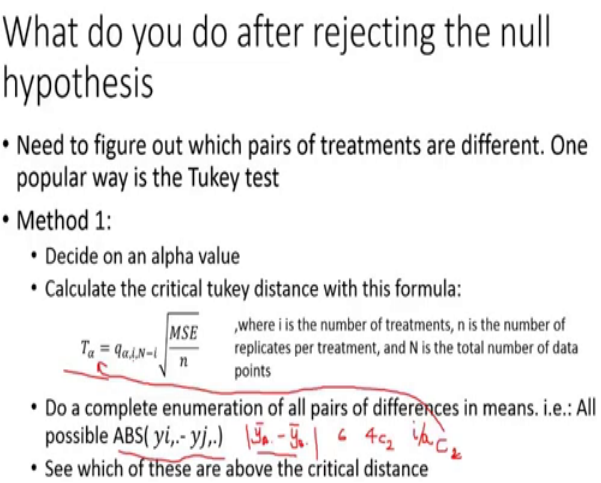
But, what happens when the null hypothesis is not true? Take this case, where the null hypothesis is not true. Now, if you take the total distribution of Y; because the null hypothesis is not true, you get some data points from distribution A, some data points from distribution B, some data points from distribution C; it is going to fall over a much larger region. Now, if you try to estimate the total variance; let us take a look at how MSE would estimate the total variance. MSE would take the variance or take the deviation of each data point with respect to its mean, not with respect to some grand mean. And, even that mean is essentially a sample mean that you are going to calculate. But, the idea is that, your estimate is going to – of variance out here is going to be of some value; your estimate of variance here is going to be another value. And similarly, out here it is going to be of some value. And, all you are doing with MSE is you are taking the average of these three numbers – of 1, 2 and 3.

Now, what is the average of these three arrows? It is an arrow with a magnitude that is much smaller than the real variance. So, when the null hypothesis is not true, the MSE method of calculating variance becomes an underestimate of the total variance. Now, let us look at what happens on the MSB side. Now, the null hypothesis is not true. And so, despite the fact that there is one source of variation, which is that the sample means are not following exactly on top of true means; but, in addition to that, the sample means are further separated from each other, because they are trying to go after a totally different line. They are true means themselves are different. So, the standard deviation or the distribution essentially of the sample means is going to be a much wider distribution,

because the means are not equal. And now, when you multiply that by n, you wind up making a huge over estimate of the total distribution of Y, because you are essentially out here you are trying to look at different – you are trying to get sample means of distribution, which have truly different means and you are looking at the standard deviation of the sample means of distributions with truly different means. It is going to be an overestimate of the total variance of the distribution of Y.

So, I hope that gives you some intuition on why the F-test for the ANOVA works the way it does. But, the important thing for you to remember is that the F-test; the F is the ratio of two variances. And, the core idea is you are using the F-test, which is used for ascertaining whether two variances are equal; you are using that in an indirect way to make a statement about the means of many distributions. And, you are doing that by saying that, if the null hypothesis is true, method A of calculating variance should be equal to method B of calculating variance. But, if the null hypothesis, which is $\mu_A = \mu_B = \mu_C$ is not true; then, these are not accurate methods of calculating variance; great.

(Refer Slide Time: 30:16)



The big question is what do you do? So, now, we know the mechanics of it; we have some intuition for the ANOVA; but, the idea is... So, you did this; you calculated an F-statistic; you then went and plugged that F-statistic and found out the probability. And, let us say your p value was really small. So, you want to reject the null hypothesis. So, what do you do after rejecting the null hypothesis? So, what is the statement you are making when you reject the null hypothesis? You are saying I am rejecting the null hypothesis that $\mu_A = \mu_B = \mu_C = \mu_D$; great. But, can you tell me which of these μs are

different from which others? So, one great way to do this is called the Tukey test. It is not the only way to do it; but, it is a fairly common way of doing it. So, that is why I have called it method 1; it is the Tukey test. But, there are other methods as well.

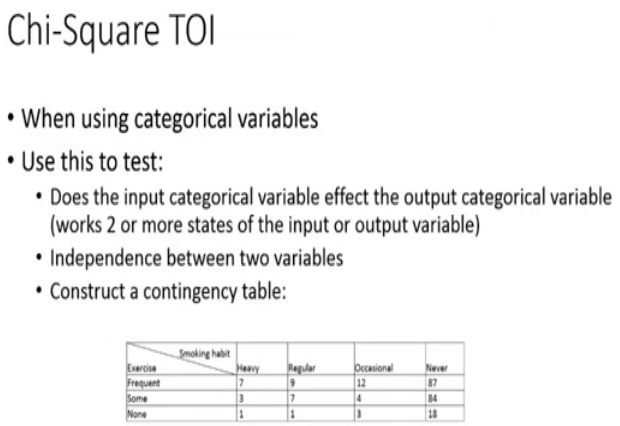
The idea here is to decide on an α value, which we do it many hypothesis tests; even before we start, we say if that p value is less than a certain value, I am going to reject it. So, let us say you start with an α value and then you calculate something called the critical Tukey distance based on this α value. So, there is a distribution called the Tukey distribution and that is again something that you can get usually from software or you can get that from the back of text books. But, you will essentially plug in the values of α, which you just decided in one step above. And, you know the value of i; and, i out here is the number of treatments. So, it is the equivalent of what we had as a. But, I did not want to confuse you in between α and a. So, I have called it i out here; N is the number of replicates. So, small n remains the same – the same concept; and, the large N remains the same.

The only thing that I have changed is I have called what we used to call a; I have called that i and the reason I have done that, so that I avoid confusion with α out here. So, we do that and we use that mean squared error formula out here and we calculate something called a critical distance. And then, you do a very simple thing; which is you do a complete enumeration of all pairs. So, for instance, this is the mathematical way of showing it; but, essentially, all I am saying is let us calculate $\bar{y}_{A.}$. And similarly, calculate $\bar{y}_{B.}$ $\bar{y}_{C.}$. And, look at the difference between each pair of $\bar{y}_{i.}$; where, the two i's are not. So, the different combinations you can come up with are – in our particular example, you will also have A , B; A , C; A , D; B , C; B , D; C , D. So, there are 6 possible combinations and because… And, you are essentially just taking the absolute value. So, b - a is the same as a - b. So, you just calculate 6 differences and you see which of these 6… It is 6 in this particular case, because we had 4 treatments; we had 4 different… We had A, B, C and D. So, we had 4 possibilities, 4 sets of data. So, that is 6 treatments, because you get the concept here is combination. So, you will have $\binom{4}{2}$ combinations.

But, how many ever treatments you have, you will similarly have the concept of whether you like to use the word i or whether you want to use a, c – two combinations; you will calculate that many combinations of distances and you will see which of these distances

are greater than the Tukey distance. And then, those are the ones that are not equal to each other. So, you will calculate a critical distance. See you just rejected the null hypothesis that $\mu_A = \mu_B = \mu_C = \mu_D$ is wrong. So, you have said that; you said that, that is wrong. But, which of these are different from each other? For that, you calculate a critical distance and see which of these combinations have a mean difference that is greater than the critical distance; and, those are different ones. So, hope that gives you an idea of what the Tukey test does.

(Refer Slide Time: 34:32)



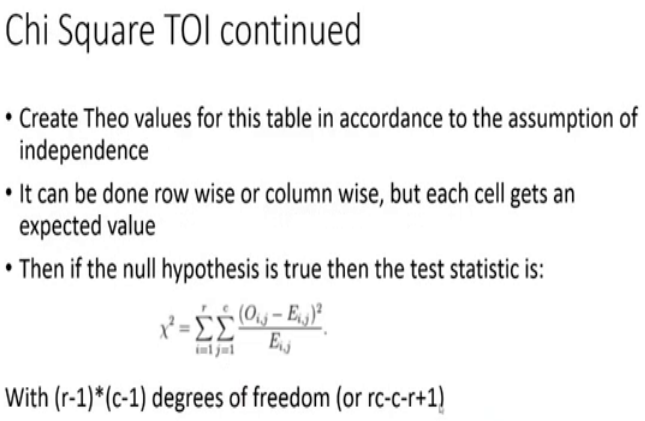
So, now we move on to the next topic, which is the chi-square test of independence. It is noteworthy at this point to say that, we have already looked at one chi-square test; we have looked at the chi-square test for standard deviation in the case of the single sample test. And, you might also encounter another test called the chi-square goodness of a test; and, that is used when you have a set of data and you want to see how well it fits to a particular distribution. But, out here what we are interested in is essentially using that chi square test of independence; and it is used here, where you have two categorical variables. So, here is the first categorical variable, which is the smoking habit. Is it heavy, regular, occasional or never. And, here is the second categorical variable exercise – frequent, some or none. And, what you are trying to do is you are creating a contingency table of how frequently various values occur. And, you are trying to see if they are independent of each other. Does smoking habit have anything to do with exercise or vice-versa. Again not implying causation, but you are taking two categorical variables and seeing if these two variables are independent of each other. So, the core

idea between that is to convert this table to a set of percentages. So, for instance, you would take each of these values and you would see how frequently they occur within each row essentially. And similarly, you would do that for each row.

Finally, what you would do is would create theoretical values for this table in accordance to the assumptions of independence. So, for instance, if I believe that these were completely independent, I would take the row-wise sums and the column-wise sums and I would for instance conclude that row 1 – there would be a 48.7 percent chance of seeing a particular value occur in a particular spot. And so, I would essentially say look since… So, this is… I am getting that 48. So, the numbers I am getting are for instance out here is 48.7. And, the number for this category is 41.5. And, for this would be I believe the remaining, which is about 9.7 percent; so, 48, 41. But, the idea is the following. The idea is that, you get percentages based of the sum of each rows. So, you do 7 + 9 + 12 + 87; and, you divide that by the total sum of all values. And, that is how you get the 48.7 percent.

Now, given that, there is a 48.7 percent chance that you will find yourself in row 1; I ask you the question that, if you are a heavy smoker and there are a total of 11 heavy smokers, the number of heavy smokers that frequently exercise that I should expect is 11 times 48.7 percent. So, just note the map that we did here; we computed the probability of being in each row. The way we did that is we took the sum of each row. So, the sum of the first row is 7 + 9 + 12 + 87; and, we divided that by the sum of all the numbers here. And, by doing that, we got a number 48.7 percent. So, we told ourselves the probability that, if I did not know anything; that I would find myself in the first row is 48.7 percent. Now, assume that I am a heavy smoker; then, there are a total of 11 heavy smokers; if I would randomly guess whether I was a frequent exerciser or not, then my probability is nothing; the number of heavy smokers that frequently exercise – the expected value of that is nothing but the 11, which I get from adding 7 and 3 and 1 and multiplying that by the 48.7. And, a small correction from before… So, you would… This I know for a fact is 48.7. This probability is also 41.5; you would just need to compute the remaining to get this value. And, that is close to about 10 percent. So, I do not believe it is exactly 9.7 percent; that is why I am correcting this up. But, that is the core idea. The core idea is you calculate percentages based on frequency of occurrence and then you come up with what is an expected value for each cell.

(Refer Slide Time: 39:43)



Once you come up with an expected value for each cell; and, you can do that row wise or column wise; I explained it to you row wise. So, I took each row, calculated a percentage; and then, I calculated an expected value. But, essentially, what you have computed is a set of expected values out here. Now, you compare each original value. So, you compare that 7. So, I compare this 7 to what I get when I multiply 48.7 by 100 – 48.7 percent times 11. I believe that is approximately like close to 5 point something. So, what I am essentially doing is I am doing that 7, which I see minus that expected value, which is the 5 … that I see; and, I am squaring that. And, I am doing that for each cell. So, I am doing that for each row and each column and thereby doing that for each cell. And, I am dividing it by their expected value. So, this 5 will also come in the denominator. And, that summation should essentially give me a chi-square distribution with (r – 1)*(c - 1) degrees of freedom.

Now, note that the chi-square distribution has only one parameter associated with degrees of freedom. So, if you just expand this, you will get this formula, which is nothing but just linear algebra rc - c - r + 1. So, you have just calculated a test statistic using this formula. And, that test statistic should be chi-square distributed with this degrees of freedom. And again, you can use that same core concept to calculate a p value associated with it and thereby either reject the hypothesis. And, what is… And, just as a refresh, what is the hypothesis here? The hypothesis – the null hypothesis is that there is no relationship between this variable and this variable; that is, these two variables are independent of each other. And, the alternate hypothesis is the rejection of that.

There is some relationship that depending on whether you are heavy regular or occasional smoker, that impacts whether you are going to be a frequent exerciser or not; I mean it is not a causal relationship; but, knowing this, helps me make – say something about this other variable. Or, it can be the other way round – knowing that you are a frequent exerciser, does that help me in anyway predict or in any way make a statement about your smoking habit. Is there some relationship or are these two variables independent of each other? The null hypothesis being that they are independent of each other and a very low p value in this chi-square test statistic will enable you to reject that null hypothesis or you would fail to reject that hypothesis. So, I hope this gives you an idea of the two tests that we have introduced in inferential statistics that have to do with multiple samples; that just do not have to do with two samples; that is the ANOVA or the analysis of variance and the Chi Square test of independence. In our next lecture, we are going to start with regression.

Thank you.