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 - 34

Lecture 34: Drawing reliable conclusions-I

Welcome back. We will continue our discussion on Performance Evaluation of Recommender System. So, in this context we have already seen while evaluating offline recommender systems how to use various matrix and how to conduct the experiments. So, now from today onward we will be talking about talking about how to draw reliable conclusions specifically in the context of offline and online experiments. So, here the idea is when we try making observations from a small set of data, how do we say that observation is correct? Similarly, when we do not know what the unseen users and unseen data is a data the system is going to face in the future. So, therefore, we must have to draw inferences from the data in an offline experiment, difference in the opinion of the subjects in user studies and change in user behavior while interacting with different kinds of recommender system specifically in case of online study.

Now this process will reduce the possibility of error that error on the conclusions that we draw on the limited sample size. Moving ahead this example I just took from some research paper. So, here you can see in a typical user study in the context of some recommender system analysis few questions are asked. The first question was taking the data in terms of 3 responses once a week, 2 to 3 times a week daily.

So, you are getting some kind of nominal variable. The second question you are getting a score between 1 to 100. Third one as well in a continuous scale. Here you get some kind of Likert scale and so on. So, this data how are you going to collect? You are going to collect from few people who participate in the user study.

You give them this questionnaire besides making observation you give them the questionnaire they will fill up this value. Similarly, from observations also you can say like their eye movement, how much time they spend on a particular page. So, those kind of values also you are going to get. So, when you get some these values from a small subset of the users that you are going to face in the future how to draw the reliable conclusion. Now, this is again another example.

Here all these questions are asked to give a rating from in a Likert scale. So, the nature of the data also varies. The nature of the data in the sense the measurement scales also vary as the kind of questions you ask. Now, the idea here is if we have a limited sample how to draw the reliable conclusion? Typically, you will take the help of statistics to draw these reliable conclusions. So, the procedures with respect to this need to be known.

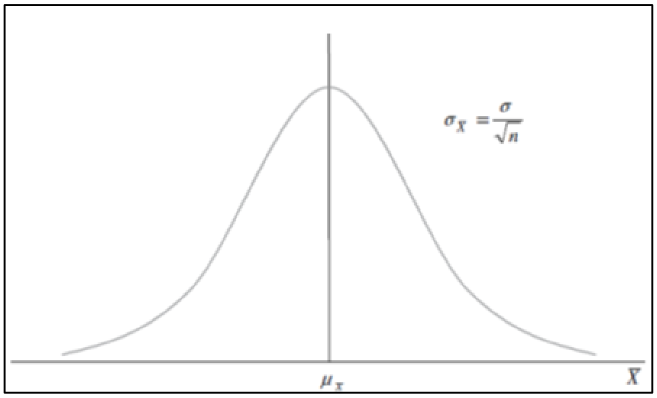
So, there are and the statistical procedure which you will be using for this purpose is called inferential statistics. These inferential statistics can be of two types parametric and nonparametric. Besides this you are also expected to do two kind of things here. One is you have to estimate the value of the statistic which can be done as finding the exact value or finding the interval within which the value lies. Similarly, you can also do some kind of hypothesis testing which basically I told you beforehand that it can be parametric or nonparametric.

Now, when it comes to parametric versus nonparametric inferential statistics, the difference lies on the fact that in case of parametric tests as the name indicate we need certain parameters and that parameters pertains to some kind of probability distribution from which the values are supposed to be drawn. Now, this distribution typically will be considered as normal distribution, but there are at times this may change as well. For example, for a small sample size when the variance is not known you will be using some other kind of you will assume that distribution to be followed is student's t distribution. Now nonparametric statistics are not based on any such assumptions. However, they also require some minimum assumptions to be followed, but not as stringent as parametric ones.

Now, these nonparametric tests sometimes are called therefore distribution-free tests. Now come to the first we will be talking about the parametric ones. And in case of this parametric ones we will be talking about few such basic concepts which can build the ground for you to study further. As you as this is not completely a part of recommender system and you need to study basically you need to study statistics basic statistics where you need to study about designing experiments and you are supposed to make how to make statistical decisions. And if these statistical decisions become multiple variables then in a multivariate setting also you are supposed to study.

So, both statistical decision modeling and multivariate statistics need to be properly understood to get a good grasp of this topic. However, for the purpose of our this course what we will do we will be simply limiting ourselves to few basic concepts and try to relate it to some of the tests which are typically used. But this is not the this is not everything about inferential statistics. So, to start with we are going to talk about the parametric parametric tests. So, before that let us talk about little bit about what is sampling distribution.

So, here our aim is to in this analysis our aim is to study some estimator which is called statistic which is the characteristic of a sample. Mean can be a statistic variance can be a statistic. So, when we try to find out the mean and variance of the population we will be taking help of sample mean. So, what is sample mean? The small group of people on which you are conducting your study make your sample. So, for example, in Netflix many kind of viewers are seeing the movies, but when it comes to doing some kind of user study or online study only a subset of this entire user will be used.



Probably they will be similar in characteristics which it is a representative of the entire population entire universe of the customers who will be seeing these Netflix movies. So, this from this small subset certain characteristics will compute and these characteristics when we take the mean of those characteristics it is called the sample mean. And for the entire people whom who are actual customers actual viewers. So, for them we call it a population and some statistic with respect to that population also we collect. So, those are called population statistic.

So, this sample population mean and population variance are those. So, we are supposed to our aim is to find out whether this sample mean is representative of population mean or the sample variance is representative of the population variance. So, for this procedure we may carry out this, but we really do not know what the population is right. Otherwise though we we could have collected the data from the entire set of people. We are we know that it is impractical to take everybody into consideration while conducting this study I mean entire customer base and moreover the customer base can change over the time.

Now, when we conduct this study with a small group of people which we call as sample we may be doing this study on multiple sample. So, when we do it with multiple sample in every sample we will be getting one mean. So, if we get let us say you make total 20 sample to conduct certain study and make one observe observe one parameter and that parameter will have how many we took 20 sample. So, we get 20 sample mean value. Now this sample mean value will follow some distribution.

So, if we it is observed that if we fit this into some distribution that typically follows normal distribution as per some theorem called central limit theorem. Now, mean of this distribution is the right estimate of the population mean. So, that distribution is called sampling distribution. Now, for a sampling these are some of the observation for a sampling distribution. So, sampling distribution of sample mean will always be normal this is as per that central limit theorem.

The mean of sampling distribution of sample mean will be equal to population mean. So, there will be many such x bars x bar is x bar 1 x x bar 2 and so on. So, there will be many such sample mean these are all sample mean. And for all this sample mean mu x bar is the mean of all those sample mean and that you plot it and you can get this one. Now, the standard deviation of the sample mean is this population mean divided by root n.

Why this root n? What is this n? n is the size sample size. And this term is also called the standard error. So, now the question is what should be the size of this sample? It depends typically it should be 30 or more to allow a close approximation of normal distribution. And whether the sample size is adequate or not for a study that also there are theories to justify that and this is not part of this discussion and you have to study you have to truly study about statistics to be strong in this particular area. And moreover, sometimes if the population distribution is symmetric and unimodal sample size of 4 and 5 also gives good estimate.

And if it is already normal population is already normal even a very small sample can give you, but which of course, you do not really many times do not know about the distribution this population distribution. Now come to the parameter estimation. I told you parameter estimation can be either point estimation or it can be interval estimation. Now in case of point estimation you estimate a single numerical value of certain statistical parameter. For example, mean or variance and when you estimate this they should be unbiased and variance should be minimum.

$$P(L \leq \theta \leq U) = 1 - \alpha$$

In a two-sided confidence interval

$$P(L \leq \theta \leq U) = 1 - \alpha$$

Now coming to the interval estimation we are supposed to suppose this is our parameter theta we are supposed to find out certain lower and upper bound to this. So, this is the lower bound this is the upper bound and we are supposed to we are supposed to determine it in such a manner. So that 1 minus alpha percent of this of the observation would be lying within this interval. For example, if we say that we are 90 percent confident that the mean of certain observation is desirable. Then if 100 such intervals are constructed one from each 100 sample we would get at least 90 of them would contain the population mean.

So, this appears to be very complicated, but there are some statistical procedure to simplify this and give you a reasonable value of this probability. Now there exist established formula for computing this confidence interval, but depending on the nature of the distribution the formula for computing this L and U varies. But anyway you really do not have to do it by hand you have to take help of certain statistical software to do it. So, while doing so you will be certainly getting some values, but what our attempt here is to know the very fundamental associate fundamental concepts associated with this so that you can conveniently interpret them. So, now come to the concept of hypothesis test hypothesis testing.

So, here we make a proposition and try to find out from the try to find out evidence from the data whether that proposition is a valid proposition or not. So, in this context we can use many statistic and call them as test statistic and use their value to draw the reliable conclusions. So, this is for Z statistic, this is for T statistic and so on. We will be knowing more about them later. So when you do so you can you make something called a null hypothesis and which and against this you make something called a alternate alternative or research hypothesis.

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \text{ where } \bar{x}, \mu \text{ are the sample and population means}$$

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \text{ where } s \text{ is the sample standard deviation}$$

So, in as the name indicates the null hypothesis is about making something null. So, what is that null here? You test whether a particular parameter equals certain value. For example, certain mu is equal to 300. What is null is happening here? Mu minus 300 is 0. So, that is how the word null comes in.

So you do this. Now while talking about this hypothesis we can test this hypothesis for both mean as well as standard deviation. And while doing so in case of parameter statistics our idea is it has to follow certain assumptions. That assumption is the population follows more or less normal distribution and this mu and variances are unbiased estimates collected from the sampling distributions which represents the to represent the population. So in this context we talk about something called a Z score. So this is the Z score for the population and to represent this may be the Z score for the sample.

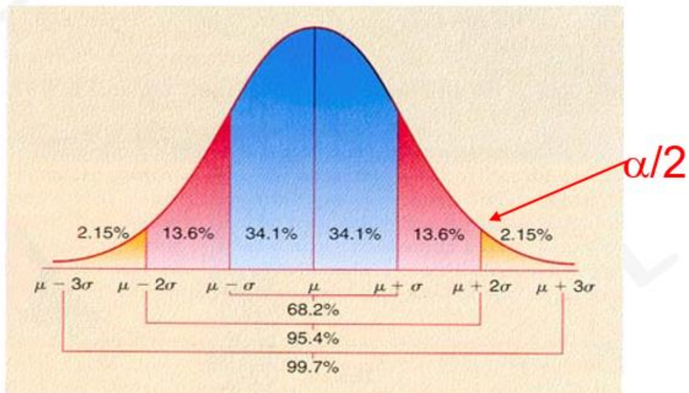
Look formula are almost same only with the difference that here you are doing it only for a small set and here from that small set from the all the x bars that you obtain from that small scale small set you will be making this Z. Now this Z score represents a value on the x axis for which we know the p value the probability value. So now I also would like to mention along with this normal distribution like this or the Gaussian distribution like this we will also have a normal distribution table from which we can get the standard Z values. Now if you look at this distribution this is the mean this is one standard deviation away this is two standard deviations away and if we look at this almost 90 percent of the observations will be within two standard deviations away from the mean value. So while testing significance that if it is lying within that 95 percent of the confidence interval we have to ensure that this is not lying in this rejection region this rejection region.

So now this rejection if we telling as from one side one side one side in the sense either left or at the right side then this is the one sided this is from some right side of this this is the left. So this can be one sided so in one tailed so the test that you consider can be either one tailed that is you check on the one side or you check on the both sides this is one side this is both sides. So if we say it is 95 percent significant this has to lie within this or this part should be 95 percent of the observation and rest 5 percent will be this side and this side.

So basically 0.25 will be here and 0.25 will be at the other side. This is one example suppose we have a situation in which some analyst is claiming analyst is claiming that the mean viewing time for recommendation is 350 milliseconds and we want to check whether his claim is reasonable or not how by taking some sample. So this is just analysts observation so in this context we can say analyst is knowledgeable enough to know that this may be the population mean observation mean viewing time. So now we are supposed to test whether this is correct or not so mu is equal to 300 is our null hypothesis and this is our alternative hypothesis that this is not equal. Not equal means in the sense we have to now conduct two sided test two tailed test it is not one tailed.

So let us proceed. So we are supposed to now gather the evidence whether to reject the null hypothesis in favor of alternative or we fail to reject. Do you remember we don't say that we accept we either fail to reject or reject because we do not have we have a limited evidence so we cannot outrightly say that we completely reject it we fail to reject it. So while doing so we can have many kind of errors. So what kind of errors basically two types of error type 1 error and type 2 error. Now type 1 error occurs when we reject a true null hypothesis we reject a true null hypothesis which means it is a good one but we reject it and type 2 error occurs when we do not reject a false hypothesis.

So a wrong thing we do not reject. Now while doing this hypothesis testing we will be taking help of certain sample. So we are supposed to see that if our mean see this is just one example situation which I am using to explain the idea but you need to study more to have different kinds of other distribution or maybe nonparametric cases but here right now the idea the basic idea however the basic idea that we are discussing remains same. So now we are supposed to find out whether some sample mean belong to this distribution or not. So if we say that if we say that our assumption is population mean is 350 and we are taking some sample and finding some other value let us say x bar is 370.16 for certain sample whether it is statistically same as that of population mean.



Population

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

Sample

$$z_i = \frac{x_i - \bar{x}}{s} \qquad \bar{x}_z = 0 \qquad s_z = 1$$

Now when we say it is statistically same as that of population mean which means they need not be actually equal in arithmetic sense. So if they are not equal in the arithmetic sense they should be probabilistically close as close as possible. So to do so we are looking at this sampling distribution and trying to find out whether it is within this sampling distribution or not. When we say within or within the got it is a very informal word got of the sampling distribution which means this part we are targeting whether it is within this part. Now what is this part as we have seen here if we consider up to this it is 68.2 percent observation if we consider from here to here it is 95.4 percent observation and so on for that particular example. Now here similarly we are supposed to find out so we have to first decide at what level of confidence we decide. So if we decide that at 95 percent of confidence we have to determine we must know this range and within this x bar should follow. Now this range or this confidence interval has to be determined.

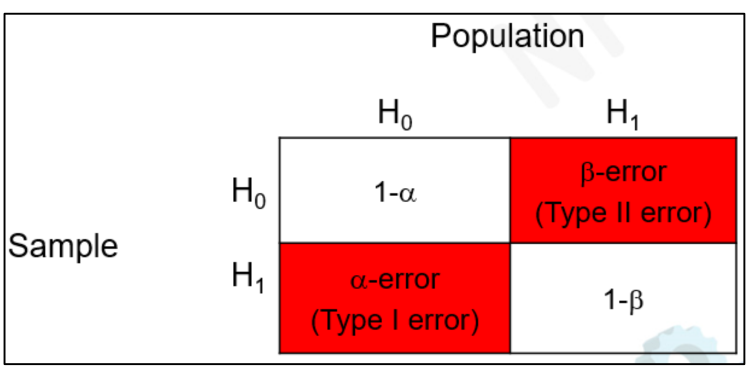
So as I told you there are many ways to determine this confidence interval but because it is a standard normal distribution with the variance as 15 and this point 1.96 we get the we get it from the normal distribution table. Then 350 plus 1.996 into 15 gives you the the upper bound of the confidence interval and this gives you the lower bound. So this is just one example we are seeing but actually if when you use a computer software it will be giving you depending on the nature of the problem and the and the distribution etc.

that you are trying to study. So here what we did we use this confidence interval. So we didn't use any test statistic as such which is standardized. So this is a procedure of hypothesis testing using only non-standardized test statistic. We can do it using a standardized test statistic. So in this case we are supposed to standardize this mean value.

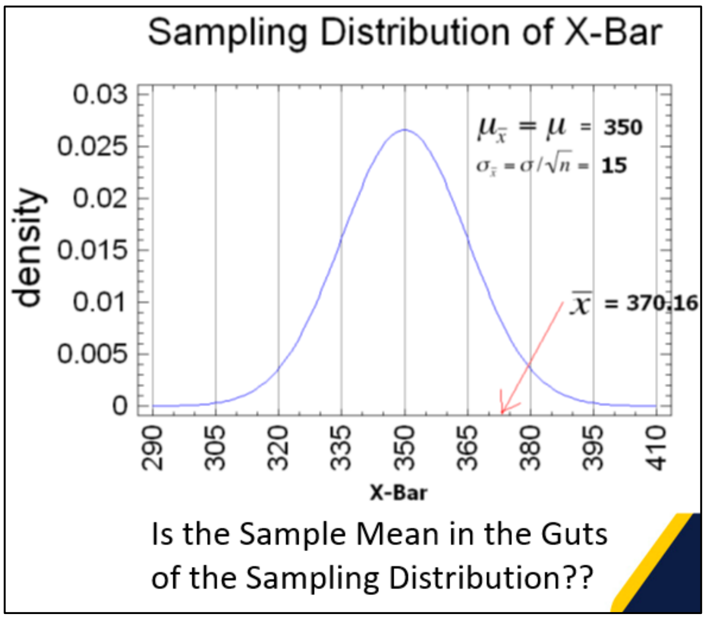
So this is the mean sample mean this is your population mean that we are getting and this is your sample standard deviation. So this value that we get we are supposed to this was our limit from this point to this point and this when we consider this one end we were supposed to go up to 1.96 and this side minus 1.96. So if our value is less than 1.96 at one side and minus 1.96 at the other side we are happy we are good. So here in this case we are good. So we can we cannot reject I mean this mean is close enough. The third way to do this is using p value. So what is this p value? This p value is basically when you this point is your see originally this was the value where this was z was 1.344. Now this was my limit point.

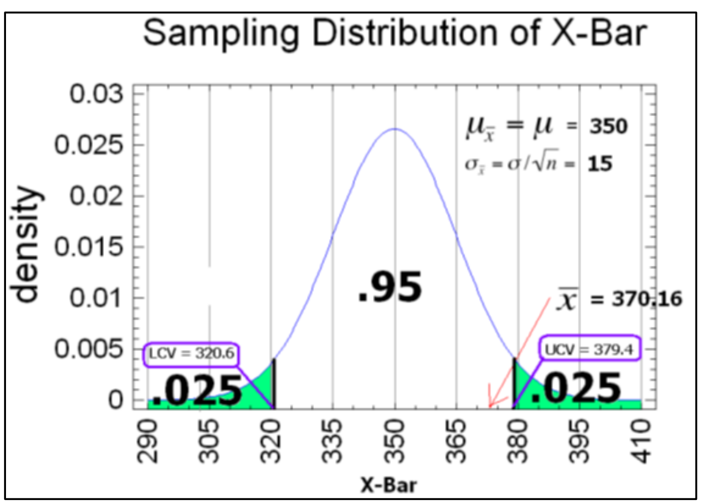
This was my limit point. Now I pull this. If I pull this the original 1.96 which was there my I mean the originally I said 25 percent observations will be at the other side but now I have increased my limit. So if I increase my limit increase my limit I am considering both from both sides I am trying to pull it this side I have pull it up to that point 0.1344 and same approach I did at the other tail.

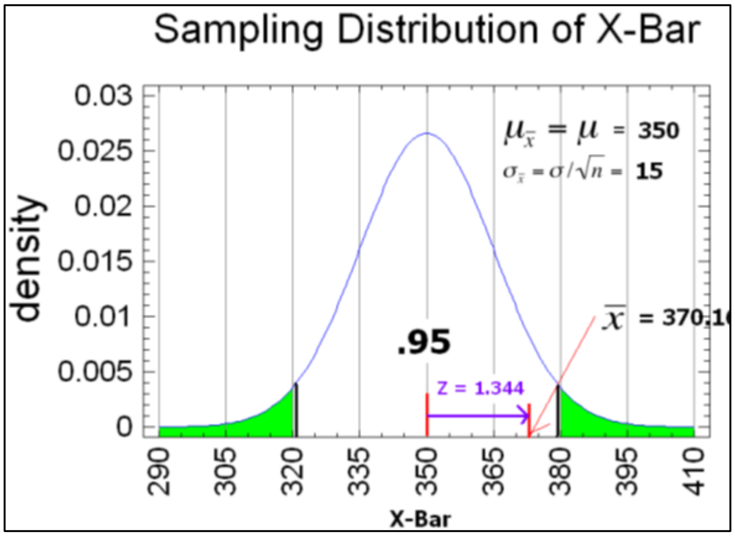
So because it is a 2 sided test I multiplied it with 2. So this becomes my p value. So this p value if I decide that p my p value is if it is if it is let us say 0.05 if we since we define our gut or the interval as the center of 95 percent that is alpha equal to 0.05 the rejection region is the other 5 percent.

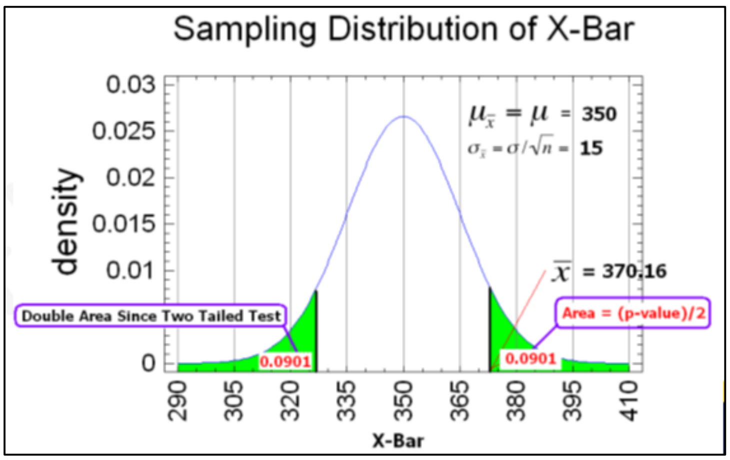


Since our sample mean x bar is in the 18.01 percent of the region it cannot be in our 5 percent rejection region. So we say that we have got sufficient evidence that this is close enough to the mean. So basically we can draw statistical conclusions specifically in case the we if we have a sampling distribution which is normal normal sampling distribution we can do it in 3 ways using a non standardized test statistics where we will be we will be deciding the confidence interval. Interval in this particular example we consider 95 percent confidence interval. We can use standardized test statistic where you have to determine the value of standardized z value. And in this case we understood we fail to reject the null hypothesis of 5 percent level of significance because it was much before. Now in the p value approach the third approach is p value approach. So these are 3 approaches. Any one if you take it is not that all the 3 you are supposed to follow anyone if you follow typically your software will be giving you p value. So as long as this p value is greater than this value of alpha what is this alpha? We talked about alpha while trying to find out the find out the types of error alpha and beta type 1 and type 2 error.









So this is that rejection region. So we can reject the null hypothesis at 5 percent level of significance. So with this we end this lecture and we continue this discussion in the next lecture. So these are my references and to draw the conclusion we saw that we can take help of inferential statistics to draw certain conclusions about the population certain parameter about the population based on the sample data. There are two main procedure procedures for this one is one is estimation another is hypothesis testing. And estimation we have point and interval estimation which depends on the nature of the data and the corresponding distribution.

And in hypothesis testing we can do it in a parametric and nonparametric way and in this particular lecture we saw a very simplistic setting of parametric inferential statistics and we draw this conclusion. We can draw this conclusion in this context using three ways using unstandardized test statistics using standardized test statistic and p-value. Thank you.