NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture-56
Hypothesis Testing III

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

So what do you do when so here we assume that we knew the population standard deviation right so what if you do not know the population standard deviation, sorry go for the t-test do you can still do the z-test? You try to estimate the sample standard deviation okay you can estimate the sample standard deviation instead of knowing the true population standard deviation assume that the sample standard deviation is correct right. And then go ahead and do the z-test result okay so that is essentially what you do.

But what you do even the mean is unknown can you say something useful if I do not know anything about the mean of the old systems is that is that even a question that you can ask right you can so you can just say that a I am going to hypothesize that the running time on this new problem should be two standard units and under is 2.8, so can I say that the new problems are things that will with confidence okay.

Can I say that they will take more than two standard units to compute and I can make an assumption about what the mean should be of what I think is the baseline case and then you can compare against the assumed mean right so you can still use the z-test I we do not be in a hurry to abandon the z-test because it is still useful right, so you do the t-test let any questions about this I mean so if you do not have the standard deviation just run some samples and deviation tests estimates.

Clip code you do not do that so essentially what would you have to do in that case is okay what is the probability that these ten different measurements I made gives me I mean I would have generated all these ten measurements from this right then I my sampling distribution becomes slightly different so I have a set often samples that I have drawn right and what is the probability

that these ten samples will turn up exactly this fashion can you imagine how horrendous that sampling distribution will be right.

So if you can if you have a easy way of computing the sampling distribution which you could write you can because with all this simulation based ideas you can set up arbitrarily complex sampling distribution the reason we have to stick to a simplify simple terms the sampling distribution because that is what central limit theorem gives us right so if you if you are happy to do this in a simulation you can set up the sampling distribution using simulation.

Assuming you have access to ways of generating many samples from the underlying data right so if you are just doing it bootstrap then you run into problems way because the sampling is no longer independent there maybe do some large n samples they are repeatedly sampling from that so the sampling is no longer independent and you gotta dress for that right but if you truly have a way of sampling from the underlying data right.

You can set up the sampling distribution you wanted right so people understood this question this question was why did I have only one $\bar{x}$ ν right why cannot a sample $\bar{x}$ ν on multiple sample so why cannot I just compute this on multiple samples and figure out right so the answer is yes you can but what do you have to do is you have to find out suppose you do this 10 times right now I will have ten numbers you have to find out what is the probability that I could have drawn all of these ten numbers under the old under the null hypothesis right.

So if I can find that out okay now I will need a different quote unquote sampling distribution for this right so if I have a way of constructing that sampling distribution then I can run the test and one way of constructing the samples sampling distribution is through simulations just keep drawing many samples sets of sets of ten samples right and then and then look at the distribution of those and then try to form industry formula estimates from that right.

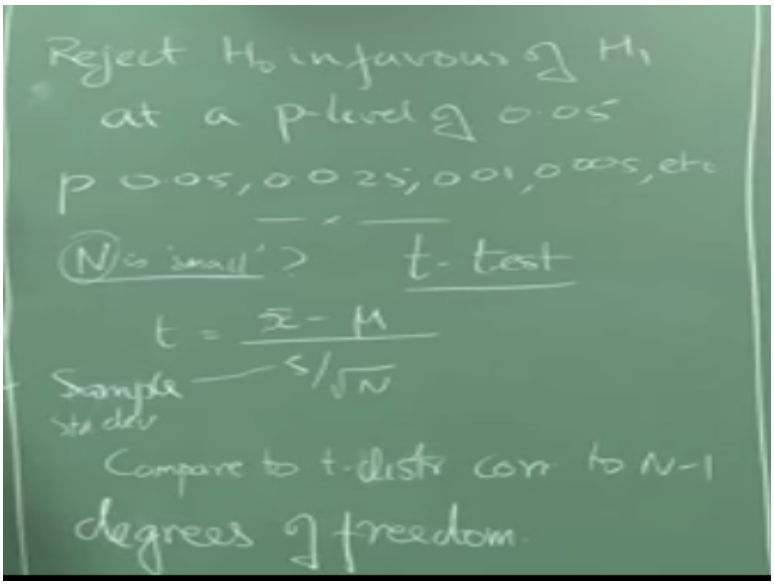
Yeah that is what I said that so you have to actually draw samples from whatever samples that you have right estimate the sample variance right so in fact if I do this right if I have this $\bar{x}$ as 2.8 okay I can estimate the sample variance of that also let and assume that variance is the variance of the population, so instead of using σ here I will use the sample variance here right and divided by √n and use that as my denominator in the Z statistic okay.

Okay this is something which have forgot to mention sorry about that in all of the things that I am talking about today right the assumption is that the means are different and I am testing for

the difference in means but I am assuming that the standard deviation is the same across the new and the old right that is why I can estimate the standard deviation on the new data and I can still use it as the population standard deviation right.

And we are assuming only that the means are different right so there are the whole class of statistical tests that you can run if you assume that variances also are significantly different and you want to estimate the variance right and this broadly fall under the class of algorithms known as anova, anova sent for analysis of variance that is what I am not going to get into anova methods just the usual case we are assuming means are different right.

(Refer Slide Time: 06:45)



So what if n is small how small is small so that goes to work 4:30 today how small is small I am just getting starting with this test will have another hours of us material how small is small less that is 10:30 okay no we are talking about the N here okay n yeah integers piece yeah right n is small so if you should think about it right it turns out that the central limit theorem works fine only if the sample sizes are reasonably large right.

Suppose my sample size to say five but my sample size is ten right then it is no longer clear that I can use the central limit theorem so the sampling distribution might not really be Gaussian it turns out that the sampling distribution is slightly different version of Gaussian they are heavier tined right there is more probability mass in the tails then you would have in the Gaussian no, so

the Gaussian is actually of a specific form right $e^{-x-\mu/\sigma^2}$ line so this is not so for the same mean and σ values this will actually be flattered okay.

So this distribution is called the T distribution or more correctly the students T distribution okay. So people know why it is called the students T distribution in a way okay. So that is a person there is a very famous statistician whose name no escapes me but he used to work in a brewery in England you know the place where they make whiskey and things like that right.

He was one of those people who was in charge of making sure that the whiskey that was being produced were of the, was of the same quality where there is not too much variance in the in the visible in the quality of the whiskey listing so there is not too much difference in the alcohol levels and things like that right and so he came up with all kinds of interesting statistical tests for figuring out known is a serious thing it is a serious application I mean infact something which will people pay you for right.

I mean for solving assignments in this class nobody is going to pay anything right but then so he was actually doing all of these things and he published serious mathematical articles based on this but if people knew that somebody from breweries publishing these articles they are not going to pay much attention to it so he wrote under the pseudonym of student right so it is called student's T-distribution.

Because the author of the paper was student his name was student a pseudonym of student that was called students T distribution right so people want to know more all of this kind of history very interesting tit-bits about history of statistics right so that is this book called the lady tasting tea so this is actually a very serious book and I recommend it to people if you are interested in knowing more of history of mathematics and stuff like that it is amazing so apparently there was this English lady who claimed that she could tell the difference.

If milk was poured into the tea or if Tea was poured into the milk okay and of course she happened to make this statement in a gathering of scientists and so on so there are a couple of statisticians who then ran one of the very first documented a case of what is known as a double-blind test okay they did not tell her what was happening right hey started giving a tea right there somewhere someone behind the screen was sitting there in some cases they were pouring milk into T some cases you are pouring tea into milk and giving it to her.

And then apparently the lady identified this correctly some X percentage of times right now the question is what she doing it by chance or what she truly able to tell the difference between milk being poured into the Tea or Tea being poured into the milk was a very valid scientific question right so they came up with significance test.

I have, go read the book right and it is actually that is this is history if this is true history right so instead of the Gaussian we use the student's t-distribution right so the thing is students T distribution is not a single distribution it is a family of distribution I just do one thing here but this is not truly just a single T distribution it is a family of distributions right one for each degree of freedom that your setup has, right.

So just like I had the Z statistic I am nothing very different right is the same thing as the set statistic the T statistics is exactly the same thing as the Z statistic except that well here I use the population standard deviation by √n here I am using the well sorry the standard since the sample standard deviation by √n right big difference right. Right this is what I was telling you so you could use the same thing in the Z statistics also right now you do not have to move to tea.

The reason you want to move to tea is if n this small right now what you do here in the z-test you compare it with the Z table right in the tea test what are you going to do tea table let us sit all right but you have to be careful about which tea table which row in the tea table that you use because there is one row for each number of samples right suppose you have n samples you have to look up the row corresponding to n-1.

If you have n samples you have n - 1 degrees of freedom okay. So that is essentially what it is so one thing about the T distribution is that it assumes so one thing about the T distribution is that it assumes that the underlying distribution from which the samples are drawn right the population distribution is normal right. So earlier we are having a sampling distribution where we did not have to worry about the underlying Sample population distribution regardless of the underlying distribution we knew the sampling distribution was normal.

But in the t distribution it assumes that the underlying distribution is normal but it turns out that in practice it is extremely robust when you can run T tests on arbitrary distributions right and still it gives you reasonable answers provided remain is not too skewed or anything right and so most

distributions that you would likely see in practice right the T the T test gives you reasonable answers okay so you can use them.

So moving on I have okay so very roughly let us look at it this way right so suppose I give you the mean okay and I give you n - 1sample you can construct 10 sample cannot you, for a given mean I give you n – 1 samples we can construct the N sample right so that is roughly that so you have only n - 1 free things that you can set in the system so but the n^{th} one will be determined so that is what it means the N – 1 degrees. That is a more formal definition of it but roughly I mean intuitively this is what the thing is so how many independent factors that you can set in the system.

IIT Madras Production

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

www.nptel.ac.in