

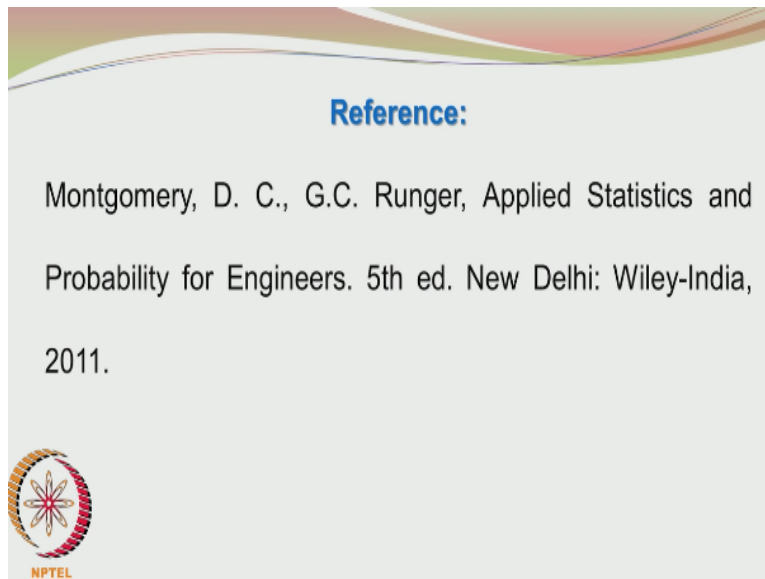
Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology – Madras

Lecture – 25
Analysis of Experiments involving Single Factor - Part A

Welcome back. We will now be starting with the second part of our course on statistics for experimentalist. In this lecture will be looking at a relatively simple situation here experiments are carried out by changing only one variable or only one factor. Usually experimentalist vary more than two factors a simple example would be we are interested in looking at the yield or conversion from a chemical reaction.

So, we may vary temperature and pressure or temperature, pressure flow rate of the reactants, temperature pressure flow rates and catalysts involved and so on. For the purpose of illustrating the basic concepts we are going to consider the variation of a single factor only. The other factors are assumed to be kept at the fixed values or at constant values they are not being changed.

(Refer Slide Time: 01:24)



The reference for this lecture is the book written by Montgomery and Runger, Applied Statistics and Probability for Engineers. 5th edition, Wiley-India.

(Refer Slide Time: 01:35)

Experiments involving ONE factor

Terminology:

Factor: Controlled variable whose effect on the outcome is being investigated

Level: Value that is assigned to the factor, and many levels of the same factor may be tested



Let us look at the terminologies first. The factor is a controlled variable whose effect on the outcome is being investigated. Level is the value that is assigned to the factor, and many levels of the same factor may be tested. We want to study the effect of temperature on the yield in a chemical reaction so the factor is temperature. We want to vary this factor to see the effect of this factor on the yield.

The levels of this factor can be different temperatures 30 degree centigrade, 50 degree centigrade, 100 degree centigrade and so on. So, we can have several levels of the same factor. Now, we are looking at another important term treatment it is somewhat very unusual term in experiments sometimes we encounter it so better to define it. It is very simple in fact treatment is each level or setting for a factor.

(Refer Slide Time: 02:56)


Experiments involving ONE factor

Terminology:

Treatment: Each level or setting for a factor

a: number of treatments that are carried out

r: number of repeats of each treatment



So, it is the value taken by a factor when it is kept at a certain level. We can have a treatments for our reactor. Example there may be a temperatures. Many times we are not satisfied with doing the experiment only once if we want to study the effect of temperature on the yield we study various temperatures like 30 degrees, 40 degrees, 50 degrees and 100 degrees so on. But we have considered each level of the factor only once that is not what I meant.

We want to repeat the experiment at the same treatment or level of a given factor and see the effect of the repetition on the reproducibility of the response. We want to carry out the experiment at the same temperature let us say 50 degree centigrade and see what are the yields when we repeat it several times? Repeats also are intuitively appealing to us because if you get more or less the same response from the experiment.

Whenever we repeat them at a given setting then we are convinced that we have done the experiments properly. The equipment or the reactor is working properly and we are having confidence in our results. So, repeats are very important from a statistical point of view also repetitions of experiments are very essential.

(Refer Slide Time: 04:45)

Experiments involving ONE factor

Hence there will be a total of " $a \cdot n$ " experiments.

Response: The outcome of the experiment for each treatment. The response for each of the " a " treatments is a random variable.



When there are a treatments and n repeats. We will have a total of $a \cdot n$ experiments. The next term which we are going to define is the response. The outcome of the experiment for each treatment what is the output from the reactor? What is the yield from the reactor? What do we get so that is what we are calling as the response and since there are several random factors that may influence the outcome of this experiment.

The response is treated as a random variable. Normally, we denote the response as y . We are going to concentrate only on one factor the reason for this is we want to establish the basic ground work introduce you to the concept of variant degrees of freedom. The means squares, the analysis of variances, f test and the conclusion you make after looking at the f statistic. This also involves hypothesis testing anyway we will cross the bridge when we come to that.

(Refer Slide Time: 06:10)

Experiments involving ONE factor

❖ The effect of changing the levels of ONE factor on the desired response is investigated i.e. effect of different treatments is observed

❖ There may be many settings or treatments of this factor as well as many replicates (repeats) of each treatment.




So, let us get on with this introduction. The effect of changing the levels of one factor on the desired response is investigated. So, we are looking at the effect of different treatments there may be many settings or treatments of this factor as well as many replicates or repeats for each treatment. Why does the experiment give different results even if we take all the precautions like keeping the factor level at pretty much a given value.

All other factors or all other variables that may influence the experiment are well controlled we are not varying them. We are making sure that the ambient conditions are not varying too much still we may get variation in the response. These are attributable to random errors when we repeat the experiments we get variability in our response and that may be attributed to the random factors or random phenomena.

(Refer Slide Time: 07:21)

Experiments involving ONE factor

- ❖ The variability in response for the same treatment when repeated is attributed to random error.
- ❖ Replicates or repeats give estimates of the experimental error.




So, in order to get an idea about the experimental error we need repeats or replicates whenever we talk about experimental error we are not acquiescing the experimentalist of doing the experiments in a bad fashion despite his best efforts to maintain proper conditions there may be variations in the response so we talk in a neutral sense whenever we refer to the experimental error in the data.

(Refer Slide Time: 07:55)

Experiments involving ONE factor

- ❖ When the level of a factor is changed, there may be a variation in the response.
- ❖ The treatment change brings forth a change in response.



When the level of the factor is changed there is going to be a variation in the response hence we think that because of the change in the level of the factor. Because of applying a new treatment there is a variation produced so let us look at crops being grow in a field and we want to test

different fertilizers. So, in plot one we put fertilizer A we look at the yield then we apply the fertilizer B that is a new treatment and we look at the yield.

If there is a change in the yield, a difference in the yield we think that it is because of the change in the treatment or the change in the level of the factor there was a difference in the crop output. So, this is what we normally think we do not think that there could have been other factors which may have caused a difference in the yield but the farmer or the person who is doing this investigation may firmly state.

Look I only varied the fertilizer, the type of the soil, the amount of watering, the length of watering all other factors were unchanged only the factor that changed was the type of fertilizer. Okay, even then we have to be careful we have to see whether the variation in the crop production, the variation the reactor yield. The variation in the response generally was due to changing the treatment or changing the level of a factor or it was because of random effects.

Random effects which were not in our control effected the experiment strongly or in whatever manner and produced a variation in the response. So the extent of this response changed may be different sometimes there may be a small change in the response sometime there maybe large change in the response. If there is a large change in the response of the experiment, then we think that it is because of the treatment change.


Sometimes there may be only a medium or small response change when you change the level of the factor or change the treatment. Then you do not know whether the response changed because of changing the treatment or it was because of random effects. So, we need to quantify this so that the results may be presented in a unambiguous fashion.

(Refer Slide Time: 10:52)

Experiments involving ONE factor

❖ Hence, compare variance due to change in treatments (called as mean square treatments), with variance due to repeats (mean square error).

❖ In other words, compare variation between treatments to variation within treatments.



So, we are going to look at variance whenever we do repeats of experiments we look at the mean outcome or the mean yield or the mean crop production. But it is not only mean which is important in addition to mean or average we also have to look at the variance so again whatever we studied in the first part of the course is becoming very relevant now. The variance is a very, very important factor.

Let not use the word factor because we are already using it for looking at the variable. Variance can create an important influence on the interpretation of the data so let us see how this happens. What we are doing is we are going to compare the variation due to change in treatments with variation due to repeats. As I said earlier repeats are representatives of the random phenomena whenever we repeat the experiments.

We may get different results and hence that variation is representative of the experimental errors that influence the process on which the experimenter usually has no control on but of course you can change the level of the factor he can go from fertilizer A to fertilizer B or he can go from 30 degree centigrade to 50 degree centigrade. So he has control over the variable or factor he is actually changing and he can maintain them at the constant value.

So, we are having change in treatment and also we are having random errors. We have to compare the two and we have to compare the variability produced by the random error with the

variability produced by the change in treatments. So, we have to compare the variation between treatments to variations within treatments.

(Refer Slide Time: 12:58)

Treatment	Repeated Observations				Totals	Averages
1	y_{11}	y_{12}	...	y_{1n}	$y_{1.}$	\bar{y}_1
2	y_{21}	y_{22}	...	y_{2n}	$y_{2.}$	\bar{y}_2
.
.
.
a	y_{a1}	y_{a2}	...	y_{an}	$y_{a.}$	\bar{y}_a
Sums	$y_{.1}$	$y_{.2}$...	$y_{.n}$	$y_{..}$	$\bar{y}_{..}$

Let us look at a table of experimental data collection. So, we are having treatments in this column. We are going from 1, 2 so on to a and for the first treatment we have carried out n repeats. So you can see that we are going from y_{11} , y_{12} so on to y_{1n} . 1 standing for the first treatment and 1,2,3 and so on to n standing for the repeats. So, we are denoting the experimental outcome as y_{ij} .

The response is given as y_{ij} , where i stand for the treatment and j stands for the repeat i is the index for the treatment and j is the index for the repeat. So, the treatments are varying row wise so i is running from 1, 2 so on to a whereas J is running from 1, 2 so on to n. So, totally we have a*n runs so all these runs are recorded as responses and we have totally n elements. Now, we can total them and for a given treatment we add all the responses.

Due to the repeats n repeats and we get $y_{1.}$. So, we are fixing 1 which is the treatment and dot represents the summation. So, it is instead of writing $\sum_{j=1}^n y_{1j}$, $y_{1.}$ we are writing it as $y_{1.}$. And when you take the average when you add up the responses for a given treatment, n responses for a given treatment you get $y_{1.}$. That you divide by the number of repeats that will be $y_{1.}/n$ which is the average response for a given treatment 1 that is represented by \bar{y}_1 .

The bar represents the averaging similarly you can do for the second treatment. You can do for the a'th treatment. So, you will get Y_1 . Y_2 . So on to Y_a . And the averages may also be denoted by \bar{Y}_1 , \bar{Y}_2 , So on to \bar{Y}_a . and similarly just as you did row wise the totalling and averaging you may also do the totalling column wise. Normally the row wise totals and averages would be used. So what I have done here is to denote the totals.

So you have Y_1 . When you go row wise for the second treatment when you add all the n treatments you get Y_2 . Because treatment 2 is fixed and so you get all these responses put in the appropriate terminology. Again I can sum up the values for the first repeat. So, there I am summing over all the treatments for the first repeats so I write it as $Y_{\cdot 1}$. Similarly, for the n-th repeat for each treatment I am totalling the responses over all the treatments for the n-th repeat.

So, I get $Y_{\cdot n}$ and when you add all these responses you get the grand total y_{\dots} and when you divide it by total number of observations which is $a*n$, number of treatments into number of repeats. $Y_{\dots}/a*n$ gives $\bar{\bar{Y}}$ which is the global average or the grand average. So, the same thing I have put in this table and I have shown the averages. So, when I am considering the first repeat and I am adding all the responses over the a treatments.

I get $Y_{\cdot 1}$ when I am averaging it out by dividing it by total number of treatments I get $\bar{y}_{\cdot 1}$. So, I am adding all these elements I will get $Y_{\cdot 1}/a$ will give me $\bar{Y}_{\cdot 1}$ or more correctly $\bar{Y}_{\cdot 1}$. Similarly, I can do the averaging for the other columns and the global average is $\bar{\bar{Y}}$


(Refer Slide Time: 18:29)

Definition of Summation Conventions

$$\sum_{j=1}^n y_{ij} = y_i.$$

$$\frac{\sum_{j=1}^n y_{ij}}{n} = \bar{y}_i.$$

$i=1,2,\dots,a \quad j = 1,2,\dots,n$



So, this is the terminology which I was explaining a couple of slides back you are adding over the index j running from 1 to n . So, I kept constant so you put Y_i . Similarly, here I am taking the same summation either this or this and that I am dividing it by the total number repeats n and I get \bar{Y}_i . Obviously, i is running from 1 to a , I am fixing i in this case and j is running from 1 to n j represents the repeats.

And i represents the treatments and when I add up all the responses over all the treatments and all the means I get Y_{ij} is equal to Y DOUBLE DOT there is a typo here I will correct the typo. Okay, thanks for waiting the terminology is you should put the i index first and the j index next. So, i running from 1 to a and j running from 1 to n , $Y_{ij}=Y$ DOUBLE DOT Similarly, I am finding the mean the grand mean. So, the grand total is divided by the total numbers of observations $a*n$ I will get \bar{Y} double dot.

This is usually found in these statistical design of experiments text books. So, it is important that become comfortable and familiar with the terminology the dot notation. So, N is the product of a treatments and the number of repeats per treatment.

(Refer Slide Time: 20:15)

Definition of Summation Conventions

❖ Here N is the product of number of treatments (a) and number of repeats per treatment (n).

❖ The dot(\bullet) represents the summation over the index it replaces.



The dot represent the summation over the index it replaces.

(Refer Slide Time: 20:24)

Definition of Summation Conventions

$$\sum_{i=1}^a \sum_{j=1}^n y_{ij} \dot{=} y_{i\bullet}$$

$$\frac{\sum_{i=1}^a \sum_{j=1}^n y_{ij}}{N} = \bar{y}_{i\bullet}$$



$$i=1,2,\dots,a \quad j=1,2,\dots,n \quad N=an$$

So, when we put $Y I$ dot, it is replacing the summation over j .

(Refer Slide Time: 20:36)

Modeling of Experimental Response

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

$$\text{or } y_{ij} = \mu_i + \epsilon_{ij}$$

Terminology



Response to the i^{th} treatment ($i=1,2,\dots,a$) and j^{th} repeat ($j=1,2,\dots,n$)

Now, let us look at the experimental response we want to model that we are not going to do any complicated modeling it is a simple linear model. But it carries a lot of punch as we will see. Why y_{ij} which is the response from the i^{th} treatment and the j^{th} repeat is modelled as a sum of 3 terms. The first is the global average μ then τ_i is the effect of the i^{th} treatment and ϵ_{ij} is the random error.

Interesting to see the different symbols μ is having no subscript because it is standing for the global average or the mean response and τ_i is the i^{th} treatment effect and it is having the index i corresponding the treatment and ϵ_{ij} is having the and this is corresponding to both treatment as well the repeats. We may write $\mu + \tau_i$ as μ_i so Y_{ij} is equal to $\mu_i + \epsilon_{ij}$. This is the simple linear model we have not put a none linear model here for example

$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$ some highly complicated model which we will find it very difficult to work with we are having only a simple linear model and we are talking about the effect of only one factor. So, we are having τ_i which is the representation of the single factor we are analyzing. So, τ_i probably stand for temperature or fertilizer this τ_i can have different levels, temperatures can take different values 30, 50, 80, 100 degree centigrade.

Fertilizer can take fertilizer A, fertilizer B, fertilizer C and so on. So, we are having only one factor so we put only one τ_i . If you are considering two factors this linear model is simply

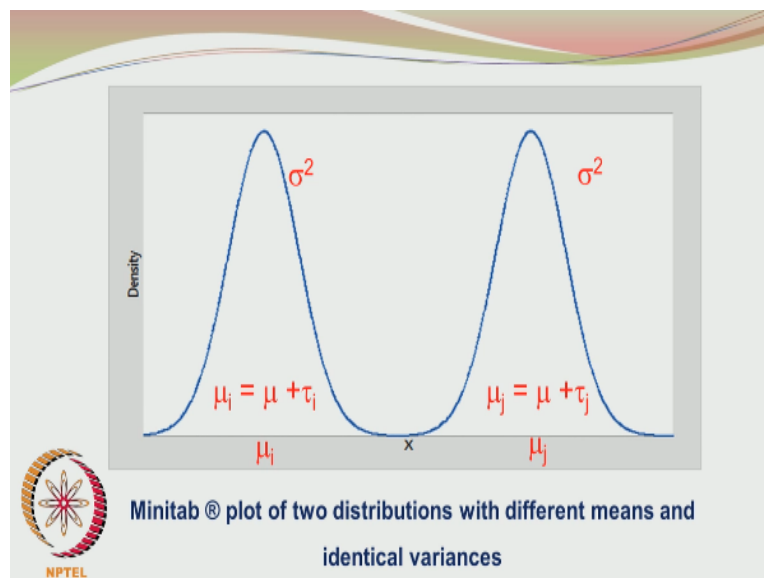
extended, we can put $\tau_i + \beta_j$ and ϵ_{ijk} because we are having now a combination of two factors i and j and then k will become the index for representing the repeats. We will be seeing these two factors shortly even more factors so we do not have to really worry about it.

Let us focus on a single factor now. Essentially μ would be the response Y_{ij} everytime when the factor is not having an effect and there is no random fluctuations we will get a unique value in our experiment or from our experiment when the treatments are not effective and random errors are not there. The next possibility is random errors are there but the treatments effects are not there then what would happen is this value of μ will get spread.

Because of the effect of the random factors or random effects. The other possibility is both of them will be present the treatment is having an effect the error is having an effect so we are now considering the variability given to a global response μ because of the treatment as well the random noise or random errors. If there is effect of treatment on μ the μ is changing because of the treatment, then it takes a unique value μ_i corresponding to the i 'th treatment.

Remember we can give a levels of the treatment so depending on what treatment you have given the μ has changed and that will become μ_i . A very interesting figure awaits us.

(Refer Slide Time: 25:31)



Here you are having μ_i and μ_j this is the response spread because of the application of first treatment or the i 'th treatment. This is the response obtained because of the application of the j 'th treatment. The middle value of this is μ_i and that is defined as $\mu + \tau_i$. If τ_i is 0 then μ_i becomes μ . If τ_j is 0 there is no effect of the j 'th treatment μ_j becomes μ . So, in both these places we will have μ and μ but if τ_i is effective μ_i will be different from μ_j .

And very interesting thing is all these spread is because of the variance σ^2 . We assume that this variance is because of this random effects, the random fluctuating components which are not in our control and the variance of the errors are constant. The errors are assumed to have 0 mean and have constant variance. So, the net sum of all these errors on the response would be to produce a spread around μ_i , around μ_j with constant variance σ^2 .

I request you to take a closer look at this figure and make sure that you have understood the concepts.

(Refer Slide Time: 27:26)

Modeling of Experimental Response

- μ : Overall mean and is a parameter common to all treatments
- μ_i : i 'th treatment mean ($\mu + \tau_i$)
- τ_i : i 'th treatment effect
- random error $N(0, \sigma^2)$

NPTEL

We are still in the process of modeling the experimental response we have the μ which is the overall mean and it is a parameter common to all the treatments. This would be the response which we will be getting if there was no effect of the treatments and there was no random error fluctuation. Everytime we do the experiment whether we put fertilizer A, fertilizer B, fertilizer C everytime the field produces one tonne per annum of the rice grains.

Or the reactor is producing exactly 30% yield irrespective of whether you put the temperature at 30 degree centigrade or whether you are putting the temperature at 100 degrees centigrade. So, this is the common uniform value if none of the treatments and the random fluctuations are influencing the process. And this is obviously not going to happen μ_i is defined as the i 'th treatment mean what is the mean response for the i 'th treatment.

When I am operating the reactor at 30 degree centigrade what is the percentage yield that is modeled as the addition to the mean μ which corresponding to the unique value unaffected by the treatment and unaffected by the random error. So, we are assuming that there is an addition to the μ . Of course there may be some cases where the effect of the treatment τ_j for instance may actually reduce the value of μ such that μ_j may be $\mu - \tau_j$.

But in general we represent μ_i or μ_j as $\mu + \tau_i$ or $\mu + \tau_j$. What I am trying to say here is τ_i may be positive or negative so we are having τ_i , we call it as the effect of the i 'th treatment and ϵ_{ij} is the random error contribution which is normally distributed with 0 mean and variance σ^2 . We are having this nomenclature to represent the normal distribution with 0 mean and variance σ^2 .

(Refer Slide Time: 29:59)

Null and Alternate Hypothesis

$H_0: \mu_1 = \mu_2 = \dots = \mu_a = \mu$

Equivalently $H_0: \tau_1 = \tau_2 = \dots = \tau_a = 0$

H_1 : At least one pair of the means μ_i, μ_j are not equal (or, equivalently)

H_1 : at least one of the τ_i values $\neq 0$

NPTEL

Now, we are coming to the null and alternate hypothesis statements which we studied very recently. We can now see the topics that we studied in the first part of the course for example the normal distribution the hypothesis testing all are nicely falling in place in the design of experiments. So, we can have the null hypothesis as $\mu_1 = \mu_2$ so on to $\mu_a = \mu$ what is the meaning of this statement?

All the responses are equal to μ whether I am applying the first treatment or the second treatment or the third treatment first temperature, second temperature or the third temperature the output is not changing there is no change. There is a status quo. There is no effect of the treatment whether I put 30 degree centigrade or 80 degree centigrade in the reactor the yield is not changing.

So, that is the skeptical view that is the neutral view and so we say that the null hypothesis indicates that there is no effect of treatment. It is a safe view. Now the alternate hypothesis is going to be in opposition with the null hypothesis. The alternate hypothesis is trying to revolt against the current status quo and say that there is a change. There will be a change upon application of the treatment.

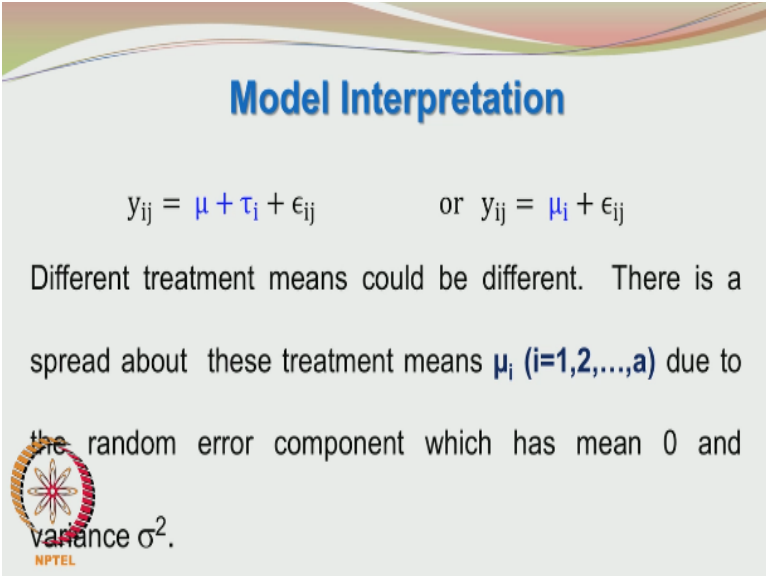
So, the alternate hypothesis is always supporting or routing for the change. It says there may be many treatments and of course I agree that there may be some treatments which are not effective but there is at least one pairs of means μ_i and μ_j which are not equal if at least one μ_i is not equal to another μ_j then there is at least one treatment which is effective and different from the others.

So at least one of the τ_i values is not equal to 0. Just going back if all the τ_i values are 0 then what will happen? μ_i will become equal to μ , $\mu_i = \mu + \tau_i$, i running from 1, 2, 3 and so on to a treatments. So, when τ_i is 0 then none of the treatments are producing the change from the global response. Okay, that is the view taken by the null hypothesis but the alternate hypothesis says among a treatments running from 1, 2 so on to a .

There is at least one treatment which is producing an effect that is different from all other treatments in this case all other treatments. In this case all other treatments are producing no effect and there is only one treatment which is producing an effect. So, the number of treatments which are actually producing effects may be different there may be one treatment which may be different from all others.

Or all the treatments may be different from each other and hence all the μ_i 's may be different from each other and from the global value μ .


(Refer Slide Time: 34:02)



Model Interpretation

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad \text{or} \quad y_{ij} = \mu_i + \epsilon_{ij}$$

Different treatment means could be different. There is a spread about these treatment means μ_i ($i=1,2,\dots,a$) due to the random error component which has mean 0 and variance σ^2 .

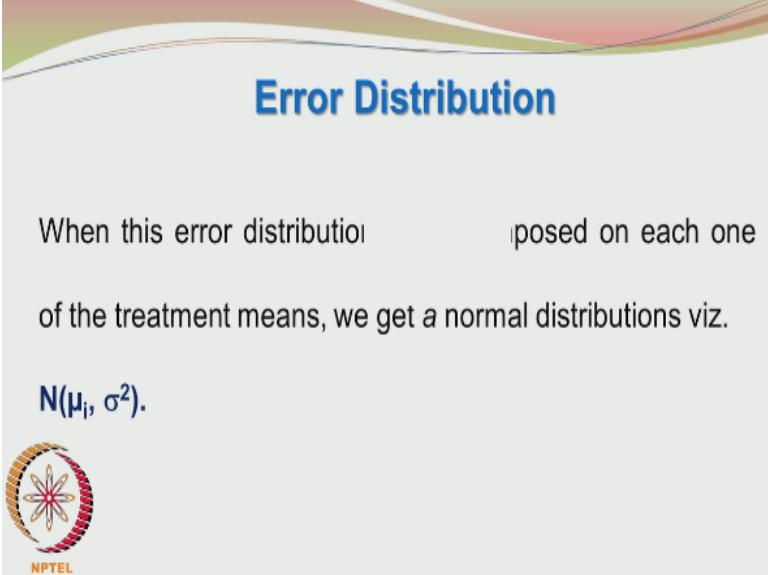


So, you are essentially having Y_{ij} which is the response and it is the combination of the treatment effect + the random fluctuations effect. If you go back to the graph I like this graph very much if there was no noise what would have happened is we would have got two values, unique values μ_i and μ_j it would have been a straight line. A direct delta impulse so that means that you have got a unique value μ_i and μ_j which are different from each other.

However, the values are spread about μ_i and μ_j because of random factors the random errors components with variance σ^2 and so that causes a spread in these deviations. The extent of the spread is the same in both these cases what I am trying to say is both μ_i and μ_j are spread in an identical fashion. Only thing is the center of this distribution is μ_i and the center of the next distribution is μ_j however the spread is the same in both these cases.

Because the error is assumed to be normally distributed with 0 mean and variance sigma square and of course then these are also normal distributions.


(Refer Slide Time: 35:52)



Error Distribution

When this error distribution is superimposed on each one of the treatment means, we get a normal distributions viz.

$N(\mu_i, \sigma^2)$.



NPTEL

So when this error distribution is superimposed on each and every one of the treatment means we get a normal distributions which are having a mean value or spread around μ_i , i running from 1 to so on to a and constant variance sigma square. This a is not we get a normal distribution we get a normal distributions. So, we have to resolve the total sum of squares. How to get the total sum of squares I will tell in the moment?

We resolved the total sum of squares into errors sum of squares and treatments sum of squares. Whenever we found the variance what did we do? We found the mean first and then we subtracted from each of the number the average or the mean value then we squared it. So, we had square of the deviations and then we divided the square of the deviations by $n-1$, where n is the number of data points.

This gave us the variance. Exactly the same concept we are going to apply here. But we are going to have different types of sum of squares and that will become obvious in a moment.

(Refer Slide Time: 37:29)

Resolution of Total Sum of Squares (SS_T)

The sum of squares of the differences between individual responses and overall treatment mean is resolved into

a. error sum of squares

and



treatment sum of squares

So, we are essentially looking at errors sum of squares and treatment sum of squares and treatment sum of square.

(Refer Slide Time: 37:34)

Resolution of Sum of Squares

Let us find out the measure of the overall variability in the experiments. It may be calculated from the following

Total Sum of Squares

$$\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = SS_T$$



The total sum of squares represents the deviation from each and every experimental response from the global average value \bar{Y} or more correctly $\bar{y}_{..}$ is nothing but the global average. So each and every experimental observation is subtracted by the global average and these deviations are square. Obviously, if we do not square them and we sum all these deviations they will become 0.

But when we square them all the negative deviations as well as the positive deviations will now be only > 0 and hence their sum will not be equal to 0 usually. Miraculously if all the observations are exactly matching with the mean value than the sum of squares will be 0 but they are very, very unlikely. So, anyway to emphasize my point $i=1$ to a , $j = 1$ to n , i index standing for treatment.

We are having a treatments, j index standing for repeats we are having n repeats and we take the square of the deviations we get total sum of squares.

(Refer Slide Time: 39:04)

Resolution of Sum of Squares

The total sum of squares may be eventually resolved into two meaningful entities viz. the treatment sum of squares and error sum of squares

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

NPTEL

Very interesting mathematical manipulations are possible unfortunately time does not permit us to get into all these nice derivations for some people these derivations may look very complex but it is very nice. It is a pity that there is not enough time to get into all these mathematical derivations which will bring out the elegance and beauty of statistics in their full glory. But we will take the main result.

And move on. $i=1$ to a , $j=1$ to n , $\sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2$ may be split into two components that is $n \sum_{i=1}^a (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{i.})^2$. Before we got next you please try to look this equation and see what they are actually representing. If I do not get distracted by n and all this summation this $\bar{y}_{i.}$ will cancelled out with $\bar{y}_{i.}$.

And so you are essentially having $\sum_{i,j} (Y_{ij} - \bar{Y})^2$ – so this is canceling out so I am getting $\sum_{i,j} (Y_{ij} - \bar{Y})^2$ which is equal to this one. You may argue that this linear combination is not possible because you are squaring the terms. If I had not put the double summation and I had not multiplied by n then I can write $\sum_{i,j} (Y_{ij} - \bar{Y})^2$ in terms of adding and subtracting the \bar{Y}_i to $Y_{ij} - \bar{Y}$ anyway so you get the point I think.

There is also another interesting interpretation to this if you know Pythagoras' theorem or remember it of course then the sum of the square of the hypotenuse = sum of the squares of the other two sides of the right angled triangle. So, the same concept is being applied here the sum of the squares may be resolved into two components one due to the treatments and other due to the error. So, if you look at this closely. Let me see.

(Refer Slide Time: 42:04)

Resolution and Interpretation of SS_T

$$SS_T = SS_{\text{Treatments}} + SS_E$$

Total Sum of Squares =

Sum of Squares due to Treatments

+

Sum of Squares due to Error

NPTEL

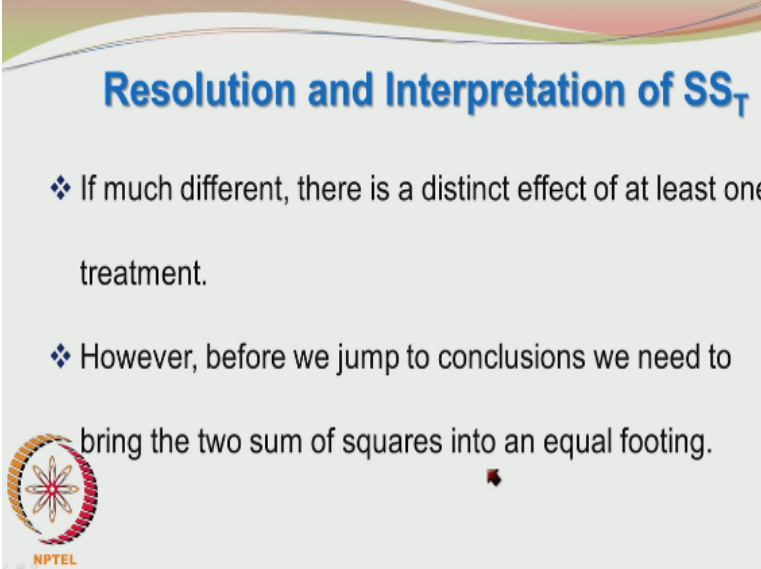
You are seeing the same thing total sum of squares is sum of squares due to treatments and sum of squares due to error. Here $\sum_{i,j} (Y_{ij} - \bar{Y})^2$ represents the deviation of the individual observation from the global mean. This is the representation of the treatment mean from the global average. This is the deviation of the individual observation from the treatment mean. So, we are doing repeats for each treatment we have done n repeats.

For each treatment we have averaged the end repeats we get \bar{Y}_i that is the treatment mean we are comparing the treatment mean with the global average. So, this is the contribution from treatment sum of squares. Here we are considering the global average sorry we are not considering the global average. Here we are considering the individual response with the treatment mean.

So, what we are doing or how we are doing rather is for a given treatment we are comparing the individual response for that treatment with the treatment average. If the error contribution was negligible or not present whenever we repeated the experiment we would have got the same response Y_{ij} in which case the Y_{ij} would have been same as \bar{Y}_i okay, they would have been the same but each repeat for a given treatment itself is producing some variation.


And that is how the error contribution comes in so we are modeling the error contribution by this sum of squares. So, with that out of the way we can now compare the treatment contribution with the error contribution and the error contribution and the treatment contributions are comparable then we can say that the treatments are not really having any effect.

(Refer Slide Time: 44:41)



Resolution and Interpretation of SS_T

- ❖ If much different, there is a distinct effect of at least one treatment.
- ❖ However, before we jump to conclusions we need to bring the two sum of squares into an equal footing.

 NPTEL


So rather than looking at the total sum of squares and comparing the treatment sum of squares and errors sum of squares we have to normalize the sum of squares for each term. Because each term in the sum of square equation have has different degrees of freedom.

(Refer Slide Time: 45:14)

Degrees of Freedom Analysis

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

Degrees of Freedom of Total Sum of Squares:

$$an - 1 = N - 1$$


Let us now look at the degrees of freedom. Here you are having a into n observations but not all of them are independent. Of course all of them are important but not all of them are independent in the sense $\sum_{i,j} (y_{ij} - \bar{y}_{..}) = 0$ if I am adding the sum of the deviations from the mean will be equal to 0. So, if I am calculating the global mean from the responses then I need to know only $n-1$ or sorry $a - 1$, Y_{ij} values.

So, knowing $a - 1$ Y_{ij} and the global average I can find out what is the remaining value. So, we have totally $a-1$ degrees of freedom. The same argument we can apply to the error sum of squares.

(Refer Slide Time: 46:10)

Degrees of Freedom Analysis

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

Degrees of Freedom of Error Sum of Squares:

$$a(n-1)$$



Forget about the treatment for the time being. Let us say we are having a particular treatment and we have found the treatment average based on the n repeats. So, there are only $n-1$ independent entities and so you have $n-1$ and then you are having treatments. So, the degrees of freedom would be $a \cdot n - 1$ it is saying that there are $a \cdot n - 1$ independent entities in this expression. So, this is also out of the way.


And since all the treatment means when averaged will give you the global average there are only $a-1$ independent treatment means. So, either you can argue on those lines or you can subtract the degrees of freedom for this expression with the degrees of freedom for this expression and you will get $a-1$. So, let us see whether it happens like that.

(Refer Slide Time: 47:11)

Degrees of Freedom Analysis

Degrees of Freedom of Treatments Sum of Squares:

(a-1)

$$N-1 = an - a + a - 1 = an - 1$$


The degrees of freedom for the treatment sum of squares just now we show as a-1. So, we have to now find the mean square treatments and the mean square error. The simple thing is the treatment sum of squares are divided by the treatment degrees of freedom the error sum of squares are divided by the error degrees of freedom that is it. We get the mean square treatment and mean square error.

Some of the squares of treatments divided by a-1 sum of the squares of the error divided by a*n-1. The expected values are pretty interesting the expected values of the mean square treatment is sigma square plus this contribution because of the treatments.


(Refer Slide Time: 47:56)

Expected Mean Squares

Expected (Mean Square Treatments) =

$$\sigma^2 + \frac{n}{a-1} \sum_{i=1}^a \tau_i^2$$

Expected (Mean Square Error) = σ^2



The expected mean square for the error is simply sigma square again I am not looking at the mathematical derivations it is quite straight forward. You are having this if the treatments were ineffective the tau i square will all become 0 or close to 0 and we have sigma squared again. So, the variance in the mean squared treatments becomes comparable to the variance with the error.

(Refer Slide Time: 48:32)

Expected Mean Squares

Expected (Mean Square Treatments) = $\sigma^2 + \frac{n}{a-1} \sum_{i=1}^a \tau_i^2$

Expected (Mean Square Error) = σ^2

The expected mean squares error is an unbiased estimator of σ^2 while the Mean Square Treatments is also an unbiased estimator of σ^2 if the null hypotheses were true.

NPTEL

So, since the expected value of the mean square error gives the error variance sigma square we can say that the expected means squared error is an unbiased estimator of sigma square. The mean square treatments also will become unbiased estimator if the null hypothesis were true. That means all the other treatment effects were negligible. All the treatment effects in fact they are negligible and so you get expected mean square treatment is equal to sigma square.


(Refer Slide Time: 49:06)

Expected Mean Squares

Expected (Mean Square Treatments) = $\sigma^2 + \frac{n}{a-1} \sum_{i=1}^a \tau_i^2$

Expected (Mean Square Error) = σ^2

If the null hypotheses were not true, then the expected value of the mean square treatment will exceed the expected value of the mean square error due to the treatment effects.




Then if the null hypothesis were not true the expected mean square treatments will exceed the expected mean square. Obviously, the effects due to the treatments will start kicking in and so this expected mean square treatment will be different from the expected mean squared error.

(Refer Slide Time: 49:28)

Expected Mean

Expected (Mean Square Treatments) = $\sigma^2 + \frac{n}{a-1} \sum_{i=1}^a \tau_i^2$

Expected (Mean Square Error) = σ^2

$$F_o = \frac{MS_{Treatments}}{MS_E}$$



So what we do is here we are looking at two statistics and what we do here is do a F test. I request you to again look at the scope of the F test what we were doing and here we are looking at mean square treatments by mean square error ratio that we relate it to F not. We are essentially looking at the ratios of two variances which is precisely what the F test was doing.

(Refer Slide Time: 50:15)

F Statistic

$$F_0 = \frac{MS_{\text{Treatments}}}{MS_E}$$

The mean square treatments in the numerator is greater than the mean square error in the denominator if at least one or some of the treatments were effective.




The mean square treatments and mean square error will be comparable if the treatments are not having an effect. But the mean square treatment would be higher than mean square error if at least one of the treatments or more of them are making a significant contribution. So, we have to see whether they are really significant.

(Refer Slide Time: 50:37)

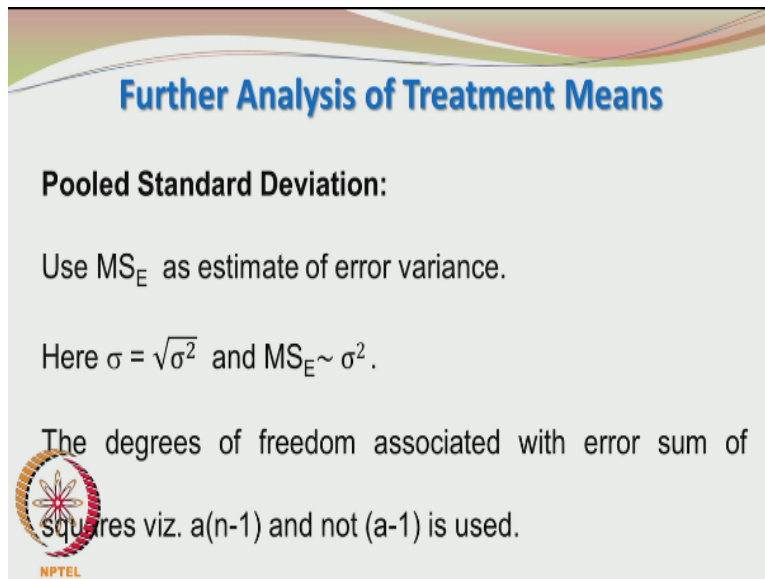
Analysis of Variance Table (ANOVA)

Source of variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$SS_{\text{Treatments}}$	$a-1$	$MS_{\text{Treatments}}$	$\frac{MS_{\text{Treatments}}}{MS_{\text{error}}}$
Error	SS_{Error}	$a(n-1)$	MS_{error}	
Total	SS_T	$an-1$		



So, we can set up the analysis of variance table where we list down the treatment error and total. So we have sum of squares due to treatment, sum of square due to error and total sum of squares the degree of freedom are $a-1$, $a*n-1$, $an-1$ is the total degrees of freedom when we divide the sum of squares by the respective degrees of freedom we get the mean square treatment and mean square error then we take the ratio of these two to find F not.

(Refer Slide Time: 51:08)




Further Analysis of Treatment Means

Pooled Standard Deviation:

Use MS_E as estimate of error variance.

Here $\sigma = \sqrt{\sigma^2}$ and $MS_E \sim \sigma^2$.

The degrees of freedom associated with error sum of squares viz. $a(n-1)$ and not $(a-1)$ is used.



So, we conclude at this point and we will continue in the next lecture. Thank you for your attention.