Data Analysis and Decision Making – I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture - 18
Hypothesis Testing

Very good morning, good afternoon, good evening my dear friends. Welcome to this Data Analysis and Decision Making - I course under the NPTEL, MOOC series. And as you know this is 12 week course, total number of hours is 30, each week we have 5 classes and each class being up for half an hour. And I am Raghu Nandan Sengupta from the IME Department, IIT, Kanpur.

And obviously, this is the 18th lecture that means, we are in the 4th week. And just to just clear up in your memory that this is DADM I, so obviously there would be DADM II which will cover maximum part being operations research. This DADM I being more from statistics and multivariate statistics, and DADM III will be more from computation perspective for both DADM I and DADM II machine learning artificial neural network and all the processes, where we use a lot of computing powers to do analysis for datas analysis for decisions and all these things.

Anyway coming back to this course of DADM I, this is the 18th lecture. So, if you remember we are discussing about we started the 16th and the 17th we dealt with method of moments, we then we dealt with maximum likelihood estimation concept, and then how the point distribution for different type of parameters for a distribution could we found out. This is just time repay recapping so please have some patience. And what we wanted to know is that given a population distribution if the parameters were known our life was fine, no worries about that we could use those parametric distribution to solve our problems.

Now, if the parameters for the distribution are not known so obviously, those parameters have to be found out. And in order to be find out those parameters we have to use some methods. So, in the method of moments what we did which is the second method which I discussed what we did is that we found out the moments the first moment, second moment, third moment. First moment being here related to the mean, second moment with along with the first moment would give you some information the variance, similarly the third moment, second moment, first moment would give you some information of the Skewness and so on and so forth.

So, we found out the relationship or the moments in the parametric form consider the parameters are not known obviously, for the proper distribution function, pdf for the population, we found out some equation relationship. Then what we did do we took up a sample of size one to n, you utilize the moments of the sample with respect to the information which is given in the sample, found out the parameters using the sample statistic of the sample information or the sample characteristics. Once they were found out obviously, you have to solve the different sets of equations they may be closed form solution, not closed form solutions that is a different question, we can do use iterative methods and other things to find out their estimates.

And those estimates were basically the best estimates with respect to the parameters which you want to find out. So, if there is only one moment one parameter I will use the first moment if there are two parameters we will use the two moments first and second and so on and so forth. So, for the problem which was shown in the or the slides which were shown we consider they were k number of parameters hence we took the k moments and considering that we found out the estimates from the sample for these k parameters and proceeded.

But obviously, we when we proceeded with double check that this estimate of the parameters which were found whether they adhere to the property of unbiasness and consistency. And if you remember I have given the concept of unbiasness and consistency using a diagram, where average values being the same with respect to the population expected value was on unbiasness and consistency technically would mean that as the sample size increases the variance basically would be almost equal to the population variance or it will basically shrink and become almost 0. So, I explained why it will become 0, just hang on I will I will give you this example.

And then in the method of likelihood estimation, we found out we took up samples of size n and we found out and the parameters was obviously, not known then the realized value depending on which sample observation you are taking. The pdf or the pmf for those particular realized values the probabilities were known; only the parameters are not known. Then we want to do it to basically find out the chances that for those parameter replace value coming from the estimated from the sample we will get the maximum value of that combined probability and they being iid, we basically multiplied the overall pdfs that means, for X 1 what was the pdf, for X 2 what was the pdf so on and so forth you multi write them.

And all the unknown values in them what the parameters because now the realized values are known because once you, see for example, once you before you roll the die the X value what is the face of the die coming is unknown, but once the experiment is done we know that what is the value in front of us. Hence, we know that what is X 1, what is X 2, what is X 3. Similarly, for tossing a coin doing any experiment whatever it is.

So, given the realized values were known we basically found out the parameters by differentiating the likelihood function, but we converted into log likelihood function because I mentioned that log is in monitoring increasing function. Partially differentiated put it to 0 and when we partially differentiate we partially differentiate with respect to each and every parameters which found the population which are not known. And once we put it to 0 if the closed form solution comes, we find out the exact formula for the estimates of those parameters considering the characteristics of the information has been obtained from the sample. If not we will use some iterative methods like sif very simply, Newton Raphson method, Runge Kutta method, whatever they are to find out this estimate.

Now, when the estimates were found again we double checked the unbiasness and concern his consistency for this MLE method which is maximum likelihood method. And basically agreed upon technically what was basically the best estimate for the population parameter.

Now, later on and so considering that we found out the point estimates for those population parameters say for example, for the exponential distribution giving a as 0, then the sample mean is the best estimate for theta or lambda whatever we say. In case of a Poisson distribution the sample mean is the best estimate for theta or lambda whatever it is the parameter we want to find out.

Similarly, for the when you are basically tossing a coin, I am not talking about the best estimates per say. So, basically we will try to prove that later on if required, but generally this course is not of proofs basically of utilization and the concepts. So, there are proves that how you can find out the minimum of a distribution that means, when you have a and b between which you are drawing the uniform discrete distribution a uniform continuous distribution the a and b values which are the parameter which are not known how you want to find out and how you want to check the unbiasness and consistency. So, there are proofs for that. So, how would you find out a using a some a hat, how would you find out b using sum b hat and all this these proofs would basically come on the maximum likelihood estimation problem the general method of moments. And that will basically be the helpful background how we find out the different type of estimates and understand the point estimation concept which is under the statistical inference.

Now, the later on part which we studied is basically the inference part. The inference part was more interesting in the sense that we want to give some range and a lower control limit and upper control limit such that the actual mean value or basically standard deviation on the variance for the population would be within that range with certain probability. When I say a certain probability if it is basically 95 percent, 99 percent, 97 percent, 90 percent whatever it is we say it as a level of confidence means how confident we are.

So, 97 percent confidence level would basically mean that if we keep doing the experiment then 97 number of times out of the 100 the range would basically contain the mean value. And then we also later on said that if you want to find out this at upper control limit or the lower control limit, we have to pick up a sample of observations n in number and basically find out two different functions t 1 and t 2, that is small t 1 and small t 2 which would basically be the lower control and the upper control. And such that the difference of this lens or technically the between those bounds of t 1 and t 2 basically will have that parameter for the population with a certain level of confidence whatever it is.

Now, in the later on problems I did mention that when you are trying to basically find out the interval estimation and this will become true for the hypothesis testing also that when you are trying to find out the interval estimation problems remember always and I am I will be repeating it time and again. That. for the mean of the and and this we are only dealing with the normal distribution per say because we would not be going into other distributions, if other distributions are consider they using central limit theorem they will convert it into a normal distribution.

Now, when we are discussing about the mean, provided the standard deviation known we will always use the z distribution way. If you want to find out something to do with the mean that is in the interval estimation case, provided the standard deviation of the variance of the population is not known we will use the t distribution with the corresponding degrees of freedom. That I will that I have repeated, I will again repeat it as we do more problems. Then we go to the second moment, second moment by the word per say I am using something to do with the variance or the standard deviation for the population.

So, if you want to find out something to do with the population we use the chi square distribution, chi square distribution with the degrees of n or n minus 1 would be applicable provided we know the mean do not know the mean for the population. Then when we want to utilize and compare the chi square distribution from two population provided the mean values for both the population are known we will use the f distribution with m comma n degrees of freedom, m being the sample size for the first population, n being the sample size for the second population. Then when we use the try to find out some information related to the standard deviation or the variance of two population and provided the mean values for the first population and the mean value of the second populations are unknown, then we will use the f distribution with m minus 1 and n minus 1 degrees of freedom. This I have been repeating I am just repeating such that things become much more clearer to you.

And we solved a few problems very simple problems and showed that how the tables will be used we utilize the we have already shown the z distribution table, we showed the t distribution table, we showed the chi square distribution table we will come to the f distribution table later on. Now, when we do this hypothesis testing please have some patience.

Now, this same concept of the chi square the f distribution for the standard deviation of the variance the z and the t, f corresponding the mean value would be utilized in the same way the conceptually the same way when we do the hypothesis testing also.

Now, remember one thing that in case and this is again I am saying with repetition for the t distribution be well aware that as the sample size increases the t distribution can be replaced by a standard normal deviate and we can utilize that distribution table in order to solve our problem. Another thing we should remember that when we are trying to basically saw utilize the problem on trying to do it for the interval estimation or the hypothesis testing the problem arises that in interval estimation our main concentration is on the interval. Such that if the level of significance is 1 minus sigma between the lower control limit and the upper control limit we will consider that that 1 minus sorry alpha my apologies or if it is 1 minus alpha between the lower control limit and the upper control limit we will consider that the rest part.

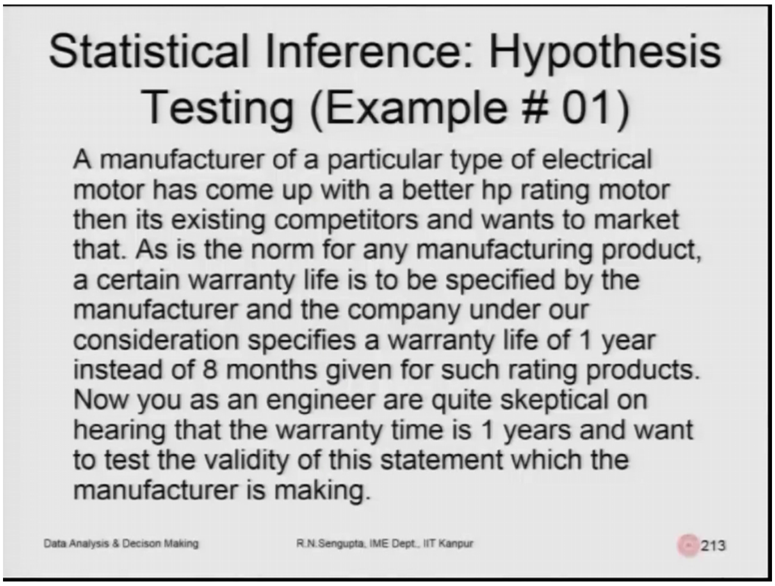
That means, if I am looking from my side the lower control limit is here at the upper control limit here then the area which is covered on to the left of the lower control limit and to the right of the lower contribute that some would basically be equal to the alpha because alpha plus 1 minus alpha will give you the total area under the curve of the distribution which would be 1.

We will see later on that for the hypothesis testing that it can be broken down the same phenomena for the interval estimation can be broken down each can be broken down into 3 parts, and why these 3 parts would be corresponding to the fact that we are interested to find out or prove some statement in the hypothesis testing where it is either less than type, greater than type and not equal to type. So, this I wanted to mention you beforehand so as we solve the problems also understand the concept they would become much clearer to you.

So, let us basically now come to the realm of hypothesis testing. So, we I did talk a long time in the introduction for the 18th lecture, but what we had done in the 16th 17th. This will become much more important as we proceed and try to solve different problems from the hypothesis testing and the multivariate statistics, and anova, manova and regression part.

So, to basically build up the environment, build up the atmosphere for the interim for the hypothesis testing which is the third part of statistical inference - first being point estimation, second been interval estimation, and third being hypothesis testing. I will discuss few very simple problems and basically build up this story. So, considers this example.

(Refer Slide Time: 14:57)



Statistical Inference: Hypothesis Testing (Example # 01)

A manufacturer of a particular type of electrical motor has come up with a better hp rating motor then its existing competitors and wants to market that. As is the norm for any manufacturing product, a certain warranty life is to be specified by the manufacturer and the company under our consideration specifies a warranty life of 1 year instead of 8 months given for such rating products. Now you as an engineer are quite skeptical on hearing that the warranty time is 1 years and want to test the validity of this statement which the manufacturer is making.

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 213

A manufacturer of a particular type of electrical motor has come up with a better hp horsepower rating motor then its existing competitors and wants to market that product into the market in the market. As is the norm for any manufacturing product, a certain warranty life or a guarantee life or in any for is to be specified by the manufacturer and the company under our consideration specifies a warranty life or a guarantee life of 1 year.

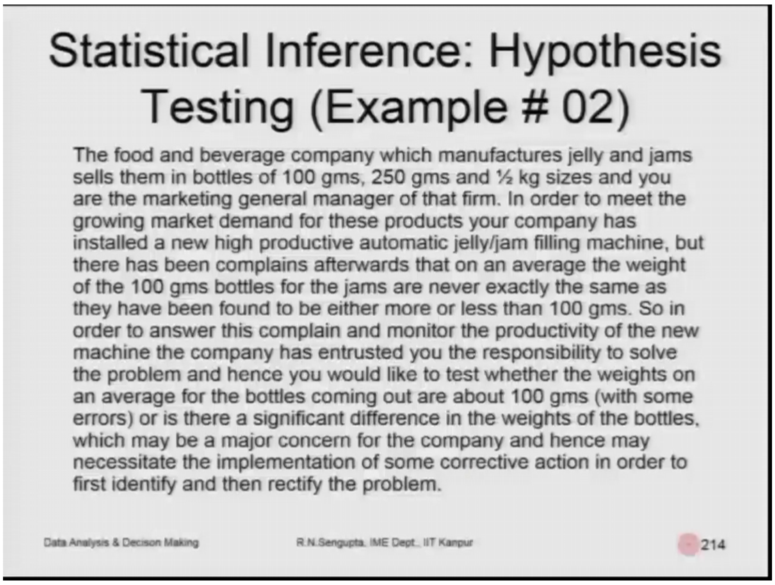
So, new product is being floated warranty life and guarantee life is 1 year which means that if the product fails before 1 year then the company will replace it and in the product fails after 1 year and if there is no say for is example EMI that means, equal monthly installment being paid for the services. Then obviously, the person has to pay from his or her pocket the customer in order to basically repair the spots or replace the machine or the motor.

Now, this 1 year is with respect to the market. So, you see what it what the line reads I will read it. So, it says that as is the norm for any manufacturing product a certain warranty life is to be specified by the manufacturer and the company under our consideration specifies a warranty life of 1 year, instead of 8 months given for such rating products.

Now, you as an engineer are quite skeptical on hearing that the warranty life is 1 year. So, you think that the warranty life is should be for the new products also which have been floated in the market the same type of products are available in the market it should be in and around 8 that 1 year, 1 warranty life is skeptical and you want to disprove it. So, on hearing that the warranty life is one year you are skeptical and want to test the validity of this statement which the manufacturer has claimed. So, what you will do? You want to basically disprove the statement, you will take some set of machines test their warranty life and depending on the outcome which you have it will basically either agree or disagree with the statement which has been made by the manufacturer.

So, this is the general story which we will be building up time and again for different examples for the hypothesis testing problems.

(Refer Slide Time: 17:14)



Statistical Inference: Hypothesis Testing (Example # 02)

The food and beverage company which manufactures jelly and jams sells them in bottles of 100 gms, 250 gms and ½ kg sizes and you are the marketing general manager of that firm. In order to meet the growing market demand for these products your company has installed a new high productive automatic jelly/jam filling machine, but there has been complains afterwards that on an average the weight of the 100 gms bottles for the jams are never exactly the same as they have been found to be either more or less than 100 gms. So in order to answer this complain and monitor the productivity of the new machine the company has entrusted you the responsibility to solve the problem and hence you would like to test whether the weights on an average for the bottles coming out are about 100 gms (with some errors) or is there a significant difference in the weights of the bottles, which may be a major concern for the company and hence may necessitate the implementation of some corrective action in order to first identify and then rectify the problem.

Data Analysis & Decision Making R.N.Sengupta, IME Dept., IIT Kanpur 214

Let us consider the second example the food and beverages company with manufactures jellies and jams sells them in bottles of 100 grams, 250 grams and half kg sizes and you are the marketing general manager of that firm. So, in order to meet the growing market demand of for these products your company has installed a new high productive automatic jelly jam filling machine, but there has been complaints afterwards that on an average the weight of the 100 gram bottles for the jams are never exactly same as they have been found to be either more or less than 100.

So, many people are complaining if they are getting less that means, the distributor he say he or she is saying that the bottles are either less than 100 grams or more than 100 grams. That means, so which means the people who will get less than 100 grams would complain and who would basically get more than 100 grams would not be complain because they are making an extra buy. So, consider that what can extra buy.

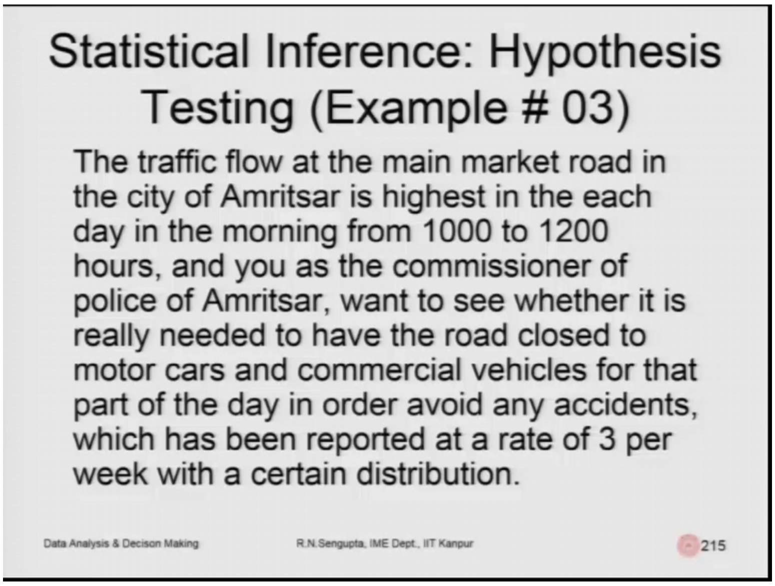
So, in order to answer his this complain and monitor the productivity of the new machine the company is interested you the responsibility to solve the problem and hence you would like to test whether the weights on an average for the bottles coming out are about 100 grams with some errors. Obviously, 100 plus minus some errors would be there or is there a significant difference in the weights of the bottles. So that means, if there is significant weights difference in the weights of the bottle you will definitely agree with the big complain being raised that the machine which is busy filling up has some problem and you will try to basically rectify that.

And in case if you find out that the weights are actually almost equal 100 grams plus minus some error which should be within the level of so called tolerances then you will say that they are the components which are being raised are not factual. And obviously, will give the data from your study in order to disprove the statement which has been raised up by the retailer or the distributor.

So, let me come continue reading it. Hence you would like to test whether the weights on the average for the bottles coming out are about 100 grams with some errors or is there a significant difference in the weights of the bottles, which may be a major concern for the company hence may necessiate the necessitate the implementation of some correct corrective action in order to first identify and then rectify the problem.

So obviously, you want to basically pick up some observation try to verify the fact which has been said. If the fact is true then obviously, take corrective actions at in the factory for the new machines or whatever it is happening, and if it is not true basically you have to basically substantiate and try to disprove the statement which is being made by the retailers and the distributors, but that should be scientific that is why you want to basically do that hypothesis testing test.

(Refer Slide Time: 20:12)



Statistical Inference: Hypothesis Testing (Example # 03)

The traffic flow at the main market road in the city of Amritsar is highest in the each day in the morning from 1000 to 1200 hours, and you as the commissioner of police of Amritsar, want to see whether it is really needed to have the road closed to motor cars and commercial vehicles for that part of the day in order avoid any accidents, which has been reported at a rate of 3 per week with a certain distribution.

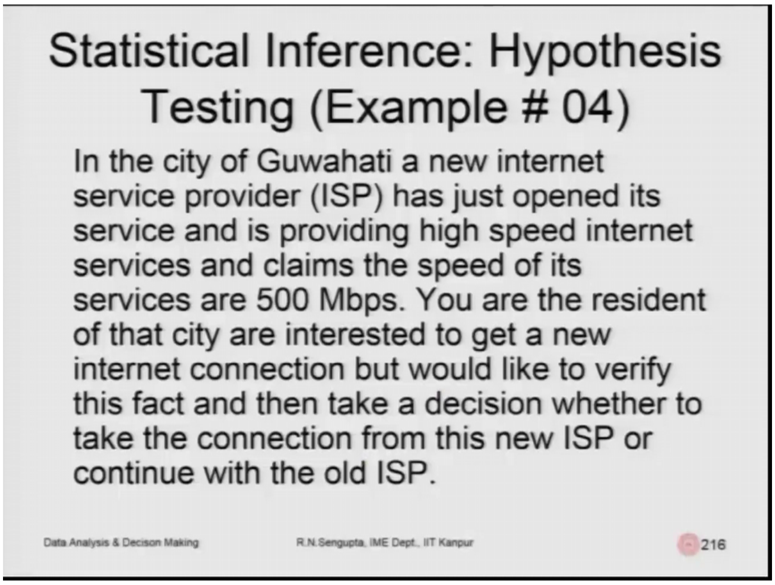
Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 215

Let us consider the third example. The traffic flow at the main market road in the city of Amritsar is highest in each day in the morning from 10 to 12 hours, and you as the commissioner of police of Amritsar, want to see whether it is really needed to have the road closed to motor cars and commercial vehicles for that part of the day that means, from 10 to 12 hours in order to avoid any accident or any major problem. And the number of accidents which has been reported is at a rate of 3 per week with a certain distribution.

So, if say for example, that problem has been say and mentioned by the newspapers by the by your staff or by the organizations or by the people or by the citizens you want to basically check that that is the number is 3 obviously, that is very high and definitely no such accidents would should happen. So, you basically try to check that whether the flow of the traffic between 10 to 12 each day is such that it basically corroborates the fact that there are 3 accidents per week.

So, if you find out that from the information which you have collected that is true then obviously, we will take action in order to make the number of accidents 0. Hence close the road as it has been said in the problem cross the road for the traffic from 10 to 12. If that is not true basically then you have to basically find out other means and try to find out why these accidents had happened, maybe it is due to the negligence of the drivers, maybe it is due to the negligence of the passengers something must have happened and if it should take corrective actions on those accounts and not close the traffic between 10 to 12 as it has been stated by either the police as I said or by the citizens or by the people who are staying in that locality.

(Refer Slide Time: 22:00)



Statistical Inference: Hypothesis Testing (Example # 04)

In the city of Guwahati a new internet service provider (ISP) has just opened its service and is providing high speed internet services and claims the speed of its services are 500 Mbps. You are the resident of that city are interested to get a new internet connection but would like to verify this fact and then take a decision whether to take the connection from this new ISP or continue with the old ISP.

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 216

Let us go to the 4th example. So, the example is like this. In the city of Guwahati a new internet service provider ISP, has just opened its services and its providing high speed internet services and claimed the speed of his services are 500 mbps; so, these bytes per second. So, you want to basically find out that whether the speed is 500 mbps or not and that is what is the claim being made by the new service provider.

You are the residents of the city and you are interested to get a new internet connection, but would like to verify this fact and then take a decision whether to take the connection from this new sp or continuing with your old internet service provider. So, person is saying the new person is saying 500 mbps, you will basically take some information’s or the speed from your friends or people who are utilizing, then check the for the statement which is being made and if you are satisfied you will basically switch over from the old to the new. If it is not obviously, we will continue with the old. That means the statement which is being made by the new service provider that it is 500 mbps has not been corroborated from the fact of the experiment which you have done.

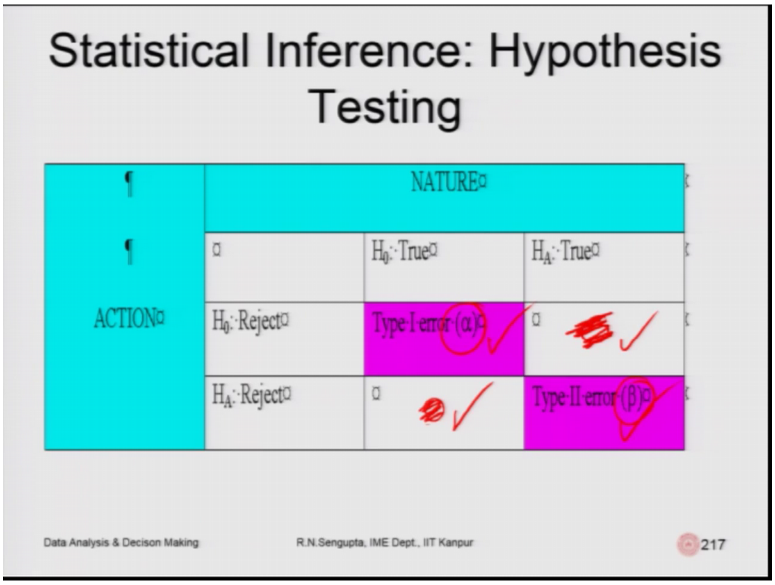
Now, in all these 4 examples the main crux of the problem is that you have an issue some information is provided with some quantitative values and you want to basically prove or disprove it. Now, what you want to do is that you want to take a set as the separate set of observations for each of the experiment conduct some statistical tests, and then prove or disprove from the with bit based on the fact that some statement has been earlier provided to you. Now, remember one thing when you are trying to prove and disprove it is not that you are trying to basically find out what is the ultimate truth, the fact is that you are trying to basically find out that whether the experiment which you have done basically collaborate that statement either agrees or disagrees.

It may be possible the actual statement which has been made by that person is not true, but you do not know you are assuming that to be true or assuming that to be not true. So, when you basically agree with that statement which is a surely not truth which you do not know then in that case you are agreeing to the non-truth statement which you are passing a judgments based on the experiment which you have done.

On the other hand if it is true and if you agreeing with that you are unknowingly agreeing to the actual truth statement, so that is two parts. That means, it is wrong and you are agreeing with the wrong it is right and you are agreeing with the right. So, you do not know what is actual truth in reality. The other two cases are the statement is wrong and you are going against it technically that means, you are agreeing with actual what is what it should be not on the fact that what has been provided to you and you are trying to basically disprove or prove.

The 4th statement would be basically a statement has been made and that is actually true what in what it may happen is that based on the experiment which you have done you disagree with it. So, in all these 4 cases you are trying to basically prove or disprove the statement which has been provided to you do not have any information what the actuality what reality is. And these concepts are what I told you in the last 5 minutes would become very apparent as you basically understand the whole problem from a very simple table format. What is the table format? I am going to come to that later.

(Refer Slide Time: 25:58)



Statistical Inference: Hypothesis Testing

¶	NATURE¤		
¶	¤	H_0: True¤	H_A: True¤
ACTION¤	H_0: Reject¤	Type I error (α)¤ ✓	¤ ✓
	H_A: Reject¤	¤ ✓	Type II error (β)¤ ✓

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 217

So, this is the table format. Now, pay attention on the blue part which is the action and the nature and especially on the pink part which is of type 1 and type 2 error. So, consider mother nature which you do not know which is an information which is being provided from outside, and consider the fact that there are two statements one is the null hypothesis which is H naught which you want to disprove this null hypothesis is stated by the person who gives you the information and you want to basically disprove that.

Another one obviously, your part would be the alternative hypothesis which you are trying to grieve such that using and alternate hypothesis you will basically disprove the person who has made the null hypothesis. Now, in this case there can be 4 different combinations, and I will basically state those combinations first and then come back to the statement which I made that they can be it the fact can be that actually what the information being given by the other person untrue and you agree with that untruths statement. Second one mean, the actual statement which is being provided by the other person is true and it agree with the truth statement. The third information is that the information being provided by the other person is untrue and you will disagree with that. And the fourth one is that the information being provided by the other person is true as per the information of with mother nature and you disagree with that. So, these are the 4 information.

So that means, actual in reality by god by nature that is true and the person says he or she does not know obviously, the person says it is true and you either agree or disagree with it. So, there are two information's. Other one is that the actual Mother Nature whatever it is there the person disagrees with it, again there are two phenomenons that means, you agree or disagree with this. So that means, in all these 4 combinations which are there will basically come out from the table.

So, consider the H naught statement is true and you basically reject the H A which is alternative one. So obviously you agree with H naught, so there is no problem. So, this is the blank which is here. I would not mark anything this is the one which is here this is the red one. The other statement is that H naught is true and you reject H A; H or H A is true and you reject H naught which means that you are in this cell.

So, this cell is the fact truth are not truth you are agreeing with that that statement which is made so that means, there are so no errors there. But the other two part is that if H naught is true and H naught is rejected then you have a type 1 error which is given by alpha. This alpha here would have some information with the one minus alpha of the level of confidence remember that. And the other part is that when you have H naught, H A as true and you reject basically H A that means, you are accepting H naught there is the type 2 error which is known as beta. So, what we are more concerned is about alpha and beta error and we will try to utilize see that how from the problem perspective these can be done in a very simple statistical format such that given the information you will agree or disagree with the statement and then basically the support H naught or not support H naught so that you will try to solve the problems accordingly.

So, with this I will end this 18th lecture and continue our discussion in the 19th and 20th more of with respect to the hypothesis testing and so on and so forth.

Have a nice day and thank you very much.