Total Quality Management-II

Prof. Raghunandan Sengupta
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 08
The Analysis of Variance (ANOVA) – I

A very good morning good afternoon good evening my dear friends and students welcome to this TQM two; NPTEL MOOC course and this is the eighth class or eighth lecture which we are going to have and I am Raghunandan Sengupta from IME department, IIT, Kanpur. So, if you remember, in the almost till the middle of seventh class or I would not use the word middle almost in the fag end of the seventh class, we were discussing about general concepts or random variable expected values then second moment which was something to do with the variance, then the distributions vary not in depth, but in generally the concept for distributions.

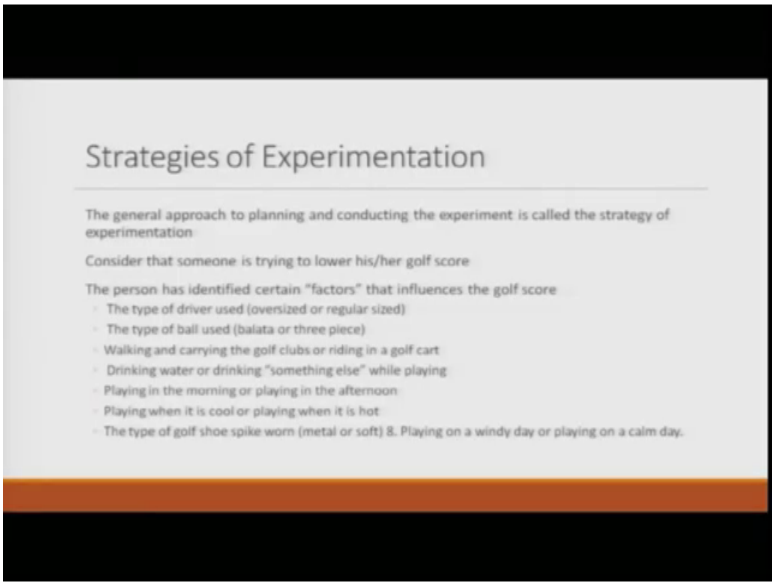
Then we went in the concept of trying to understand fear the inequalities Tchebyshev's and Markov inequality, then invent took the concept of an estimation provided we pick up a sample from a population, what are the 2 characteristics of unbiasedness and consistency, then we consider the other 3 distributions based on normal distribution was which was chi f and t and then we went into discussing, what are the different estimates further on we ventured and discussed and had a look at interval estimation what is the concept of level of confidence.

Then the concept of degrees of freedom came up and how these three distributions were formed considering the normal distribution and the Baye's distribution, then we went to in the concept of what do you mean by hypothesis testing, what is alpha and beta errors, how we discussed and gave a very nice example in the area of banking where a loan has to be given and there is a cut off score.

Then we discuss the different type of hypothesis testing rules and I did mention time and again something to do with the mean or the first moment you will definitely use the z or the t and something to do with the variance of the population you want to find out something some statements you will definitely use the chi and f and; obviously, we understood what was X bar the sample mean S without the dash, S with the dash and how the degrees of freedom could reduce by 1 or 2 depending on the example, like for the f distribution, it was lost 1 1 in both the populations. So, it was m minus 1, n minus 1 and in the case of t distribution, we saw I mean when it was just to do with one population, it was one degrees of freedom lost and then we are further on saw that it was m plus n minus 2 this minus 2 coming because it was being lost one from the first population 1 from the second population, then we in the in the last 5 minutes of the seventh lecture, we discussed the what is an experiment what are the errors and those were qualitative fields and slowly, we will start off the discussion this with this words let us start with full we got the lecture number 8.

Now, the strategies for when you are doing an experiment because you want to test something and trying to find out what is the best set of input variables as the white noises or the uncontrollable so called errors can be minimized.

(Refer Slide Time: 03:40)



Strategies of Experimentation

The general approach to planning and conducting the experiment is called the strategy of experimentation

Consider that someone is trying to lower his/her golf score

The person has identified certain "factors" that influences the golf score

- The type of driver used (oversized or regular sized)
- The type of ball used (balata or three piece)
- Walking and carrying the golf clubs or riding in a golf cart
- Drinking water or drinking "something else" while playing
- Playing in the morning or playing in the afternoon
- Playing when it is cool or playing when it is hot
- The type of golf shoe spike worn (metal or soft) 8. Playing on a windy day or playing on a calm day.

So, the general approach to planning and conducting the experiment is called strategy of X of for experimentation consider that someone is trying to lower his or her golf score, this may be either you want to basically try to reduce your blood pressure, it may be that you want to basically increase your input in trying in doing some work.

Or it may be that you want to jog or walk more in order to reduce your fat or basically be fit on all this is so, in this example, considering game of golf anyone to reduce the score. So, the deduction the score basically means that you want to be at par or below par such that your score is much better lower the score better the type. So, you will try basically try to find out the type of drive or the club which you use whether they are oversized undersized whether you want one use the club of size 1 or 2 or 3, then you only basically try to find out what type of ball you will use.

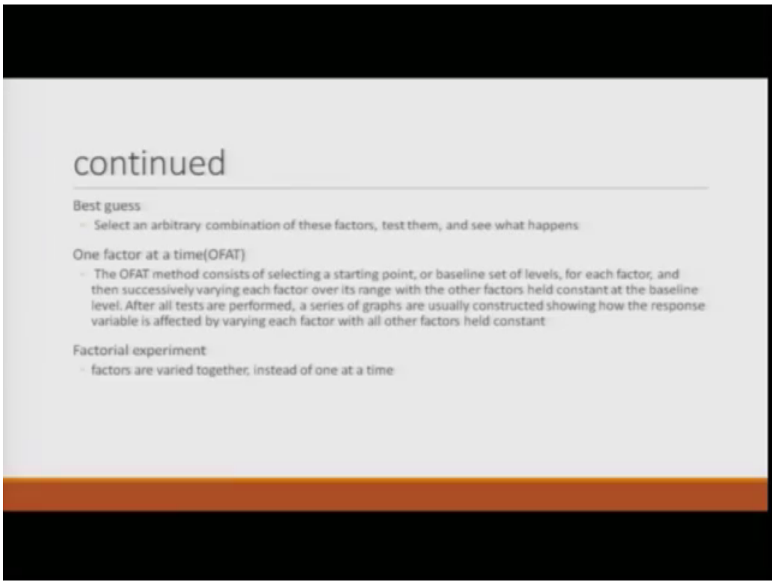
So, in India; obviously, we generally are not very much aware of golf as a game, but there are different type of balls which are there and if you have seen a golf ball, it is basically or not almost, but its likewise the size of a of a table tennis ball, but with dimples. So, inward dimples, which basically reduces drag and let us the ball have a smooth flight for a long distance these are just the information which I am giving if somebody is interested to check.

So, you will also try to find out what is your are working and the carrying in the golf ball or adding the golf cart, what you would try to do and whether you will try to take a caddy caddies, those persons would you carry your golf bag with all the clubs. So, we want to have a pull cart or go in a golf cart or have a caddy and whether the caddy who is allowed to give you some advice I change the golf the club and so on and so forth. So, whether you would be basically drinking water or several drinking juice when you are playing the game whether that is actually allowed and you want to basically refresh up yourself.

Because it is it continues for long and you have to walk for a long time and for quite a distance because if there are 18 holes, the total distance travelled would be in the tune of about 8 to 9 kilometres. So, you would like to play in the morning or the afternoon depending on the weather you want to play in, it is cool or hot and the type of golf shoe, the spike one whether they are metal or they are non metallic soft, whether they, you get a good grip on when you are basically trying to take a swing. So, all this basically would be studied in order to understand whether you can reduce your score.

So, whether you want to play on a windy day or a calm day. So, windy day basically means there is lot of wind factor and the deviation of the ball would happen in such a way that will be of the dark it larger than for the case when it is a cool and a calm day I would not use the word cool day, but calm day means not a windy day.

(Refer Slide Time: 06:56)



continued

Best guess

- Select an arbitrary combination of these factors, test them, and see what happens

One factor at a time(OFAT)

- The OFAT method consists of selecting a starting point, or baseline set of levels, for each factor, and then successively varying each factor over its range with the other factors held constant at the baseline level. After all tests are performed, a series of graphs are usually constructed showing how the response variable is affected by varying each factor with all other factors held constant

Factorial experiment

- factors are varied together, instead of one at a time

So, best guess would base basically you will see, select an arbitrator set of combinations of these factors and then try to find out; what is the best set of combinations for the external factors which will help you to reduce your score and basically pull you up with the ranking with your other players or you want to basically perform better. So, what is generally understood is basically one factor at a time.

So, this method consists of selecting a starting point or a baseline from where you will basically try to compare yourself or compare the overall effect of an experiment or hour or an existing condition for each effect, you take a baseline and then successfully vary each factor over its range with the period for a particular period of time such that the all the combinations would give you some results which will basically be able to based on which you will be able to judge whether the planning which you are doing for the experimentation or the word experimentation which we were using in the normal sense is really making sense some scientific deductions can be made.

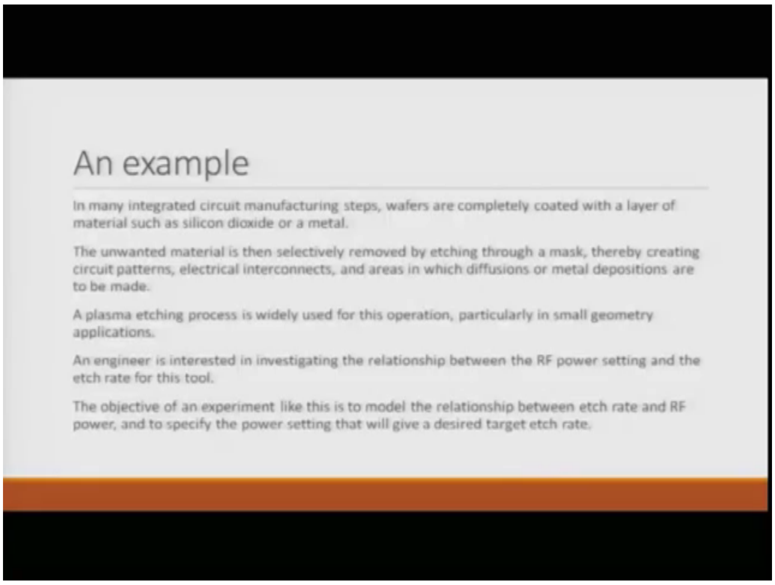
So, as you change each factor over this range; obviously, will keep the other factors constant at the baseline value whatever you started, say for example, you think that you will change the your timing of the play you want to basically play in not in the hot, but in the cool weather so; obviously, everything being constant you will continue practicing for and playing for a long on a cool day and then try to check how your performance is then say for example, you will keep the other factors one factor changing other factors being constant and try to basically compare how your performance is.

After all tests are performed in different combinations a series of graphs or basic analysis are usually constructed showing how the response variable has, is effectiveness most variable is basically the output which you want you are trying to understand that whether your score is reduced or whether you have reduced weight if you are basically jogging, say for example, jogging on walking you think, I am not shifting the whole focus of the discussion I am giving a different example say for example, you think that walking brisk walking in the morning is better or else you may think that you cannot work very basically because in the morning or little bit tired. So, if you take some time before you really get in the mood.

So, hence you will basically try to try to or like to switch on to the evening time maybe you would like to basically change your path or change your shoes or basically wear a knee cap or in the winters of the summers you would like to change your timing accordingly such that you basically give out your maximum amount of energy because loss of calories would basically burn out your fat. So, you are able to reduce a weight and be much healthier in the general sense. So, when we do that. So, you will basically constructing showing how the response variables are affected by varying each factor with all the factors constant.

So, when you do that it is a basically a factorial experimentation which is being done when factors are very together instead of one at a time. So, they are very together in such a way that you are able to understand the collective overall effect on the on the response. So, in one factor at a time you will basically change one keep other fixed and in when you are doing a factorial explanation you will basically do a combination of them. So, what is the important factor which is basically giving the results if at all you may not be able to judge, but considering the overall different combinations you may make some good guesstimate or estimate such that you can take corrective actions in this direction.

(Refer Slide Time: 10:45)



An example

In many integrated circuit manufacturing steps, wafers are completely coated with a layer of material such as silicon dioxide or a metal.

The unwanted material is then selectively removed by etching through a mask, thereby creating circuit patterns, electrical interconnects, and areas in which diffusions or metal depositions are to be made.

A plasma etching process is widely used for this operation, particularly in small geometry applications.

An engineer is interested in investigating the relationship between the RF power setting and the etch rate for this tool.

The objective of an experiment like this is to model the relationship between etch rate and RF power, and to specify the power setting that will give a desired target etch rate.

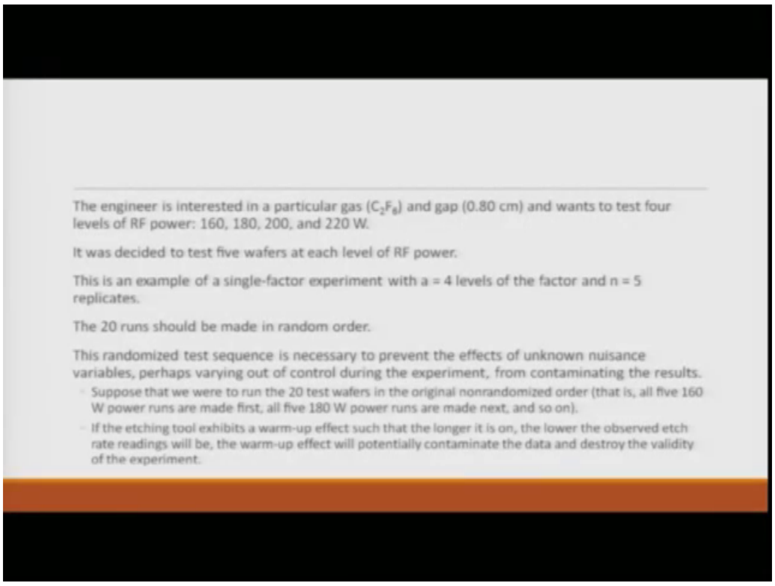
So, let us consider an example or in inequality with sense in many integrated circuit manufacturing steps wafers are completely coated with layers of materials such as silicon dioxide or a metal then the unwanted material is then selectively removed by etching. So, this can be chemical etching and or electrical matching through a mask thereby.

So, basically you have a wafer you put this the very thin layer in microns of a silicon diet and then you put a mass like mass is basically you want to write something on the wall. So, and you have a stencil is basically we generally use the right word when we generally use in classes you want to use a stencil. So, the writing is very nice. So, the mask would be put on the wafer which is already coated as I mentioned and then you etch out using some chemical process or a micro nano electrical process.

So, that the circuit is basically made. So, that is how the thing is done. So, on unwanted material is then selectively removed by etching through a mass thereby creating circuit patterns electrical interconnections and connections are made and areas in which diffusions or metal depositions are to made are done a plasma etching process is widely used for this operation particularly in small geometry applications and engineer is basically interested in to investigate the relationship between the RF factor setting and the etch rate for this tool says that the etching is not too much or not too less the objective of an experiment like this is to model the relationship between each etch and the power.

So, as to specify the power setting that will give a desired target rate for these date such that if there is more etching or less etching; obviously, the circuit basically gets damaged it is not off no you. So, you have to basically have a particular rate of etching depending on which the circuit is absolutely perfect. So, there is no so called short circuit and or loops are not there where the current does not flow. So, this is the main aim when you are doing integrated circuits IC manufacturing is being done.

(Refer Slide Time: 13:08)



The engineer is interested in a particular gas (C_2F_6) and gap (0.80 cm) and wants to test four levels of RF power: 160, 180, 200, and 220 W.

It was decided to test five wafers at each level of RF power.

This is an example of a single-factor experiment with a = 4 levels of the factor and n = 5 replicates.

The 20 runs should be made in random order.

This randomized test sequence is necessary to prevent the effects of unknown nuisance variables, perhaps varying out of control during the experiment, from contaminating the results.

- Suppose that we were to run the 20 test wafers in the original nonrandomized order (that is, all five 160 W power runs are made first, all five 180 W power runs are made next, and so on).
- If the etching tool exhibits a warm-up effect such that the longer it is on, the lower the observed etch rate readings will be, the warm-up effect will potentially contaminate the data and destroy the validity of the experiment.

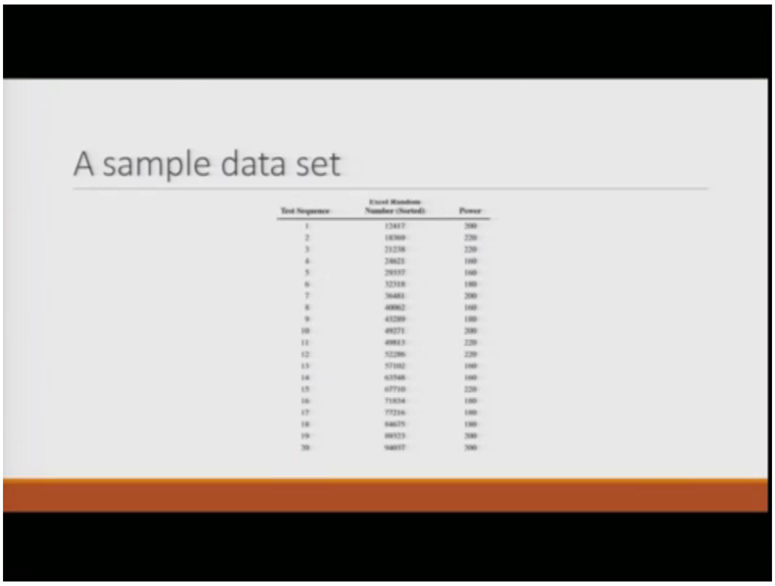
So, the engineering interested in a particular gas and the gap say for example, is 0.80 centimetres and want to test for levels of powers which are 160, 180, 200 and 220 depending on that you will want to find out what is the rate of etching it was decided to test 5, it wafers at each level of this power. So, for 180; 60, we will test 5 of them and find out the average performance is similar to do for 5 different other IC circuits which has been in the raw stage like wafer with a coating then again for 2 hundred you and I have another 5, then for 220, you have another 5.

So, this is example of a single factor; that means, you are keeping everything constant or you are only changing the power to find out the performance. So, this 4; 4 different wattages into 5 wafers in each this 20 run should be made in a random order. So, you need not do 160, 160 all at one would can be 160, 200, 220, 160, 180 and in this sequence any of them, this randomized test sequence is necessary to prevent the effects of unknown nuisances.

So, variables which you cannot control perhaps varying out the control during the experiment is best from contaminating things in the results suppose that we were to run twenty test wafers in the original non randomized order that is all five 1 to 60 first, then 180 and so on and so forth if the etching example exhibits a warm up its. So, in say for example, when you are doing it 160, one at a time first go then 180. So, what you are doing you this slowly increasing the power. So, that may happen have an effect undesirable which should not be there in the experiment and it would contaminate I am using the word contamination in a very general sense that will get contaminate your results.

So, the etching tool exhibits a warm up effect such that longer, it is on the lower the observed h ratings will be the warm up effect will potentially contaminate the data and destroy the validity the experiment based on which you can find some meaningful conclusions.

(Refer Slide Time: 15:18)

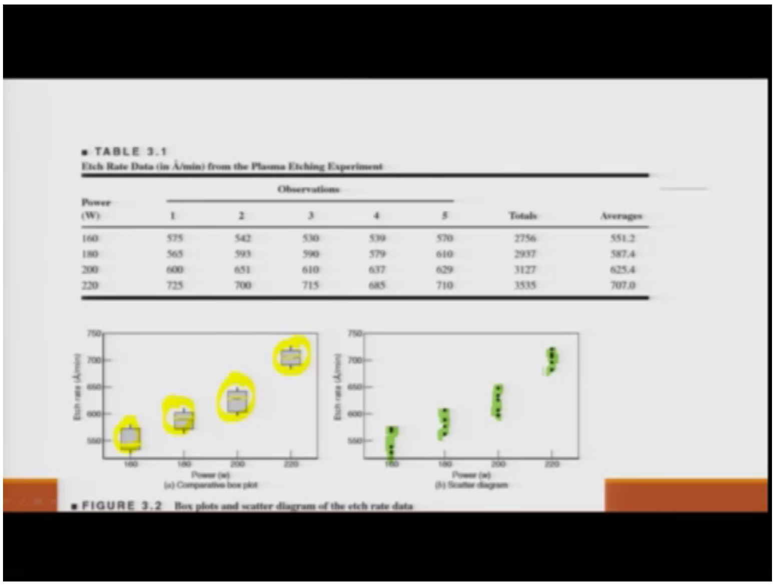


A sample data set

Test Sequence	Excel Random Number (Sorted)	Power
1	12417	200
2	18369	220
3	21238	220
4	24621	160
5	29337	160
6	32318	180
7	36481	200
8	40062	160
9	43289	180
10	49271	200
11	49813	220
12	52286	220
13	57102	160
14	63548	160
15	67710	220
16	71834	180
17	77216	180
18	84675	180
19	89323	200
20	94037	200

So, this is a set of example. So, 1 to 20 are given for the example the random number sorted out are given random numbers based on how whichever the sequence of etching you are doing and the power sequence are given. So, may pay attention to the last most column which is the third column you see the powers are changing from 200, 220, then again 220, 160, then 160, 180. So, this is a random test based on which you are trying to do. So, there is no such biasness which you willingly try to basically force into the system. So, so taking this information from as I said it is basically the Montgomery book.

(Refer Slide Time: 16:11)



■ TABLE 3.1
Etch Rate Data (in Å/min) from the Plasma Etching Experiment

Power	Observations						
(W)	1	2	3	4	5	Totals	Averages
160	575	542	530	539	570	2756	551.2
180	565	593	590	579	610	2937	587.4
200	600	651	610	637	629	3127	625.4
220	725	700	715	685	710	3535	707.0

■ FIGURE 3.2 Box plots and scatter diagram of the etch rate data

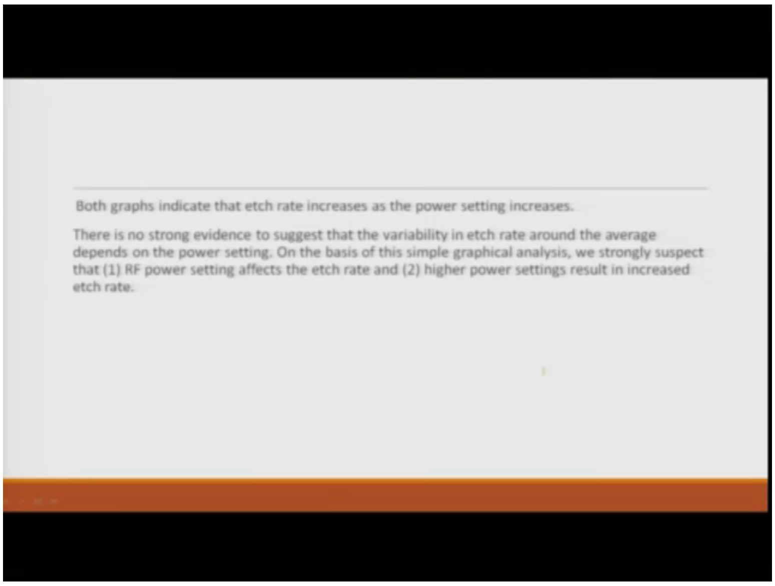
So, etch rate the table basically discusses the etch rate in angstrom per minute from the plasma etching experiment which you have done. So, you have basically observations which I marked as 1, 2, 3, 4, 5 and the powers are given on the leftmost column that is starting from 160, 180, 200, 220 and the totals are given from where you find out the averages. So, they there are say for example, for 160, the observations are in that is angstrom per minutes.

So, they are if you consider the row corresponding to 160. So, they are a 575, 542, 530, 539, 570. So, the total is 2 7 5 6 u divided by 5 when you get the averages. So, if you look at the averages which is again giving on the rightmost column they are those corresponding to 551.2 so and so forth till the last one which is 707.0 are for power values starting from 160 to 220.

So, if you if you basically plot them on a scatter plot on with basically power on the x axis and the etch rate on the y axis. Similarly, you do a competitive box plot with power rate on the x axis and an etch rate on the y axis. So, you basically get the box plot which you will I try to highlight. So, these are the box plots for 160, 180, 200 and 220. So, I will just highlight it first here and then go into the next diagram which is for the power with respect to the etch rate and you are trying to basically do the scatter diagrams. So, what I want to do is like this let me check the colour yes. So, if you see this middle line which is the median and this box.

So, the part above vertically above the median and below the median are the respective quintiles greater than or less than with respect to the median because the values are increased along the y axis as you go vertically up. So, this gives you the values of with respect the median and the quintile plots or the competitive box plots. So, if I go to the case where I want to understand, let me use the colour with respect to the scatter plot. So, the scatter plots are given for 160, 180, then 200, 220. So, they give me the points, but they do not give me where the median is or the median mean is. So, we can find it out and over study about that.

(Refer Slide Time: 19:09)



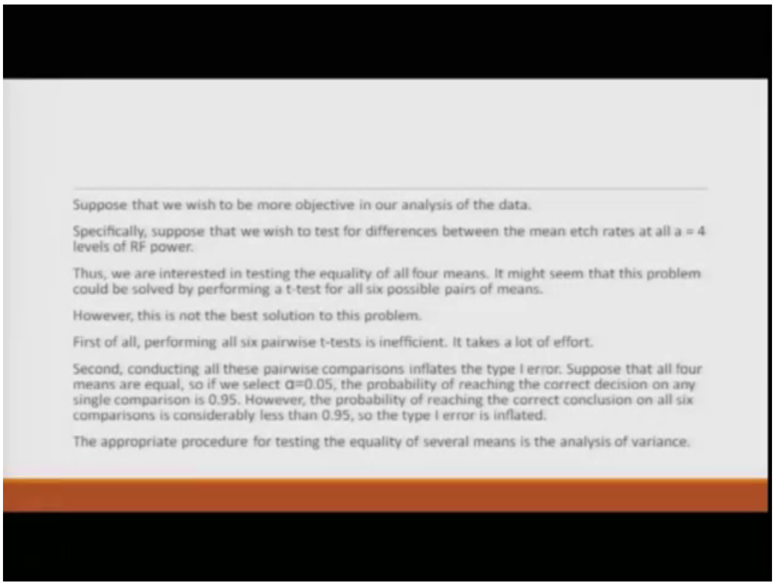
Both graphs indicate that etch rate increases as the power setting increases.

There is no strong evidence to suggest that the variability in etch rate around the average depends on the power setting. On the basis of this simple graphical analysis, we strongly suspect that (1) RF power setting affects the etch rate and (2) higher power settings result in increased etch rate.

Both graphs indicate that the etch rate increases as power setting increases there are no strong evidence to suggest that the variability in etch rate around the average depends the power setting. So, for average; obviously, would be changing, but considering there is huge amount of fluctuation in and around the mean value with respect to the power we cannot generally find out. So, on the basis of the simple graphical analysis, we strongly suspect that there bar setting affects the etch rate. So, so there is no strong evidence to suggest the variability in each etch rate depends on the power setting on the basis of the simple graphic and current condition we strongly suspect that the power settings effects the etch rate and high power settings results in increase etch rate so; obviously, I am not talking about the variability is the average value which is changing.

So, average values are the median value I said so; obviously, they would another horizontal line would be mean value so; obviously, that would change depending on the power setting which is their which I am putting for this example suppose that we wish to be more objective analysis of this data.

(Refer Slide Time: 20:18)



Suppose that we wish to be more objective in our analysis of the data.

Specifically, suppose that we wish to test for differences between the mean etch rates at all a = 4 levels of RF power.

Thus, we are interested in testing the equality of all four means. It might seem that this problem could be solved by performing a t-test for all six possible pairs of means.

However, this is not the best solution to this problem.

First of all, performing all six pairwise t-tests is inefficient. It takes a lot of effort.

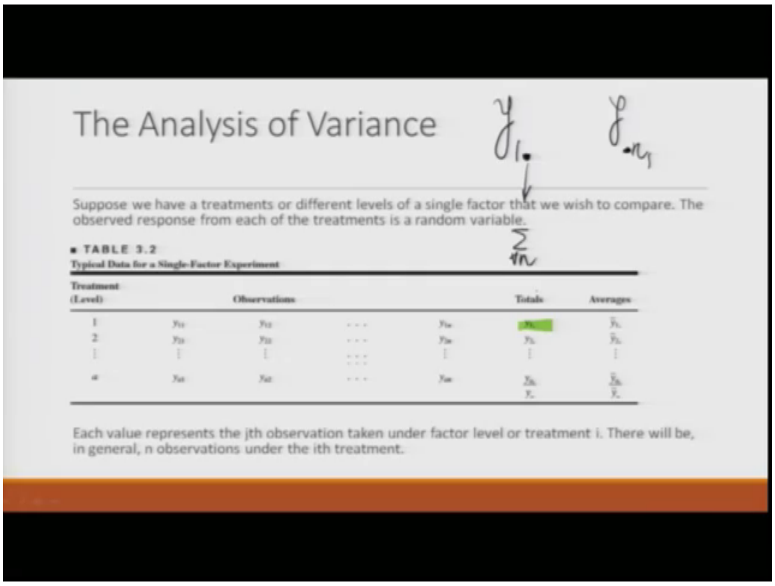
Second, conducting all these pairwise comparisons inflates the type I error. Suppose that all four means are equal, so if we select α=0.05, the probability of reaching the correct decision on any single comparison is 0.95. However, the probability of reaching the correct conclusion on all six comparisons is considerably less than 0.95, so the type I error is inflated.

The appropriate procedure for testing the equality of several means is the analysis of variance.

Specifically, suppose that we wish to test for different, but for differences between the mean h rate at all levels. So, all these levels are 4 levels which I am considering for 160, 180, 200, 220, thus we are interested in testing the equality of all the 4 means, it might seem that this problem could be solved by performing a simple t test for all the 6 possible pairs of means; however, this is a no it is not the best solution to the problem. So, t test is basically remember, now things would be very clear why we a did discuss and spend a decent amount of time about 7 lectures live aside the 5, 5 minutes at the end of the seventh lecture to discuss about different type of concept of hypothesis testing and so on and so forth hypothesis testing; obviously, was in the later part, but before that we just build up the theory for probability distribution then estimation problem point estimation interval estimation.

First of all, performing all 6 pair wise t test is inefficient. So, it will take a lot of effort. So, in that case what should be the answer second conducting all these p competitive pair wise comparison influences are type one error because you are adding up and you are trying to basically compare them so; obviously, it will it will be on a on a additive mode also consider that all 4 means are equal. So, if we select alpha is 0.05 level of confidence or one minus alpha, whatever it is, the probability of reaching the correct decision on any single component on comparison would be a would be 1 minus 0.05 which is 95 percent; however, the probability of reaching the correct conclusion on all 6 comparison is contributed be less than 0.95. So, the type one error has been infinity because if you are continuing one or the other they would be if there may be some effect so; obviously, those concept or interrelationship would be coming up, but which is not being considered the appropriate procedure.

(Refer Slide Time: 22:27)



The Analysis of Variance

$y_{1\cdot}$ $\quad y_{\cdot n_1}$

Suppose we have a treatments or different levels of a single factor that we wish to compare. The observed response from each of the treatments is a random variable.

$\sum_{n}$

■ TABLE 3.2

Typical Data for a Single-Factor Experiment

Treatment (Level)	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}	$y_{a.}$	$\bar{y}_{a.}$
					$y_{..}$	$\bar{y}_{..}$

Each value represents the jth observation taken under factor level or treatment i. There will be, in general, n observations under the ith treatment.

So, now, here comes our technical solution the appropriate procedure for testing the quality equality of several means is the analysis of variance. So, now, what is the general background of analysis of variance, suppose, we have a treatment on different levels of a single factor. So, they can be single and slowly we will go to the multiple factor that we wish to compare the observed response from each of the treatment is a random variable and it is given in the table three point 2 again it has been taken from Montgomery as I did mention and for the other 2 figures. So, the treatment levels are given on the first column starting from 1, 2, 3, 4, till 8 and the observations for each treatment.

So, for say for example, so, for the time being, we will consider the number of observation for each treatment are equal so; obviously, they may change, but for the time being we are considering them to be equal for treatment for level one the corresponding values which we have observed they are n in number n is in Nagpur.

So, these values are 1 y suffix 1 1 y suffix 1 2 till y suffix 1 n so; obviously, the first one if you remember y the suffix one is basically related to the treatment number and the second number is basically related to the observation. So, it will be y 1 1 y 1 2 till dot dot till y 1 n similarly if I go to the last row corresponding with the treatment level a it will be y a a is for the treatment for next number is basically further reading.

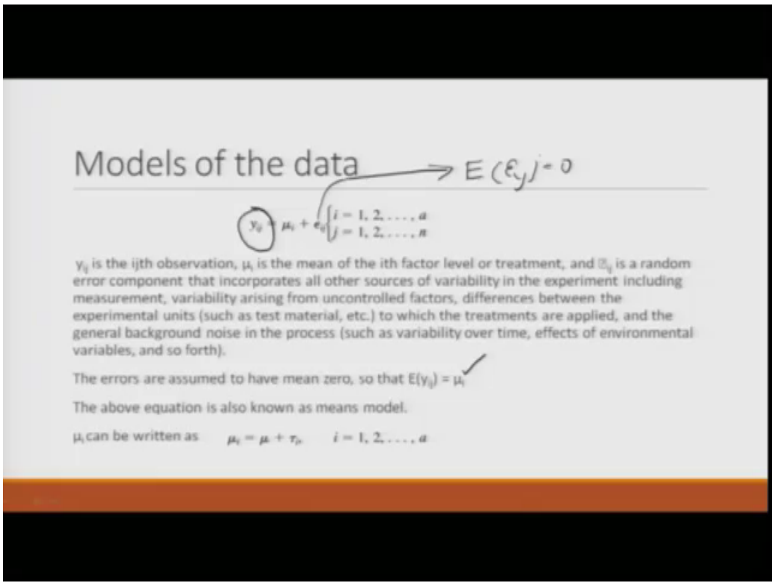
So, it is one then 2 three 4 till n. So, you have basically the totals totals are given by because you would be adding up all the values corresponding to the number. So, it will be y suffix one is for the treatment and then it will be sum up summed up. So, that is why he's given a dot. So, I will try to highlight this for your convenience. So, if you see this there is a dot here which means let me change the colour. So, it would be easier.

So, what I am doing is that it is y 1 dot. So, when I am using this dot it basically means I am summing up all the observation pertaining to n and in case if I write as y dot say for example, n one it would mean that I am basically. So, so the first one which I wrote y one dot suffix is basically sum up along the row and it is y dot and one is basically for any particular observation we are summing up all the one which are along the column. So, they may be relevant later on, but I just wanted it to tell you; what is the nomenclature we are trying to follow.

So, similarly; so, as we do that we have the totals y 1 dot y 2 dot till y a dot if we sum them up; obviously, summation would give you the idea that both would be dots dots basically very simply give you the information they have been some not both along the rows and on the columns we find out the averages where averages nomenclature or would also be very scientific based on what we are doing.

So, it will be y one dot with the bar corresponding to the fact that all the sums have been done for the particular reading number and not the experiment number or the treatment level so; obviously, it can be also put as y bar then it is dot comma n one corresponding the case I am trying to find out the average for a particular reading, but similarly, for all the treatment and numbers. So, each value represents the jth observation taken under factor level or keep in i. So, there will be in general n observations under the ith treatment as I did mention it very clearly.

(Refer Slide Time: 26:22)



Models of the data

$E(\varepsilon_{ij}) = 0$

$$y_{ij} = \mu_i + \varepsilon_{ij}\begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases}$$

y_{ij} is the ijth observation, μ_i is the mean of the ith factor level or treatment, and ε_{ij} is a random error component that incorporates all other sources of variability in the experiment including measurement, variability arising from uncontrolled factors, differences between the experimental units (such as test material, etc.) to which the treatments are applied, and the general background noise in the process (such as variability over time, effects of environmental variables, and so forth).

The errors are assumed to have mean zero, so that $E(y_{ij}) = \mu_i$ ✓

The above equation is also known as means model.

μ_i can be written as $\mu_i = \mu + \tau_i$, $i = 1, 2, \ldots, a$

So, now, what we will try to understand is basically build up the model. So, y i j which is basically for the treatment number I and the jth observation it will basically would have a a some mean value with some error. So, that mean value; obviously, would be mu suffix i plus epsilon i j is basically corresponding to the experiment number and observation number.

So, here I is changing from one to a and j is basically changing from one to n. So, y i j is the i jth observation mu i is the mean of the ith factor level of treatment and and corresponding epsilon is the random error component that incorporates all the other source of variability in the experiment including measurement variability in arising for uncontrol factors different combinations.

It will be divisions between the experiment units which are being used such as the test element to which the treatments are being applied there is a general background noise in the process such as the variability over time effects of environmental variables humidities, they may be related to temperature pressure to whatever it is.

So, the errors have certain assumptions which is the errors are assumed to have a mean value of 0. So, hence the expected value of if I consider this one. So, which I am going to circle now this average value would be basically mu I as rightly mentioned and moreover the errors have an expected value a 0 so; obviously, that would not come the above equation is known as the means model and the mu value can be written as basically corresponding to average mean value with some tau I corresponding to say for example, the etch.

So, generally what I am considering that if you go back to the experiment of 180, 160, 200, 220, they would be one mean value combined for all of them and plus and minus would be the value which will come corresponding to the experiment which are performing corresponding to the fact that I am changing one of the main factors for the experiment. So, this tau value can be plus and minus depending on how we have set the examples.

So, with this I will end the, this lecture eight lecture and continue discussing more about and our models in the ninth and so on and so forth and I am sure that things are becoming number now right in the right frame of mind considering that we are slowly trying to basically go into the design of experiments and consider total quality management in the right perspective.

Thank you very much for your attention, have a nice day, bye.