NPTEL

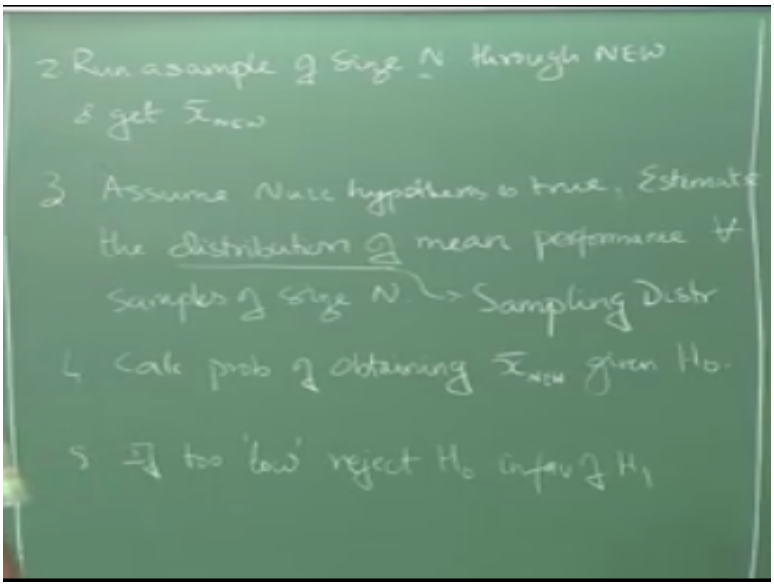
NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture-55
Hypothesis Testing – II – Sampling Distributions & the Z test

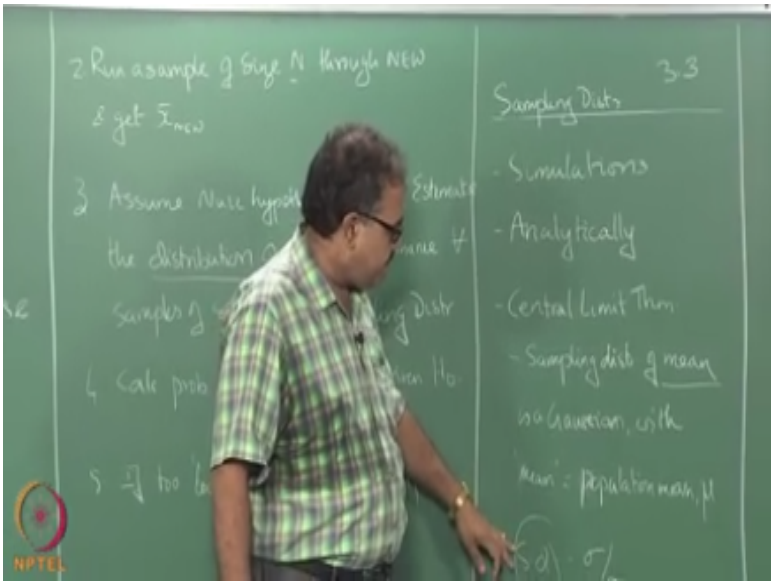
Prof. Balaraman Ravindran
Computer Science and Engineering
Indian Institute of Technology Madras

(Refer Slide Time: 00:16)



So I see as people what is your answer to that, well as yeah mostly partly see as people from one, what is your answer to that. Bootstrap right, I can do with some kind of simulations bootstrap does not really give me sampling distributions per se right. Yeah, bootstrap is one way of doing it, yeah you could do bootstrap, but you have to be careful about it.

(Refer Slide Time: 00:57)



So some kind of simulation based methods right, or I can do, I can try to find the sampling distribution analytically provided I have simple enough underlying data distributions and I know something about the underlying distributions, then I can find the sampling distribution analytically. For example, so if I want to look at let us say, let us take as a case, where I am tossing coins okay.

So I toss a coin 20 times and end up with 14 heads okay. So is the coin likely to be fair or not, 20 coins, 14 heads is it fair or not? You do not know, 20 coins 18 heads okay. See all of these are introducing, how will you make it, now all of you know how to do this right. Now tell me how will you do it formally, think you will setup a null hypothesis where you will say that the probability of the coin coming up heads is 0.5, that is the null hypothesis.

The alternate hypothesis will be probability of the coin coming up which is greater than 0.5 okay. Now what I do is, I run a sample which is 20 tosses, I have found out that it is 14, so the statistics is 14 not xbar it is not the average, but the statistics is 14 right. This whole process can be done for anything not just for mean right, so the statistic is 14. Now I am assuming the null hypothesis is true, which is 0.5 okay.

And I have to figure out what will be the distribution of the number of heads right. So what is the probability that I will get, 1 head, what is the probability I will get 0 heads if I toss 20 times, what is the probability I will get 1 head if I toss 20 times, what is the probability I will get 2

heads bla.. bla… bla.., like that right, I compute all the probabilities right. Now I will look at the probability of obtaining 14 according to this distribution.

And if I do not like that probability right, then I can say that no, no rejection the null hypothesis right. If I like the probability I can say no, no accept the null hypothesis. There is no alternate here, alternate just say this is greater than 0.5 right. So in this case yes, you can accept, see this is why you have to be very careful about formulating the alternate hypothesis, because at the end of the day I am going to say accept the alternate hypothesis rejecting the null hypothesis.

But if the alternate hypothesis was it is less than 0.5 that is not a valid alternate hypothesis to make given the data that you have right. So it is nothing wrong with the whole hypothesis testing process it is something wrong with the where you setup the problem right. So if you are sure if it is higher or lower then there is a different issue, then you can say not equal to 0.5, and then run the test right.

But then it is up to you, if you have to use your understanding of the domain to come up with appropriate alternate hypothesis, there was another question here. Then the, yeah, yeah that is not, I mean if you do exploratory experiments you remember that, I told you in the previous class before you start your actual experiments before you do your actual experiments you do some exploratory analysis of the data right.

When you do the exploratory analysis you will get some idea okay, you start suspecting that okay, this coin is actually biased towards H, and then you will setup this experiment. So this is one very specific statistic that you gather while you are running the experiment, but before you setup the alternate hypothesis you should do some amount of exploration right. So you cannot walk in to an experiment blind about the domain.

So this is the practical issues that you should be aware off, you remember last class I told you very much about the need for exploratory analysis right, so we have to do exploratory analysis before you setup your actual experiment right. Yeah, so that I will talk about, it is just the basic structure I will talk about how about the confidence with that I will come to in a minute yeah right.

So but you know how to do this right, you can analytically you can figure out what is the proportion of heads you will get in 20 tosses right. So all of you know how to do the binomial

and then you can figure out what the probability is right. So the another way of doing it right just to give you another example of that right so in another way is to make use of specific properties of the parameter that you are trying to estimate in fact if you are looking to estimate mean right so we have one big advantage what is that okay something called central limit theorem so what does the central limit theorem say

Right so since is samples I draw or all independent samples right I draw samples of n variables I mean the size n right a draw a samples of size n and these are drawn independently right I do not have any bases of I am sorry any bias from the previous samples I have drawn I am going to draw a lot of independent samples of n variables so if you think about the performance of the algorithm on this okay it is essentially independent samples I am drawing from the same similar distributed random variable right.

So essentially what central limit theorem tells us is regard less of the underlying distribution from which the data is drawn okay the sampling distribution will be Gaussian right will be normal distribution right it tells us that the sampling distribution will be a normal distribution and anything else exactly so the mean of the sampling distribution will be the mean of the population from which the samples are being drawn right so central limit theorem tells us that okay.

So I am call it the population mean as mu right and the standard deviation is s σ is the population standard deviation right so the thing to note is that this does not depend on the underlying distribution right regard less of what the pollution distribution is right the sampling distribution will be a Gaussian the mean will be the same as the population mean right and the standard deviation will; be σ √10 where σ is the population standard deviation right standard deviation will be the population standard deviation by√10.

Right so the larger the sample size the narrower the sampling distribution does it make scene so the sampling distribution will always be centered around the population mean right so one thing which we can control is how wide is the sampling distribution right so if n is small then the sampling distribution is wide n Is large the sampling distribution is narrow okay.

So unfortunately we only have a central limit theorem for mean right so this is what said the sampling distribution of the mean okay so I just want to stop once and just sorry if it is a getting

top reparative just once more to emphases what I mean by sampling distribution of mean what does it mean to many means remains we have hold kamala Hassan movie distribution of the sample means of the data so that is what I mean by sampling distribution of means's okay is that clear.

Right because I have seen people get confused and give all kinds of different interpretation of that so I have the sample data so I have taken I readily sample data of size n from the population right and at compute the mean of this samples and distribution of that means is the sampling distribution of means I see great so this standard deviation of this sampling distribution is also sometimes called the standard error of the mean is the standard deviation of the sampling distribution.

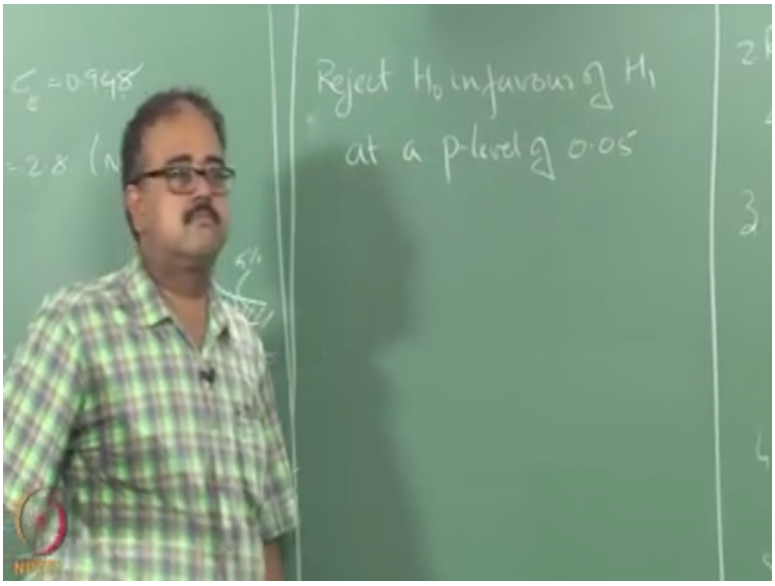
Okay so empirically right we can say that n greater tan equal to 30 indicates n is large enough right your standard deviation standard error becomes small right the standard error becomes small what happens so any sample of size n I can take and estimate the statics and I am more or less correct with the very high probability I will be correct so that is what it means. But okay, you have to be careful about what is correct.

Okay, so couple of savages here so if you are, if your population has a high standard deviation so what you have to do, you have to make your sample size very large so that your standard error comes down right, further if you variance of your thing is very high so underlying population is very high right, that means you require very large samples right, so what you should be thinking at about at that point is to see if there are some other way you can step up the test, right.

So you should not come to a point where I need millions of sample just to reduce my standard error. I am sorry, yeah is it that is what I am saying you have to think of some other way of going about doing this rather than just saying that increase the number of samples right, so what would be other ways of doing it is to try to bin data right, so if you bin data what happens is so small variations in the data will go away right so something like 3., 2.3, 3.4, 5,6 all of them will can be bin to say 3.5 right, that means a small variations will go away so lesser amounts of noise right.

So the variance will also drop a bit, so you can do things like this they can do some kind of noise reduction techniques, try to do is see if we can reduce the variance in the data without actually dropping anything important so that is curial, right so that is a kind of things you have to try so that is one caviet that I wanted to say.

(Refer Slide Time: 14:15)



So the next we look at is specifically using the sampling distribution of mean right, and come with something call the Z test okay, the people know about the Z test right, okay, so again okay let us do another example here okay, so I am giving you, you know how to solve a certain kind of problem right, so you have been train to solve of solve these problems for you know years and years and shuffle like that, like you do in your JE preparation kind of things right, and then I know how much I would expect student to score on a specific set of problems, right.

Then suddenly I find that a new kind of problem comes up right, and the students are taking longer to solve these problems right, let us say that the students takes a unit time to solve problem traditionally now they are taking say 2.8 times that to solve the problem right. So my claim is that this new problems are harder than usual right, okay how will I verify that right, so again I am just going to walk you through this setting up this hypothesis right.

So basically I am going to start off with saying that so I have the easy problems, I have the hard problems now null hypothesis is there are no different okay, I am going to say they both take unit time to solve, right. So on the second is the alternate hypothesis right, so the easy problems take lesser time than the hard problems to solve, is it fine. So what I know from previous data is that okay, just some number say the actual numbers are less important but just the process here, right.

And then so what now as it is said set is up I take 25 hard problems I ask the students to solve it and I get my okay, so we are looking at mean so the sampling distribution is going to be

Gaussian right, what will be the mean of the sampling distribution. You do not have to do that yeah okay, so now what I am going to do is to a little trick now I have the sampling distribution right so I need to know what is the probability of seeing 2.8 under the sampling distribution.

So the mean is one the standard deviation is 0.19 right, and there is 2.8 I need to know what is the probability of c 2.8 under the sampling distribution right. So I can do that lot of you know how to find the probability here right so this is mean is one and this is so many standard deviation above the mean right.

So the probability is actually very small okay, but you can also do this in a straightly both convenient fashion so you basically points and something call the z score let us say essentially assuming that you know sampling distribution is going to be 0 mean and unit variants so what you do this? This is called the standard normal so the standard Gaussian so I am going to convert my sampling distribution in to standard Gaussian and then try to find the probability of 2.8 in that standard Gaussian right.

So essentially the z score is right so that will be the z score so essentially I take my actual statistic subtract the mean from that right and then divide by the variants, so this case will be unit variants that gives me the 0 mean right. So it is 9.47 so essentially what it says is this guy is 9.47 standard deviations above my mean right so what is the probability of that happening? Very, very small right so what we do? Well we do not know that, so en we do is essentially looking at some standard values that we do now right.

What is the probability of something lying greater than 1.645 σ above the mean, this number you should by order it if you know at some point 1.65 is in front right, so implies the probability of that happening is yeah, so basically what it means is right so this side is 95% right so if we take the Gaussian take this is mean right take 1.65 σ above the Gaussian right the area to the curve to the right of that is 5% of the total area right.

Likewise right area here is 5% right so it is number we need to know, so why you are interested in 0.0 to 5 2 tail right sometimes we might what do? Not look at greater than we might want to look at not equal to right. Now I have make the, my hypothesis was the second set of problems was harder than the first set of problems, if my hypothesis had been if the second set of problems are different from the first set of problems then I will have new is not equal to μ old right.

So in such cases I should have been looking at 0.196 because I could go on either side right so I have to be make sure that my statistic is greater than either +0.196 or lesser than -0.196 for me to be sure that I will be wrong only 5 % of the times right. So in the olden days we actually use to have a z table that use to tell you for one side a test what should be the z statistic you should look at for 2 sided test what is should be z statistic which you should look at so on and so forth.

If you people have actually looked at clacks table ever you actually have a Z table as part of the clacks table and this is essentially what the Z table is telling you μe is the, the easy task we had right that is under the null hypothesis that is the population mean yeah. It is 9.47 right I mean 9.47 is way, way higher than my 1.65 right so therefore I can reject the null hypothesis right at a confidence level the way we straight this like this I can reject the null hypothesis at a confidence level of at least .05 so that means that probability of making me error in rejecting the null hypothesis is less than .05 so the conclusion is that yes it is harder.

So I accepted the alternative hypothesis right in rejecting the null hypothesis and the probability of making me an error is less than .05 okay yeah it could say that but confidence level means something else okay so that I why the statisticians are very careful about the conclusion that they will draw from this.

I will write down the conclusion okay you can say p level or p value right but the sensity the p value of something means the probability of P being wrong in that conclusion okay so probability is wrong in that conclusion given all the assumptions I have made see this is the reason why we want a very high things right we are assuming our sample size is large enough for central limit theorem to apply right.

And then whole bunch of things that we are doing right the point is here right there is no notion of the accuracy here so the probability of making me an error in concluding that h0 is not write is 5% okay so if you want the whole answers of confidence is 95% I recommend you to read the book so it is not you cannot really say confidence level is 95% right.

So I can just say the probability of error is I will come back later end tell you the confidence when I talk about confidence intervals at this point I have just leave it out right so therefore we can reject the null hypothesis so this is good right yeah this is Z test yeah and in the null hypothesis they are the same right.

There is a whole idea right under the null hypothesis what the sampling distribution is? That is ht we are trying to under the null hypothesis is being true what is the sampling distribution is what we are trying to find right so they are the same level when null hypothesis is true right so typical p value is that run these things alright fine so if we are looking at more critical domain like medical domain and I would expect the p value of .001 right not.005 however so you have to be more careful really sure what, what kind of.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in