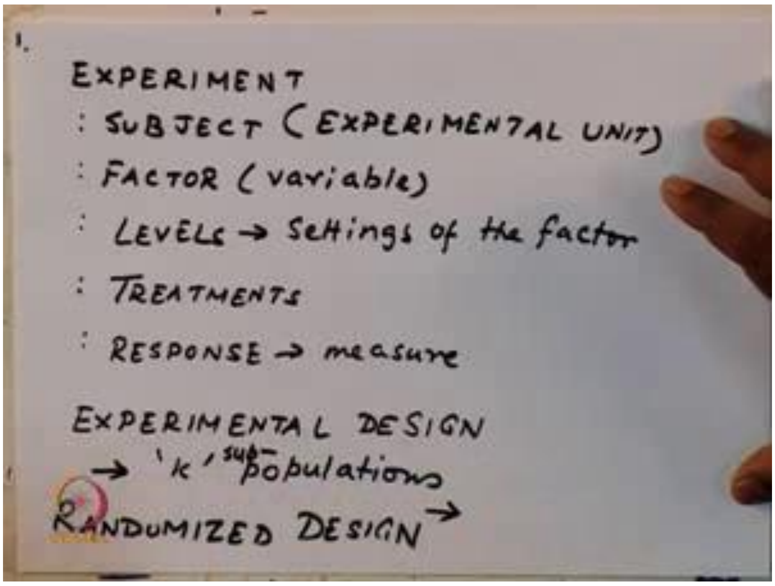
Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 39
ANOVA

Hello and welcome to today's lecture. In the last lecture we had introduced 2 important concepts of analysis of variance and experimental design. So, let us briefly to a recap of these ideas.

(Refer Slide Time: 00:32)



1.

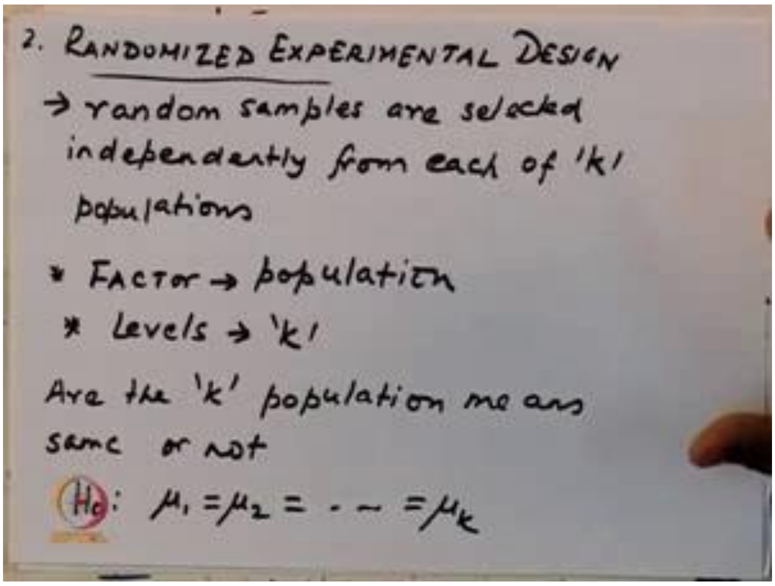
EXPERIMENT
: SUBJECT (EXPERIMENTAL UNIT)
: FACTOR (variable)
: LEVELS → Settings of the factor
: TREATMENTS
: RESPONSE → measure

EXPERIMENTAL DESIGN
→ 'k' sub-populations
RANDOMIZED DESIGN →

So, for an experimental design there are various components of an experiment. The subject on whom you do the experiment, this is also referred as experimental unit, the factor which you are changing. So variable, the levels, levels refer to settings of the factor. You can have another thing called treatments. If there are more than one factor, then you might have more than one treatment and your final response which is what you measure. So, in experimental design there are various ways of doing it.

So, imagine you have k populations, or k sub populations. For a randomized design you choose random samples, this will be right it in the next page.

(Refer Slide Time: 02:29)



2. RANDOMIZED EXPERIMENTAL DESIGN

→ random samples are selected independently from each of 'k' populations

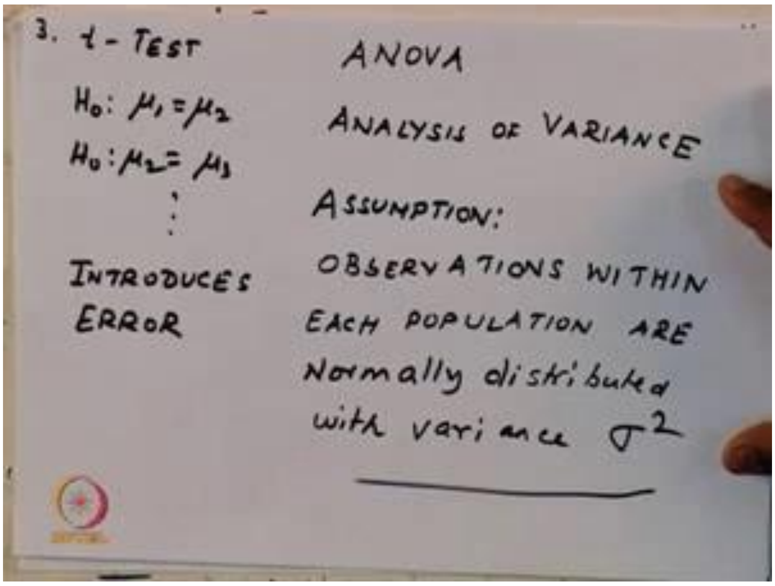
- FACTOR → population
- levels → 'k'

Are the 'k' population means same or not

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$

So, random samples are selected independently from each of k populations. So, this is called a randomized design. So, in this of course, your factor is the population itself. Your number of levels are 4 are k as there are k sub populations. So, the question; with these designs with these experiments you want to addresses are the k population means same or not. So, essentially you want to test the hypothesis, that if you have k populations you want to test mu 1 equal to mu 2 equal to mu k. So, as we discussed in last class you can do this using the t test.

(Refer Slide Time: 04:21)



3. t - TEST

$H_0: \mu_1 = \mu_2$

$H_0: \mu_2 = \mu_3$

⋮

INTRODUCES ERROR

ANOVA

ANALYSIS OF VARIANCE

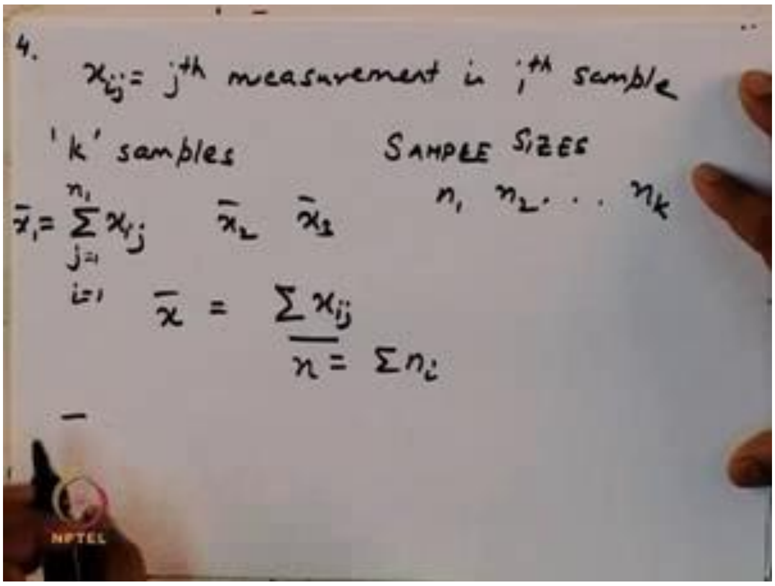
ASSUMPTION:

OBSERVATIONS WITHIN EACH POPULATION ARE Normally distributed with variance σ^2

So, you can do this using the t test, but in that case you will have various hypothesis you have to test. So, for example, you have to test hypothesis mu 1 equal to mu 2, hypothesis mu 2 equal to mu 3 and so on and so forth which is tedious and it also introduces error. So, ANOVA sort for analysis of variance allows you to answer this question that are the k population means same or not in a single a test.

So, what do you exactly do in ANOVA and what are the assumptions? So, the primary assumption of this approach that the observations within each population are normally distributed with variance sigma square, essentially the variance for each population is the same.

(Refer Slide Time: 06:12)



4.

x_{ij} = j^{th} measurement in i^{th} sample

'k' samples

SAMPLE SIZES

$n_1, n_2 . . . n_k$

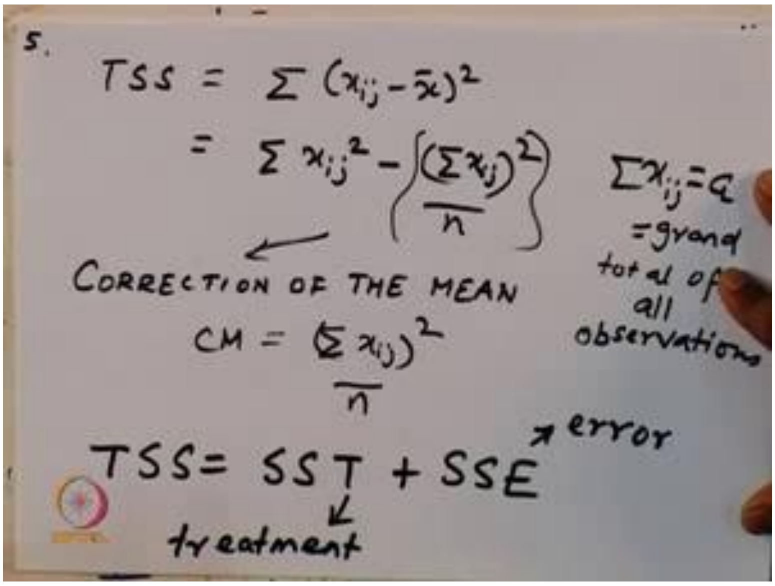
$\bar{x}_1 = \sum_{j=1}^{n_1} x_{ij}$ $\bar{x}_2$ $\bar{x}_3$

$i=1$

$\bar{x} = \frac{\sum x_{ij}}{n = \sum n_i}$

So, in ANOVA let say x ij corresponds to jth measurement in ith sample. In short you have let say k samples or k populations k samples corresponding to k populations. So, you can do summation of x ij for all j and give it i is equal to 1 or 2. So, then you will have the value of x 1 bar. So, similarly you have x 2 bar x 3 bar and you can also calculate x bar which is essentially summation x ij for all. So, let say this is j equal to 1 to n 1 where n 1 is the total number of the sample size of population 1. Similarly, you can do all. So, x ij by n where n is defined as summation of n i, you can define this x bar and what you do. So, your sample sizes are n 1, n 2, n k where k is the total number of populations.

(Refer Slide Time: 07:44)



5.

$$TSS = \sum (x_{ij} - \bar{x})^2$$

$$= \sum x_{ij}^2 - \left\{ \frac{(\sum x_{ij})^2}{n} \right\}$$

$\sum x_{ij} = G$ = grand total of all observations

CORRECTION OF THE MEAN

$$CM = \frac{(\sum x_{ij})^2}{n}$$

$$TSS = SST + SSE$$

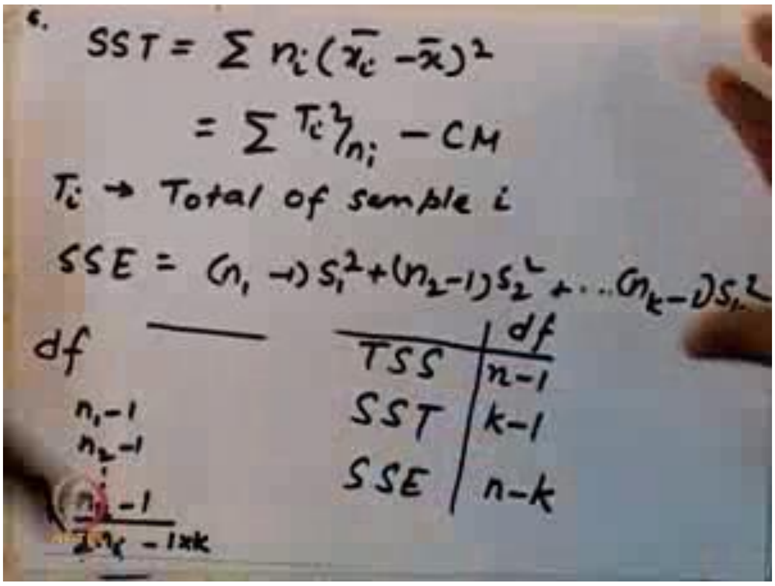
error

treatment

So, you calculate few matrix let me write in next page. So, you calculate something called TSS or total sum of squares and you define this as x ij minus x bar whole square. So, x bar is the average of the all the samples together. So, you can expand this and write this as x ij square minus summation x ij whole square by n. So, summation x ij is equal to G. So, G denotes the grand total of all observations and this term. So, this term is referred to as correction for the mean and we refer to as CM. So, your CM is given by summation of x ij whole square by dispositional foot.

Now, this total sum of squares is divided into 2 components. So, you write TSS is equal to 1 component called sum of squares for treatments and sum of squares for errors. So, e is for error and t is for treatment.

(Refer Slide Time: 09:34)



$$SST = \sum n_i (\bar{x}_i - \bar{x})^2$$

$$= \sum T_i^2/n_i - CM$$

T_i → Total of sample i

$$SSE = (n_1 - 1) s_1^2 + (n_2 - 1) s_2^2 + \ldots (n_k - 1) s_k^2$$

df

$n_1 - 1$

$n_2 - 1$

$\vdots$

$n_k - 1$

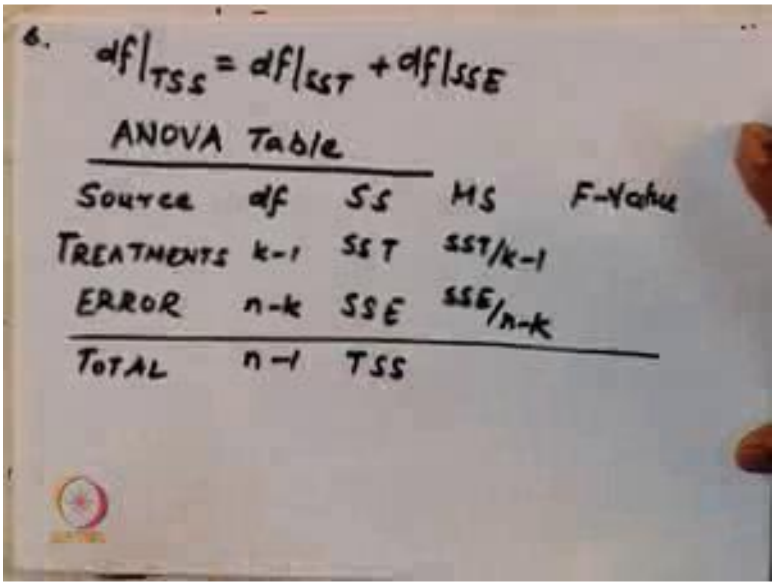
$\sum n_i - 1 \times k$

	df
TSS	$n-1$
SST	$k-1$
SSE	$n-k$

So, you can compute you can compute SST as and this can be expanded to be written as ti square by n i minus CM where ti corresponds to total of sample i, and you have the sum of squares of errors and this is given by the sum.

So, the degrees of freedom or df corresponding to, So, you can write for TSS SST and SSE you can write the degrees of freedom. So, TSS it has a total of in terms contributions of n square terms. So, degree of freedom of TSS is n minus of 1, SST accordingly you can see has k terms. So, degree of SST is k minus 1 and degree of SSE is going to be n minus k. And that you can make out from this term, you have n 1 minus 1 plus n 2 minus 1 dot dot n k minus 1. So, n 1 minus 1 plus n 2 minus 1 dot dot n k minus 1, you have summation in i let say and minus 1 into k. So, since summation n i is n. So, this comes to be n minus k. So, what you do you can see that n minus 1 minus k minus 1 equal to n minus k.

(Refer Slide Time: 11:46)



6. $df|_{TSS} = df|_{SST} + df|_{SSE}$

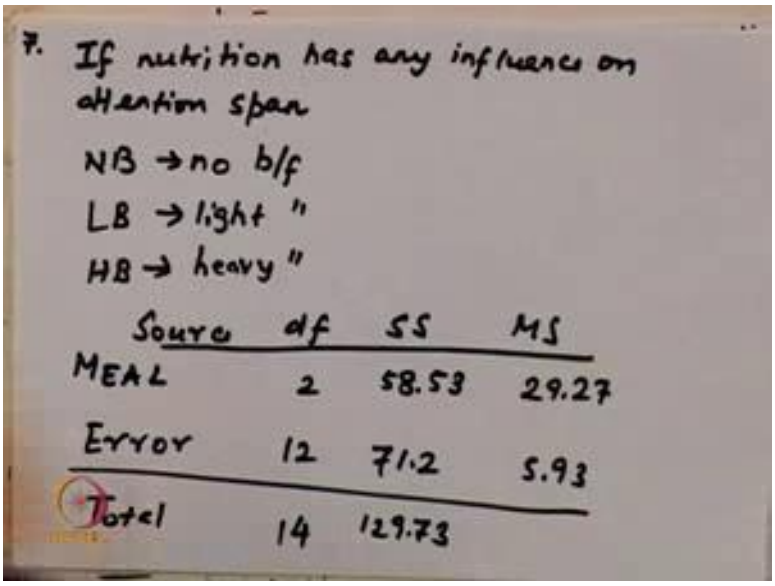
ANOVA Table

Source	df	SS	MS	F-Value
TREATMENTS	$k-1$	SST	$SST/k-1$	
ERROR	$n-k$	SSE	$SSE/n-k$	
TOTAL	$n-1$	TSS		

So, you can write down the expression that df of TSS is equal to df of SST plus df of SSE. And you can formulate the ANOVA table. This given by you have source, you have the degrees of freedom the total sum of squares the mean square and you can compute the based on this you can value the f value.

So, what are your source? The source includes treatments error and in sum this is total. So, this is k minus 1, this is n minus k, this is n minus n minus 1. So, you can calculate this as. So, just write SST or SSE and this is your TSS this is given by SST. This is a mean square either the mean square of treatments are mean square of error and this is given by SST by the corresponding degree of freedom.

(Refer Slide Time: 13:30)



7. If nutrition has any influence on attention span

NB → no b/f

LB → light "

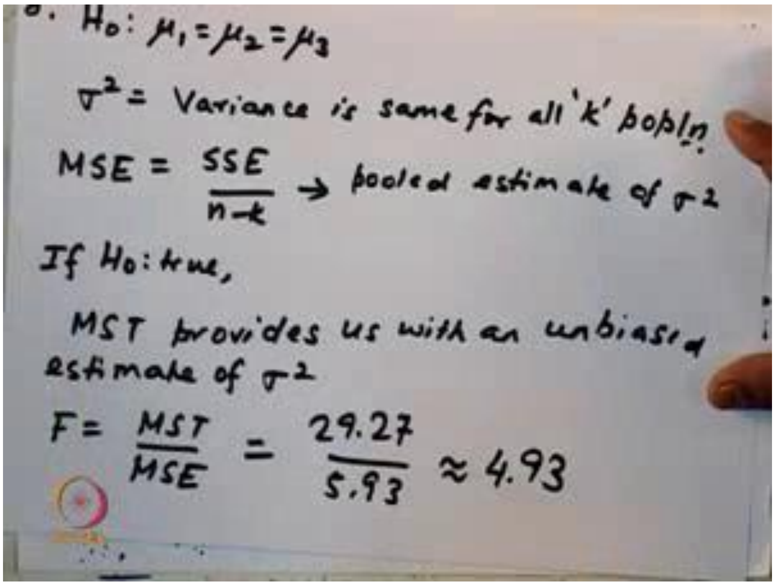
HB → heavy "

Source	df	SS	MS
MEAL	2	58.53	29.27
Error	12	71.2	5.93
Total	14	129.73	

So, this gives you the mean square 1 value. So, let us revisit the problem we had solved yesterday, where we had tried to correlate if we ask asked the question if nutrition has any influence on attention span. And what we had was you had 3 conditions no breakfast light breakfast or heavy breakfast. So, you can accordingly calculate. So, you can find out. So, in this case our treatment for this problem or treatment is nothing, but having the meal or you can write it as meal.

So, I will just jot down the final values that we had obtained, meal is your treatment error and total. So, for this case, let say if you wanted to test the hypothesis, if you wanted to test the hypothesis, whether nutrition has any influence on attention span, how will we find that out we will test the hypothesis?

(Refer Slide Time: 15:19)



o. $H_0: \mu_1 = \mu_2 = \mu_3$

σ^2 = Variance is same for all 'k' popl'n

$MSE = \frac{SSE}{n-k} \rightarrow$ pooled estimate of σ^2

If H_0: true,

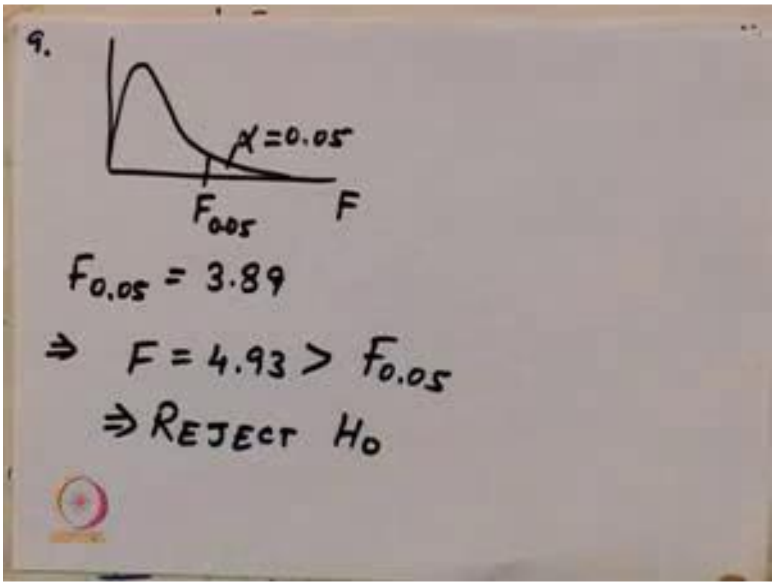
MST provides us with an unbiased estimate of σ^2

$$F = \frac{MST}{MSE} = \frac{29.27}{5.93} \approx 4.93$$

We will test the hypothesis, mu 1 equal to mu 2 equal to mu 3. We have 3 conditions and for you assume that there is no difference on of nutrition on attention span. And so, if are doing this you need to calculate. So, you have to calculate some test computes some test and what you do. So, the inherent assumption of sigma square, that this is the variance is same for all k populations. So, our MSE which is given by SSE by n minus k, this provides us with the pooled estimate of sigma square. So, if H naught is true then MST also gives us an estimate. So, MST also provides us with an with an unbiased estimate of sigma square.

So, you can calculate the test statistic as F you can use the F test where you define f by MST by MSE. So, depending on your significance level, if this value is very large then you can rule out H naught. So, in our case let say for this example that we did we can calculate the F value is 29.27 by 5.93 and it comes out to be 4.93. So, for this case F turns out to be 4.93.

(Refer Slide Time: 17:39)


9.
$\alpha = 0.05$
$F_{0.05}$
F

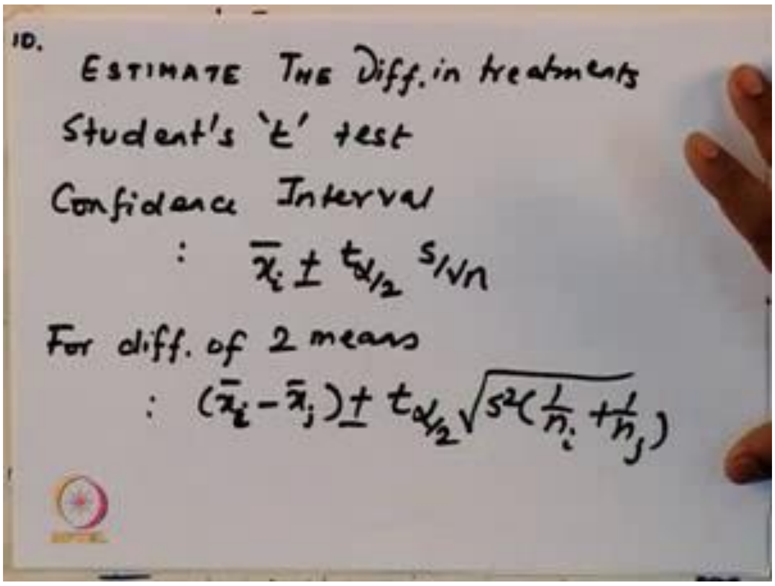
$F_{0.05} = 3.89$

⇒ $F = 4.93 > F_{0.05}$

⇒ REJECT H_0

So, you can to ask. So, you can calculate from the from the F curve you can ask what is F 0.05. So, then that case alpha is equal to 0.05 and what you see is f of 0.05, if you look up the tables it comes out to be 3.89. So, this implies since our f value of 4.93 is much greater than f of 0.05 this implies I can reject H naught. And if I reject H naught then the next question comes is can I come up with a confidence interval for means or the difference of means.

(Refer Slide Time: 18:35)



10. ESTIMATE THE Diff. in treatments

Student's 't' test

Confidence Interval

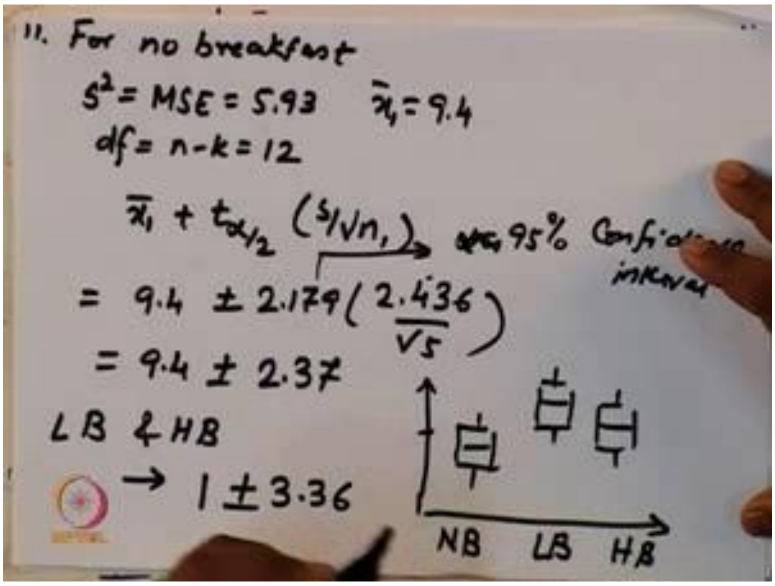
: $\bar{x}_i \pm t_{\alpha/2}\, s/\sqrt{n}$

For diff. of 2 means

: $(\bar{x}_i - \bar{x}_j) \pm t_{\alpha/2}\sqrt{s^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$

So, the question becomes, can I estimate the difference in means or in treatments. So, for this you can use the students’ t test, where your confidence interval is given by x i bar plus minus t of alpha by 2 s by root of n and for difference of 2 means it becomes.

(Refer Slide Time: 19:59)



11. For no breakfast

$s^2 = MSE = 5.93 \quad \bar{x}_1 = 9.4$

$df = n-k = 12$

$\bar{x}_1 + t_{\alpha/2}(s/\sqrt{n_1}) \rightarrow$ 95% Confidence interval

$= 9.4 \pm 2.179\left(\frac{2.436}{\sqrt{5}}\right)$

$= 9.4 \pm 2.37$

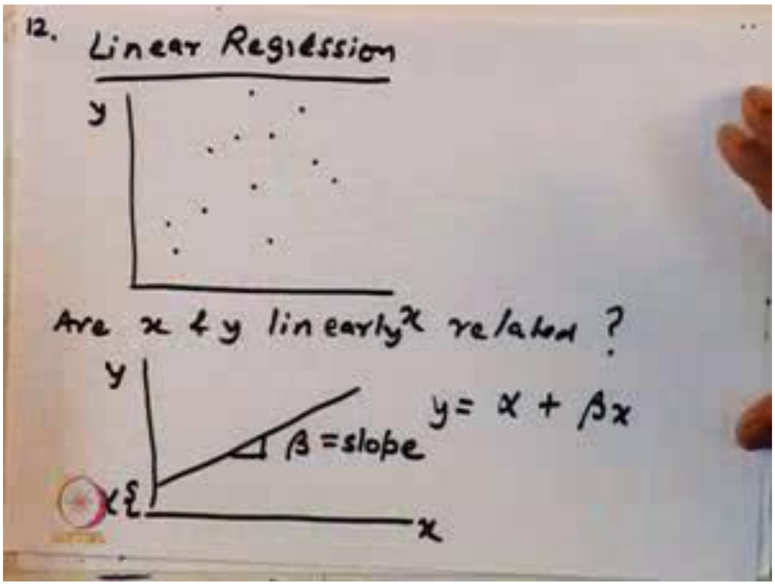
LB & HB

$\rightarrow 1 \pm 3.36$

So, for the case for the previous case, for let say for no breakfast you know the value of s square came out to be 5.93. And it is df was n minus k equal to 12. So, the confidence interval becomes x 1 bar plus t of and for this your x 1 bar was 9.4. So, you can calculate the confidence interval, as this comes out to be 9.4 k plus minus 2.1 7 9 into2.436 by root of 5. So, this is square root of 5.93. This is s square. So, you can compute the value of s as root of this value and n 1 was equal to 5. So, this is correspondence interval corresponding to alpha equal to, so, correspond into 95 percent confidence interval. So, this would come out to be 9.4 plus minus 2.37. You can similarly estimate the difference of means between light breakfast and heavy breakfast and that comes out to be one plus minus 3.36.

So, if I were to plot the 3 conditions and the distributions this is what the curve would look like. So, this was no breakfast, light breakfast, heavy breakfast. So, what you can clearly see if you see the trend if you see the trend of this. It gives us the feeling that this is like a saturation curve. So, there is a statistically significant not difference between these 2 points or between these 2 points. These 2 points may or may not be different, but between these 2 points there is very little difference. So, we can also use ANOVA.

(Refer Slide Time: 22:32)



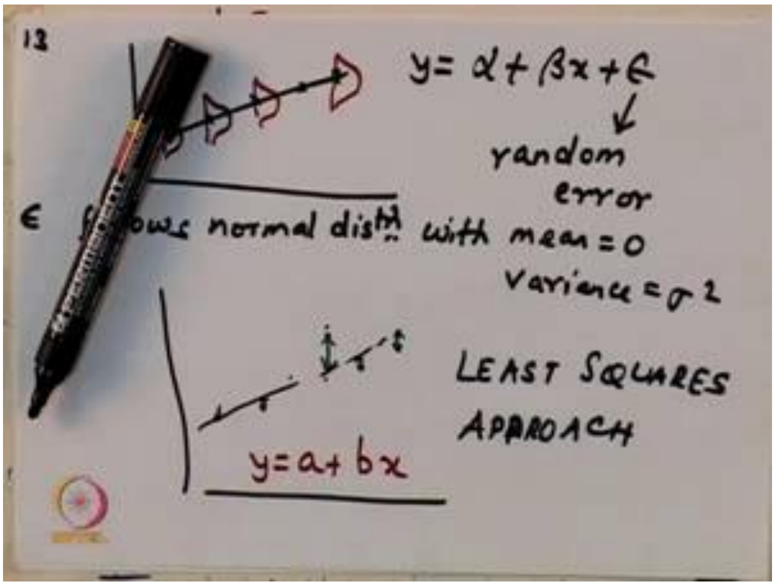
12. Linear Regression

Are x & y linearly related ?

$y = \alpha + \beta x$

We can also use ANOVA for linear regression, which we had done long back. What is linear regression imagining you have 2 variables x and y and you plotted the points together. Let us this is the data raw data of x and y. And the question you wish to ask is are x and y linearly related. So, if they were really linearly related in a deterministic sense, if there was in general, if you had all these points fall on an line which is exactly following a straight line equation, the equation of this line would have been something like alpha plus beta x, where this intercept is alpha and the slope is beta. So, beta is slope, but as you can clearly see from these data points there is a scatter.

(Refer Slide Time: 24:04)



13

$y = \alpha + \beta x + \epsilon$

random error

ϵ follows normal distn with mean = 0

variance = σ^2

LEAST SQUARES APPROACH

So, they; obviously, do not follow fall on the given line. So, to get an idea of this, what is of an done is to say that on top of this line imagine at all individual points I add an oyster. So, what I have drawn here. So, this is as if your equation of the line was y equal to alpha plus beta x plus some error and e epsilon is a random error with mean of 0. So, random error and follows the normal distribution. So, now, follows normal distribution. It follows epsilon say follows random normal distribution with mean of 0 and variance of sigma square. So, it is possible. So, it is possible. So, on top of this line, you can add a random error at every point. Let say you add a random error to generate this random points these points around at the mean along this line and hence by you generate some randomness.

So, these points no longer fall exactly on this line, but there is some scatter. And how do you find out, So, that also brings us to question that what should be the final values of alpha and beta and you as you have done earlier you know that we use the least squares approach. So, what least squares approach does it essentially minimizes these errors. It minimizes these errors on both sides of this line. It minimizes these errors or rather the square of these errors thereby findings some estimate of alpha and beta and this least square line is written as y is equal to a plus b x. So, in next class we will discuss how ANOVA can be used to test the possibility that this linear regression makes sense or not.

Thank you for your attention.