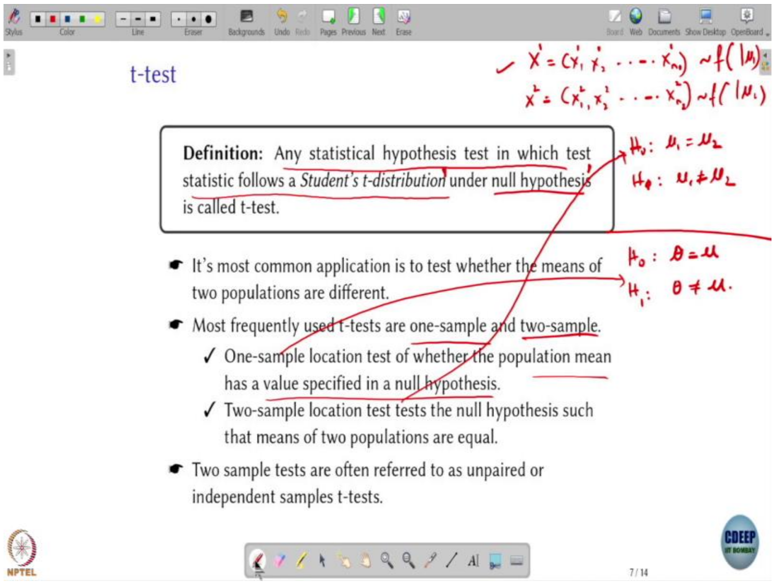
Engineering Statistics

Professor Manjesh Hanawal
Department of Industrial Engineering and Operations Research
Indian Institute of Technology Bombay

Week 11
Lecture 54
t-test & F-test, ANOVA

(Refer Slide Time: 00:14)



t-test

$X^1 = (X_1^1, X_2^1, \dots X_{n_1}^1) \sim f(\cdot|\mu_1)$

$X^2 = (X_1^2, X_2^2 \dots X_{n_2}^2) \sim f(\cdot|\mu_1)$

Definition: Any statistical hypothesis test in which test statistic follows a *Student's t-distribution* under null hypothesis is called t-test.

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

- It's most common application is to test whether the means of two populations are different.
- Most frequently used t-tests are one-sample and two-sample.
  - ✓ One-sample location test of whether the population mean has a value specified in a null hypothesis.
  - ✓ Two-sample location test tests the null hypothesis such that means of two populations are equal.
- Two sample tests are often referred to as unpaired or independent samples t-tests.

$H_0: \theta = \mu$

$H_1: \theta \neq \mu.$

Now let us move on to the t test. By the name, we see that t distributions are associated here. Any statistical hypothesis test in which test statistics follows student's t distribution, under null hypothesis is called t test, the most common use of t test comes then the means of two populations are different.

So notice that in the previous example, when we computed about the P values and P value test, we, were there dealing with one sample I mean one population x1 x2. Now, we may be dealing with, so I am going to write them like this one population, I am going to superscript them one to indicate that this corresponds one, you may also have another population x2, which I am going to superscript with 2.

Now, under this numbers, like the number of samples could differ. Now the question here is, I want to check, let us say, the hypothesis is this is whether they are coming from a population, which has the same mean support, let us say, I am assuming they say like f given under by mu 1 and they say let us say mu.

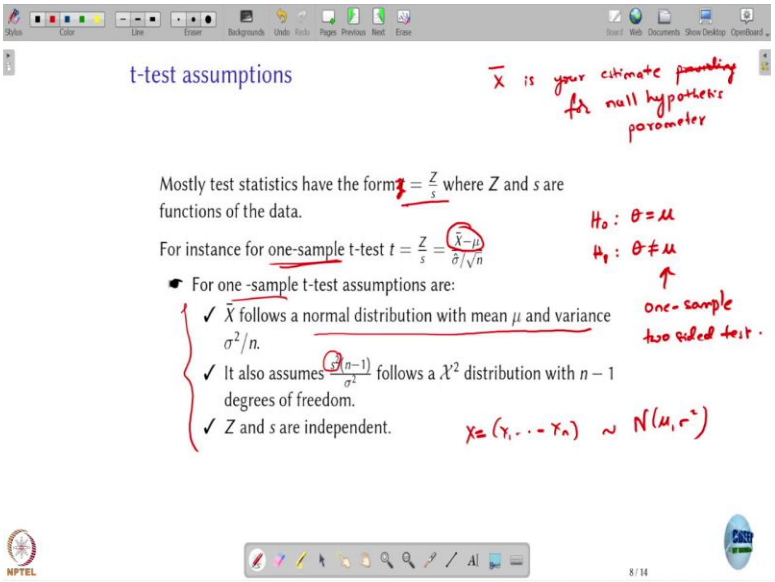
Now, my null hypothesis is mu 1 is equals to sorry equals to mu 2, and my alternate hypothesis could be like mu 1 not equal to mu 2. So, this is the common usage where we want to this t test comes into picture, but it is not necessarily that we have to deal with the 2 population in using 2 tests it could be used on a single population also.

So, let us now look into this care like I mean, we have whatever I explained it is just written here, the most frequently are like one sample or two sample in the one sample test, we will check whether the population mean has the same specified value in the null hypothesis that is they are having the same value like let us say to something mu common and to the two sample tests says that whether null hypothesis such that the means of the two populations or equal sorry, like, let me rewind this.

So, this is now, let me rewind this. So, in the one sample location test, we have just to one population samples given to us and they are my hypothesis is to adjust whether my parameter is equals to mu, which is the underlying specified parameter for the null hypothesis mu.

And, of course, the null hypothesis here could be like a is not equals to mu, and this is for the one case and for the two location tests. I have already given here in the two location, I have two sets of sample and there I want to check whether this two samples sorry, two populations have the same mean, or they are different. Often these two sample tests are referred to as unpaid or independent sample t test we will just see this.

(Refer Slide Time: 04:20)



t-test assumptions

$\bar{X}$ is your estimate ~~providing~~ for null hypothesis parameter

Mostly test statistics have the form $z = \frac{Z}{s}$ where Z and s are functions of the data.

For instance for one-sample t-test $t = \frac{Z}{s} = \frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}}$

$H_0 : \theta = \mu$

$H_1 : \theta \neq \mu$

one-sample two sided test.

- For one -sample t-test assumptions are:
  - ✓ $\bar{X}$ follows a normal distribution with mean μ and variance σ^2/n.
  - ✓ It also assumes $\frac{s^2(n-1)}{\sigma^2}$ follows a $\mathcal{X}^2$ distribution with $n-1$ degrees of freedom.
  - ✓ Z and s are independent.

$X = (X_1, \ldots X_n) \sim N(\mu, \sigma^2)$

To apply t tests. We make some assumptions. Which by the way, will hold for the Gaussian distributions by default, Gaussian samples by default, but by making these assumptions, maybe we can say something more general. The most test statistics are like in general, this when that statistics, I am interested will have this form Z by s.

So there is a little notational change, I would like to continue to denote this by Z, where Z and s are function sub data, so Z and s are themselves random variables. Now first for, let us look into one sample t test. And in this one sample, the numerator Z can be simply the sample mean. So here notice that in the one sample test, I am basically testing the hypothesis that my parameter theta is mu or not. It is a one sample test.

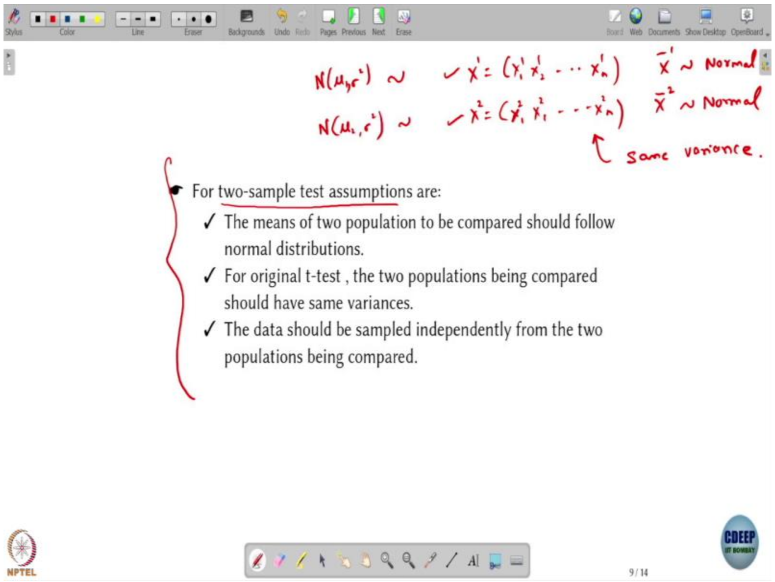
And let us say this is a one sample two sided test like one sample two sided test. So, here the numerator is the sample mean centralized by subtracting the true the parameter of your null hypothesis and the denominator is your estimate of your standard deviation s.

Now, in this one sample test, we are going to assume that this X bar follows a normal distribution with mu and variance sigma by n and no always in this recall that X bar is your estimate, which estimate providing estimate for null hypothesis parameter you are claiming by this us claiming that X bar is giving me a good representation of mu and that is what like X bar have this X bar is the average.

So, it will have the variance sigma square by n. And we are also going to assume that s square recall that s square is the estimate unbiased estimate of the variance when you multiplied by n minus 1 divided by sigma square, it follows a Chi square distribution with n minus degrees of freedom and Z and s are independent.

Notice that even though we put this as an assumption, when my x my random sample is coming from Gaussian distributed with parameter let us say mu and sigma this assumptions naturally hold which we have already seen, when we are talked about sampling from random distribution sorry, when we talked about sampling, and studied properties of random samples.

(Refer Slide Time: 08:13)



$N(\mu_1, \sigma^2) \sim \quad ✓\ x^1 = (x_1^1\ x_2^1 \cdots x_n^1) \quad \bar{x}^1 \sim$ Normal

$N(\mu_2, \sigma^2) \sim \quad ✓\ x^2 = (x_1^2\ x_1^2 \cdots x_n^2) \quad \bar{x}^2 \sim$ Normal

same variance.

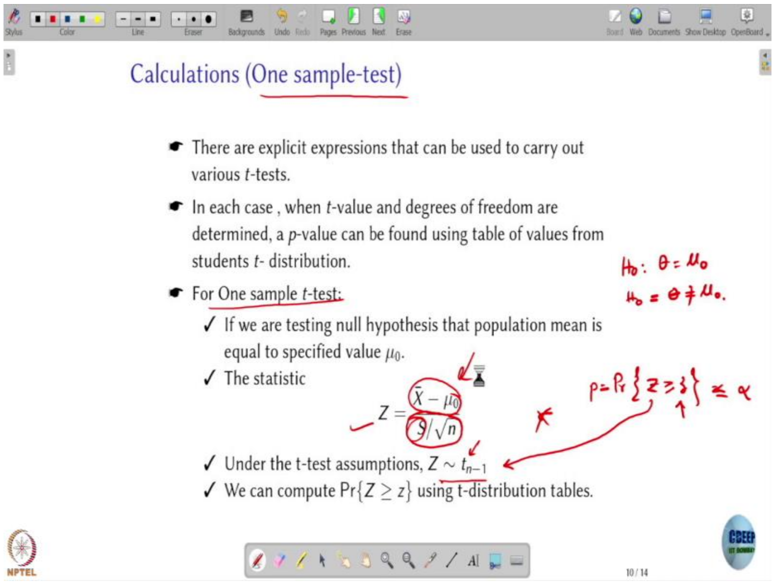
- For two-sample test assumptions are:
  - ✓ The means of two population to be compared should follow normal distributions.
  - ✓ For original t-test , the two populations being compared should have same variances.
  - ✓ The data should be sampled independently from the two populations being compared.

In the two sample test where let us say we have one set of random sample like this we are going to say that this means for the two populations to be compared should follow normal distribution. That is, if you are going to compute this, this is going to be normal. And also if you are going to compute the mean of this there should be normal.

And both of these having the same variance the samples are coming from a population distribution having same variance and the data this set of samples, these random samples should be independent are the sample independently. So, we can just say that they are going to have the same variance the samples are generated from a underlying population which have the same variance and we want their sample mean to follow normal distribution.

Notice that all distribution again holds when your samples are drawn from Gaussian like if they are drawn from let us say some question mu 1 and sigma square. And this one let us say is coming from and let us say mu 2 and sigma square it is all this properties naturally holds good.

(Refer Slide Time: 10:07)



Calculations (One sample-test)

- There are explicit expressions that can be used to carry out various *t*-tests.
- In each case , when *t*-value and degrees of freedom are determined, a *p*-value can be found using table of values from students *t*- distribution.
- For One sample *t*-test:
  - ✓ If we are testing null hypothesis that population mean is equal to specified value μ_0.
  - ✓ The statistic

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

  - ✓ Under the t-test assumptions, $Z \sim t_{n-1}$
  - ✓ We can compute $\Pr\{Z \geq z\}$ using t-distribution tables.

$H_0 : \theta = \mu_0$

$H_0 = \theta \neq \mu_0.$

$p = \Pr\{Z \geq z\} \leq \alpha$

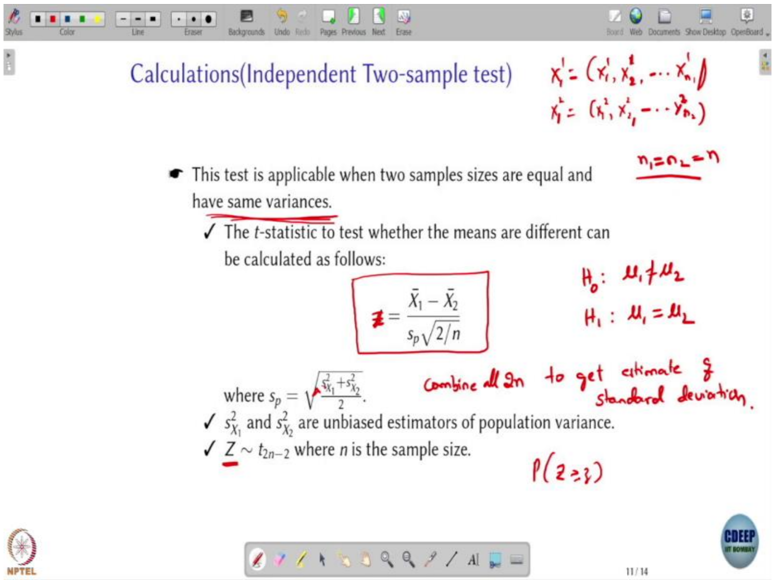
Now, let us say what would be the statistics for me here to compute the p value. Now that let us focus on one sample test here, now let us directly look into this. And here in the one sample test, my I am basically testing the hypothesis whether mu naught or naught be let us say take the two sided case for this, I can have a statistics which is X bar minus mu naught divided by s by square root n notice that earlier it was sigma sorry sigma.

Now, I have replaced it by s, because I do not know sigma also in my case. So here, we are basically saying, I do not know none of this both this mean and variance are unknown. If we knew that, we could have gone with the t test, which we already did before. Now, if you recall by our assumption that this is normal distributed and this is chi square distribution with n minus degrees of freedom, if you look into their ratio, the ratio is distributed as student t distributions with n minus 1 degrees of freedom, which you do not like this.

Now, with this I can again go back and compute my p values, what is the probability that Z is greater than equal to z. Now here my Z is t distributed with n minus 1 degrees of freedom. And from the t distribution table, for any given Z, which is coming from my data, I can readily compute this value and compare it against a given significance level and decide whether my claims are statistically significant or not.

And this is where the I hope it is clear how the t distribution came into picture here, because we do not know the variance that is why we used unbiased estimator for variance here or like rather unbiased estimator for standard variance here. And once we do that, we know already that this statistics follows t distribution, and we can use the properties of t distributions here.

(Refer Slide Time: 13:10)



Calculations(Independent Two-sample test)

$X_1^1 = (x_1^1, x_2^1, \ldots x_{n_1}^1)$

$X_1^2 = (x_1^2, x_2^2, \ldots y_{n_2}^2)$

$n_1 = n_2 = n$

- This test is applicable when two samples sizes are equal and have same variances.
  - ✓ The *t*-statistic to test whether the means are different can be calculated as follows:

$$\boxed{Z = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{2/n}}}$$

$H_0: \mu_1 \neq \mu_2$

$H_1: \mu_1 = \mu_2$

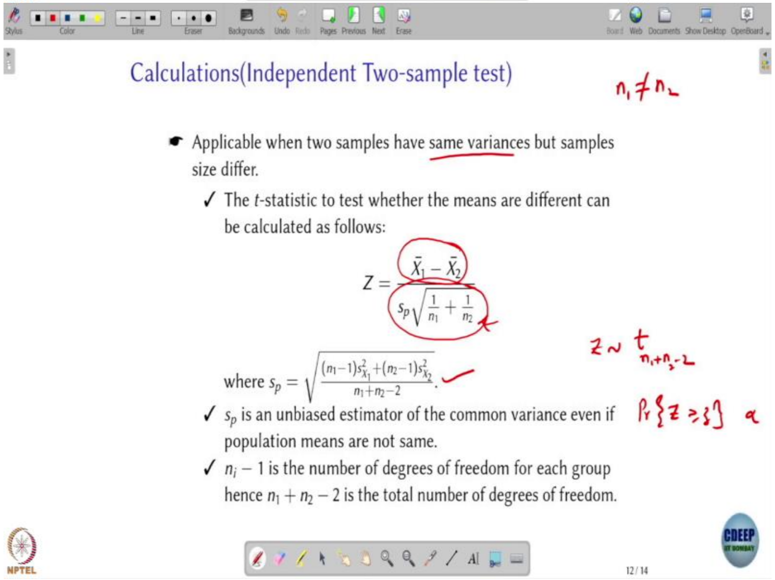
where $s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}$.

Combine all 2n to get estimate of standard deviation.

  - ✓ $s_{X_1}^2$ and $s_{X_2}^2$ are unbiased estimators of population variance.
  - ✓ $Z \sim t_{2n-2}$ where n is the sample size.

$P(Z \geq z)$

11 / 14



Calculations(Independent Two-sample test)

$n_1 \neq n_2$

- Applicable when two samples have same variances but samples size differ.
  - ✓ The *t*-statistic to test whether the means are different can be calculated as follows:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$Z \sim t_{n_1+n_2-2}$

where $s_p = \sqrt{\frac{(n_1-1)s_{X_1}^2 + (n_2-1)s_{X_2}^2}{n_1+n_2-2}}$.

  - ✓ s_p is an unbiased estimator of the common variance even if population means are not same.

$\Pr\{Z \geq z\}$ α

  - ✓ $n_i - 1$ is the number of degrees of freedom for each group hence $n_1 + n_2 - 2$ is the total number of degrees of freedom.

12 / 14

Now, find this is clear. I hope for the case of one, one sample test. Now how to check this for the two sample case and recall that for the two sample case, I am going to assume the variances are same. I am also going to assume the two samples have the same number of samples. They could be different, but I am going to start with the case where they have the same number of samples. So, here X n 1 so let us case where n 1 equals n2. Now in this case I am interested in the hypothesis that whether the, their means are equal or not.

One can argue that after little bit of manipulation this could be taken as statistics. Where X1 bar as usual is the sample mean coming from the first population and X2 bar is the sample mean coming from the second population. And Sp here is the value of the standard deviation estimator unbiased value that I got and this is again here s square x1 square is the unbiased estimator of your variance and sx to squared here it is going to be unbiased sample estimates of your variance of sample two.

So, notice that we have assumed in this case they are same actually you could combine these two samples and get one value for your common variance or sorry estimate of the variance or your estimate combine all the samples like in this case maybe let us call this n we can combine all the samples.

We are going to end up with the 2n samples to get combined all to get your estimate of standard deviation even though here have written that as if this is computed from this and one sample separately from the population of the first population and x 2 is coming from the second population.

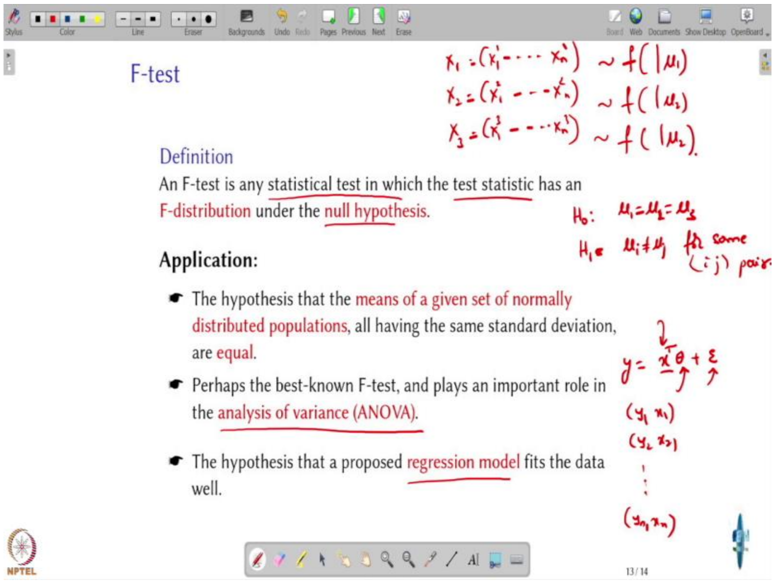
But for the variance since they have the same common variants, and all the samples are independent of each other, you can just use all the 2n samples to get one standard sorry estimate of your standard deviation. And now, one can show that or it is actually straightforward to observe that the Z here is going to be a t distribution with 2 n minus 2 degrees of freedom where n is the common sample size.

Again now, once you have this for a given Z, you can compute the p value and then compare against the given significance level to see whether you want to accept or reject or like accept or reject the alternate hypothesis this could be also extended to the case when the number of sample size are not equal n1 is not same as n2, but still under the same variance case.

Again here one can show that the statistics that is of relevance here can be given as the difference of the sample means of the populations divided by their standard estimated standard deviation and that could be computed in this fashion, I am just living this calculations, but you can verify that indeed the standard deviation the estimate of the standard deviation can be given that. And here now, we can see that this Z the denominator here is actually a chi square distribution with n 1 plus n 2 minus 2 degrees of freedom.

And now, you can argue that the Z is also have t distribution with n 1 plus n 2 minus 2 degrees of freedom. Now, once you know this is a t distribution with a certain degrees of freedom you can again go and compute your P value and compare against your significance level and decide to accept the alternate hypothesis or reject your alternate hypothesis or not accept the alternate hypothesis.

(Refer Slide Time: 19:56)



F-test

$X_1 = (x_1^1 - \cdots x_n^1) \sim f(\,|\mu_1)$

$X_2 = (x_1^2 - - - x_n^2) \sim f(\,|\mu_2)$

$X_3 = (x_1^3 - - \cdots x_n^3) \sim f(\,|\mu_2)$.

Definition

An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis.

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

$H_1: \mu_i \neq \mu_j$ for some $(i\,j)$ pair

Application:

- The hypothesis that the means of a given set of normally distributed populations, all having the same standard deviation, are equal.
- Perhaps the best-known F-test, and plays an important role in the analysis of variance (ANOVA).
- The hypothesis that a proposed regression model fits the data well.

$y = \underline{x}^T \theta + \varepsilon$

$(y_1\, x_1)$

$(y_2\, x_2)$

$\vdots$

(y_n, x_n)

13 / 14

Now, this is where whenever we have our test statistics following the t distribution, we can use all this but it may happen that every time we may not start statistics may not be just like t distribution you may end up with some situation where your statistical tests will involve test statistic which has F distribution under the null hypothesis and you may have to use properties of the distribution to compute your p value.

So, this F distribution we will not go into detail here like I just want to give you an idea of what is going to how this F distribution can possibly arise suppose let us say you have now more than two populations that is call this as to sorry X2 1 all the way up to x2 n, and now X3 is X1 3 all the way up to Xn 3 let us say this is with some distribution with parameter mu 1, this one with some distribution with parameter mu 2, and this one distribution with mu 3.

So, while you may be interested in testing the hypothesis that whether all these values are equal or mu i equals to mu j for some ij when you can go on like you can have n number of such are n number of such populations for some ij pair.

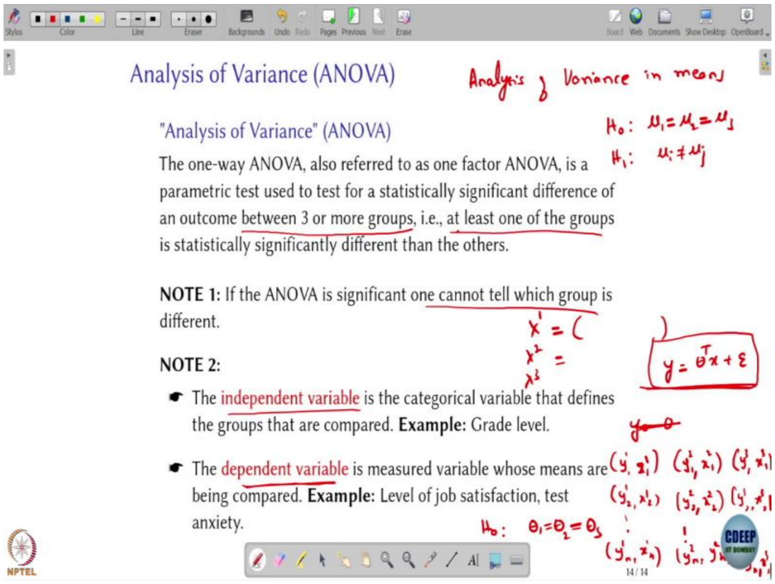
Now, when you are going to answer such question, you may have to construct certain statistics and after some analysis one may end up with statistic which actually satisfies a F distribution and in fact, this is the case in analysis of variance. So, in analysis of variance you are interested in exactly this question whether the underlying parameters or the population parameters are all same or they differ.

And when you are going to construct test statistics for that and when you have to suitably make the test statistic so, that you are able to say something about your claim then that statistics will have F distributions we will not get into the details here, but that is something you can just keep in mind and study more and also that arises in regression models.

Like for example, when you want to have let us say I hope all many of you might be already knowing, linear regression. So, linear regressions texts of given data point x, this is like your input it is trying to find a relation between x and y. And this relation between x and y happens through this parameter theta and there is something motivation are noise here. So, given x, we want to find out what is the best y here and initially the theta is unknown.

We want to find what is that theta and given a set of observation like let us say y1 x1, y2x2 like that, you have some n observation and based on that you want to find a best estimation or representation of that theta to test whether whatever you are done are good, you need to have a statistics and when you find statistics there the F Distribution arises.

(Refer Slide Time: 25:10)



Analysis of Variance (ANOVA)

Analysis of Variance in means

"Analysis of Variance" (ANOVA)

The one-way ANOVA, also referred to as one factor ANOVA, is a parametric test used to test for a statistically significant difference of an outcome between 3 or more groups, i.e., at least one of the groups is statistically significantly different than the others.

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

$H_1: \mu_i \neq \mu_j$

NOTE 1: If the ANOVA is significant one cannot tell which group is different.

$x^1 = ($

$x^2 =$

x^3

$$\boxed{y = \theta^T x + \varepsilon}$$

NOTE 2:

- The independent variable is the categorical variable that defines the groups that are compared. **Example:** Grade level.
- The dependent variable is measured variable whose means are being compared. **Example:** Level of job satisfaction, test anxiety.

~~$y \; \theta$~~

(y^1_1, x^1_1) (y^2_1, x^2_1) $(y^3_1, x^3_1$

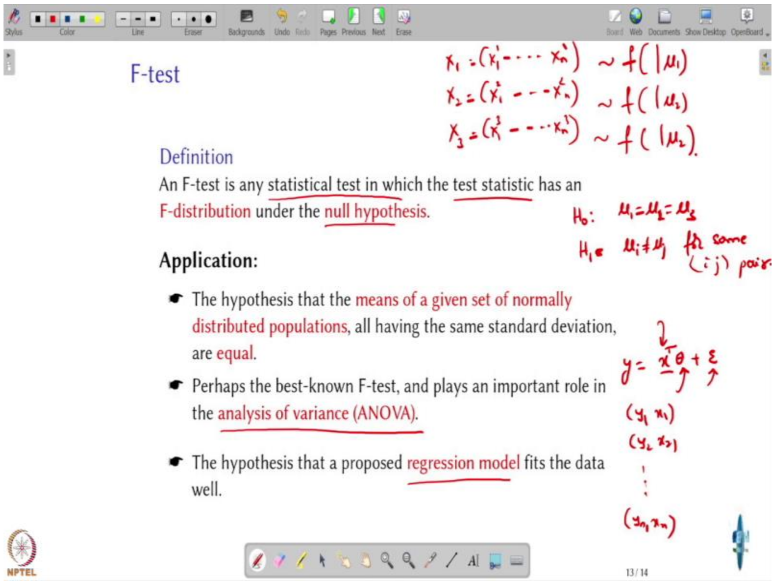
(y^1_2, x^1_2) (y^2_2, x^2_2) $(y^3_2, x^3_2$

$\vdots$

(y^1_n, x^1_n) $(y^2_n, y^2_n$ x^3_n

$H_0: \theta_1 = \theta_2 = \theta_3$

14 / 14



F-test

$X_1 = (x^1_1 - \cdots x^1_n) \sim f(\,|\mu_1)$

$X_2 = (x^2_1 - - - x^2_n) \sim f(\,|\mu_2)$

$X_3 = (x^3_1 - - - \cdot x^3_n) \sim f(\,|\mu_2)$

Definition

An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis.

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

$H_1: \mu_i \neq \mu_j$ for some $(i\,j)$ pair

Application:

- The hypothesis that the means of a given set of normally distributed populations, all having the same standard deviation, are equal.
- Perhaps the best-known F-test, and plays an important role in the analysis of variance (ANOVA).
- The hypothesis that a proposed regression model fits the data well.

$y = x^T \theta + \varepsilon$

$(y_1\; x_1)$

$(y_2\; x_2)$

$\vdots$

(y_n, x_n)

13 / 14

To just briefly say a little more about analysis of variance. So, the one way ANOVA also referred to as one factor ANOVA is a parametric test you to test for a statistically significant difference of an outcome between 3 or more groups. So, here you had be interested to consider when 3 or more groups are there, when it is less than 3, we already know how to use significance tests using our p values computed based on our t distributions.

So, here, we would be interested in checking I want to challenge saying that at least one of the groups is statistically significant sorry, statistically significantly different than the other. So, actually, the name ANOVA here it talks about analysis of variance.

But it is not actually about the variance it is actually analysis of the variance in the means. Analysis of variance in means like for example, as in the previous example, I said like your null hypothesis is checking, whether all the parameters are same. And your alternate hypothesis is are the differ.

So, the variance in the mean is what the ANOVA test is trying to identify whether that parameters are let us say the mean values of all these distributions are the same or not, is what we are going to take as a null hypothesis and try to validate it. So, if it so happens that when I do ANOVA test, the p value happens to be statistically significant, then one cannot tell which group is different, like among this, which groups are different like maybe possibly they are all the same, you do not have enough evidence to say that, there are different.

So, this we already this is like, brief things, which we already talked about, like support, let us say we have this X1 X2 X3 like that, like we had this samples, like as I written here, like so these are like all independent variables, and we want to define, that defines the groups that are to be compared, let us say these are all the values here all the grades of three bunch of students, and we want to see that, whether their average scores are same, or I want to validate the hypothesis that, there every test will be the same.

And where maybe we can use ANOVA test here, and there, the f test arises, or like, as I said, it could be also can come in, let us say, in kind of regression models where you have a bunch of like, you are going to observe data, which are of the form y equals to theta transpose x plus noise. And this, maybe you may be observing this data for a bunch of let us say, you can observe, let me call this y1 1 x1 1, and y1 2 x1 2.

Let us say like this, you have some bunch of data. And this is one population and another could be like, let us say y2 x2 1, y2 2 x2 2, then let us say y2 n, and y2 n. And the third could be y3 1 x3 1, y3 2, x3 2 like this, these are like three bunch of data’s you are observed, which are like where y is our dependent on your x.

And now you want to claim whether these parameters the three parameters associated with this through this linear relation that is whether the theta 1 and theta 2 and theta 3 are same or not then you want to again want to use this ANOVA test, where F test can arise or F distribution can arise where you can calculate again your p value using those F distribution tables.

So, with this I hope you people got a summary of what is p value what is p test one can use your when we know the variance and we are dealing with one population and then you got some exposure to t values t test where we do not know the variance and we have to find the population mean or the parameter of a single sample or two samples.

Then, we also talked about when we have to look for more than two samples, whether they have the same parameters or not either in the independent case or a dependent case, how F distribution can help us compute a p value and decide our significance of the statistical test. So with this we will stop here. Thank you.