Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 07
Lecture - 35

Lecture 35: Drawing reliable conclusions-II

Welcome back. We will continue our discussion on Evaluation of Recommender System that too drawing reliable conclusions and this is the part 2 of this series and today is the last lecture on the module 7. So, this is the content and last lecture we are talking about hypothesis testing. So, these are the generic steps. First we have to formulate the null and alternative hypothesis, then we have to determine our test statistic. Last example that we saw in the last lecture it our test statistic was the Z statistic and we were trying to test the hypothesis with respect to mean and specifically we are trying to find out whether the population mean was same as that of the sample mean.

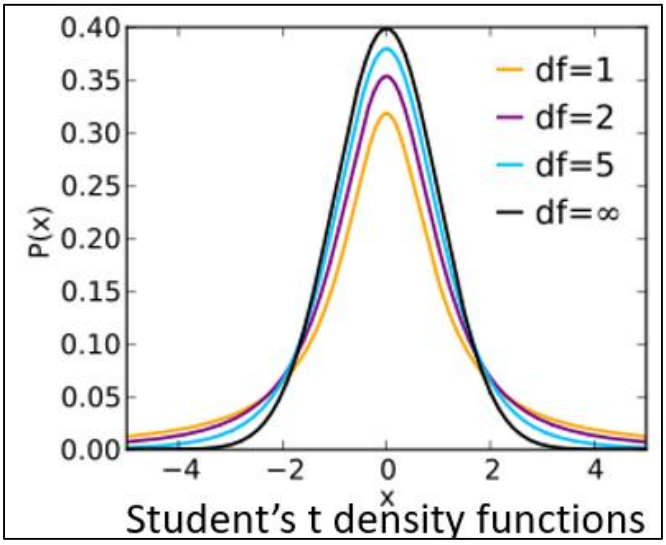
Now, we have to determine the rejection. Next step is to determine the rejection region of null hypothesis choosing a level of significance alpha. Now, we make the decision if the test statistics lie in the rejection region, reject the null hypothesis otherwise do not reject the null hypothesis. We saw we did it in 3 approaches looking at the confidence interval that is non-standardized test statistics, then we found out the Z value and compared with its table value it is available in the table that is that was standardized test statistic and third one was using p value.

And depending on this value of the level of significance if this value was higher than this then we were not able to reject the null hypothesis. Now, there are various tests about the Z test which you already know. So, this Z test is about testing the hypothesis testing the hypothesis with respect to the mean. Here this test requires the knowledge of the standard deviation of sigma and we follow this formula where n is the sample size, x bar is the sample mean, mu is the mean of the population. How do you discover it from the sampling distribution? Mean of the population and sigma is the standard deviation of the population which is known.

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

If this population mean and standard deviation are not known what do you do? Only thing that you have in your hand is your sampling is the sample whose standard deviation and mean you know. So, in such cases you do something called a t test that is quite similar. Now this t test uses sample standard deviation and has a similar formula, but it

also has another parameter called the degree of freedom. So, this degree of freedom basically tells relates to the sample size it is actually sample size minus 1. So, why it is sample size minus 1? Because mean is also used here.



So, if you keep mean as fixed then you can vary only n minus 1 number of values. So, here it uses again the p-value concept, but the nature of the it looks like normal distribution only, but the nature of this distribution nature of this distribution that is the width and the height. So, those things actually depend on the size as the number of as the sample size increase it almost approaches normal distribution. So, when the sample size is small we use it and you can see this is a family of student t density functions and all this change as we change their degree of freedom. So, there are basically I am not going to for discuss formula etcetera right now I am I will just make you familiarize with the concept.

So, that you can use some kind of software to help you out. However, knowledge of theory is equally important and you can refer some good book for this purpose. So, you can have three types of t tests. You test a sample mean with respect to pre-existing population value. Then you can have paired sample in which you have two samples there only one sample where you are comparing it with population mean.

Now you have two sample and we are comparing whether two values mean of two are same. Let us say for you are testing certain algorithm with group of young people and a group of middle aged people are they giving the similar kind of are you getting the similar kind of observation. So, when you do so, you compare to a single group sorry the example that I gave is for the other one. So, in this case you compare two different conditions by a single group of participant. So, let us say with recommender system yourself how and with without recommender system the cell that is happening the people putting things in their shopping cart.

Now in case of independent sample test you have two independent populations. The example that I was telling you have completely in the young people let us say who were in 20s and let us say some middle aged people 40, 30 to 50 and the first one is from let us say from 10 to for 15 to 30, 15 to 25. So, so the first the first one when a single group is given exposure to two different situations is also called within subject t test. And when you have different broad population from completely different category of people then you say it between subject t test. We can also check equality of two variances for this purpose we use something called a f test.

1. Formulate the null and alternate hypotheses.

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

   H_a: $\sigma_1^2 > \sigma_2^2$

   [Note that we might also use $\sigma_1^2 < \sigma_2^2$ or $\sigma_1^2 \neq \sigma_2^2$]

2. Calculate the F ratio.

   $F = s_1^2/s_2^2$

   [where s_1 is the largest or the two variances]

3. Reject the null hypothesis of equal population variances if $F(v_1\text{-}1, v_2\text{-}1) > F_\alpha$

   [or $F_{\alpha/2}$ in the case of a two tailed test]

	Sample 1	Sample 2
n	25	25
s^2	1.04	0.51

So, in this f test the distribution that we use is f distribution and because your variance is a square term it is always possible always positive. So, f distribution which is from only for positive values unlike a normal distribution which starts with 0 and both side it goes. So, this is on the positive part. Now we draw our conclusion from f distribution and we here we consider the ratio between two variances. So, this f distribution again has two degrees of freedom with respect to each group.

First group has n candidates, n subjects, second group have n 2 subjects, n 1 subject, n 2 subjects. Now here the basic assumption is it is randomly drawn they are independent sample from two normal populations. So, here the sample hypothesis is both are equal and if we go by one sided it is greater than if we go by both sided it can be not equal and we calculate this ratio. And we reject the null hypothesis if this value is greater than the f value at alpha level of significance and this becomes alpha by 2 in case of tool tailed test. This is one example to of comparing two variances both have same sample size variances we are supposed to check whether the variances are same or not.

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

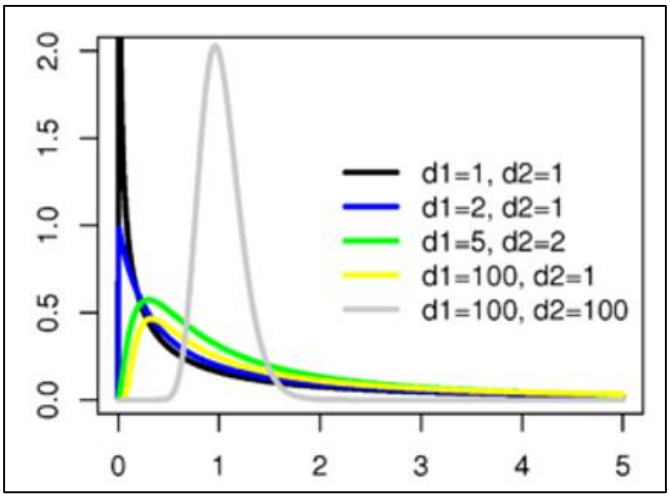
H_a: $\sigma_1^2 > \sigma_2^2$

$F(24,24) = s_1^2/s_2^2 = 1.04/.51 = 2.04$

Assuming $\alpha = 0.05$

Critical value = 1.98 < 2.04

So, we found out this ratio and got this one assuming this we got this one at this level of significance we got this value from table related to f distribution. And since sorry sorry sorry sorry we got this value from f distribution and because this value is less than that of this value you say variability in the new process that sample 2 is less than the variability in the original process. So, this is a one sided test. Now if we have two sample we can compare their mean using t statistic, but if we have more than one sample and we are trying to compare their mean we use something called analysis of variance. Here we have two types of variable one is the response variable which is the quantitative variable another is a factor variable which is used for defining various groups.



So, you have a number of groups who give their responses on quantitative in terms of certain quantitative variable. Now we have to test if the mean of the quantitative variable depends on the group. If each group mean of each group is different from each other. Now the if we have only two groups then we can use two sample t test, but if we have three or more we use ANOVA. Now the question is why do we do so? Actually if we could have compared pair each pair at let us say we have three.

So, 1 2 2 3 1 3 we could have compared, but that gives you that gives you certain wrong inferences that bias that inference becomes biased. So, those theories etcetera you can

refer to some statistic book and study. But a more unbiased situation which gives a better which gives a more reliable conclusion is your ANOVA. Of course after doing ANOVA because it is on all the sample taken together you can do some kind of post hoc analysis by comparing every two at a time. Now this is one typical sample ANOVA situation you are supposed to compare the perceived user perceived diversity let us say in a continuous scale based on three algorithms.

	User Perceived Diversity			
Algorithm 1	10.2	11.8	9.6	12.4
Algorithm 2	12.8	14.7	13.3	15.4
Algorithm 3	7.2	9.8	8.7	9.2

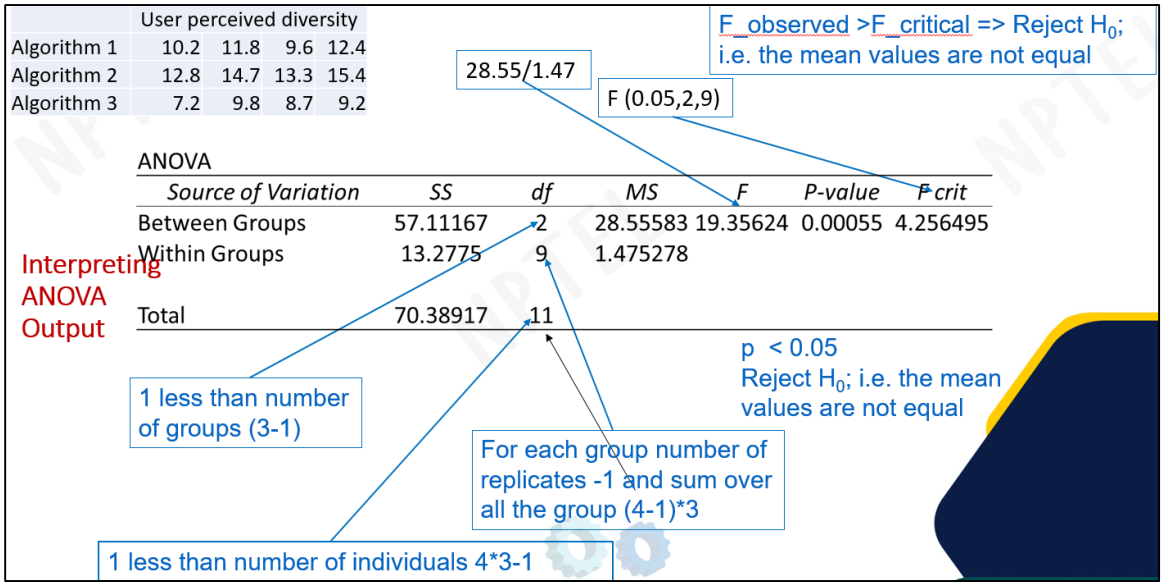
So while doing so what is your hypothesis? Your hypothesis is all group means are equal and alternative hypothesis is they are not. So, once again to determine the p values it uses f statistic. Here the basic assumptions of ANOVA each group is approximately normal you can check this by there are many ways to check whether a particular distribution is normal or not by making histograms or making Q Q quantile plots and so on. And if it is not normal then if there are outliers etcetera that you can find out and try removing them and give one second check. And to check the normality there are many tests like Kolmogorov Smirnov test then you can also do certain chi-square analysis and so on.

Now standard deviation of each group are approximately equal that is the second assumption. So this is a I am not going to into the theory and the basic ANOVA situation we just discussed. There are multiple versions of ANOVA that we are of course, not introducing. But suppose this is one if you do the ANOVA kind of ANOVA that we studied just now and this was our example data. This is this is the kind of table you will be getting.

So from this table what you do you will be finding this some squares between the groups and within the group. What are the groups? This is the first group, this is the second group, this is the third group. And when we find out this between the groups we consider this entire as one population and do the find out their mean and calculate this value. And within the group take individually mean of this, mean of this, mean of this find out then within group and take this value and within the group you find out. Then degrees of freedom to this way because here you are testing three algorithms so 3 minus 1 2.

Here you have total 4 here 4 minus 1 3 and 3 this way, this way 3 and 4 this way. So 4 number of this is 1 2 3 4 4 number of observations minus 1 into 3. And total it is 4 into 3 4 into 3 minus 1 that makes it 11 4 into 3 in the absence this is 4 row wise 3 rows and 4

columns. And you find out this mean square values then f statistic then p value then f critical. We saw there are three ways to remove this if this critical value is this observed value is greater than critical value you have to reject null hypothesis the means are not equal.

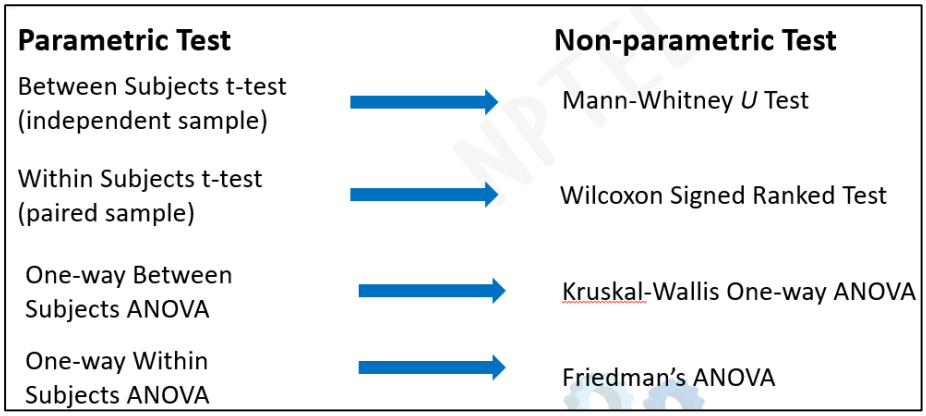


So, this is one approach you compare the observed value that is this one and the critical value. Observed value is greater than critical value so you reject. So then you can also do so by this p value method if p is less than this then you reject H0 the mean values are not equal. These are two important forms of ANOVA so there are actually many forms of ANOVA but the one that we saw just now this is oneway ANOVA is used when assessing for differences in one continuous variable between one group variable. So in our example there is one dependent variable that is user perceived diversity and one dependent variable type of algorithm.

Now within subject ANOVA is another type in which within in which the it is we are we it is appropriate to apply such kind of ANOVA when we examine for the difference in a continuous level variable over time. So the you have to repeat that experiment in number of times the that is saw one times but if it is repeated for number of times at least two over the time then we say within the same subject we repeat the experiment and see the difference. So we won't be talking much about nonparametric tests. Now parametric test as we know parametric tests have requirements about nature and shape of the population involved. These criteria include satisfying the assumptions of outlier, linearity, normality, homoscedasticity to name a few.

T-test and ANOVA are parametric and they need to fulfill these criteria but nonparametric tests which are typically called distribution free they make fewer assumptions and there is a trade-off of course. The nonparametric tests are having more

general ability criteria than the corresponding parametric ones. So typical situations when such tests can be applied are when the outcome is an ordinal variable or rank, when there are outliers in the data and when outcome has clear limits of detection. So when we have this nonparametric so we have talked about t-test and ANOVA on a nonparametric test setting. Independent sample can be replaced by Mann Whitney U-test.



For paired sample t-test we can have Wilcoxon signed rank test and for one way between subject ANOVA we can have Kruskal-Wallis test and one way within subject ANOVA we can have Friedman's ANOVA. So, these are some of the references and we now complete our discussion on inference in drawing inference drawing reliable inference and this was a part of evaluating recommender systems. So, today is the last lecture of evaluating recommender system. So these are our conclusions. We started talking about the steps important steps in hypothesis testing which includes formulating the null and alternative hypothesis, determining the test statistic, determining the rejection region of null hypothesis by choosing a significance level, then making the decision if the test statistics lies in the rejection region or not.

Then we understood that z-test and t-test can be used to test the hypothesis on mean. Z-test is used when the population mean and standard deviation are known otherwise you use t-statistics when the sample size becomes very large t-statistics automatically becomes equivalent to z-statistic. Now f-test is used to compare variances. ANOVA is used to test the claim that mean values of 3 or more populations are equal. Now nonparametric tests do not require the sample that samples come from a population with normal distribution or have any other particular distribution they have fewer assumptions and there are equivalent kind of tests which are comparable with the parametric ones.

But the general liability criteria power of such nonparametric tests are less than the parametric ones. Thank you.