

Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology - Kharagpur

Lecture – 31
Statistical Inference (Part 8)

Hi guys. So, basically to the lecture, we will continue again with statistical inference.

(Refer Slide Time: 00:35)



And in fact, this is the last topic on comprehension on two population inferences on comprehension of two populations. In the first lecture, when we try to compare 2 different populations, first lecture, we have seen, we try to find out the inference of mean and two populations for that as I told you, there can be 2 different types of samples the way we collected data one is independent sample one is dependent samples.

So, remember what we have discussed? We have discussed for independent samples, we have discussed for independent sample, we have discussed 3 different cases when the samples are independent, but whether the variances are known or not known, if it is not known what came in what may be the case there are 2 cases again if it is not known, you can we assume it equal if we assume it equal it was fine we can we could use t distribution if it is not equal then there are some other techniques.

So, that was for independent samples. Now, what we will see is that if the samples are dependent like I have given the example, when we have discussed on two population remember I have given the example on if I remember correctly, we have given the example of migraine medicine blue pill and red pill. So, what is dependent samples, the same group of people same set of people are taking both the tablets it is that it is among the same set of people some people are taking blue pill first and red pill second other set of people is taking red pill first blue pill second.

But the people are the same. So, it is the same when under this condition, when the samples are the same on the same samples are what to say modified for 2 different purpose which we call it as a factors basically. So, this type of data collection we call it as dependent samples. So, we will try to infer do inference on mean for dependent samples of two populations.

(Refer Slide Time: 02:25)

The slide is titled "Dependent Samples" in red text. Below the title, a green box contains the text: "Example: Consider two different methods for conducting a experiment to find the effect of a special diet on weight gains".

There are two columns of text:

- Left Column (Independent Samples):** A blue box contains the text: "Randomly divide a sample of subjects into two groups and give the special diet to one of these groups and then compare the weights of the individuals from these two groups." Below this, an orange box states: "This method of data collection is called independent samples".
- Right Column (Dependent Samples):** A blue box contains the text: "Weigh a random sample of individuals before they go on the diet and then weigh the same individuals after they have been subjected to the diet." Below this, an orange box states: "This method of data collection is called dependent samples".

The slide also features a small atom icon on the right side and a video inset of a woman in the bottom right corner. At the bottom, there are logos for IIT Kharagpur and the name "Monalisa Sarma IIT KHARAGPUR".

So, similarly, now, I have one more example like one is migrant tablet what we have seen one more example is that consider 2 different methods for conducting an experiment to find the effect of a special diet on weight gains. Some company has developed some particular what to say some special diets like we have we do not have special diets for weight gain, but we have special diet for weight loss. Is not it? And similarly, say assume that some company has developed certain diets for weight gain.

So, we want to find out whether this diet what they have found out for weight gain it is really effective or not. So, how we can find out? One way is that, again let us people who for me, my population is that people who wants to gain weight, that is my whole population from this population. I will take a sample one sample of students, people, students or whatever it is I will take one sample say maybe around say 20 people and to them I have given the special diet for 10 days.

It is this company's what to say claiming that after 10 days people will gain weight. So, I have given them this special diet and for 10 days and again I picked another sample for whom I did not give this diet and this people are also having their normal diet. So, I will try to find out the weight gain between these 2 groups of people, whether this will show whether the special diet is effective or not. But question is can I do this experiment in this way.

Like same is the example for the migraine medicine also, if I do the example, if I take one group of person and I have given them blue tablet and another group of person I have given the red tablet. So, why this way of taking sample is sampling is not good for this type of x and y suppose in this take consider this migraine tablets. Suppose in one group that is Group A suppose there is some people who are suffering from some other disease, which may have an effect on the tablet.

And that is what might be their efficacy I am not getting the correct efficacy or maybe there are some other people who are already taking some other medicines and under top of that, when they are taking this blue pill, maybe the effect is becoming more prominent on the other hand is Group B group B people very healthy people maybe so they do not have any ailments so whenever they are taking a tablet the effect is slowing.

So, this side Group A maybe somehow when I am picking randomly, there may be chances that here I have picked some people who are already suffering from ailments there are more number of people who are suffering from some other elements other than migraine because people have many diseases is not it? There are all of us at work no one can say I am totally healthy, some of the other things are there. So, this set of people who may be comparatively more, healthier than this set of people so the effect what the group A sold on the tablet.

That may be different from group B because of the maybe physiological structure maybe the state of the health at that moment there are different factors that is why this way of taking sample is not effective for this particular experiment, similarly for weight gain, if we try to see the weight gain now says this group of people one group of people which I have given a special diet, and when I am trying to find out the I have given a special diet for today's group of people.

And after this given the special diet I am trying to find out the mean weight of these people after 10 days and 2 other group I have given the normal diet and then trying to find out a mean weight of this people see here while maybe while I have picked this group A group I have you been special diet group B I give the normal diet, so now this group A people while picking maybe the group A people as competitively more healthier than group B, competitively, it is more healthier maybe.

So, in whether this weight gain has any effect or not, if the people are more, healthier, definitely my mean weight will be much more than this group B, is not it? So, there maybe even if there is one outlier with who is quite healthy, that will change the result. So, this way, for this type of experiment if I do independent study, then definitely I will not get good results. Basically, what I want to say is that when we study 2 different populations, the within population difference should not overwhelm what to say the record is difference what we want to see.

The already existing difference should not overwhelm the difference what we want to see here we want to see the difference of weight gain, because of this special weight, because of this special diet, but there may be inherent difference in the both the sample itself or there may be inherent difference in the within the sample itself. If we try to see and find out a variance, suppose this remember when I try to compare the 2 different populations, comparing the 2 weight of 2 different populations.

In case of independent when the variance is not known, but when I use t distribution I use the pooled variance. So, what I have done? I have taken out the individual variants of both the sample and then I found out the pooled variance, when I am trying to find out the individual

variances of one sample then what happens the people who might have taken for one sample there their weight may also be very much varying with one maybe 41, maybe 45, another maybe 51, another maybe 60 that that variance within sample difference maybe also quite high, is not it?

So, this will overwhelm the difference what we want to see. So, in such cases, taking independent sample is not at all a solution. In such case we should already always go for a dependent samples dependent sample means here in this case, when I want to find out the efficacy of the special diet, let us see what we will do. So, the first case what I have already discussed randomly divide a sample of subject into 2 groups and give a special diet to one of these groups and then compare the weight of integers from these 2 groups.

After some days maybe this is one way this is called independent sample which is not a very viable option in this case. This method of data collection is called independent sample. Second is way a random sample of individuals before they go on a diet I have taken a random sample before they go on that special diet for septic and a weight of this people say I have taken the weight of each and found out the mean so I found the mean maybe \bar{x}_1 then I have subjected to this damn to the special diet for after 10 days then I have taken a weight.

I say I got this \bar{x}_2 . Now if I try to find out $\bar{x}_1 - \bar{x}_2$ that makes sense, because the same set of people, they the same set of people they may be they may have some other elements. They may be taking some other medicines while they are introduce to this special medicine. So, whatever effect they will have it here only is not it? So, here my the variance within the sample will not affect the results what I want to see.

But in the first case, in this case, the variance within the samples or variance between the samples will affect the results what I want to see will overwrite the results what I want to see. So, this data independent sample is not a feasible solution not a viable solution for this type of test. Similarly, the example what I have given remember when this we are trying to find out the efficacy of 2 types of fertilizer in a single land so, here I have diagonally.

Suppose it is a land is a very used land and suppose I have divided the land diagonally into 2 parts in one part I have to use one fertilizer and another part I have used another one fertilizer then what happens what if I try to use your independent sample suppose in one part of the land suppose there are lots of trees side by there are lots of trees. So, because of the trees whatever; suppose it is not getting much sunlight and because of if there are big trees are there then soil quality also gets affected.

And because of the roots of the trees and the other part it is getting proper sunlight and there are no soil quality is also good if I considered as an independent sample this if I take it, it becomes an independent sample then I will not get the correct results of the efficacy of the 2 types of fertilizer. So, what I in that case what to get a good if the same characteristics suppose there is no tree side by side whatever if it is the equal amount of sunshine the whole land is getting then of course, I can consider independent samples.

Otherwise, if this case is not same, then we cannot go for independent sample then we will have to go for dependent sample dependent sample means meaning some for one season we will use one fertilizer then for the next season maybe we will use next type of different types of fertilizers and then accordingly we will see so, that becomes the dependent sample.

(Refer Slide Time: 12:16)

Dependent Samples

Independent Samples vs Dependent Samples

Independent samples	Dependent samples
For independent samples, the difference in weights among individuals in each sample is probably larger than those induced by the special diet.	For dependent samples, the <u>individuals' differences in weight before and after the special diet are then a more precise indicator of the effect of the diet.</u>

Manalisa Sarma
IIT KHARAGPUR

So, independent samples the difference in weight among the individual in each sample is probably larger than those induced were a special diet for the case which I already mentioned difference in weights among individual in each samples is probably larger than those in us that a special diet difference in weight among the sample is only larger than this way diet is having an effect. So, it will model results.

Dependent samples that for dependent samples the individual differences in weight before and after the special diet are more a more precise indicator of the effect of the diet there is a more precise indicator. So, in this sort of example, it is always better we go for dependent samples.

(Refer Slide Time: 13:02)

Dependent Samples

- The two sets of weights from dependent samples are no longer independent, since the same individuals belong to both.
- For two populations, such dependent samples are called "paired samples" because the analysis will be based on the differences between pairs of observed values.
- This procedure can be used in almost any context in which the data can physically be paired.

Manalika Sarma
IIT KHARAGPUR

So, the 2 sets of weights from dependent samples are no longer independent since the same individual belongs to both, both the test border is taking the special diet also same individual that is not taking the diet same individual we are doing. Here basically we are weighing we are weighing a person before going for the diet and we are being the person after taking diet and again for the migrant tablet also we are for each people we are giving both the tablets.

Maybe the one set of people is taking one tablet before another set of people is taking the other tablet before but it is the same set of individuals. So, they are called dependent as there no longer independent since the same individual belongs to both it is not only applicable for individual case as I told you give you the example of the land for trying to find out the efficacy of the

fertilizer that is applicable for many cases many applications for two populations as dependent samples are called paired samples.

Because the analysis will be based on difference between pairs of observed values so, we will do analyses based on the difference between this pair so that is why the sample is called the paired sample. This procedure can be used almost in any context in which the data can be physically be paired. So, when we try to compare 2 different populations, you may tell that this is why not use always; use the dependent samples why go for independent samples? Because independent samples the within sample variance is very negligible, is not it?

There is no within sample variances because we are not I should not say negligible that is because there is no variance we have within sample variance because we are taking the same set of people for the different experiment. Then the always it will be people met the human think whether survey this is better to go for this type of data collection, we will see why we should not go?

(Refer Slide Time: 15:07)

Dependent Samples: Inference on difference between Means

Inferences on the difference in means of two populations based on paired samples use as data the simple differences between paired values.

- For example, in the diet study the observed value for each individual is obtained by subtracting the after weight from the before weight.
- The result becomes a single sample of differences
- The result can be analyzed in exactly the same way as any single sample experiment
- Thus the basic statistic is $t = \frac{\bar{d} - d_0}{s_d / \sqrt{n}}$
- The t statistic is usually called the "paired t statistic."

Monalisa Sarma
IIT BHARATPUR

So, inference on the difference in means for of two population based on a paired sample uses data the sample difference between the paired values, so, what we do what will afford us to find the inference what we will do? What data we will use? The data that will use this one single set

of data what is that single set of data the difference between 2 values the difference of weight between before going to diet and after the diet.

So, when we trying to compare two population always we see till now, we saw we got 2 values 2 data's for x_1, x_2 are $\sigma_1 \sigma_2$, but here, when we are trying to compare 2 different population and when we are using dependent sample, we will get just one set of data, what is that set of data that is the difference between these 2 samples, difference of value between these 2 samples and we got just one set of data and that means, as if we are trying to infer about a single population.

So, whatever we have used for single populations same thing, same way we can use it here. So, the result becomes a single sample of differences, the result can be analyzed in exactly the same way as in a single sample experiment. So, basically, we will use the t statistics. So, the t statistic is what is the difference? The difference between the 2 samples as I told $\bar{d} - \lambda_0$ whatever it is whatever we have what to say hypothesis.

And this is the as this square is the variance of this difference what will be definitely there will be one sample size only one sample is not it? So, s^2 / n so, this is my t value same as what we have done for single x for single population. This t statistic is called paired t statistic like previously what we have seen that was pool t statistic this t statistics is called paired t statistic and the test we do is called a paired t test.

(Refer Slide Time: 17:02)

Example

Problem

A taxi company manager is trying to decide whether the use of radial tires instead of regular belted tires improves fuel economy. Twelve cars were equipped with radial tires and driven over a prescribed test course. Without changing drivers, the same cars were then equipped with regular belted tires and driven once again over the test course. The gasoline consumption, in kilometers per liter, was recorded as follows:

Car	Kilometers per Liter	
	Radial Tires	Belted Tires
1	4.2	4.1
2	4.7	4.9
3	6.6	6.2
4	7.0	6.9
5	6.7	6.8
6	4.5	4.4
7	5.7	5.7
8	6.0	5.8
9	7.4	6.9
10	4.9	4.7
11	6.1	6.0
12	5.2	4.9

So, we learned 2 different t test one is pool t test and another is called pair t test, pool t test is sometimes in formula it is also called simple t test, and this is the paired t test like you will see an example which will make things more clear. So, a taxi company manager is trying to decide whether the use of radial tires instead of regular belted tires improves fuel economy, there are 2 different types of tires radial tires and regular belted tires.

A company taxi company manager is trying to decide which use of which will improve the fuel economy. So, to find out the fuel economy definitely if you take some Maruti car and suppose Maruti 800 and if he takes a Verner, so, he definitely he cannot compare the fuel economy suppose in Maruti 800 here use radial tire in other one Verner use a regular belted tires definitely by that he cannot come to the conclusion of fuel economy forget about Maruti and Verner even same car also, we will we may get different mileage.

Because of the different condition of the car at that moment so, here definitely independent sample is totally out of question, we will have to look for the dependent samples. So, 12 cars were equipped with radial tires and driven over prescribed test course taken a sample size of 12 and equipped with radial tires and driven over prescribed test course, without changing drivers, even the driver also have an effect how you are driving the car. If you are using too much of brakes, then what happens our fuel efficiency goes down is not it?

So, each driver has a specific style of walking. So that is why we are not changing the driver also if we change the driver that that will also happen that will also given variance to the data. So, this variance will overwhelm the required difference what we want to see. So, here so without changing drivers the same cars within equipped with regular belted tires and driven once again over the test course the gasoline consumption in kilometers per litre was recorded.

Petrol consumption basically and kilometers per hour record so this is recorded same car same driver, first this one is run and this one is run. So, we want to find out whether which tire is better. I will if we try again better fuel economy, fuel or what to say economy. So, what we will do is basically now we will have a separate table for that, this is basically the difference 4.2 – 4.1 whatever difference we get, we will find out the difference. So, we will get one set of data. From this one set of data we can find out whatever we need to find out.

(Refer Slide Time: 19:50)

Example

Problem	Question
A taxi company manager is trying to decide whether the use of radial tires instead of regular belted tires improves fuel economy. Twelve cars were equipped with radial tires and driven over a prescribed test course. Without changing drivers, the same cars were then equipped with regular belted tires and driven once again over the test course. The gasoline consumption, in kilometers per liter, was recorded as follows:	Can we conclude that cars equipped with radial tires give better fuel economy than those equipped with belted tires? Assume the populations to be normally distributed. Use a P-value in your conclusion.

Monalisa Sarmah
IIT KHARAGPUR


So, question is, can we conclude that cars equipped with radial tires give better fuel economy than those equipped with belted tires assume the population to be normally distributed use a p value in your conclusion.

(Refer Slide Time: 20:03)

Example 1: Solution

We need to compute the difference of Gasoline consumption when the belted and radial tires are used for all the cars.

A taxi company manager is trying to decide whether the use of radial tires instead of regular belted tires improves fuel economy. Twelve cars were equipped with radial tires and driven over a prescribed test course. Without changing drivers, the same cars were then equipped with regular belted tires and driven once again over the test course. The gasoline consumption is given in Table. Can we conclude that cars equipped with radial tires give better fuel economy than those equipped with belted tires? Assume the populations to be normally distributed. Use a P-value in your conclusion.



What we need to compute? We need to compute the difference of petrol consumption when the belt and radial tires were used.

(Refer Slide Time: 20:11)

Example 1: Solution

We need to compute the difference of Gasoline consumption when the belted and radial tires are used for all the cars.

From table, we can get,


$\bar{d} = 0.1417, s_d = 0.198$


$t = \frac{0.1417}{\frac{0.198}{\sqrt{12}}} = 2.48$

The p-value is $0.015 < p\text{-value} < 0.02$, with 11 degrees of freedom.

A taxi company manager is trying to decide whether the use of radial tires instead of regular belted tires improves fuel economy. Twelve cars were equipped with radial tires and driven over a prescribed test course. Without changing drivers, the same cars were then equipped with regular belted tires and driven once again over the test course. The gasoline consumption is given in Table. Can we conclude that cars equipped with radial tires give better fuel economy than those equipped with belted tires? Assume the populations to be normally distributed. Use a P-value in your conclusion.

Handwritten notes: $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 > \mu_2$, $n_1 = n_2 = 12$





So, from the table we found the difference we found the s d that is the standard deviation of the sample than the we found what is the t value t value, how do we get calculate same technique we can calculate the t value t value is 2.48. So, for corresponding to t value 2.48 what is my p value? I can see in the table for the 11 degrees of freedom and my p value is p value lies within this range. So, basically p value is lying between this range.

So, accordingly, whatever significance level you want to consider, if you are very much concerned about type one type of error, then definitely this 0.01 And 0.02 is a very less significant level then we will what to say reject the null hypothesis here the null hypothesis I did not so, specifically what is the null hypothesis What is a random hypothesis? What will be the null hypothesis can you tell me yes see the equation while forming the hypothesis always try to see the question.

Can we conclude that cars equipped with radial tires give better fuel economy than those which equipped with belted tires that is what is my null hypothesis here my null hypothesis is the fuel economy of both the cars the same that means my $\mu_1 = \mu_2$ and that my this is my null hypothesis basically I can say $\mu_1 - \mu_2 = 0$ or $\mu_1 = \mu_2$ my alternate hypothesis is μ_1 greater than μ_2 that is one tail hypothesis because I am interested in finding out it can be radial tires gives better fuel economy.

I am interested in finding out if it gets better I am not interested in finding out does the fuel economy differ in both occurs both the tires type of tires as a field is a difference in fuel economy in both the type of tires, I am not interested in finding out that so it is definitely not a 2 tailed test I am interested in finding out does it give better fuel economy. So, this is my alternate null hypothesis is $\mu_1 = \mu_2$ my null hypothesis is μ_1 greater than μ_2 .

So, it is a single tail single tail whenever I got a compare corresponding to 2.48 whatever p value, what I will get is that is only the whatever probability I will get corresponding to 2.48 that is my p value for double tail I would have added it twice remember, because it is probability of in the left side and probability in the right side put together becomes a p value. So, it is so, since the p value is a very less significance level if we consider means if I draw the finger the distribution is something of this sort.

So, p value since it is 0.0 on it is very less this area this point greater means we will see definitely this area. So, it is very less that means we can reject the null hypothesis that we can accept the alternate hypotheses that we will offer μ_1 means radial tires really gives better fuel

economy than the other one what type of tire was that? Whatever it is regular belted tires, radial tires gives better fuel economy than regular belted tire because null hypothesis.

So, such a less p value definitely falls in a rejection region and we can say that means the alternate hypothesis is true that is fuel the economy's better in the case of radial tires. So, now the question is see since for paired sample, the inherent variance within the sample is not there. So, paired sample will always give good results then why not always use paired sample why should we go for independent sample always use dependent sample in whatever type of things we want to consider.

Like suppose this fertilizer case why we will divide the land into 2 parts and then make it independent sample? Why instead of doing that when one season I will use one fertilizer and the second season I will use another fertilizer. This both becomes the means why instead of dividing let us use in one simple season I will use one fertilizer in the second sense and I will use the second fertilizer that way my sample will be dependent There are many such cases how my sample can be dependent.

So, why not use the dependent sample in all the cases why use independent samples there are many reasons. One reason is that all type of population cannot be paired. There are some populations which you cannot pair it you will have to take it as independent samples, but there is some population which you can consider independent as well as dependent like the example I have given for fertilizer just trying to find out the efficacy of the fertilizer. So, now is when you can use both independent and dependent question is which one you will use?

We will use independent or dependent both has its own pros and cons. When we use paired samples, though, we within sample variance is less, but what we have we sacrifice on the degrees of freedom, sacrifice on the degrees of freedom means see since my sample size is n my degrees of freedom is $n - 1$, say my sample sizes n_1 my sample sizes n_2 means my degrees of freedom $n_1 - 1$ in case of a paired sample because I take this one sample, if my sample size is n_1 , then my degrees of freedom is $n_1 - 1$, this is the case of my dependent sample.

Because I have taken the same sample I have used it twice. But if I use independent samples, so, one sample I have taken n_1 another sample I have taken n_2 I have used t distribution of it the variances of course equal, if it is not equal, then I will use normal distribution whatever if it is if the variance is equal, one sample is where size n_1 and other sample size n_2 , when I use the here, what is my degrees of freedom?

My degrees of freedom is $n_1 + n_2 - 2$. This is much more than $n_1 - 1$. So, what happens when I am using dependent samples so listen very carefully when I am using dependent samples, I am sacrificing on the degrees of freedom I am compromising on the degrees of freedom because for the independent sample my degrees of freedoms are more because I am taking the sample size of 2 samples is not it? Adding it up I am getting a more better bigger number, but for dependent sample my sample size is just one sample minus 1 that is my degrees of freedom.

If you see a t table you just open any textbook and go to the backside of the book and you will see different tables if you see the t tables and you find out that t values for different degrees of freedom when my degrees of freedom is less my t value is more, when my degrees of freedom is less my t value is more what happens.

(Refer Slide Time: 27:38)

$$E = (k_{\alpha}) \pm \sqrt{\frac{\sigma^2}{n}}$$

$\frac{\sigma}{2} \alpha$

$\alpha \gg y$

When my t value is more remember when we talk this error of estimation what is my error of estimation? Error of estimation is $t(\alpha/2) \pm \sigma^2 / \sqrt{n}$ this is my error of estimation. So, if this

value is more what happens my precision becomes less we have explained this is not it? So, what happens so, t of $\alpha / 2$ so, here what is my t of α when my degrees of freedom is less this value becomes more you see I need any table any t table you see for the same value of α .

For the same value of α for different degrees of freedom and for same value of α say α let us take it single tail if it is single tail it is $t \alpha$. So, for same value of α 0.05 for same value of α you see for degrees of freedom say 20 another you see degrees of freedom say 9 for degrees of freedom 9 for same values of α you will get a value x for days for the same amount of α for degrees of freedom you will get a say value y when x is much greater than y .

So, when this value is more what happens my error becomes more the case my precision decreases same degree of confidence for same degree of confidence for the my precision reduces quantity the degree of confidence remains same if it is 5% means 95% my degree of confidence is 95% with the same degree of confidence remains 95% only, but here if my degrees of freedom is less my E is more that means my weight is more my precision reduces. So, that way I use it when I use the paired sample.


So, definitely if the requirement is search that way the inter sample variance will not overwhelm the what we want to study than definitely will not go for dependent sample because here we are sacrificing on the degrees of freedom, though the dependent sample has its advantage it have its own share of disadvantage as well.


(Refer Slide Time: 30:00)

Summary


Comparison between the pooled t-statistic and paired t statistic

The pooled t statistic	The paired t statistic
<ul style="list-style-type: none"> ✔ The two samples are independent. ✔ The distributions of the two populations are normal or of such a size that the central limit theorem is applicable. ✔ The variances of the two populations are considered equal. 	<ul style="list-style-type: none"> ✔ The observations are paired. ✔ The distribution of the differences is normal or of such a size that the central limit theorem is applicable.





Monalisa Sarma
IIT KHARAGPUR



So, now, we have come to the end of statistical inferences for single population and two population definitely will have to see for more than two population more than two population treatment is a bit different basically, they will be discussing ANOVA let us determine is a bit different. So, before going to that, let us summarize whatever we have studied till now it let us quickly go for the summary. So, compare in between the pooled t statistics and paired t statistics whatever we have seen in the pool t statistics we have seen the 2 samples are independent.

The distribution of the two population or normal parent population all have such a size that a central limit theorem is applicable. I have the pair t statistics the observations are paired and the distribution of the differences is normal or have such a size that a central limit theorem is applicable. See the difference here. Under in pool t statistics we consider the variance of the two populations are considered equal.

(Refer Slide Time: 31:06)

Summary

Inferences on binomial populations vs Inferences on variances

Inferences on binomial populations	Inferences on variances
<ul style="list-style-type: none"> Observations are independent. The probability of success is constant for all observations. 	<ul style="list-style-type: none"> The samples are independent. The distributions of the two populations are approximately normal.

Monalisa Sarma
IIT KHARAGPUR

Now, we have done also influences on binomial population and influence on variances also, let us see this quickly. Just a quick recap in inference on binomial population whatever what we have seen observations are independent. Here, let us go one by one, the probability of success is constant for all the observation for binomial population we have seen the probability of success is constant all observation that is why we use binomial distribution and for inferences on variants the samples are independent and the distribution of the two population are approximately normal.

If you are trying to compare two population if single population then the distribution of the population is approximately normal.

(Refer Slide Time: 31:47)

Summary

Normality of the Sampling Distribution of the Sample Mean

- The sampling distribution of the mean is reasonably close to normal.
- It was assumed for the discussion on Hypothesis Testing as well as Estimation

- The sampling distribution of the sample mean is normal, if
 - The population itself is normal
 - Or if, the sample size is large enough to satisfy the central limit theorem.

The normality of the sampling distribution of the mean is not always assured

- For relatively small samples, especially those from highly skewed distributions
- Or where the observations may be dominated by a few extreme values.

Monalisa Sarma
IIT KHARAGPUR

Then while trying to find out the sampling distribution of the mean, we have assumed that a sampling distribution of mean is reasonably close to normal. We have assumed that it was for the discussion on hypothesis testing as well as for estimation confidence interval estimation for both the cases we have assumed that the distribution of the mean is reasonably close to normal. If actually the sampling distribution is not normal, if we find the sampling distribution is not normal, because that depends on the population also.

The population is very away from the normal and if we do not take a bigger sample size, then what happens the sampling distribution may not be normal, when the sampling distribution in the normal distribution as when I discussed the distribution I have discussed see, when we are talking about normal distribution, the 2 parameters are mean and a variance is not it? When the population is not normal that mean and variance it does not remain as the variance of the desert remained a parameter of the particular distribution.

So, unnecessarily we are trying to infer something with a population is not normal and unnecessary, we are trying to find a mean and a variance makes no sense. So, the sampling distribution of the sample mean is normal, if the population itself is normal or the sample size is large enough to satisfy the central limit theorem. But however, the normality of the sampling distribution of the mean is not always assured for relatively small samples, especially those with highly skewed distribution, it is distribution of very much away from the normal.

And we have taken a very small size and sampling distribution it will not be normal, then when the sampling distribution will not be normal mean and variance makes no sense always the observation may be dominated by view few very extreme values, then in that case also sampling distribution is not normal.

(Refer Slide Time: 33:34)

Summary

If the Assumption of Normality does not hold??

- When the assumption of normality does not hold, use of methods requiring this assumption may produce misleading inferences.
 - ⇒ The significance level of a hypothesis test or the confidence level of an estimate may not be as specified by the procedure.

Example

- The use of the normal distribution for a test statistic may indicate rejection at 0.05 significance level, but due to nonfulfillment of the assumptions, true protection against making a type I error may be as high as 0.10.
- Unfortunately, we cannot know the true value of α in such cases.

Monalisa Sarma
IIT KHARAGPUR

If the assumption of normality does not hold then what because till now, whatever we have studied we have assumed that as our population is normal or in for z distribution what we have used population if the population is not normal we have taken a bigger sample size by how our t distribution chi square distribution we have assumed that the parent population is normal based on that we have done all the calculation if it is a slight away from normal.

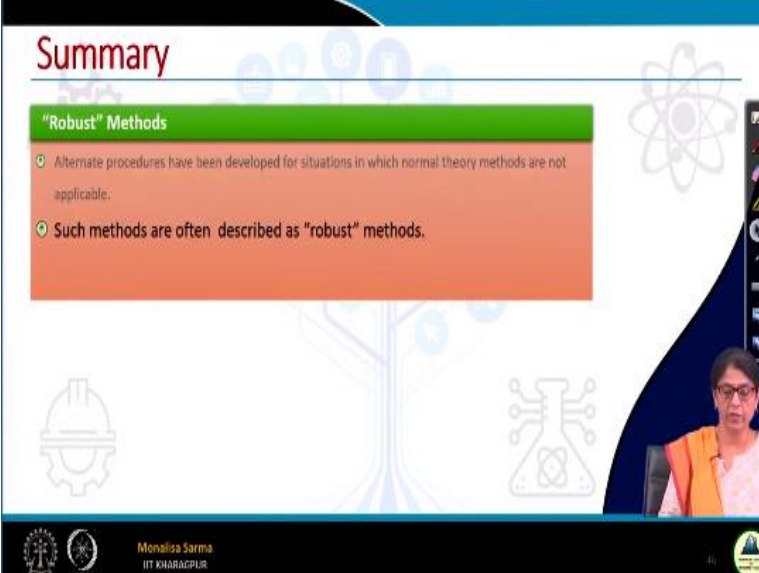
But if it is very much away from the normal it is detailed that way that way is robust slightly away from normal it still considers, but chi square and f is not very overstays. So, normally the assumption of the parent population is very necessary. So, if the normality assumption does not hold them, what then use the methods required this assumption they produce misleading inferences, then the results what we may get actually those are those may not be the correct results.

Maybe the significance level of a hypothesis test or confidence interval of an estimate may not be as specified by the procedure, we have specified the significance level based on the significance level we have calculated everything, but our population itself is not normal, that is where the sampling distribution we are not getting the normal in case of mean in case of inference and in case of variants and in case of variants only variants and mean where parents is not known.

That is where we use t distribution chi square distribution, the parent population is not normal and we have used a significance level in particular using a particular significance level we have calculated the hypothesis test we have done confidence intervals, but our parent population is not normal and then the whole test goes heavy so some example, the use of normal distribution for test statistics may indicate rejection at 0.05.

Suppose, if a consider 0.05 significant level. So, we will reject if it falls in this 0.05 significance level, but due to non fulfillment of the assumption true protection means against making a type one error maybe as high as 0.10 the distribution movement is not normal actually the significance level what we got is actually as high as 0.10. Unfortunately, we cannot know the true value of α also in that cases we may not know so, what is the true value of α because we have done everything as α everything as normal.

(Refer Slide Time: 35:48)



Summary

"Robust" Methods

- Alternate procedures have been developed for situations in which normal theory methods are not applicable.
- Such methods are often described as "robust" methods.

Monalisa Sarma
IIT KANPUR

So, in such cases we have to use some alternate methods. So, alternate procedures have been developed for situation in which normal theory methods are not applicable. Such methods are often described as robust method.

(Refer Slide Time: 36:03)

Summary

Nonparametric Methods to Develop "Robust" Methods

- However, most of these robust methods have wider confidence intervals and/or have power curves generally lower than those provided by normal theory methods when the assumption of normality is indeed satisfied.
- A widely used method for developing robust methods is Nonparametric methods.
- Nonparametric methods avoid dependence on the sampling distribution by making strictly probabilistic arguments (often referred to as distribution-free methods).

Monalisa Sarma
IIT KHARAGPUR

However, most of these robust methods have wider confidence interval and have power curves generally lower but it is reverse inverse has one con that is its confidence interval is quite wide that means precision is low, when we have a robust interval wider in turn that means our precision is low and what happens here power is also low what is remember what is power? Power is rejecting a false null hypothesis, rejecting a false null hypothesis that is power, power is $1 - \beta$ remember.


So, in case of robust is most robust method it we have a power curve also which is generally lower a widely used method for developing robust method is nonparametric methods. So, after we complete ANOVA now, next we will go to ANOVA after we complete ANOVA next we will be taking nonparametric methods nonparametric method avoid dependence on sampling distribution by making strictly probabilistic arguments.


There we make strictly probabilistic agreements about anything whatever parameter we have to make any arguments may not be mean variance whatever we have to make any arguments, we will not take help of any distributions. Just one distribution means we are taking this distribution comes parameter. So, we are not taking help of any distribution, we will just make probabilistic arguments, so often referred to as distribution free methods.

(Refer Slide Time: 37:27)

CONCLUSION

- ⊕ In this lecture,
 - ⊕ We covered the topic of inferences on mean for dependent samples of two populations
 - ⊕ We also introduced the idea of using non-parametric methods for those scenarios where the assumption of normality does not hold
 - ⊕ In next lecture, we will discuss about inferences on more than two populations.




 Monalisa Sarma
 IIT KHARAGPUR

So, in this lecture, we cover the topic of inference and mean for dependent samples of two populations. We have also introduced the idea of Bayesian nonparametric method for those scenarios, where the assumption of normally does not normality does not hold. In the next lecture, we will discuss about the inferences on more than two population basically we will be discussing ANOVA.

(Refer Slide Time: 37:54)

REFERENCES

 <p>Probability and Statistics for Engineers and Scientists (4th Ed.) By Ronald E. Walpole, Raymond S. Wilson, John G. Linton, Thomas Y. Yip Wiley, Hoboken, NJ, 2012</p>	 <p>STATISTICS 10th Edition Robert S. Yarnold Wiley</p>	 <p>Probability and Statistics, 2nd Edition R. S. Yarnold Addison-Wesley</p>	 <p>Statistical Methods 2nd Edition Freund, R.J. & Wilson, W.J. Academic Press, San Diego, 1997</p>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------


 Monalisa Sarma
 IIT KHARAGPUR

So, get up to study ANOVA from our next lecture thank you guys.