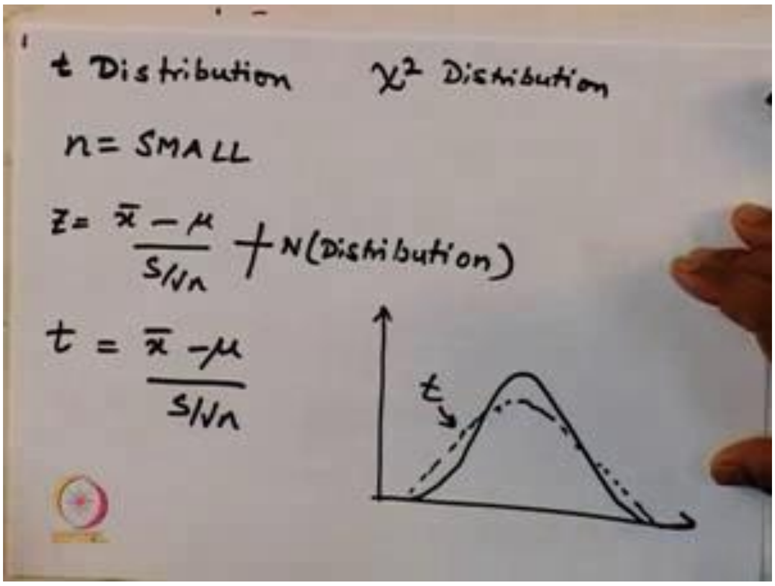
Introduction to Biostatistics

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

Lecture - 37
ANOVA

Hello and welcome to today's lecture. So, in the last 2 lectures we had briefly discussed about 2 distributions the t test, the t distribution and the chi square distribution.

(Refer Slide Time: 00:28)



t Distribution χ^2 Distribution

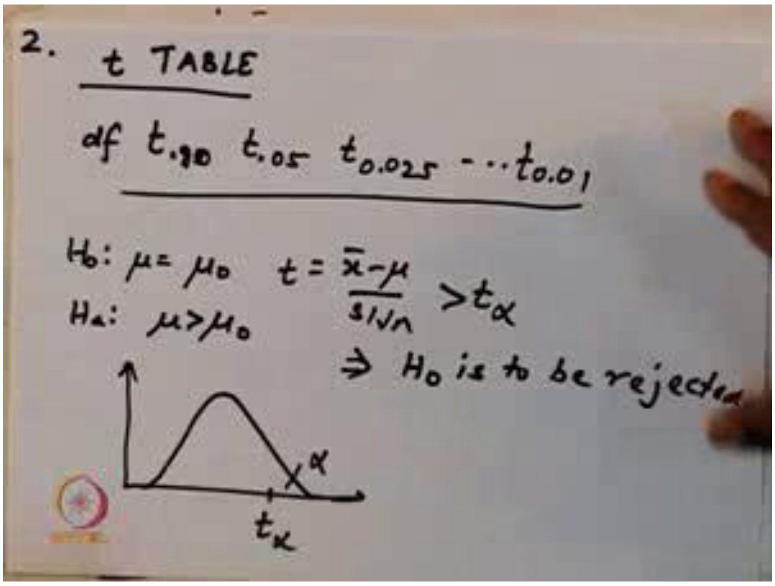
n= SMALL

$z = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ $\nsim$ N(Distribution)

$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$

So, the t test is used for when n is small. So, your z equal to x bar minus mu by s by root n, does not follow normal distribution. So, instead we calculate, the same value calculated at s bar minus mu by s by root n is called the t value. So as opposed to the normal distribution the t distribution is much flatter here with longer tails. So, this is your t distribution.

(Refer Slide Time: 01:48)



2. t TABLE

df $t_{.90}$ $t_{.05}$ $t_{0.025}$ ... $t_{0.01}$

$H_0: \mu = \mu_0$ $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} > t_\alpha$

$H_a: \mu > \mu_0$

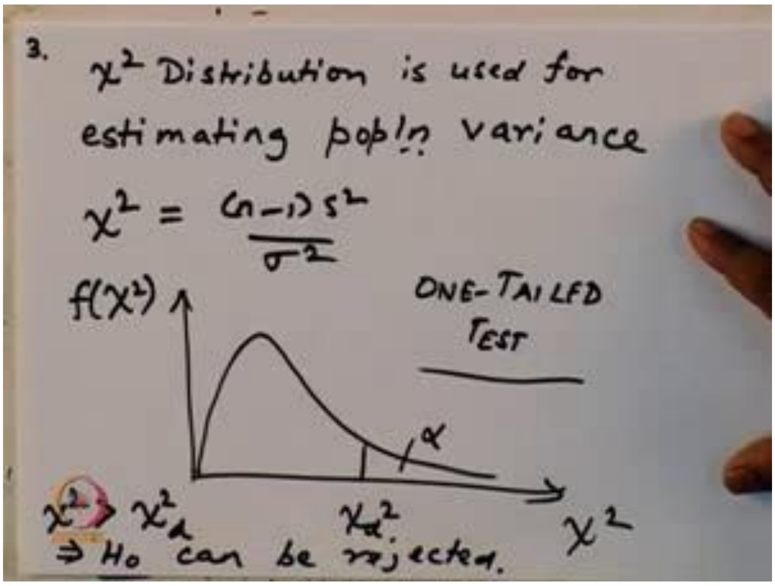
$\Rightarrow H_0$ is to be rejecte

α
t_α

So, the way of using a t test, is to look up the t table where you have d f and then various t values t of 0.1. So, t of 0.95, 0.05 so on and so forth. So, p of 0.1, it is other way round it is actually starts from 0.1 t of 0.05 t of 0.025 up to t of 0.01. So, the t test actually gives you the area for the upper tail only. So, for a t test your hypothesis is mu equal to mu not right and let us say your alternate hypothesis is mu is greater than mu naught.

So what you will do is you will calculate the t, you will have for given value of alpha you will calculate t of alpha. And if your t value is x bar minus mu by s by root n is greater than t alpha. So, this would imply that H naught is to be rejected. So, t is used for small sample estimating the population mean or differences in population mean. The chi square distribution is used for estimating population variance.

(Refer Slide Time: 03:16)



3. χ^2 Distribution is used for estimating pop!n variance

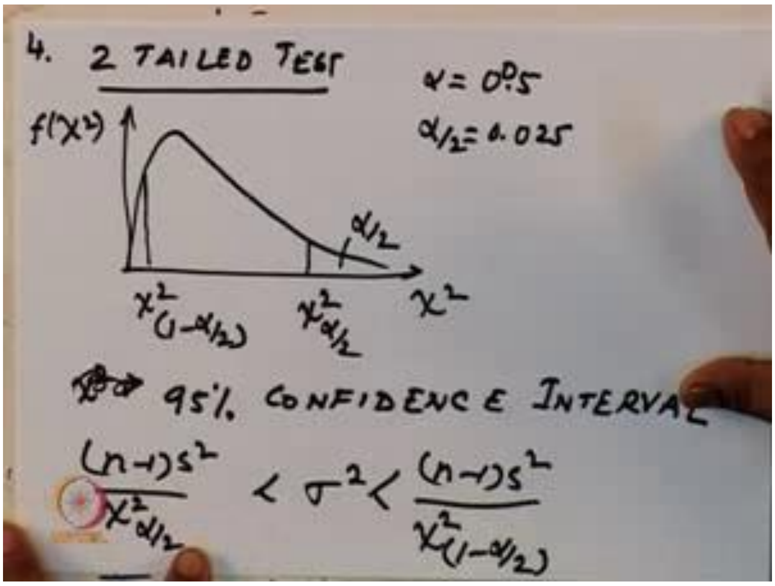
$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

$\chi^2 > \chi^2_\alpha$

⇒ H_0 can be rejected.

And the statistics that you calculate is used the called the chi square value, is given by n minus 1 into s square by sigma square. So, the chi square distribution is a symmetric. It looks something like this. So, this is of f of chi square this is your chi square value. So, for a one tailed test, you for a given value of alpha, you calculate lambda square alpha. So, if your lambda square is greater than lambda square alpha implying your hypothesis can be rejected.

(Refer Slide Time: 04:49)



4. 2 TAILED TEST

$\alpha = 0.05$

$\alpha/2 = 0.025$

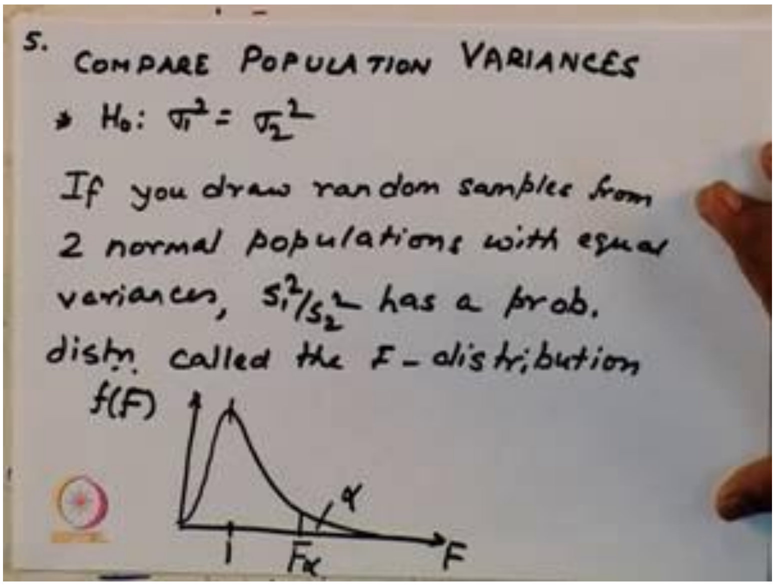
95% CONFIDENCE INTERVAL

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}$$

So, you can also use the chi square distribution. You can use the chi square distribution for a 2 tailed test. In this case you calculate chi square of alpha by 2 and chi square of 1 minus alpha by 2.

So if you choose, so if your lambda square is either greater than chi square of alpha by 2 or less than of chi square 1 minus alpha to you accept your alternate hypothesis. So, the confidence interval for lambda square given for a 2 tailed test. So, your 95 percent confidence interval is given by n minus 1. So, if you choose alpha equal to 0.5. So, your alpha by 2 is going to be 0.025. So, if you choose alpha equal to 0.025, then this value is going to be chi square of 0.025 and this is chi square of 1 minus 0.025. Based on this you can calculate the 95 percent confidence interval.

(Refer Slide Time: 06:55)



5. COMPARE POPULATION VARIANCES

- $H_0: \sigma_1^2 = \sigma_2^2$

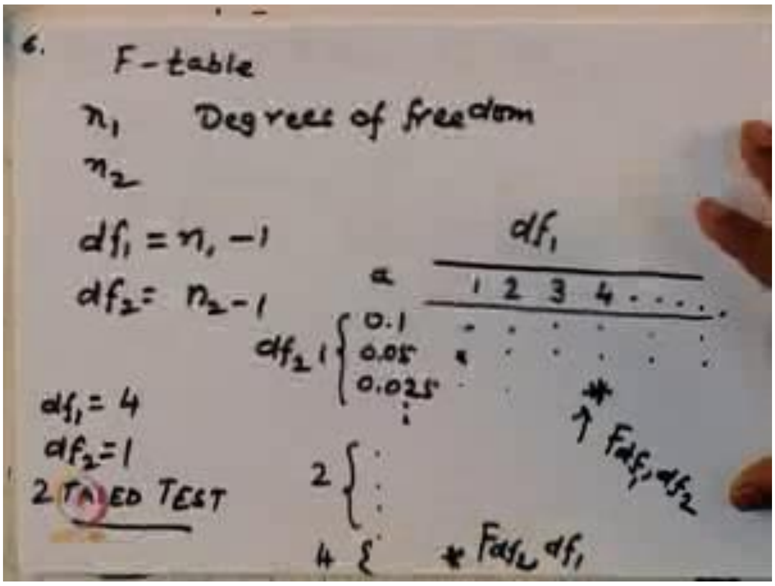
If you draw random samples from 2 normal populations with equal variances, s_1^2/s_2^2 has a prob. distn. called the F-distribution

$f(F)$
α
1
F_α
F

Now, let us say instead of estimating a population variance, you want to estimate or probe the difference of 2 population variances. So, in this case, so you want to compare population variances. So, your null hypothesis is sigma 1 square equal to sigma 2 square. And in this case you draw, so if you draw random samples from 2 normal populations, with equal variances, in the metric s 1 square by s 2 square, has a probability distribution called the f distribution.

So your f distribution this is the metric f. And the f distribution looks something like this. And the peak is corresponding to f equal to 1 that is when both the population variances are same. And you would do a single tail test for an f distribution just like you have done for all the other case. You will take the value of alpha which is the area under the curve and find out what is f of alpha, you will find out the value of f of alpha. So, how does a f table look like.

(Refer Slide Time: 09:07)

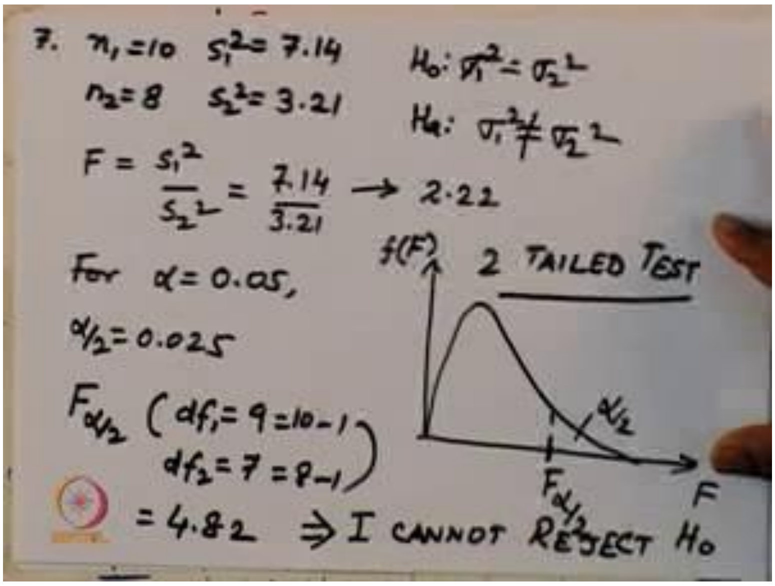


So, in an f table you have 2 metrics. So, you have drawn a sample from. So, you have sample sizes of n 1 and n 2. So, you have 2 degrees of freedom, one is d f 1 which is n 1 minus 1 and d f 2 is n 2 minus 1. So, in an f table you have the where the values of d f 1 where the value of significance and d f 2.

So, for example, in d f 1 you will have various values let us say 1 2 3 4 so on and so forth. And a will have various significance levels let us say 0.1 0.05 0.025 and so on and so forth. And for each of these you will have d f 2. So, this is one value corresponding to d f 2equal to 1. You have these values of a similarly corresponding to d f 2 you will have various other values of a. So, these are your significance levels. So, you the way to look up the table, is find out what is d f 1 find out what is d f 2 and see if corresponding to the value of a at the desired level that you want what is the value of alpha.

So, in for example, if I d f 1 was 1 and d f 2 was 1 and my significance level was 0.05. Then this is the value I will choose. So, you have values written for all of this. So, all of this you will have corresponding to each combination you will have values specified. So, let us take a sample example.

(Refer Slide Time: 11:08)



7. $n_1=10$ $s_1^2=7.14$ $H_0: \sigma_1^2=\sigma_2^2$

$n_2=8$ $s_2^2=3.21$ $H_a: \sigma_1^2 \neq \sigma_2^2$

$F = \frac{s_1^2}{s_2^2} = \frac{7.14}{3.21} \rightarrow 2.22$

For $\alpha = 0.05$,

$\alpha/2 = 0.025$

$F_{\alpha/2}$ $(df_1 = 9 = 10-1$, $df_2 = 7 = 8-1)$

$= 4.82 \Rightarrow$ I CANNOT REJECT H_0

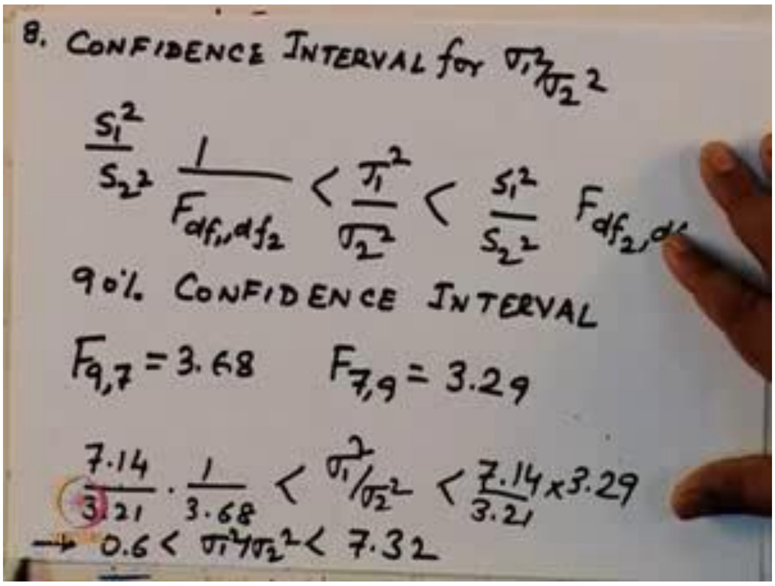
$f(F)$
2 TAILED TEST
$\alpha/2$
$F_{\alpha/2}$
F

So, imagine I have n 1 equal to 10, s 1 square equal to 7.14 n 2 equal to 8 and s 2 square equal to 3.21. So, what is the value of f and whether I want to calculate whether f is significant or not.

So, I have defined f as s 1 square by s 2 square is my test metric. So, in this case it is simply 7.14 by 3.21 and you get a corresponding value of f equal to roughly 2.22. Now for significance level alpha equal to 0.05. So, if I am when I am doing a single tail test. So, if let us say my hypothesis. So, my H naught is sigma 1 square equal to sigma 2 square. And H a is sigma 1 square not equal to sigma 2 square. Then what I have to do for this curve f, I find out alpha and the corresponding value of f of alpha. Sorry. So, I have 2 values. So, if my alternate hypothesis is sigma 1 square not equal to sigma 2 square. I want to correspond to find out the value of f of alpha by 2 corresponding to alpha by 2. So, this is for 2 tailed test.

So, for alpha by 2 in this case alpha by 2 is equal to 0.025 you will look up the value of f of alpha by 2 corresponding to d f 1 equal to 9 and d f 2 equal to 7, 9 is 10 minus 1 this is 8 minus 1. So, this f of alpha by 2 is 4.82. So, I can reject my hypothesis, if this value f calculated as 2.2 is greater than f of alpha by 2, but in this particular case 2.22 is less than 4.82. So, this means I cannot reject H naught.

(Refer Slide Time: 14:03)



8. CONFIDENCE INTERVAL for σ_1^2/σ_2^2

$$\frac{s_1^2}{s_2^2}\frac{1}{F_{df_1,df_2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} F_{df_2,d}$$

90% CONFIDENCE INTERVAL

$F_{9,7} = 3.68$ $F_{7,9} = 3.29$

$$\frac{7.14}{3.21}\cdot\frac{1}{3.68} < \sigma_1^2/\sigma_2^2 < \frac{7.14}{3.21}\times 3.29$$

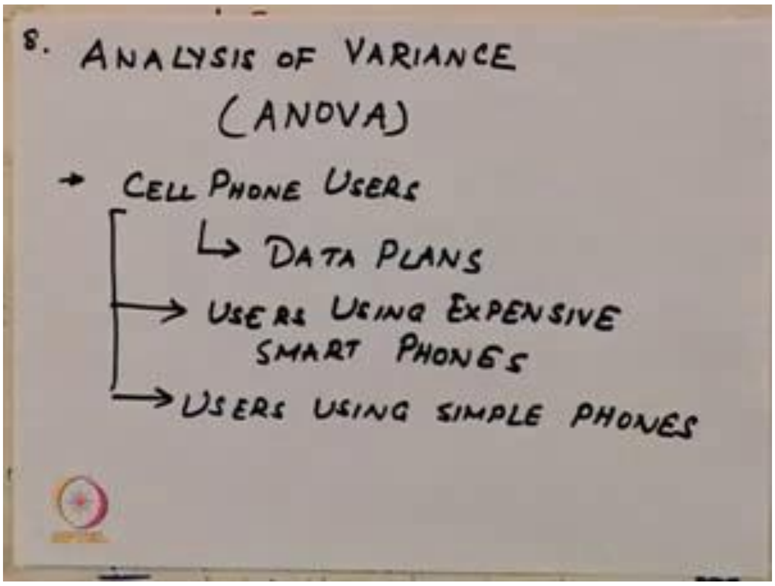
$\rightarrow 0.6 < \sigma_1^2/\sigma_2^2 < 7.32$

So, you can also find out the confidence interval, for sigma 1 square by sigma 2 square. And that is given by the following bounds, you have sigma 2 square. So, what you see at 2 different terms s 1 square into s 2 square into 1 by capital f of d f 1 comma d f 2. And the other one is f of d f 2 comma d f 1. So, if we go back to the table let us say if I had a d f 1 say for example, if my d f 1 was equal to 4 and d f 2 equal to 1. So, 4 comma 1 with a significance level of 0.05 and if I am doing a 2 tailed test, this is the value I would be looking for.

So for 2 tailed test this is my f of d f 1 comma d f 2. So, correspondingly I can find out f of d f 2 comma d f 1. So, d f 2 is the corresponding value. So, in this case d f 2 is 1 and 1 comma. So, I can accordingly write let us say for 4 I will have various values. So, this will be f of d f 2 comma d f 1. So, you can have 2 different values, so on the expression gives you 1 by f of d f 1 comma d f 2 or f of d f 2 comma d f 1. So, for example, for the exam for the case that we discussed right now, for a 9ty percent confidence interval, you can look up f of 9 comma 7. This is 3.68 and f of 7 comma 9 this is 3.29. So, you are ranges for this case you had s 1 square, for this case your s 1 square is 7.14 and s 2 square 3.21 right. So, that is why you got d f 1 is 9 and d f 2 is 7. So, for this particular case you can calculate the lower and upper bounds of sigma 1 square by sigma 2 square as 7.14. So, this comes out to be point 6 less than sigma 1 square by sigma 2 square less than 7.32. So, you see that your bound is reasonably wide, that is why you have been unable to reject H naught.

So, even though you have been unable to reject H naught your variance is the ratio is quite variable. So, that brings us to end of our discussion of on t test and f test for difference of population means and population variance what we will initiate.

(Refer Slide Time: 18:06)



8. ANALYSIS OF VARIANCE

(ANOVA)

→ CELL PHONE USERS

↳ DATA PLANS

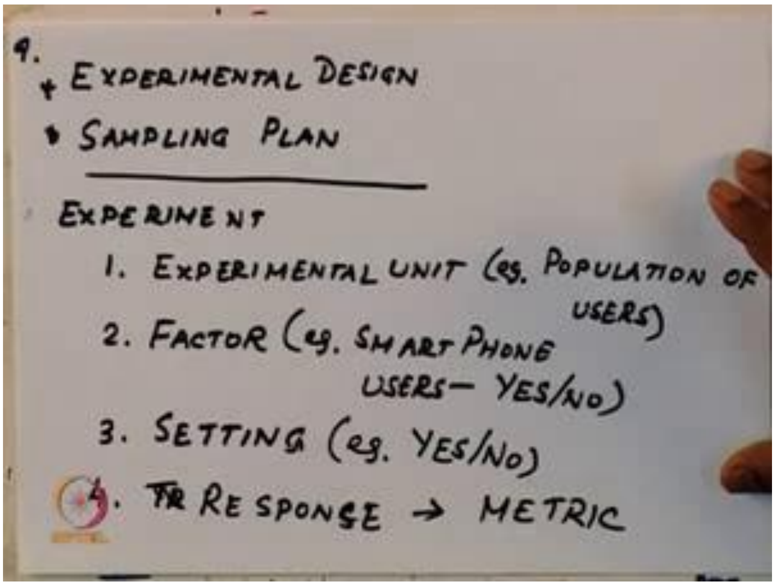
→ USERS USING EXPENSIVE SMART PHONES

→ USERS USING SIMPLE PHONES

Now, is our discussion on a very important aspect of statistics called analysis of variance, or in short it is referred to as ANOVA. So, in ANOVA why are we interested in analysis of variance? So, let us say you want to calculate you have cell phone users, and you want to look at their data plans that they sign up for. So, cell phone users can themselves be distributed as users using expensive smart phones and users using simplest phones which are not smart phones.

So, if you were to calculate the variability in the data plans that people search up for you can clearly see that there will be reasonable variability because people who use smart phones, they sign up for data plans they have expensive data plans, versus people who use simple phones they are usability is only dictated by having a phone connection. So, you will have huge amount of variability when you compile the statistics for data plans. So, to address it how can you make a plan or when a company is thinking of what should be it is pricing strategy. It should take into account the inherent variation in the population. And analysis of variance is that metric is that that procedure which allows us to incorporate what is the effect of a sub population on the variability of the overall population. So, you try to assign the variability into individual factors which are likely to have distinct effects on the overall variability of the statistics that we are measuring.

(Refer Slide Time: 20:36)



9.

- EXPERIMENTAL DESIGN
- SAMPLING PLAN

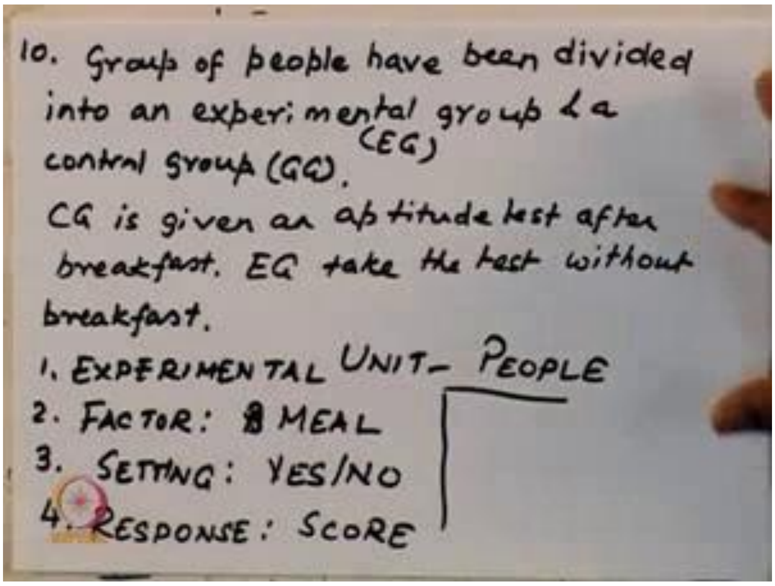
EXPERIMENT

1. EXPERIMENTAL UNIT (eg. POPULATION OF USERS)
2. FACTOR (eg. SMART PHONE USERS— YES/NO)
3. SETTING (eg. YES/NO)
4. ~~TR~~ RESPONSE → METRIC

So as a consequence of which when you want to compile the statistics, you have what is called an experimental plan. Experimental plan or experimental design or it is also called the sampling plan. So, this is to account for how should you actually set up the experiment and when you are thinking of experiment, so for any experiment. So, there are various aspects of the experiment.

First is who being you measuring it on or what is your experimental unit. So, if for example, in the case of cell phone users, your experimental unit can be just the population of users. Second is the factor. So, in the example that we discussed the factor can be whether they are smart phone users or not, yes or no you will have a setting. So, setting is a value are the intensity scale of the factor. In case for this particular example my setting is simply yes or no and you have your response, which basically whatever is your metric that you calculate from the statistics.

(Refer Slide Time: 22:46)

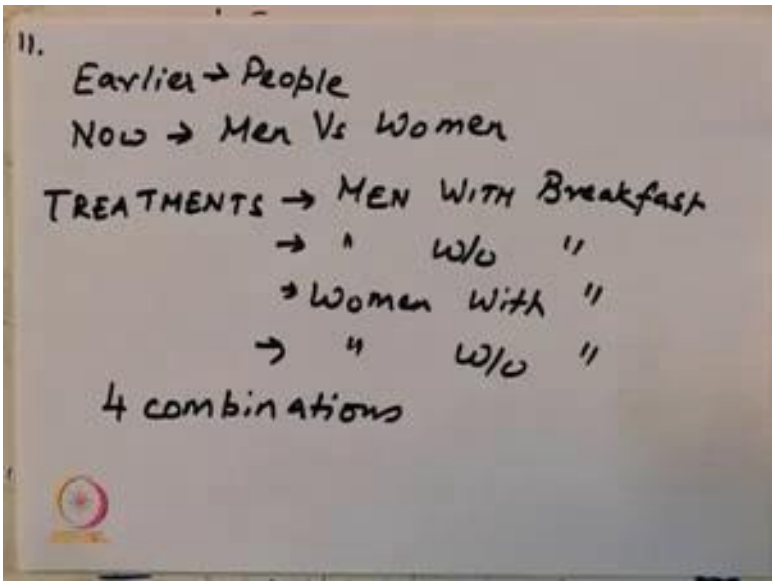


10. Group of people have been divided
into an experimental group (EG) & a
control group (CG).
CG is given an aptitude test after
breakfast. EG take the test without
breakfast.
1. EXPERIMENTAL UNIT- PEOPLE
2. FACTOR: MEAL
3. SETTING: YES/NO
4. RESPONSE: SCORE

So let us take another 2 examples to discuss to clear these ideas. So, imagine that a group of people have been divided into an experimental group and a control group. So, the control group, so I will write CG for control group and EG for experimental group. So, your CG is given an aptitude test after breakfast versus the experimental group take the test without breakfast.

So, let us identify each of the things that we said. So, what is my experimental unit? The experimental unit in this case you say people is nothing people what taking the test both the experimental and the control group. What is the factor? Factor is your breakfast or meal because you want to proof the effect of meal. What is your setting? Whether it is yes or it is no. So, your treatment or setting is both the same because treatment is either they have had the breakfast or they have not had the breakfast and what is your response is your score aptitude score.

(Refer Slide Time: 25:21)



11.

Earlier → People

Now → Men Vs Women

TREATMENTS → MEN WITH Breakfast
→ " W/O "
→ Women With "
→ " W/O "

4 combinations

Now, imagine in this same population. So, in the same population; so earlier I just had people now I have men versus women. So, now, in terms of treatments can be men with breakfast men without breakfast, women with breakfast and women without breakfast. So, now, you have instead of 2 conditions you have 4 cases 4 combinations. So, depending on your experimental plan you might have many more experimental combinations.

With that I thank you for your attention. And we will stop here in the next class we will continue with our discussion on ANOVA.