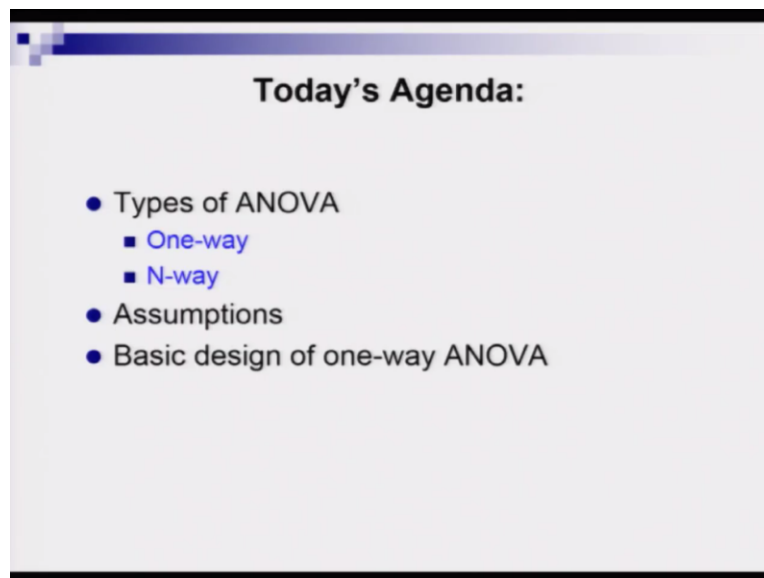


Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Lecture-17
Analysis of Variance

Hello, friends. Welcome back to the lecture series on Applied Statistics and Econometrics. We are done with the core part of module one, which is devoted to statistical methods. So, now, I am going to start discussion on some special topics. And the first topic I have chosen is analysis of variance.

Now, this analysis of variance is an interesting area, where you will see bits and pieces of all the techniques that we have learned so far are being applied. And some scholars say that analysis of variance or the abbreviated term ANOVA, this particular technique has been used severely in social science disciplines like psychology, education research and marketing research. So, we are only going to look at very basic models of ANOVA. ANOVA is a vast field, I am not going to give you a very broad coverage of the topic, but we are going to just look at the most fundamental ANOVA models. So, let us look at today's agenda items.

(Refer Slide Time: 1:27)



So, we are going to start with the definition of ANOVA, and then we are going to differentiate between one-way and n-way ANOVA. And then we are going to lay out the basic assumptions,

if you want to conduct ANOVA. And then we will end today's lecture with basic designing of one-way ANOVA.

So, it was British bio-statistician Ronald A Fisher, who developed ANOVA full-fledged back in early part of 20th century. But even before him, there were mathematicians and statisticians who used ANOVA in the context of linear regression analysis, and Laplace was one of them. But it was Ronald A Fisher, who actually worked heavily on that particular topic, this explanation or the decomposition of the variation in data set and it was he who coined also this term analysis of variance. So, the credit goes to him.

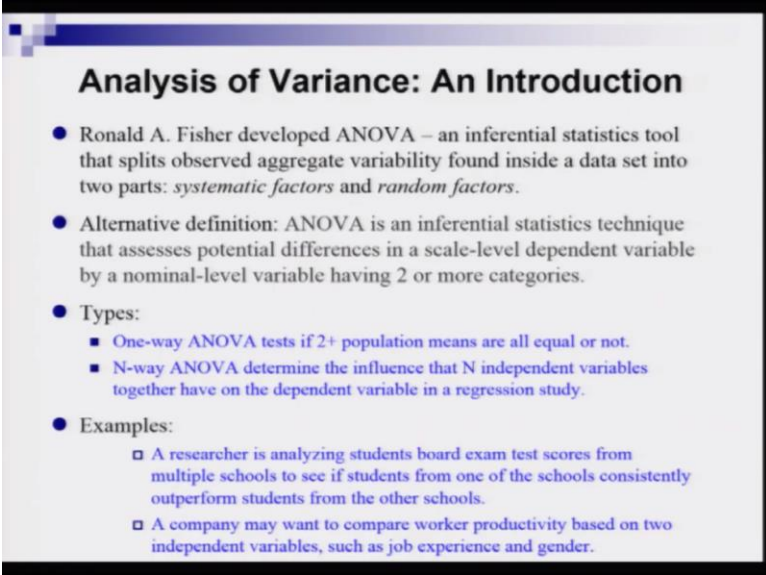
So, what does ANOVA do? ANOVA can be seen from two different perspectives, one is basically linked with the hypothesis testing that we have covered. And if you remember, we talked about comparing population means across groups, and that was basically between groups. So, we had two groups in mind and then we talked about how to compare population mean, whether they are equal or greater than or lower than that. But you think about a situation where you have multiple population groups. So, maybe you are dealing with 4 or 5 such populations, and you are interested to make a comparison of population means across the groups.

Now, if you want to apply that tool of the means test, which is basically t test or the z test that I have taught you previously. Then think about the complexity of the problem. So, you have to take two populations at a time and then conduct a test and then draw a conclusion. Then you take another two populations, conduct the test and then make a conclusion. But note that you can do this kind of pairwise t test or you can compare two population means at a time, but then you will have several t tests or z tests in the process, and then how to combine? I mean, can you really draw an inference that this population means are different from each other or at least one population group mean is different from all other group means, that you can't say from the t test or the z test that I taught you.

So, ANOVA actually helps you in doing this kind of analysis. So, that is one way of looking at ANOVA. And the other way of looking at ANOVA is from the linear regression perspective, that also I have taught you. So, ANOVA is basically a tool which helps you to explain the variation in the y variable, which you can call response variable or outcome variable or your dependent variable through the variation in the independent variables or regressors axis. So, later

we will also establish the linkage between regression and the ANOVA, but for today's lecture I am going to focus on the first angle of ANOVA, which is basically comparing population means.

(Refer Slide Time: 5:10)



Analysis of Variance: An Introduction

- Ronald A. Fisher developed ANOVA – an inferential statistics tool that splits observed aggregate variability found inside a data set into two parts: *systematic factors* and *random factors*.
- Alternative definition: ANOVA is an inferential statistics technique that assesses potential differences in a scale-level dependent variable by a nominal-level variable having 2 or more categories.
- Types:
 - One-way ANOVA tests if 2+ population means are all equal or not.
 - N-way ANOVA determine the influence that N independent variables together have on the dependent variable in a regression study.
- Examples:
 - A researcher is analyzing students board exam test scores from multiple schools to see if students from one of the schools consistently outperform students from the other schools.
 - A company may want to compare worker productivity based on two independent variables, such as job experience and gender.

So, now, I am going to talk about what ANOVA does. So, basically ANOVA splits the observed aggregate variability found in a data set into two parts, one is called the systematic factors and the other one is called random factors. So, what do I mean by systematic and what do I mean by random, that will be much clearer as we progress in the lecture.

Now, there is also an alternative definition of ANOVA. So, ANOVA can be called an inferential statistical technique that assesses potential differences in a scale level dependent variable by a nominal level variable having two or more categories. So, you see, this alternative definition of ANOVA actually takes care of both the angles that I have spoken a bit earlier.

So, angle number one was, of course, the comparison of population means, unknown populations means, when you are involving more than two populations, so that you can figure it out towards the last part of the sentence where I am saying that having two or more categories. And first part of this definition actually is related to the regression concept, where it is being told here that the ANOVA actually helps me to assess the potential differences in a scale level dependent variable. So, basically the scale level dependent variable, I mean, here, a kind of continuous y that we have seen in the regression setup, and difference is basically the variability.

So, this alternative definition is more compact definition of ANOVA as it explains what ANOVA does from two different perspectives. ANOVA could be of two types, and the first one is called one-way ANOVA and that basically helps us to test when we encounter with two plus population means, and it helps us to test whether these population means are equal or not. And the second type is n-way ANOVA. And of course, you can guess that n-way means it's a general case. So, basically here in this case, ANOVA determines the influence that n independent variables together may have on the dependent variable in a regression study.

Now, we are going to look at some examples of one-way ANOVA and n-way ANOVA. And after that we are going to discuss one-way ANOVA at length. So, let us look at the first example. Let us assume that there is a social science researcher who is interested to analyze the students' board exam test scores. And he or she comes with a hypothesis that in a city or in a suburban area or wherever, in some locality, some region, there are multiple schools, but not all the schools are equally good. So, basically there is a school effect on the students' exam performance.

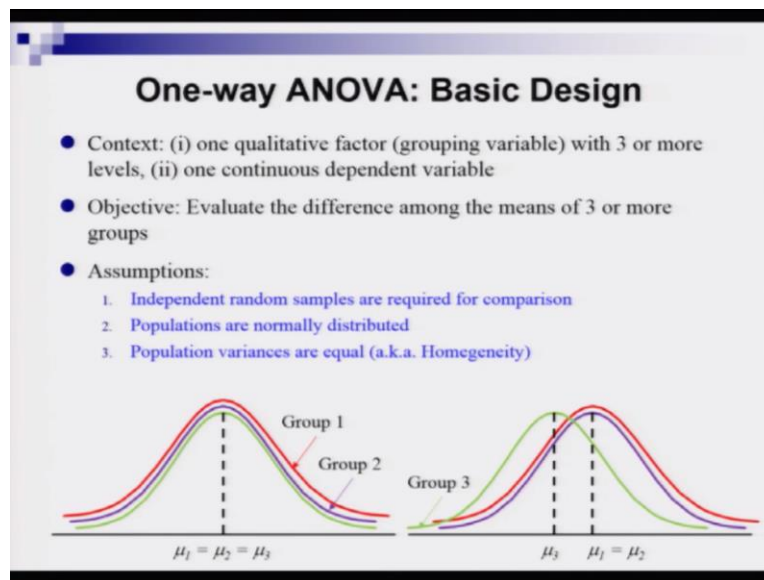
So, suppose he or she wants to see if the students from one of the schools consistently outperform the students from other schools. So, if this researcher has data for three or four schools from a locality, then one-way ANOVA could be helpful to him or her, because then this will help him or her to test more than three population means are equal or not simultaneously.

So, now let me talk about the second example, and this example is going to be on n-way ANOVA which has the regression dimension. So, here let us assume that there is a company and the company's manager or the authority, they want to know whether worker productivity or labour productivity varies across certain factors. So, one can also say that they want to compare the workers' productivity based on two independent variables, if you want to attach the story more close to the regression concept that you have learnt earlier.

So, here, let us assume we are talking about two independent variables. So, here you know you have job experience as one independent variable and second independent variable could be, note that here the worker productivity can be seen as a continuous variable, it can be measured as value of output per hours of labour, input for a particular worker, that is the way you can define

worker productivity. Now, in the following slide, we are going to talk about the basic design of one-way ANOVA.

(Refer Slide Time: 10:12)



So, what is the context? So, here, under one-way ANOVA we are going to explain the variation of one continuous dependent variable in terms of one qualitative factor or grouping variable with three or more levels. And what is the objective? Well, we have already laid out the objective in the previous slide, but let me repeat again. So, here, under one-way ANOVA, our objective is to evaluate the difference among the means of three or more groups. Now, note that if you want to conduct one-way ANOVA you have to start with some basic assumptions. And let's now have a look at those critical assumptions.

So, here we have three assumptions. And the first one says that you must have independent random samples for comparison of population means. Now, what do I mean by independent random samples? It means that when I draw a sample from one population, that should not have any impact on the samples that I am going to draw from the other populations. Now, we move on to the second assumption, and that is relatively simple. I have to assume that populations are normally distributed. Then comes the third assumption and that tells us that population variances are equal, and this is basically the homogeneity condition.

So, now, let us have a look at these assumptions by looking at a diagram. So, now, let us concentrate on the two panels of diagrams at the bottom of the slide. In the left-hand side, you see I am measuring the random variable which is the dependent variable, it is a continuous one. And I am plotting the probability density functions of that continuous dependent variable for 3 different groups.

So, first, we start with how to form a null hypothesis. So, it is easy to begin with a null hypothesis that there is no difference between the population means. So, the population mean of group 1 which is μ_1 population mean of group 2 which is μ_2 and population mean of group 3 which is μ_3 , they are all equal, okay. So, that is basically mark here as $\mu_1 = \mu_2 = \mu_3$. So, you see these 3 distribution functions, they share the same mean and they are showing more or less same population variance as we have to assume homogeneity to conduct ANOVA.

Now, let us talk about what could be the alternative scenarios through a diagram. So, here there is a diagram which is on the southeast corner of the slide, and you note that although I am assuming that my population group 1 and population group 2, they have more or less equal population mean, and let me assume that we know they are identical. So, then I am showing that thing by $\mu_1 = \mu_2$, and they also have the equal variance with the standard assumptions.

But let me assume that group 3 is different from the population group means μ_1 and μ_2 . So, here you see that the green PDF is shifted towards left and it is showing a lower population mean μ_3 , but the variance is still remaining the same. So, variance of population group 3 corresponding to mean μ_3 is equal to the variances that we observe in population groups 1 and 2. So, this could be one you know alternative situation.

The most extreme alternative situation could be that three means they are all not equal to the other. So, in that case you will see that there will be some overlap between the probability distribution functions for this continuous dependent variable, but μ_1 , μ_2 and μ_3 will lie, if not way apart, but there will be different points on these horizontal axes. So, now we are done with setting the basic assumptions and we have roughly spoken about the null and alternative hypotheses. So, its time to formally write down our ANOVA model, okay.

(Refer Slide Time: 15:10)

The Model

- Hypothesis formation:
 - $H_0: \mu_1 = \mu_2 = \dots = \mu_k$
 - H_1 : Not all population means are the same
- The model for the observed response is given by: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$
 - Y_{ij} = Variable value for item number j in group number i
 - α_i = effect of being in group number i
 - ε_{ij} = idiosyncratic error $\sim N(0, \sigma^2)$
- Need to estimate all μ and error variance σ^2
- Total variation in Y or Total Sum of Squares (SST) can be split in:
 - ♣ Sum of Squares Between/Among groups (SSB)
= dispersion between the factor/group sample means
 - ♣ Sum of Squares Within groups (SSW)
= dispersion that exists among the data values within a particular factor level

So, here, in this slide we will start with the hypothesis formation first. So, from the previous slide's diagram, hopefully you can imagine how we are going to frame our null and alternative hypothesis in this case. So, we are going to start with null which says that, if I am dealing with k groups, then basically I have $\mu_1 = \mu_2 = \dots = \mu_k$. So, basically all population group means are equal.

Now, the alternative says that not all population means are the same, so it implies that at least one group's population mean is going to differ from the other population group means. So, now, we are going to write a mathematical expression an equation to present our model. So, the observed response is given by y_{ij} , where i is basically one particular group and j is basically one particular entity in that particular group. So, y_{ij} will be equal to μ , which is the grand mean or the overall mean in the data plus, α_i . So, that is going to be the effect of being in group number i .

And then the last component in the equation is ε_{ij} , that is the idiosyncratic error. So, that means that this is the error for the group member j in any particular group i . And let us assume that this idiosyncratic error follows a normal distribution with 0 mean and constant variance σ^2 . So, this is very important assumption that we must make about the idiosyncratic random error.

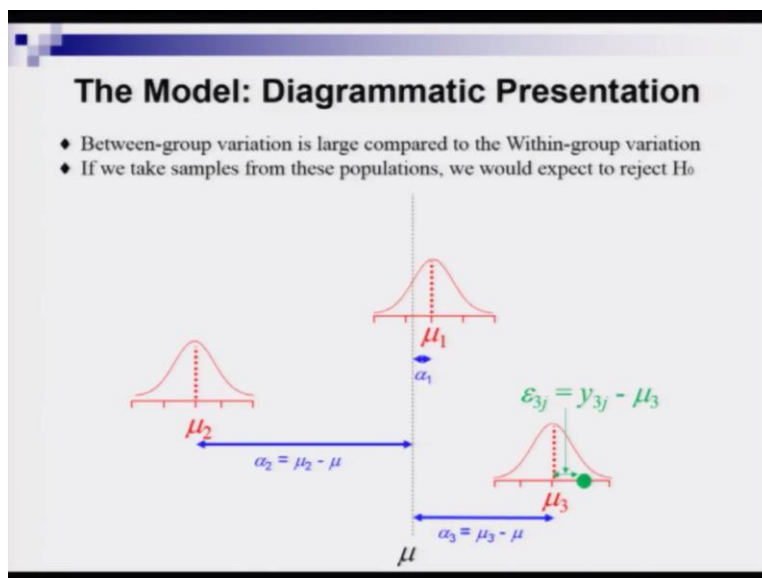
Now, note that in this model, what are unknowns. The unknown parameters are all mu's and the error variance sigma-square. So, we need to estimate them. And after we have some kind of estimates for all these mu's and error variances, then actually some related concepts can be proposed and they are very pivotal.

So, in ANOVA, we say that the total variation in the dependent variable y or the total sum of squares, which is abbreviated as SST, can be split in two components, one is called the sum of squares between groups and that is abbreviated as SSB. So, that is basically the dispersion between the factor or group sample means, that we calculate from data.

Now, why I am writing between/among, because ANOVA can also be applied two population groups, but if you have more than two then you know we say that it is among the groups. And then the second component of this decomposition of total sum of squares SST, is called sum of squares within groups, or the abbreviation is SSW.

So, this is the dispersion that exists among the data values within a particular factor level. So, these ideas of SSB and SSW are going to be much clearer as we move along in this particular lecture. And in the next lecture, I am also going to show you a numerical exercise, so hopefully then it is going to be even more clear.

(Refer Slide Time: 18:45)



So, now in this slide, I am going to give you a diagrammatic presentation of one-way ANOVA model. And of course, it is very difficult to explain every dimension of a one-way ANOVA model in one diagram. So, I am going to have a representative diagram. But let us talk about this diagram and try to understand how actually ANOVA works.

So, I will now start talking about that diagram. But first I am going to bring one story in front of you so that it becomes much more clearer to understand the diagram. So, we all know that average height of human beings in our country, it's not uniform, it varies from north to south, east to west. So, do you know what is the average height of an Indian male? So, it is roughly 5 feet 5 inches. And you go to, say, Punjab, Haryana, Rajasthan, towards the northwest part, and of course Jammu and Kashmir, if you go to the northwest part of the country, you see that average height of a male are much above than the Indian standard or Indian average, which is 5 feet 5 inches.

So, some demographers have also found that the average height of a male in that northwest part of our country, it can reach up to 5 feet 10 inches, so it is significantly different from the Indian average of 5 feet 5 inches. Now, if you move towards northeast, then generally you see that average height actually falls. And it's not only in the northeast, but in the eastern part of our country also, like Bengal, Bihar, Orissa and then you move towards northeast like Assam and Manipur and all, you see that average height actually is below 5 feet 5 inches, so it may be somewhere around 5 feet 3 inches, 2 inches. So, you see that the population average height for our country, it varies across regions.

So, if you want to test whether the average height of a male in, say, state Punjab and Haryana and Rajasthan, and some states from the eastern side, like West Bengal, Orissa and, say, Assam, are they different? If you want to test that thing from some randomly drawn sample, then actually you have to conduct a one-way ANOVA.

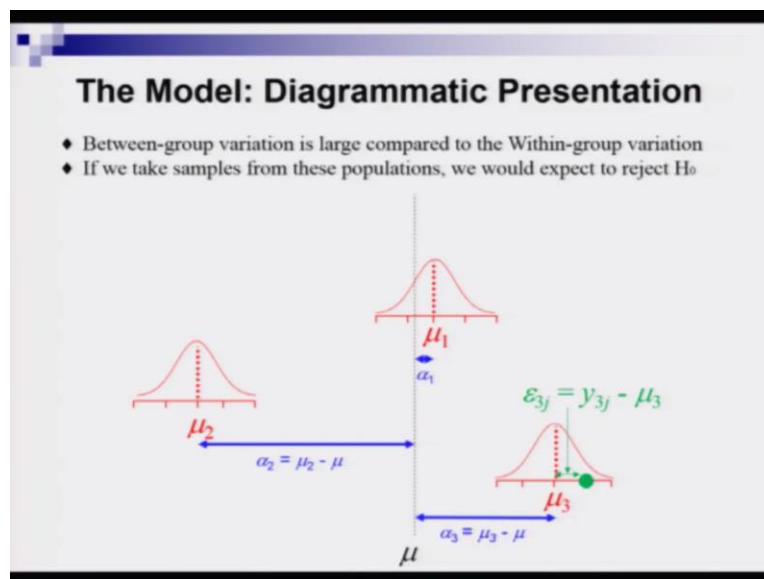
And to conduct that model, you have a hypothetical model in your mind first. So, you can say that, will, particular persons or an individual male's height can be broken down into three different components. So, the first component is the overall mean, which is basically the average height of a male in our country. Then there could be a factor, and that factor is basically the region factor, and that is basically that alpha that I have shown you in the previous slide.

So, that alpha region factor is going to have an impact on the individual. And with that factor the heights are going to change a little bit, both sides, so positive or negative, upper or lower, whatever you want to call it, in both regions it is going to impact. And not only that, there will be some personal idiosyncrasies which will also have impact on the height of a person.

And what could be those idiosyncrasies? It could be the person's status in the economy or the society. So, if the person belongs to a rich or well to do family, then the nutrition that he or she is getting from his or her food is better, and that will have a positive impact on the height of that individual. But if the person belongs to lower middle class or poor family, then that person may not have that much of nutrition from food, and then that may have an impact on his height.

There could be many such issues which are like personal issues and we may not have data on these factors. And we club every sort of unknown or unseen factors under this mask of idiosyncratic or random error, epsilon. So, that is basically the story. And now, let's have a look at the diagram.

(Refer Slide Time: 24:03)



So, here, let us look at the diagram, then we will concentrate on the bullet points. So, here, you see that I have a broken line, which basically denotes my average height of an Indian male. So, this is the overall mean height for an Indian male from the entire country's data. And then, you see, I have chosen 3 different states and I have collected random samples. So, my state one could

be, say, Madhya Pradesh, exactly center of our country. So, now, of course, if I assume that these different population groups have different means, if I assume, they know let us assume that there will be separate symbols. So, for my population group one, which is Madhya Pradesh, I am assuming the population mean height for male is μ_1 .

So, if you look at my population group 3, you see there is a huge difference between μ_3 and μ , and that is denoted by α_3 . So, that is the group effect, α_3 , for group number 3. And as you see in the case of population group-1, there is not much difference between the overall mean μ and the average for the population group-1, μ_1 . So, this α_1 , this group effect is tiny compared to α_2 and α_3 .

So, now let's look at the case of one particular individual who is from, say, my population group-3. And that individual is marked as a green circle and you see that that person's height is above the average height of that population group 3, so the individual's height is higher than the group mean μ_3 . So, here, I am showing you this ϵ_j equation for this particular individual in the graph.

Now, of course, this is a hypothetical picture. What could be the most extreme scenario where you can assume that there is no difference between the population means across groups? So, in that case, the μ_1 , μ_2 and μ_3 will all lie on that vertical broken line. So, basically, you can say other way around, μ_1 is equal to μ_2 is equal to μ_3 , equal to μ , the overall mean. So, that is basically the most extreme case. But there could be other possible cases also where μ_1 is also not lying close to that overall mean μ .

So, now, what do we learn from this hypothetical diagram? So, we observe here that the between group variation is large compared to the within group variation, because the within group variation is given by the dispersion corresponding to this bell-shaped curve for each group, and you see that is fairly low. Now, the second point that we can observe from this diagram is that, if we now take samples from these populations, then we can expect to reject our null hypothesis, which is basically saying that the population means equal or identical across groups.

So, far we have spoken about one-way ANOVA in terms of population, but population of course, you cannot access entirely so you have to draw random samples. And then based on the random

samples, you need to calculate certain statistics and then based on that you have to draw inferences.

(Refer Slide Time: 28:23)

Estimation from Data

- In the One-way ANOVA case, we estimate:
 1. The grand population mean μ , with the grand sample mean $\bar{y}_{..}$
 2. Individual population means μ_i , with sample means $\bar{y}_{i.}$
 3. The effect for group i , $(\mu_i - \mu)$, with $(\bar{y}_{i.} - \bar{y}_{..})$
 4. Population errors ε_{ij} , with sample residuals $e_{ij} = (y_{ij} - \bar{y}_{i.})$
- The deviation of any observation from the overall mean can be partitioned as :

$$\boxed{(y_{ij} - \bar{y}_{..}) = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})}$$

\uparrow
 The effect for group i

\uparrow
 Sample residual

So, now, let us see if we have access to random samples from different populations, then what type of calculations we are going to conduct. So, to conduct one-way ANOVA we have to estimate four things, and we will first start with estimating the grand population mean or the overall population mean μ , and for that we have to use the statistic called grand sample mean, and that is denoted here by y -bar double dot.

Now, note that this is kind of a new notation for you. So, I have used y -bar to denote arithmetic mean before, but these double dots are probably a bit new. Although we have spoken about similar concept in the case of contingency table, but let me explain it to you once more here.

So, here double dot means that we are varying both dimensions of our data. So, both dimension means that we have here two dimensions, one is basically the group dimension, which is basically i , and the other is individual dimension that is basically denoted by j . So, here both i and j are varying and that is why when we are computing arithmetic mean by varying both dimensions, we are talking about taking the sample mean of the entire data set. So, that is why it is called grand sample mean our overall sample mean.

Now we move down to the second item in the list and we need to estimate individual population means, μ_i with sample means. So, here I introduce a similar notation, \bar{y}_i . So, what do I mean by this? So, here you see I am fixing i -th group, I am interested in deriving the population mean for the i -th group. So, I have to fix one particular i and then I have to calculate the sample mean of my variable for all individuals which are member of that particular group i . So, here I am varying over individuals, which is basically denoted by j , but I am fixing i , so that is why it is \bar{y}_i .

Then the third in the list is the effect for group i . So, that is basically that α_i that I have shown you in the slide. So, that is basically $\mu_i - \mu$ in other words. How can we find an estimator to measure the effect for group- i ? So, that is basically by taking the difference between these two means that we have already calculated.

Now, we move to the fourth and last item in the list, and that is basically the population errors, ϵ_{ij} . Now, we can estimate population errors by sample residuals which is defined as e_{ij} , and that is basically equal to the individual observations $y_{ij} - \bar{y}_i$. So that is basically the individual sample mean. So, the sample residual is the difference between the individual observations and the group specific sample means.

Now, we can talk about the deviation of any particular observation from the overall mean, and I am going to show you how that deviation can be broken down or partitioned into two concepts. So, here, look at the left-hand side of this equation in this red box. So, that is basically talking about the difference between individual observation and grand sample mean or the overall sample mean. And you see that in the right-hand side now, I have two different components in two far bracketed or parentheses items.

So, here, in the first parenthesis, you see that I am talking about the difference between the grand mean and the group specific sample mean. So, first component in the equation is going to give us the effect for group- i , that we have discussed in this slide only. And now the second component is basically that e_{ij} component, so that is basically the sample residual component, as it is defined as the individual observation and the group specific mean. So, here, I am talking about that particular group where this j individual belongs to.

So, now, from this equation, we can derive the ANOVA identity. So, how do we derive the ANOVA identity? So, first, these individual components of this partition or decomposition formula, we can actually take squares, and then we can sum overall data points of these squared items, and then that will give us the famous ANOVA identity. So, now, let's have a look at that.

(Refer Slide Time: 33:39)

Partition of Total Sum of Squares

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

↗ ↘
Between Group Sum of Squares Within Group Sum of Squares

- ♣ This equation is the *ANOVA Identity*: $SS_T = SS_B + SS_W$
- ♣ This identity partitions the **Total Sum of Squares** into two components of interest for our hypothesis

So, here you see that expression or equation that I am showing here at the top, so that is basically the ANOVA identity. So, in the left hand side you have this expression involving double sum. You know how the double sum is to be computed, because we have spoken about this double sum concept in the case of contingency tables and all. But let me repeat once more for a couple of seconds.

So, basically, here you were summing over all data points that you have. So, basically, you have k number of groups to start with, and that is why the first sum ranges from i equal to 1 to k. And then within these k groups, you have n-1, n-2, dot dot dot n-k number of observations in different groups. So, that is what the second sum is talking about.

So, here, the individual ranges from j equal to 1 to n-i, which is basically the number of observations in a particular group. So, when you are taking double sum, you are basically taking the sum over these squared differences for all data points that you have. Now, this is called the total sum of squares.

And now, this is broken in two different components. So, the first component is the between group sum of squares, and that is basically given by $SS-B$, and the second component in this equation or identity is basically the within group sum of squares, and that is abbreviated $SS-W$. So, this ANOVA identity partitions the total variation in your dependent variable y . And that partition actually is being done in terms of different groups and this has two components in the identity or equation, and they are called the between group sum of squares and within groups sum of squares. So, this ANOVA identity will now lead us to hypothesis testing. So, this ANOVA identity will now lead us to hypothesis testing and that we will cover in the next lecture. Thank you.