

Example–2

Problem	Solution
An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.	<p>Here, a sample of 16 bulbs is drawn from the population.</p> <p>Sample mean = Population mean = 800 hrs</p> $S.D \text{ of sample} = \frac{S.D \text{ of population}}{\sqrt{\text{sample size}}}$ $= \frac{\sigma}{\sqrt{n}}$ $= \frac{40}{\sqrt{16}}$ $= 10$ <p>$P(\text{average life of given sample} < 775)$</p> $= P(\bar{x} < 775)$ $= P(z < -2.5)$ $= 0.0062$

Monalisa Sarma
IIT KHARAGPUR

So now we will just do 1 problem, definitely, we will be doing more problems in our tutorial class also, just 1 problem so as you understand the things a bit so see the problem. An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with mean equals to 800 hours and a standard deviation of 40 hours. So what it is given an electrical firm, it manufactures light bulbs.

So and its life, it is normally distributed. It is this firm, it is claiming that the life of the bulbs is normally distributed. And it is also claiming that its mean is 800. That means my μ is 800 and standard deviation that is my σ is 40 hours. This farm is claiming mind it. So suppose I am the owner of the farm, I am claiming that the mean life is 800. And the standard deviation is 40 hours, what I am claiming may not be true.

Because I am just it may be an educated guess, basically, I am predicting something out of some past experience or whatever it is, but it is not possible for me to tell that is exactly true. Why? Because how can I find out the life of all the bulbs? I did not check off all the balls, how can I tell? So this is just a manufacturer that is claiming. Now find the probability if that is true assuming whatever he is telling is true.

Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours. So look the question other way around. I have gone to purchase some electric bulk from a manufacturer and the manufacturer has told me this my life is normally distributed. I am educated person I know about distribution and all. So, that is why he talked in terms of technical terms, he told me this, life for the bulb is normally distributed.

And its mean is 800 and standard deviation is 40. They told me that, but I am not such a gullible person that I will believe everything I wanted to check what did I do? I took a sample of 16 bulbs and I have used some technique to find out the life of those bulbs, immediately I could not find definitely took time for me before purchasing a whole lot. So, I took a 16 bulb and from that I found try to find out the life of those bulbs the life of those bulbs, what I got the average life of those bulbs I got 775 hours.

Now, from this, this is what from the sample what a good that is correct value, is not it? I have checked it I have checked it and I found that it is 775 hours. So now assuming that what the manufacturer have claimed that mean is 800 and standard deviation is 40. Is it really possible for me to get an average level of 775 hours that I will see if it is possible? If assuming that is true, if it is possible from a sample to get an average level of 775 hour if it is true.

Then what the manufacturer is claiming I can tell that is true. If it is not possible, then I can tell that what the manufacturer is claiming that is not true. H is trying to fool me. Now how it is true or not true how can I tell? When I will, find out the probability of this having a life of less than 775 hours. If this probability is very, very small, then that means the manufacturer is going to fool me that is not true.

If this probability is quite high, because while doing the calculus I am assuming that whatever he told me is correct with that assumption when I found that this sample having an average life of 775 hours, if it is probability is quite good as a good quality then. The manufacturers his claim is correct. So, how we will do that? So, that means, we will be using the sampling distribution of mean, is not it?

So, sampling distribution of mean for that using sampling distribution I mean, what will be the mean of the sample distribution population mean that is 800 what will be the standard deviation of the population mean because I need to find out the probability of less than 775 hours that is our sample is not it? I need to find a probability of less than 700 from a sample. So, for to find a probability; I need to probability distribution what probability distribution?

Sampling distribution of mean. So, now, when we talk about normal distribution, when I discuss normal distribution, I think you can remember there are 2 parameters which can describe a normal distribution if I know those 2 parameters, I can easily find out the normal distribution what are those 2 parameters? Mean and standard deviation or mean and variance whatever it is now, for me in the sampling distribution.

I know my mean of the sampling distribution is population mean 800 standard variance of the sampling distribution is σ^2 / n what is my sample size? Sample size is 16 right see here. So, what is my sample mean? Sample mean is 800 hours then what is my standard deviation or variance where basically whatever whichever you find out then a standard deviation of the sample standard deviation of the population by square root of sample size if I find out sample variance.

So, variance by sample size is not it? So, the σ / \sqrt{n} what is σ ? Σ is 40, 16 is my sample size. So, I got my standard division is 10 so, this is normally distributed this is basically 800 you can see the figure this is 800. This is the mean and if this is correct, I want to find out what is the probability that it is less than 775 it is definitely it will be to the left of the mean because mean is 800, 775 which is less than.

So, I need to find out the probability of having this value 775. How do I find out? So, I need to find out this probability we have seen normal distribution while we have solved problem using normal distribution first, we will calculate the take the value to the standard normal value that is the z value from z value we can construct the table I have done many problems on this the I do not think you will forget that because we have done many number of problems.

So, now, first thing is I have to find out the z value of this. So, how do I find out the z value of this? How does that remember z equals to? What is the z equals to? $x - \mu / \sigma$ is not it? So, now, what is my x ? x is 775 μ is 800 the σ this is σ . So, that means, I need to find a probability x bar less than 775 this I have converted to z value. So, what is z ? z is equals to here it is my x is x bar, x bar - μ / σ .

So, what is x bar? x bar is 775, $775 - \mu$ is $800 / 10$. So, z value is less than probability that z value is less than - 2.5. Here, I am not showing the table I have shown in many classes, how

to consult the table. So, from the table we can find out so, what is the z value? What is the probability corresponding to - 2.5? It is 0.0062. So, this is a very, very less probably not even 1% probability is not it?

However the question is not asking that, but we can consider that we can tell that we can claim that the manufacturer whatever manufacturer is claiming is not correct, the life of the bulb is not 800 if the life of the bulb is 800 with the standard deviation of 40 hours, I would not have from the sample result I would not have got a sample life of 775 hours, because probably this is true, probably of getting this I found it to be very small that means this is not true. Understood the use of sampling distribution at least we will do many more examples.

(Refer Slide Time: 24:27)

Sampling Distribution

Theorem: Sampling Distribution of the Difference Between Two Means

If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 , respectively, then the sampling distribution of the differences of means, $\bar{X}_1 - \bar{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Monalisa Sarma
IIT KHARAGPUR

So, there is one more theorem. This is again an important theorem used in many like when I try to compare 2 different populations we will be; needing it again and again. See, go through the theorem if the independent sample of size n_1 and n_2 are drawn, size n_1 and n_2 are drawn at random from 2 populations, n_1 I am drawing from population A n_2 I am drawing from population B to different population, whether discrete or continuous.

And say first population means μ_1 second population means μ_2 and first population variance is σ_1^2 second population the variance is σ_2^2 . Now, if I want to find out the if I want to compare these 2 populations, whether I am interested in finding out the difference of to support difference in means, suppose there are 2 different companies are producing this LED bulbs 2 different company A is producing LED bulbs.

Company B is producing LED bulbs, suppose, I want to find out the mean lifetime of these 2 bulbs the mean lifetime of this how much is a difference in the mean lifetime on these 2 bulbs? Then what I will do I will find out it will be substrate. So, then the sampling distribution of the difference of means, that is $\bar{x}_1 - \bar{x}_2$ is approximately normally distributed with mean and variance given by mean will be $\mu_1 - \mu_2$.

And variance I already talked variance, we never subtract variance always get added up we have also done that variance results on that one we have solved this is not it? Variance, if you find out the difference of 2 variances, it is always added up. So, this is the if we are interested in finding out the sampling distribution of the difference of mean this sampling distribution will have a mean of $\mu_1 - \mu_2$ and variance will be an addition of both the 2 variances.

(Refer Slide Time: 26:43)

Theorem: Sampling Distribution of the Difference Between Two Means

- ④ If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 , respectively, then the sampling distribution of the differences of means, $\bar{X}_1 - \bar{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Hence,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

- ④ Z is approximately a standard normal variable.

Monalisa Sarma
IIT KHARAGPUR

So, now, to find out the problem, we definitely we will have to find out the z value. So, what will be my Z value? Z value will be $\bar{x}_1 - \bar{x}_2$ bar what is the basically that means the x value and this is my μ is $\mu_1 - \mu_2$ and this is minus this is my standard deviation. Now, this z is approximately a standard normal variable.

(Refer Slide Time: 27:12)

Theorem: Sampling Distribution of the Difference Between Two Means (contd...)

Reproductive property of normal distribution: If X_1, X_2, \dots, X_n are independent random variables, having normal distribution with mean $\mu_1, \mu_2, \dots, \mu_n$ and variance $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ then the random variable $\bar{X} = a_1X_1 + a_2X_2 + \dots + a_nX_n$ has normal distribution with

mean, $\mu_{\bar{X}} = a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n$
variance, $\sigma_{\bar{X}}^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots + a_n^2\sigma_n^2$

Monalisa Sarma
IIT KHARAGPUR

So, this is basically a corollary of the previous theorem we can say see if X_1, X_2, X_n are independent random variables having normal distribution if X_1 is a normal distribution X_2 is a random variable having normal distribution X_3 is a random variable having normal distribution all independent random variables and all have normal distribution. And each has different means, suppose X_1 has mean μ_1 X_2 has mean μ_2 like likewise X_n has mean μ_n and variance.

Similarly, $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ then if we are interested in finding out an \bar{X} . What is \bar{X} ? \bar{X} if I am telling it this is a relation of \bar{X} to all the other random variable is some linear we are using some constant a_1, a_2, \dots, a_n these are all there is a linear relationship between among the random variables. So, if my \bar{X} is $a_1X_1 + a_2X_2 + \dots + a_nX_n$ then this definitely will because all X_1, X_2, X_n are normal distribution then definitely \bar{X} will also be normal distribution and what will be the mean of \bar{X} ?

Mean of \bar{X} will be $a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n$. So, like if we consider the previous theorem it is the same here a_1 and a_2 is basically a_1 is 1 a_2 is -1 so, what it will be $\mu_{\bar{X}}$? $\mu_1 - \mu_2$ variances variance all gets added up.

(Refer Slide Time: 29:04)

Example-3

Problem

Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0.

Assuming that the mean drying time is equal for the two types of paint, find $P(\bar{X}_A - \bar{X}_B > 1.0)$, where \bar{X}_A and \bar{X}_B are average drying times for samples of size 18.

So, to use this theorem a small example, just go to this problem 2 independent experiment are run in which 2 different types of paint are compared we want to compare 2 different types of paints we have taken 18 specimen from type A how we are comparing we are comparing based on the drying times which drying time which paint takes for drying times we are comparing trying to compare these 2 population.

Based on the drying times and the drying times in hours is recorded for each the same is done with type B the population standard deviation are both known to be 1 we know the population standard deviation of both the population is 1 A as well as B assuming that the mean drying time is equal for the 2 types of paint. Suppose what to say producer has claimed that mean drying times.

Producer of type A has claimed a particular mean drying time same claimed that the drying time is say y_1 the producer of paint B it has also claimed that mean drying time is y_1 both that means both the mean drying time is same. So, now, we want to compare whether this is really true than what we have taken we have taken a specimen of both types of specimen of size 8 and we found that the mean drying time of both on an average it is greater than 1.

So, $X_A - X_B$ is greater than 1. Now, if this is true, if the mean drying time of population, we consider the population mean drying time of both the population is same and both the population standard deviation is same is 1 how much probability is that we will get distinct probability that $X_A - X_B$ is greater than 1 because this is something which

we have taken the sample and we got the result this is something which the result which he got.

Now, we will try to find out same as the previous question, we will try to find out what is the probability of this occurrence assuming this is true, what the producer is claiming assuming that is true, same way, if this probability is high, then the producer is whatever the producer is claiming that is true, that means the mean drying time of both the paint is same and it is standard deviation is 1 that is true what they are claiming is that if this probability is very low that means what they are claiming that is not correct.

Anyway, this question is not asking that. This question is just asking what is the probability of this if we calculate this that is done, but try to understand what is the meaning of this problem? That is why I have given all those explanation.

(Refer Slide Time: 31:46)

Example-3

Solution

From the sampling distribution of $\bar{X}_A - \bar{X}_B$, we know that the distribution is approximately normal with mean $\mu_{\bar{X}_A - \bar{X}_B} = \mu_A - \mu_B = 0$ and variance $\sigma^2_{\bar{X}_A - \bar{X}_B} = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = \frac{1}{18} + \frac{1}{18} = \frac{1}{9}$.

Corresponding to the value $\bar{X}_A - \bar{X}_B = 1.0$, we have

$$z = \frac{1 - (\mu_A - \mu_B)}{\sqrt{1/9}} = \frac{1 - 0}{\sqrt{1/9}} = 3.0$$

$$P(Z > 3.0) = 1 - P(Z < 3.0)$$

$$= 1 - 0.9987$$

$$= 0.0013$$

Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0. Assuming that the mean drying time is equal for the two types of paint, find $P(\bar{X}_A - \bar{X}_B > 1.0)$, where \bar{X}_A and \bar{X}_B are average drying times for samples of size 18.

Monalisa Sarma
IIT KHARAGPUR

So, similarly, we have found what is this? As I as you have seen here, how do we find out the z variable this is the z variable $X_1 \bar{X}_1 - X_2 \bar{X}_2 - \mu_1 - \mu_2$. So, $\mu_1 - \mu_2$ is 0 because both the population mean is same is not it? And what is the standard deviation of this one $18 + 1 / 10$ root over because we have taken a sample of 18 and $X_1 \bar{X}_1 - X_2 \bar{X}_2$ it is we are trying to find for greater than 1 so we will find out a z value keeping it 1.

So, this is my this is equals to 0. This is my what to say σ^2 then what will be my Z value? Z value is this, so, I got this a probability of Z greater than 3, because I am interested in finding out P of Z greater than 1. So, probability of Z greater than 3 it will be nothing but 1 minus.

So, because normal distribution remember we always normal distribution is always specified in terms of cumulative distribution, cumulative distribution means it is possible from $-\infty$ to X that particular value area interest.

So, when we interested in finding out greater than 3, so, it will be $1 - P$ of Z less than equals to 3. So, from this value, we will get it from the normal table. I am not saying the normal table again and again I have shown it in many classes. So, this is the value we will get. So, this is the probability this is again a very low probability what does this indicate this is a low probability means what does this simple value cannot be wrong.

Because we have taken this we have checked it. So, what can go wrong what the producer is claiming? Because that is based on guesswork that is guesswork, maybe an educated guess maybe based on past experience, but there is no proof to that, but this we have done it and we have got the results.

(Refer Slide Time: 33:48)

The slide has a dark blue header with the word 'CONCLUSION' in large yellow capital letters. Below the header is a white section containing a bulleted list of learning outcomes:

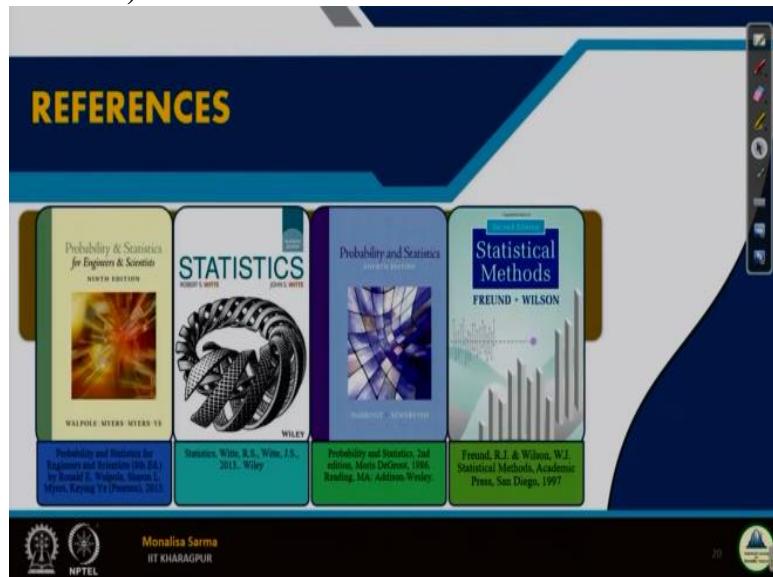
- ④ In this lecture we learned the basics of sampling distribution.
- ④ Next, we learned three very important theorems of statistics:
 - ④ Theorem on sampling distribution
 - ④ Central limit theorem
 - ④ Sampling distribution of the difference between two means
- ④ Learners are instructed to understand these theorems theoretically and also via practice problems.
- ④ In the next lecture, we will look into more concepts of Sampling distribution.

On the right side of the slide, there is a video feed of a woman with glasses and short hair, wearing a white shirt, speaking. At the bottom of the slide, there is a footer bar with the NPTEL logo, the name 'Monalisa Sarma', and the text 'IIT KHARAGPUR'. There is also a small number '19' in the bottom right corner of the slide area.

So, and I should not say in this lecture in this coming 2 in the last 2 lecture, this lecture and the last 1 lecture, we have learned the basics of the sampling distribution, we have learned 3 important theorems of statistics like theorem of sampling distribution, sampling central limit theorem, and sampling distribution of the difference between 2 means, that also we have seen.

So, as always, I will again ask all the learners to understand this theorem theoretically as well as via practical problems. In the next lecture, we will look into some more concept of sampling distribution.

(Refer Slide Time: 34:25)

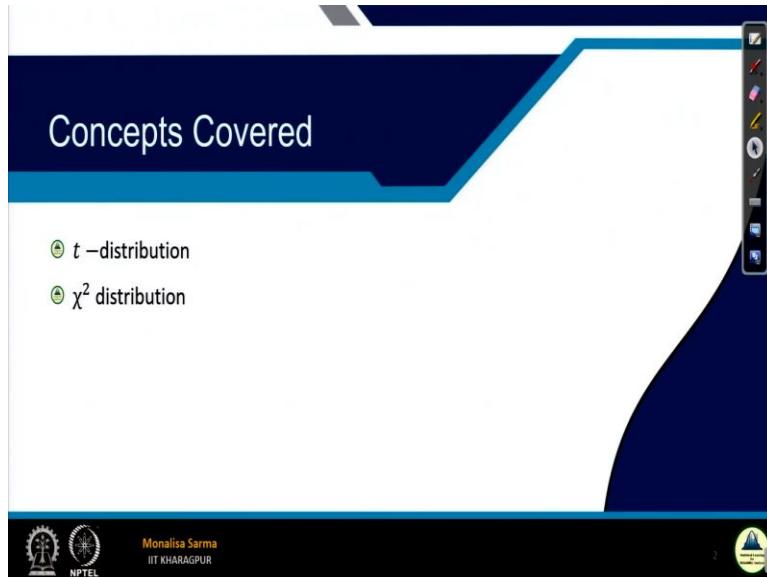


So, these are my references and thank you guys.

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

Lecture - 18
Sampling Distribution (Part 3)

(Refer Slide Time: 00:31)



Hello, welcome back. So, in continuation of our discussion on sampling distribution, today we will be seeing few other distribution some other sampling distribution along this. On the last class we have discussed the sampling distribution of mean for which we have used that is the normal distribution that is the Z distribution, sampling distribution of mean it is nothing but it has approximately it is normally distributed. So, basically we have considered the normal distribution that is, I can also say it is not normal basically the standard normal distribution.

So, in this lecture, there are some other distribution for sampling distribution I should not say sampling distribution of mean there are some other distribution which basically, which we use to find out the sampling distributions, that is the t distribution, chi square distribution, F distribution to name a few which we will be discussing and coming in this lectures as a listener coming through lectures.

(Refer Slide Time: 01:25)

Apart from the normal distribution to describe sampling distribution, there are some other quite different sampling distributions, which are extensively referred in the study of statistical inference.

t –distribution: Describes the distribution of normally distributed random variable standardized by an estimate of the standard deviation.

χ^2 –distribution: Describes the distribution of sample variance.

F –distribution: Describes the distribution of the ratio of two variables. This also has applications to inference on means from several populations.

So, now firstly, the thing is that what we have seen in the last class, we were interested in finding out the population mean if you are interested in finding out the population mean than the sampling distribution, which will make some we will form the sampling distribution of mean. If we have to infer about the population mean, we for that we will need sampling distribution of mean similarly, if we have to infer about the population variance, we will need sampling distribution of variance.

So, likewise to infer about the different things of the population accordingly, we need the sampling distribution so based on that we will have different distributions now like here as I have mentioned apart from normal distribution to describe sampling distribution that was the sampling distribution of mean. There are some other quite different sampling distributions which are extremely referred in the study of statistical inferences.

There is one says distribution is t distribution, t distribution is also used for the sampling distribution of mean, sampling distribution of mean we use Z distribution we have seen or normal distribution I can say that similarly, t distribution is also used for sampling distribution of mean, but here, sometimes in the population, we cannot say anything about a population standard deviation, like when we talk about the sampling distribution of the mean when we use normal distribution.

That time, you have seen that we have estimated the population mean as well as the population standard deviation, is not it? So, sometimes population standard deviation, it is really not possible to know the possible I mean population standard deviation, when we do not know the population standard deviation, then we cannot possibly use normal distribution, then we will have to use t distribution. Another is the chi square distribution, when we are interested in the distribution of sample variance that is called a sampling distribution of variance.

Then the distribution that we will use is chi square distribution. So, then again, there is another one distribution that is F distribution, when we want to compare the variance of 2 different populations, like it is mainly used in what to say this food and beverage industry, where in food and beverage in this industry, like suppose, there are 2 different companies which produces cold drinks now, let us suppose a cold drinks of volume say 1 litre.

So, we will be definitely when we are buying a cold drinks bottle of 1 litre we will expect that he exactly has 1 litre of cold drinks, but if it has less cold drinks, then the customer will get cheated if it has more than it will be customer will be benefited, but it will be loss to the company. So, basically what is necessary is that it should strict to that 1 litre volume means it has 1 litre with the slight variations. So, like so if we want to find out so, whether this is sticking to the required specification.

Or it is more or less then we will be using chi distribution when we want to compare the variances. Now when we want to compare 2 different populations like similarly for food industry only we are trying to compare for 2 different populations based on their variance. Then we will be using F distribution, of course, F distribution has other application as well, which we will be seeing later in this lecture series.

That is, we have something called ANOVA when we will be discussing a number then we will see F distribution is used however, distribution is used in ANOVA. So, now for your till now, it is sufficient to know that we use F distribution when we try to compare the variance of 2 different populations.

(Refer Slide Time: 05:31)

The χ^2 Distribution

Definition: The χ^2 distribution

If a random sample of size n is drawn from a **normal population** with mean μ and variance σ^2 and the sample variance is computed, we obtain a value of the statistic S^2 . Sampling distribution of S^2 can be described using **χ^2 distribution**.

NPTEL

Monalisa Sarma
IIT KHARAGPUR

Now, first we will discuss is the chi square distribution. So, first if a random sample of size n is drawn from a normal population with mean μ and variance σ^2 the sample variance is computed and we obtain the value of the statistics S^2 the sampling distribution of S^2 can be described using the chi square distribution as I told you sampling distribution of variance to x what to say to represent the sampling distribution of variance we use chi square distribution.

Now, what is the sampling distribution of variance? Sampling distribution variances we will call it as S^2 because we are finding the variance of sample so, it is not σ^2 that S^2 and one more thing like when we are doing sampling distribution of mean we were not very much interested in the parent population, I have parent population can be any distribution. It can be normal, it can be non normal, it can be slightly away from normal, it may be very much away from the normal.

Accordingly this we could solve it by adjusting the sample size into it the population is normal we take a some smaller sample space, They have more near normal we take a bit more sample size, but not very big also. But if the sample size is totally not normal then we take a bigger sample size. But chi square distribution, it is very much sensitivity to normality assumption of the parent population, this is a very important point. Chi square distribution we cannot use with the parent population is not a normal population.

Because in such situation if you use chi square distribution, then our result will not be a reliable result our results will not be precise results. There are in those cases basically, we have other

techniques so this techniques of inferring about the populations and is called parametric methods, this methods and there are 2 different methods of is in what is a say one is parametric, what is one is nonparametric. So, what we are discussing this probability distribution sampling distribution, all this comes under parametric method.

So, when the population is not normal, and we want to compare the variation of 2 different populations, then definitely we cannot use chi square distribution, then definitely, we will have to use some other methods, those methods we will be discussing later. Even now, what is it? I will repeat this again if a random sample of size n is drawn from a normal population that is a normal I have marked it red.

If a random sample of size n is drawn from a normal population with mean μ and variance σ^2 and the sample variance is computed, what is sample variance is S^2 . The sampling distribution of S^2 is described by chi square distribution.

(Refer Slide Time: 08:21)

The χ^2 Distribution

Definition: The χ^2 distribution

If x_1, x_2, \dots, x_n are independent random variables having identical normal distribution with mean μ and variance σ^2 , then the random variable

$$Y = \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2$$

has a **Chi square distribution** with n degrees of freedom

Monalisa Sarma
IIT KHARAGPUR

Now, question is what is chi square distribution? Earlier was sampling distribution of mean we use Z distribution we knew that this was a normal distribution, but was that we know, but chi square distribution we do not know. So, what is chi square distribution first let us see that, see this is the definition of chi square distribution if x_1, x_2, x_n are independent random variables

having identical normal distribution, these are different independent random variables having normal distribution with mean μ and variance σ^2 then the random variable Y .

What is Y ? Y is nothing but a summation of $Z^2 x_i - \mu / \sigma$ what is that? $x_i - \mu / \sigma$ is Z is not it? So, Y is nothing but a summation of Z^2 where this x_i is a different random variables having normal distribution that means I can say normal random variable x_i is a normal random variable. And μ and variance μ is the mean of that random variable and σ is the standard deviation of those random variables.

This here I told it is identical normal distribution why identical? That means, each has mean μ and variance σ^2 . That means chi square Y is nothing but the summation of $Z^2 Z_i^2$. So, this is called it has a chi square distribution with n degrees of freedom. So, the n degrees of freedom mean how many what to say logically independent unit there are totally n logically independent unit when you consider a chi square distribution.

So, the degrees of freedom are n so, when we consider a chi square distribution, the only parameter that we have to worry is the degrees of freedom. That is one parameter that is the degrees of freedom. That means for chi square distribution also we will have a table where we can find out the value. In the table, we will have values for different degrees of freedom. Remember, for normal distribution, when we are considering normal distribution, we have the values for different values.

Since it was not possible to have different values of μ and σ that is why we have converted it in to Z value. So, we have different values corresponding to different Z value. Similarly, when chi square distribution will have different values corresponding to different degrees of freedom that is, the only single parameter of chi square distribution that is degrees of freedom, that is the number of logical independent units. So, it is a summation of Z_i^2 and i go from 1 to n . So, there are n logical independent unit.

(Refer Slide Time: 10:54)

The χ^2 Distribution

Definition: The χ^2 distribution

If X_1, X_2, \dots, X_n are mutually independent random variables that have, respectively **Chi-squared distribution** with v_1, v_2, \dots, v_n degrees of freedom, then the random variable

$$Y = X_1 + X_2 + \dots + X_n$$

has a Chi squared distribution with $v_1 + v_2 + \dots + v_n$ degrees of freedom.

Monalisa Sarma
IIT KHARAGPUR

There is one more theorem here. So, if X_1, X_2, X_n are mutually independent random variables that have respectively chi square distribution, if X_1, X_2, X_n all have chi square distribution, X_1 as chi square distribution X_2 has chi square distribution with v_1, v_2, v_n degrees of freedom, then the random variable Y , $Y = X_1 + X_2 + \dots + X_n$ this Y is also a chi square distribution. What will be its degrees of freedom? Degrees of freedom will be the addition of all these degrees of freedom.

X_1 has degrees of freedom v_1 , X_2 has degrees of freedom v_2 , so, Y will have a degrees of freedom $v_1 + v_2 + v_n$ that will be the degrees of freedom of Y . And Y is also a chi square distribution Y will also have a chi square distribution, this is the important theorem I should say corollary of this previous definition of the chi square distribution.

(Refer Slide Time: 11:54)

The χ^2 Distribution

Definition: The χ^2 distribution

Each of the n independent random variable $\frac{(x_i - \mu)^2}{\sigma^2}$, $i = 1, 2, 3, \dots, n$ has Chi-squared distribution with 1 degree of freedom.

Now we can derive χ^2 -distribution for sample variance.

We can write

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n [(x_i - \bar{x}) + (\bar{x} - \mu)]^2$$

$\sum_{i=1}^n (x_i - \bar{x})^2$

or $\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$

or $\frac{1}{\sigma^2} \sum (x_i - \mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{x} - \mu)^2}{\sigma^2/n}$



Now, here when we consider the chi square distribution and there are as I told you in a chi square distribution, there are totally n independent units. So, each independent unit what are the each independent unit is Z_i is one independent unit Z_i is nothing but $(x_i - \mu)^2$ Z_i^2 this one independent unit but in Z_1^2 is one unit and Z_2^2 is one unit. So, this is one independent unit $(x_i - \mu)^2$ this each unit has a degrees of freedom of 1.

So, each unit has a degree of freedom 1 that is why chi square distribution as a degrees of freedom of n . Now, we got 2 different definitions one is this definition that is this definition and another we got this definition we also got this definition with these definitions with these 3 definitions. Now, we can derive this chi square distribution for sample variance here see, I already mentioned we use chi square distribution to find out the sampling distribution of variance means sample variance.

See in this chi square definition what we have seen till now, there is no mention of S^2 anywhere somehow we will have to bring the quantity S^2 to it then only then only that will be a distribution of because if that term is only missing, how can it be a random variable of that term. So, somehow we will have to bring S^2 to it, how do we get that. So, basically, first just simple consider this term $(x_i - \mu)^2$ that I have done is that.

Here I have just done some sort of manipulations here this can I write it in this way $x_i - \bar{x}$ bar + \bar{x} bar means \bar{x} bar I have substituted and added it. So, I got this now, this is $(a + b)^2$ what is $(a + b)^2$? $(a + b)^2$ is $a^2 + b^2 + 2ab$. So, I am writing it here only it will be easier. So, what do I get from this summation of first is a square. So, this is my first term. This is a square. Similarly, b^2 what I will get b^2 I will get is $i = 1$ to $(n \bar{x} - \mu)^2$ but now way i is there.

So, what it will be a ? It will be just n into $\bar{x} - \mu$. So, this is the term I got this is my b^2 . Next what is the remaining is $2ab$, summation of $2ab$ so $2ab$ is what is this?

(Refer Slide Time: 15:00)

The χ^2 Distribution

Definition: The χ^2 distribution

Each of the n independent random variable $\left(\frac{x_i - \mu}{\sigma}\right)^2$, $i = 1, 2, 3, \dots, n$ has Chi-squared distribution with 1 degree of freedom.

Now we can derive χ^2 -distribution for sample variance.

We can write

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n [(x_i - \bar{x}) + (\bar{x} - \mu)]^2$$

$$\text{or } \sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

$$\text{or } \frac{1}{\sigma^2} \sum (x_i - \mu)^2 = \frac{(n-1)s^2}{\sigma^2} + \frac{(\bar{x} - \mu)^2}{\sigma^2/n}$$

Monalisa Sarma
IIT Kharagpur

2 summation I am putting it inside say a summation of $x_i - \bar{x}$ bar into $\bar{x} - \mu$ is not it? This is my $2ab$ now, see this term summation of $x_i - \bar{x}$ bar. While calculating variance I think you have already learned in class 9 and 10 while calculating variance why we use square $x_i - \bar{x}$ bar 2 you know that can you remember why we? Because if we do not use square because this minus plus minus plus updating will be done and we know what the end result will be we will get 0.

When we because some value will be greater than mean some value will be less than mean what are the different x_i , x_i are the different values \bar{x} bar is the mean when you try to calculate the variance when we subtract each value from the mean some value will get greater than mean some value will get less than mean and everything sum together when you say we will get 0.

That is why to nullify the effect of minus sign we have used square that is why in variance we use square there that is the reason.

So, now, this term that means, this term will be equal to 0. So, this term is gone. So, what remaining is this 2 term only. So, what remaining is this 2 term. Now, what I have done now, this left hand side and right hand side I have divided by σ^2 . So, this term divided by σ^2 this term divided by σ^2 now, if this term I divided by σ^2 what do I get. What is my S^2 ? S^2 is nothing but variance only is not it? Variance of the sample so, what is my S^2 I am not using the board I am writing it here because then it will be easy for you to relate.

(Refer Slide Time: 16:55)

The slide title is "The χ^2 Distribution".

Definition: The χ^2 distribution

Each of the n independent random variable $\left(\frac{x_i - \mu}{\sigma}\right)^2, i = 1, 2, 3, \dots, n$ has Chi-squared distribution with 1 degree of freedom.

Now we can derive χ^2 -distribution for sample variance.

We can write

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n [(x_i - \bar{x}) + (\bar{x} - \mu)]^2$$

$$\text{or } \sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

$$\text{or } \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{x} - \mu)^2}{\sigma^2/n}$$

A handwritten note on the slide shows the derivation: $\sum_{i=1}^n (x_i - \bar{x})^2$ is circled in red, and the equation $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \mu)^2 - n(\bar{x} - \mu)^2$ is written next to it.

At the bottom of the slide, there is a logo and the name "Monalisa Sarma" followed by "III Kharagpur".

So, what is my S^2 ? $S^2 = 1/n - 1$ it my sample size is n summation of this is irritant actually i is equals to 1 to $(n x_i - x \bar{})^2$ this is my s^2 is not it? Somehow why, we are doing this manipulation. So, as we have somehow we need the S squared term to what to say, so, that we can find out the sampling distribution of variance. So, this is my S^2 now, this term is nothing but this term. So, what is this term this term is equals to S^2 into $n - 1$ will give me this term.

So, that is why I said instead of this term $n - 1 S^2 / S^2$ because σ^2 we have divided both left hand side and right hand side. So, this is what I got. So, this is clear. So, from this that is why I got this. Now, what is this? This is a chi square distribution. What is chi square distribution? Chi

square distribution is this summation $\sum_{i=1}^n X_i^2$ is chi square distribution this is summation $i = 1$ to n it is implicit not written here.

So, this is a this term is a chi square distribution with what degrees of freedom n degrees of freedom $i = 1$ to n it is n degrees of freedom and what is this? This is also a chi square distribution, but just 1 degree of freedom just 1 term. So, this is a chi square distribution with n degrees of freedom this is a chi square distribution with one; degrees of freedom. So, that means, this has to be a chi square distribution with $n - 1$ degrees of freedom.

(Refer Slide Time: 18:43)

The χ^2 Distribution

Definition: The χ^2 distribution

If X_1, X_2, \dots, X_n are mutually independent random variables that have, respectively Chi-squared distribution with v_1, v_2, \dots, v_n degrees of freedom, then the random variable

$$Y = X_1 + X_2 + \dots + X_n$$

has a Chi squared distribution with $v_1 + v_2 + \dots + v_n$ degrees of freedom.

This theorem if this all is a chi square distribution, this will be a chi square distribution so, now, I have this as chi square distribution this may be this as chi square distribution definitely this will be also a chi square distribution.

(Refer Slide Time: 19:00)

The χ^2 Distribution

Definition: The χ^2 distribution

Each of the n independent random variable $\left(\frac{x_i - \mu}{\sigma}\right)^2$, $i = 1, 2, 3, \dots, n$ has Chi-squared distribution with 1 degree of freedom.

Now we can derive χ^2 -distribution for sample variance.

We can write

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n [(x_i - \bar{x}) + (\bar{x} - \mu)]^2$$

or $\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n, (\bar{x} - \mu)^2$

or $\frac{1}{\sigma^2} \sum (x_i - \mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{x} - \mu)^2}{\sigma^2/n}$



So, this is a chi square distribution with n degrees of freedom this is a chi square distribution with 1, degrees of freedom. So, definitely this is a chi square distribution with $n - 1$, degrees of freedom. So, now I got it. So, this means my this value will have a chi square distribution and sampling distribution of mean my x bar has normal distribution is not it? x bar is approximately normally distributed, What is x bar? x bar is sample mean now, here I have this value.

This value is chi square distributed this is not only S^2 but something added some linear combination $n - 1 S^2 / \sigma^2$. So, now this is my sample statistics, when while finding out a sampling distribution of mean my sample statistic was the sample mean, here, my sample statistic is not only sample variance, but as well as I am multiplying sample variance with the size of the sample -1 $n - 1$ into S^2 and divided by the population variance.

Now, this population variance, and the sample size these are constant values, if we take different values different sample also, if we take different samples, this $n - 1$ and σ^2 will remain constant. So, this is the constant factor. So, what will vary S^2 will vary, so that means S^2 is chi square distributed. So basically, I can say this whole thing is chi squared distributed.

(Refer Slide Time: 20:49)

The χ^2 Distribution

Definition: χ^2 distribution for Sampling Variance

If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2 , then the statistics

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma} \right)^2$$

has a chi-squared distribution with $v = n - 1$ degrees of freedom

NPTEL Monalisa Sarma IIT KHARAGPUR 15

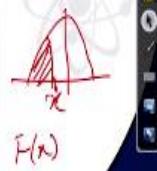
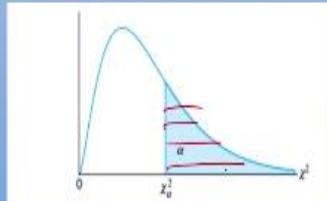
So, if S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2 then the statistics this statistics, these statistics I am calling it chi square, this whole statistics $n - 1 S^2 / \sigma^2$. This statistics has a chi square distribution with what is the degrees of freedom? Degrees of freedom $n - 1$ degrees of freedom. Why I am telling this statistics, because I am using the sample data to calculate this. What this sample data x squared calculated in from the sample is not it?

So, I am taking the sample data to calculate it. So, I am calling it sample statistics. So, this is my sample statistics, this have a chi square distribution with degrees of freedom $n - 1$. So, now this chi square distribution, I will be using it to predict about the population variance. Sampling distribution of mean I have used to predict about the population mean similarly sampling distribution of variance I will be using to predict about the population variance.

(Refer Slide Time: 22:00)

The χ^2 Distribution

The probability that a random sample produces a χ^2 value greater than some specified value is equal to the area under the curve to the right of this value.



Monalisa Sarma
IIT Kharagpur

Now, this chi square also like normal distribution how we can refer the table to get the value. Similarly, for chi square distribution, also, we do not have to calculate it ourselves, we can very well have the standard all the standard books have this statistical table available chi square table, from the statistical table, we will be able to calculate the chi square value, but there is a difference in the normal table, we had that cumulative distribution value.

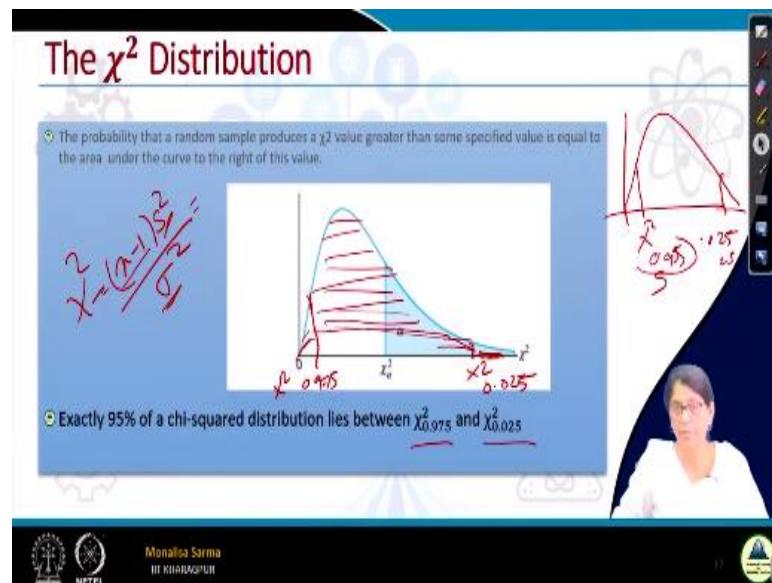
Remember, if it is a normal table, if this is as the normal Z table Z value, so when we if we are interested in finding out of f of x, so that means suppose this is x value, we got this value ∞ to that value. That was in case of normal distribution, we got the in binomial distribution also, we got the cumulative distribution value from $-\infty$ to that particular value x value, whatever that. But in chi square distribution, we just get the opposite table just give the opposite value.

So, if you have a chi square value, so see the probability that the random sample produces a chi square value greater than some specified value is equal to the area under the curve right of this value. So, in normal distribution, when we are interested in finding out a greater than value, then we have done 1 minus of less than, now, because in the table, we have the less than value, but in the chi square the table gives us the greater than value.

So, if you are interested in finding out the probability that a particular sample will have a particular value greater than a particular chi square value, so it will be this area. From the table,

we will be getting this greater than value, if you are interested in finding out the less than so we will have to do 1 minus of that.

(Refer Slide Time: 23:52)



And one more important point to note here in a chi square distribution is a very much you see it is a skewed distribution that is not a symmetric distribution it is very much a skewed distribution. In this distribution, what we have seen is that 95% of the value falls within this range. So, this is chi square 0.975 this is chi square 0.025, 95% of the value lies in this range. So, well calculating the probability if we find our probability values lies in this range. That is not what to say not more than 0.975.

If we can decide it will be more than 0.975 is not it because we consider the area to the right, it should not be more than 0.975. And moreover, it should not be less than 0.025 if a chi square value falls within this range then it is an acceptable way. See how chi square we write here it is for α we write is chi square α we get it a subscript. So, similarly, so, this is how we are writing in this. So, if 95% of the value lies in this.

Now, while calculating the probability if we find our value lies in this region, then we can say this is our whatever we have our estimate is correct, if under what case you see we will see our value will lies in this range or in this range under what situation. So, if you see the value will lie

in this recent situation means, this is the chi square, so, this is chi square 0.975 corresponding to this say particular value chi square 0.975 a corresponding to this suppose, if it is starting from 0 suppose this value is say 5 let us say any value any unit say 5.

And suppose this value that is equals to 0.025 let us say does this value is it 25. If we get a area greater than 0.975 when we will get a getting? Get area greater than 0.975 when my chi square value will be less than 5 suppose, if this is equals to 5 when my value will be less than 5 then I will be getting area greater than 0.975. Similarly, 0.025 similarly, I will get a suppose this is area corresponds to 0.025 is suppose as I told you it is 25 when I will get a value area less than this when my value will be more than 25.

Under what condition see what is my chi square? chi square is equals to $n - 1 S^2 / \sigma^2$ is not it now, my S^2 is something that I have calculated from the sample I have already taken a sample from the sample I have calculated the S^2 value. What is $n - 1$ the sample what I have taken that is the sample size - 1 there cannot be any error in that. So, I am getting a value which is greater than 0.75 under what situation I will get a value greater than 0.75?

When this σ^2 the σ^2 is what? Does a population σ^2 is we have just guessed, it is an estimation, it may be an educated guess is we have just predicted when I will get a value greater than 0.975 When this σ^2 value is very less if denominator is very less than what happen, I am sorry, if this denominator value is very less than I will get a value chi square value more bigger value that means maybe more than 25.

When my chi square value is very big, then I will get a chi square very small value, if my chi square σ^2 value is very small, I will get a bigger chi square value. In both the case there is this is because maybe my I have estimated my σ^2 value wrong why? If the σ^2 value is very big, getting a very big σ^2 means I am getting area greater than 0.975. Getting a σ^2 value very big that means my population is very much variant population.

When the population is very much variant, usually it is like this sort of statistical inference is usually not done in most of the case when the population is very much variant population that

means that is not a stable population. So, before coming to this sort of study only steps are already taken there it may be correct, there may be case when my population variance is very big that may be correct for some populations, but usually in general and the population variance is very big this sort of study is not done at all.

Because in that case, the population is very unstable, some steps has to be taken first to stabilize the population, then talking about statistical inferencing and all those stuff. So, that is one thing, second case, when my population variance is very small then what will happen my this value will be very big that means my area will be very less it will be in this area. So, that is also very unlikely usually in a population the population cannot be so stable that variance is so less very unlikely it may happen it is not that it will not happen.

It may happen. But that probability of that is very, very unlikely. So that is why it is considered that if my chi square value falls in this range, then that is that means that estimate what we have estimated what we have estimated? We have estimated the population parameter rest are things we have not estimated we have got it from the sample that means, if we got it if our probability we got in this range that means, our estimation was correct, if we get it in this range the side drains then maybe our estimation is wrong.

(Refer Slide Time: 30:24)

Some facts about χ^2 distribution

- The curves are non symmetrical and skewed to the right
- χ^2 values cannot be negative since they are sums of squares
- The mean of the χ^2 distribution is v , and the variance is $2v$
- When $v > 30$, the Chi-square curve approximates the normal distribution.

Then, you may write the following

$$Z = \frac{\chi^2 - v}{\sqrt{2v}}$$

NPTEL

Monalisa Sarma
IIT KHARAGPUR

21

So, some facts about chi square distribution the curves are non symmetrical and skewed to the right we have seen that chi square value cannot be negative since their sum of squares obvious. The mean of a square distribution is v and a variance is $2v$, v is the degrees of freedom why does mean and variance why it is necessary? If when v is greater than 30, the chi square curve approximates the normal distribution. When our sample size is more than 30 then what happens instead of using the chi square curve, because chi square curve has very limited table.

In the table, we get very limited value for the degrees of freedom. So, when the sample size is big, instead of using chi square distribution, we can also use normal distribution to approximate the probability we are using all this probability distribution to find out the probability that is the only objective. So, when the sample size is greater than 30 instead of using chi square, we can very well use the normal distribution.

So, in that case, so, normal distribution means we need the value of Z . So, Z means what is Z ? Z means $x \bar{ } - \mu / \sigma$, so, what is $x \bar{ }$ here $x \bar{ }$ is my chi square value this is not x squared this is chi square, this is $x \bar{ }$ is my chi square value $x \bar{ } - \text{mean}$, this is the mean of the chi square distribution is v and the variance is $2v$, v is the degrees of freedom and $2v$ is the variance, then $\sqrt{2v}$ is the standard division. That is how I find what to say that is the Z value once we know the Z value we can find out the probability distribution.

(Refer Slide Time: 32:14)

The slide is titled "The χ^2 distribution: Example–1". A blue box contains the "Problem" statement:

A manufacturer of car batteries guarantees that the batteries will last, on average, 3 years with a standard deviation of 1 year. If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year? Assume that the battery lifetime follows a normal distribution.

The slide features a video player window in the bottom right corner showing a woman speaking. There are decorative icons of a gear and a flask on the left side. The footer includes the NPTEL logo, the name "Monalisa Sarma IIT KHARAGPUR", and the IIT Kharagpur logo.

So, now one minute quickly we will solve 1 problem, or in fact, instead of doing it quickly let us stop this class here. I will start it from next class. I will start it from this point. Let me stop it here. And then in the next class first we will definitely solve this problem and then we will go to the next distribution. With this I end this lecture. Thank you.

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

Lecture - 19
Sampling Distributions (Part 4)

Welcome back guys. So, in continuation of our discussion on sampling distribution last class in my last lecture, if you remember.

(Refer Slide Time: 00:37)



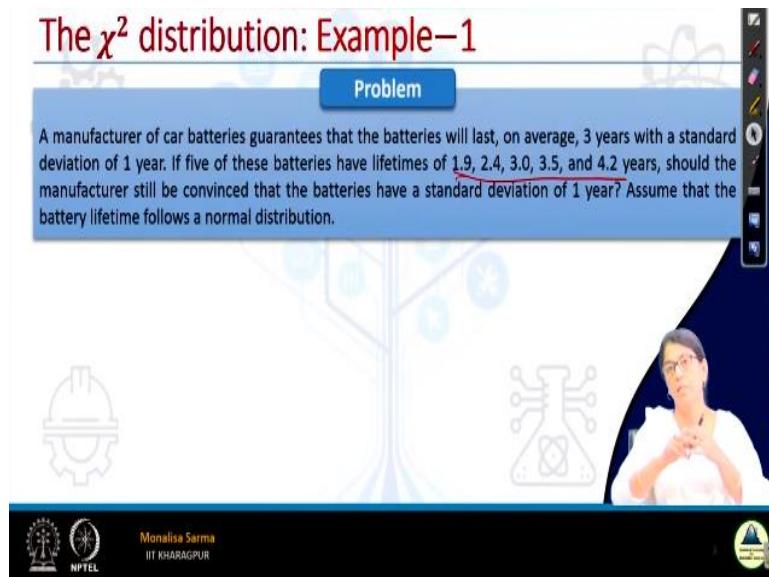
We discussed chi square distribution, and just we ran short of time in discussing a problem. The means related to chi square distribution first I will start with that and then I will go to t distribution.

(Refer Slide Time: 00:51)

The χ^2 distribution: Example–1

Problem

A manufacturer of car batteries guarantees that the batteries will last, on average, 3 years with a standard deviation of 1 year. If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year? Assume that the battery lifetime follows a normal distribution.



So, I think you guys can remember what is chi square distribution? Chi square distribution, we use for sampling distribution of variance, like sampling distribution of mean we have sampling distribution of variance, sampling distribution of variance, why it is used? We used to infer if we were to infer something about the population variance, usually it is more mostly used in food and beverages industry, where quality is very much a paramount importance.

And then we need to find out how much variance is it from the specified volume or the specified weight. In those cases, we need to find out the; infer the volume I mean the variance of the volume, for that, we need the sampling distribution of the variance. And for sampling distribution of the variance, we use chi square distribution. So we have discussed what is chi square distribution in my last class.

So today, we will be solving 1 problem, of course, I will be solving some other problem in my tutorial classes also, this is a problem so that you can understand the topic better, basically. So now, what is the question here? A manufacturer of a car batteries, guarantees that a battery will last on average, 3 years with a standard deviation of 1 year. So, manufacture it guarantees that on an average battery lasts 3 years with a standard deviation of 1 year.

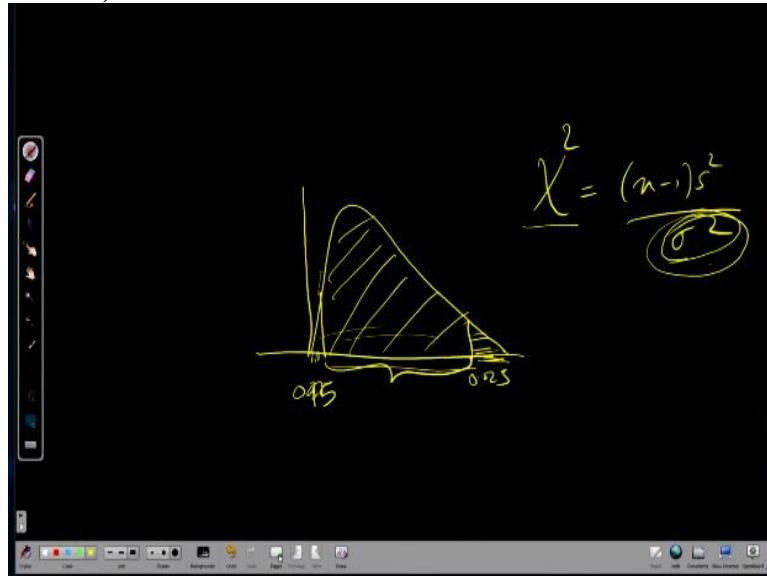
That is the manufacturer's what to say his claim. So now, if 5 of these batteries have lifetime of 1.9, 2.4, 3.0, 3.5 and 4.2 years, so what happened a person he bought this, that reorder might be you can take in this way, like he wanted to do a bulk purchase of batteries. Before that he is

might want to check it whether what the manufacturer is claiming is correct or not, maybe that is the reason or maybe he has bought it for his own need.

And he found that lifetime of this battery is this 1.9, 2.4, 3.0, 3.5 and 4.2 years. So, the manufacturer still be convinced that the batteries are the standard division of 1 year. So, with this, we are not interested in finding them average lifetime, just what we are interested? So, the manufacturer still be convinced that the batteries have a standard division of 1 year that means the variance among them is 1 year or not.

That is the manufacturer whatever his claim, is it correct? That is what he wants to find out. So, remember, when we discussing chi square distribution.

(Refer Slide Time: 03:27)



We have seen that let me take this chi square distribution is something of some skewed sort of distribution, remember, so it was something this portion at both the position is same, but my figure is not very good. So it is like 0.9975 and this is 0.025, this area is 0.025 and this whole area is 0.975. So, if the probability lies in this range, yesterday we have discussed then whatever its claim, our claim is correct that is what we were.

And that is a reasonable value the probability basically, if it lies within this range that is a reasonable values of probability, we have seen yesterday there in last class basically, if the probability lies in this range which is very less probability under what situation the probability

will lie in this range. Remember what was the chi square value? Chi square value is $n - 1 S^2 / \sigma^2$ probability will lie in this range.

When chi square value is quite high, under what situation chi square values will be quite high, when this value will be very less when sigma square value is very less than we will get a very high chi square value so σ^2 value does a variance value very less variance value that is very much unlikely it is not that it cannot be but it is very much unlikely that it will have a very less chi square value.

At the same time, we will have a chi square value is very lesser chi square value if we have something in this region a chi square value will vary less, is not it, because it is starting from 0. So, under what situation we will have when chi square value is when σ^2 is very large. So, that is also quite unlikely why when the σ^2 is very large that means, this process is not stable when the process is not stable, why at all will go for doing this statistical inference is not it?

It is an unstable process. So, that is the reason that the value lie in this range is very, very unlikely, but it is not that it may not happen it may happen, but if it happens in this range, it might probability lies in this range then it is. Now, this is the claim is correct. So, now, essentially this problem we need to find out whether our; what to say probability lies in that acceptable range are not.

So, how do we solve this question first is that we need to find out the chi square value for chi square value, we need to find out the S^2 that is the variance of the sample.

(Refer Slide Time: 06:06)

The χ^2 distribution: Example–1

Solution

We first find the sample variance

$$S^2 = \frac{5 \times 48.26 - 15^2}{5 \times 4} = 0.815$$

A manufacturer of car batteries guarantees that the batteries will last, on average, 3 years with a standard deviation of 1 year. If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year? Assume that the battery lifetime follows a normal distribution.



Monalisa Sarma
IIT KHARAGPUR



So, variance of the sample this is variance you know that formula for variance or we can use one more, there is one more formula for variance which are.

(Refer Slide Time: 06:18)

$$\frac{1}{n(n-1)} \left[n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right]$$

Let me give you there is 1 more formula for variance like it is $1 / n x n - 1$ this is an $n x n - 1$ than sum n into $\sum_{i=1}^n y_i^2 - i = 1$ to $n y_i$ sorry it is not y_i always x_i have x_i whole square. So, this is also 1 formula for let me rub it. So, this is also one way of calculating variance. Variance formula I think you know the one the standard formula which you know you can calculate it using that formula also or you can calculate it using this formula also.

(Refer Slide Time: 07:15)

The χ^2 distribution: Example–1

Solution

We first find the sample variance

$$S^2 = \frac{5 \times 48.26 - 15^2}{5 \times 4} = 0.815$$

Then, $\chi^2 = \frac{4 \times 0.815}{1} = 3.26$

It is a value from a chi-squared distribution with 4 degrees of freedom.

Monalisa Sarma
IIT KHARAGPUR

Here, I have used the other formula which I have just now written. So, we calculate the variance from the sample. Sample variance I already told you we specify it by S^2 . So, we found S^2 square sigma square is given we know what is n , n is 5 so, it is $n - 1$. So, we found the chi square value, chi square value we got is 3.26. Now, we need to find out for what is the degree of freedom here? Degree of freedom is $n - 1$ that is 4.

So, we need to find out for 4 degrees of freedom what is the acceptable range that we will for that we will consult a chi square table. So, I told you we have a chi square table so, as to consult the value.

(Refer Slide Time: 07:57)

The χ^2 distribution: Example–1

d.f.	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01
1	0	0	0	0	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.1	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.3	0.48	0.71	1.06	7.78	9.49	11.143	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.2	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.7	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21

Since 95% of the χ^2 values with 4 degrees of freedom fall between 0.484 and 11.143, the computed value with $\sigma^2 = 1$ is reasonable, and therefore the manufacturer has no reason to suspect that the standard deviation is other than 1 year

Monalisa Sarma
IIT KHARAGPUR

So, basically you see this is the chi square table and chi square table you will see 4 degrees of freedom. This is the area corresponding to this is the value of chi square corresponding the value area of 0.975. So, this maybe this area sorry this area if this area is 0.975 this value is 0.484. This is for 4 degree of freedom and then it is 0.025 maybe this region 0.025. This and the value of this maybe 11.143.

So, if my value lies between this and this than it is acceptable. So, what value I got definitely it lies within this range. So, what value I got? I got 3.26, chi square value 3.26. So, 3.26 lies very much within this range within this 4.484 to 11.41, it lies within this range. So, that is why since, you have written it since 95% of the chi square values with 4 degrees of freedom fall between 0.484 and 11.143, the computed value with $\sigma^2 = 1$ is reasonable.

It is a reasonable value because it follows an acceptable is not it is which is falls it in 95% of the values then therefore, the manufacturer has no reason to suspect that a standard deviation is harder than 1 year. So, what the manufacturer claim is it we can say that it is correct the he has no reason to suspect it, but always all these things when you talk about probability always all these things is a probability concept problem is the uncertainty it is already always there. It is not a never it is a definitive answer. So, that is the chi square distribution.

(Refer Slide Time: 09:57)

The **t** Distribution

- 1** To know the sampling distribution of mean we make use of Central Limit Theorem with $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$
- 2** This require the known value of σ a priori.
- 3** However, in many situation, knowledge of σ is certainly no more reasonable than the knowledge of the population mean μ .

NPTEL Monalisa Sarma IIT KHARAGPUR

Now, coming to t distribution like chi square distribution and we use for sampling distribution of variance. For we have seen for sampling distribution of mean we have seen that we use Z

distribution. So, symbol for sampling distribution of mean we also used t distribution. Now, under what situation we use t distribution that we will see first you see, when we have used sampling distribution of mean when we have used Z distribution, say this was our value.

This is the value, where we use to find out the Z value. Z is X bar what is X bar? X bar is the mean that we get from the sample $X \bar{ } - \mu \sigma / \sqrt{n}$ here you see when we compute Z while explaining sampling distribution of mean I have told again and again this μ value what is our objective? From sampling distribution of mean our objective is to infer about the population mean.

So, you may think our we have to infer about the population mean, then we already have mu and what we have to infer it is already given no it is not given it is I told you it is not it is estimated it is or I can say it is an educated guess it is in maybe the producer or manufacturer or whoever the person maybe he is just claiming from the sample we have to find out whether what he is claiming is true or not, we have just taken it as estimated value or predicted value.

Now, the question is when we do not know the mean of the population having an idea what the variance of the population is very, very remote, we do not have mean only, how can we know the variance of the population isn't? If the population is something very known population or from past experience, then we can say variance of the population maybe so, and so, the easily in most of the cases when mean is not known having a knowledge about the variance is difficult.

In such cases when we do not know the variance of the population μ^2 have guessed it or we have predicted it that we will find it out using the sampling distribution of the mean whatever we have predicted is true or not that will find it out, but the variance what we have predicted variance what we have predicted that is usually it is very remote that you can predict a variance. In that case, when you do not want a variance we cannot use Z distribution because Z distribution the formula says directly says that we needed.

The sigma here directly says then how can we use Z distribution. So, instead of σ we will need to use S that is the standard deviation of the sample or rather we will have to use the variance of the

sample. So, when we use instead of σ when we use S then it becomes then it no longer fits to a normal distribution. So the distribution that we use here is the t distribution. So, that is what, so, whatever I told here this required a known values of σ a priori.

However, in many situation knowledge of σ is certainly no more reasonable than the knowledge of the population mean μ .

(Refer Slide Time: 12:56)

The slide is titled "The *t* Distribution". It contains three numbered points:

- 4 In such situation, only measure of the standard deviation available may be the sample standard deviation S .
- 5 It is natural then to substitute S for σ . The problem is that the resulting statistic is not normally distributed!
- 6 The *t* distribution is to alleviate this problem. This distribution is called student's *t* or simply *t* - distribution

At the bottom left is the NPTEL logo, and at the bottom center is the text "Monalisa Sarma IIT KHARAGPUR".

So, in such situation only measures are the standard deviation available, maybe the sample standard deviation S . So, if we substitute σ / S , the problem is that a resulting statistic, so, what is the statistics now the sample statistic is a *t* value. So, sample statistic is now not normally distributed, earlier my sample statistic was *Z* value for sampling distribution of men that was normally distributed in sampling distribution of variance my sample strategy statistics was the chi square value.

So, that chi square value was chi square distributed. So, now, this distributed empirically it is found that it fits a distribution which is called a *t* distribution, it is called student's *t* distribution or simply we can say this *t* distribution.

(Refer Slide Time: 13:49)

Now, what is t distribution? Any distribution means, we need to know its PDF probability distribution function, if it is continuous then call the density function whatever is we need to know that. So, t distribution, this is the PDF of t distribution. Here also there is only 1 parameter that is a degrees of freedom, t distribution with v degrees of freedom it takes Z by chi square with v degrees of freedom divided by the degrees of freedom that is a PDF of t distribution.

When if you just Google t distribution you may find a different expression, but one simplified form this is a simplified form of t distribution. Similarly, if you Google chi square distribution you may find a different form whatever form I have given that is again a simplified form different simplification different way of simplifying because chi square we use in different application in different application we will use that its different representations.

So, similarly, t distribution we will be using this representation. So, what is this Z by square root of chi square with v degrees of freedom divided by the v degrees of freedom. Now, we know what is that well, we know what is chi square with v degrees of freedom what is that if we substitute this then maybe we will get it whatever we need we need something in terms of sample parameter, a sample statistics. For in case of sample I should never talk as parameter, parameter is always for population.

(Refer Slide Time: 15:28)

The *t* Distribution

Let X_1, X_2, \dots, X_n be independent random variables that are all normal with mean μ and standard deviation σ .

Let $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

Using the definition of *t*-distribution, we can develop the sampling distribution of the sample mean when the population variance, σ^2 is unknown.

That is,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has the standard normal distribution.

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

has the χ^2 distribution with $(n - 1)$ degrees of freedom.

So, now, see here now isn't here? So, we have Z is equals to this, chi squares this, we have seen chi square is this we have seen Z is equals to this and from the sample if X_1, X_2 are the independent random variable and with the mean μ and standard deviation, then I can find \bar{X} is this way, and S^2 is just that is the sample variance this already I know. So, now, in this *t* distribution formula if I replace substitute the value of Z and chi square, what do I get?

(Refer Slide Time: 16:05)

The *t* Distribution

That is,

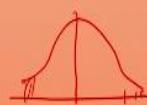
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has the standard normal distribution.

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

has the χ^2 distribution with $(n - 1)$ degrees of freedom.

Thus, $T = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}}$



or $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$

has the *t-distribution* with $(n - 1)$ degrees of freedom.

I am replacing this Z by chi square value is just a simplification after upon simplification, I will get this as my this is the PDF takes this form $\frac{1}{\sqrt{(n-1)S^2/\sigma^2}} e^{-\frac{(n-1)(\bar{X}-\mu)^2}{2S^2}}$, S is the sample variance. Now, I do not need to know the population variance, even if I do not know population variance, I can still make a guess on that I can still make an inference on the population mean, but of course, I

will have to first take an educated guess or I have to take some estimated value of the population mean or I have to predict some population mean.

So, this is the thing so, this is the statistics, this t value is the statistics that we will have to find out from the sample given a sample I like to find this t value what is how do I find the t value? t value is $X \bar{ } - \mu / S / \sqrt{n}$ and then accordingly now, the way we do for Z distribution the way we do for chi distribution, chi square distribution. Similarly, for t distribution, we have also lookup table from the lookup table we can get the value.

But t distribution is also one of the characteristics of t distribution this is also this distribution also very much symmetric like a normal distribution. So, since the symmetric so on the table, we have only the values for the upper tail, but right tail in the lower tail values are not given in the distribution because if we know the right tail values, we can go get the values for the low left as well because it is symmetric, is not it, like for them distribution, the whatever value we get for Z - 2, the same value will be getting for Z equals to + 2, is not it?

So because it is symmetric, similarly for t also first so that is why in fact, in Z we have seen we have given values for all the upper tail and a lower tail, but t distribution in a table only in most of the table only the upper tail is given, but we can definitely find it out from the lower from the upper tail we can find out the lower tail. So, now t distribution, some more characteristics of t distribution t distribution is like if I draw the figure for t distribution, like it is very much similar to normal distribution.

But it says a fatter tail, this tail portion is fatter compared to the normal distribution, fatter means probability here, the probability of occurrence random variable is quite more compared to the normal distribution. And one more thing, when my sample size is big, then this difference does not make this if I take whether I am taking S square or whether I am taking σ^2 it is the difference does it make have a much impact and that is why for greater sample size instead of t distribution.

I can very well use the normal distribution, the result will not be unreliable, it will be with good precision only instead of t distribution I can very well use the normal distribution for a bigger sample size.

(Refer Slide Time: 19:12)

The **t** Distribution: Example – 2

Problem

A chemical engineer claims that the population mean yield of a certain batch process is 500 grams per milliliter of raw material. To check this claim he samples 25 batches each month. If the computed t-value falls between $-t_{0.05}$ and $t_{0.05}$, he is satisfied with this claim. What conclusion should he draw from a sample that has a mean $\bar{x} = 518$ grams per milliliter and a sample standard deviation $s = 40$ grams? Assume the distribution of yields to be approximately normal.

Monalisa Sarma
IIT Kharagpur

So, now the question a chemical engineer claims that population mean yield or a certain batch is 500 grams per milliliter of raw material from per milliliter there is some raw materials are there from and that the mean yield of the population mean yield will be it is 500 grams. He claims that to check his claim, this person is suspicious of his own claim. So, he wants to check it whatever he is to check his claim he samples 25 batches each month. That means it the sample size is 25 now.

If the computed t value falls between this t of minus 0.05 to t of 0.05, he is satisfied with this claim. What conclusion should he draw from the sample that has a mean of $X \bar{=} 518$ grams per milliliter and a sample standard deviation of $S = 40$ gram. see here population standard division he does not know he is just having a guess of mean, mean means he is guessing that it is around 500 grams, but he himself is not convinced.

So, there so, he wants to check for more checking what he has done he has taken a sample batch of 25 that is a sample size and from there since he has to infer on the population mean, so, it will be sampling distribution of mean now sampling distribution mean we have learned to we can

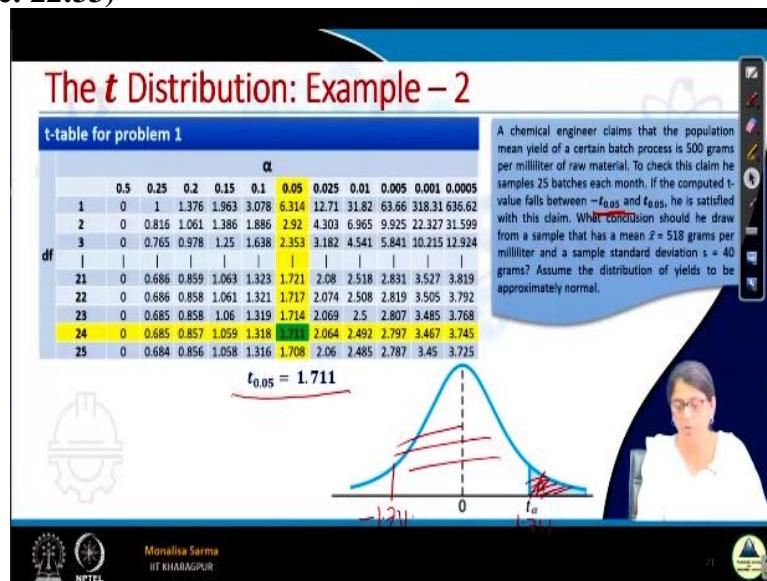
either use Z distribution, we can use t distribution. Now, here we cannot use Z distribution because we do not know the population standard deviation.

So, we will have to use t distribution because we do not know the population standard deviation and we have here we have the sample standard deviation definitely if it is not there also given the data we will be able to collect the sample standard deviation. So he will use calculate the t value now what he is claiming that if the computed t value falls within this range this range means what t of that let me draw the figure it will be something. So, this is I am very bad in drawing actually.

So, it goes this way, this way it goes. So, he is satisfied with this claim if it is falls in t of minus 0.05 means this portion plus 0.05 means this portion. So, he is satisfied, if it is falls within this range, that is quite reasonable that means it is falling within this range means probability of occurrence is quite high means, if this is true probability that I will be getting this 518 value is a high probability because this is very less probability is not it?

That is what his whatever explain he is taking quite justifiable experiment, because if this is true than from the sample, what he is getting probability of occurrence it should have a higher probability, if it has a very remote probability that means what he is claiming is not true. So, that is why he has taken this range.

(Refer Slide Time: 22:33)



Now, let us see how he has done it. So, we have seen from the t table as I told you, in the t table only the upper tail values are given lower tail values are not given. So, in the t table, and one more difference like in Z distribution as I told you, we get the CDF cumulative distribution from minus infinity to x, but chi square is different chi square we get the value the right side of the value in fact 1 minus of the CDF that we get in chi square.

Similarly, in t distribution, t distribution alpha value is the area we got this is the area rather than this portion, we have this portion in a table. So, that is the reason t value and chi square is same we have the right side value. So, here or what is it how many degrees of freedom? Degree of freedom is 24. Because there is 25; sample size is 24 degrees of freedom for 0.05 what is the value? 1.711.

For 24 degrees of freedom t 0.05 is this value corresponding to this is 1.711. So, what will be the value corresponding to minus t of 0.5? It will be because it is symmetric it will be minus 1.711 this minus value is lower values are not given but it is symmetric. So, I can even if I know that I will be able to know this, is not it? So, if it falls it might calculate a t value falls within this region calculated t value falls within - 1.71 to + 1.71 that means, this is whatever his claim is justifiable. So, what did he get?

(Refer Slide Time: 24:10)

The **t** Distribution: Example – 2

Solution

From t-Table, we get $t_{0.05} = 1.711$ for 24 degree of freedom (v).
Therefore the engineer can be satisfied with his claim if a sample of 25 batches yields a t-value between -1.711 and 1.711.
If $\mu = 500$, then

$$t = \frac{(518-500)}{\left(\frac{40}{\sqrt{25}}\right)} = 2.25 > 1.711$$

The probability of getting t-value greater than or equal to 2.25, for $v=24$, is approximately 0.02.
if $\mu > 500$, the value of t computed from the sample is more reasonable.
Hence, the engineer is likely to conclude that the process produces a better product than he thought.

A chemical engineer claims that the population mean yield of a certain batch process is 500 grams per milliliter of raw material. To check this claim he samples 25 batches each month. If the computed t-value falls between $-t_{0.05}$ and $t_{0.05}$, he is satisfied with this claim. What conclusion should he draw from a sample that has a mean $\bar{x} = 518$ grams per milliliter and a sample standard deviation $s = 40$ grams? Assume the distribution of yields to be approximately normal.

NPTEL
Monalisa Sarma
IIT KHARAGPUR

So, he calculated his t value, this is how we calculated his t value. Using the t value formula what is the formula for the t you remember, $X \bar{S} - \mu S / \sqrt{n}$ this is the formula for the t value. So, he

calculate the t statistics from the sample he got 518. This is the, here is the predicted this population mean 500. Then sample standard deviation 40 divided by root n that is 25 and what we got is 2.25.

2.25 means if I draw the figure if this portion is say 1.711. And then he got it somewhere here that is 2.25, somewhere this portion this is minus 1.711. He did not get within this mean, what the value of what you got is beyond this that means probability of that occurrence is very less that is very remote probability that means, whatever he is claiming that mean yield 500 grams, that is not true.

If that is true we are from the sample not bigger sample is something which we have tested it cannot be wrong, it is we have this collected and tested this cannot be wrong, it is giving us a true picture. So, if that is true, then this should have been true, but what I got this is from this taking that is true, I got a very less probability that means what I have assume and whatever estimated the population mean 500 that is not correct.

So, now, you can see him that is not correct. So, what may be the probable values of 500? You see I got a value of 2.25 If I got if I would have got a lesser value than 1.711 lesser than under what situation I will get if this 500 would have been if this is instead of 500 if this value would have been more then I would have got a lesser value is not it? So, first thing the probability getting a t value greater than 2.254 days is approximately 0.02 very remote probability that is my first observation.

Second observation mu is better than 500 value from t sample is no more reasonable. Third, the engineer is likely to conclude that a process produces a better product than he thought actually, he was very pessimistic instead of doing this optimistic, he was telling that my thing mean yield is 500. But he did not know that his process yield is much more than that. That if he would have predicted his process yield is much more than 500, then his value would have been within this range.

(Refer Slide Time: 27:21)

CONCLUSION

In this lecture we learned

- χ^2 distribution
- t –distribution

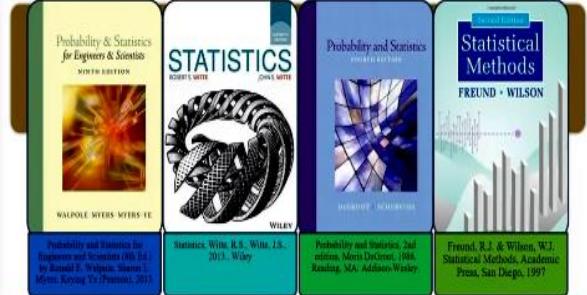


NPTEL Monalisa Sarma IIT KHARAGPUR

So, that is all in this lecture so we have till now we have learned chi square distribution, we have learned t distribution and few more sampling distribution which will be learning and my next lecture.

(Refer Slide Time: 27:36)

REFERENCES



Walpole, Myers, Myers, Ye
Probability and Statistics for Engineers and Scientists (9th Ed.) by Ronald E. Walpole, Raymond H. Myers, and Sharon L. Myers, Pearson Education, 2012.

Statistics, Witter, R.S., Witte, J.S., 2013, Wiley

DeGroot, M.H., Schervish, M.J., 2012, Addison Wesley

Freund, R.J. & Wilson, W.J., Statistical Methods, Academic Press, San Diego, 1997

NPTEL Monalisa Sarma IIT KHARAGPUR

Then, these are the references guys. Thank you.

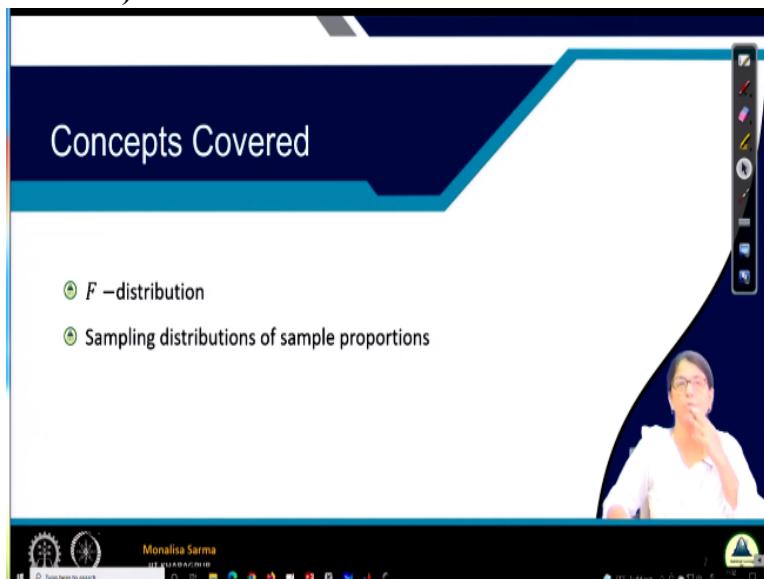
Statistical Learning for Reliability Analysis
Dr. Monalisa Sarma
Subir Chowdhury of Quality and Reliability
Indian Institute of Technology – Kharagpur

Lecture – 20
Sampling Distributions (Part 5)

Welcome back again. So, this is again in continuation of our earlier lecture on sampling distribution, where we have learned till now we have learned sampling distribution of mean that using both Z distribution and T distribution, then we have learned Z distribution and T distribution and we have learned chisquare distribution, this distribution we have learned is not it? Remember what is for sampling distribution of mean when the gradient population variance is known.

And we Z distribution when population variance is not known, we use T distribution and for sampling distribution of variance when we infer something about the population variance, then we use chisquare distribution.

(Refer Slide Time: 01:05)



Today, we will be learning 2 more distributions basically which we will be used our objective here is not for learn the distribution objective here is to learn that this distribution which is necessary for our sampling distribution. So, next to which we will be learning here is F distribution and sampling distribution of sample proportion.

(Refer Slide Time: 01:27)

The *F*-distribution finds enormous applications in comparing sample variances.

Definition: The *F*-distribution

The statistics *F* is defined to be the ratio of two independent Chi-Squared random variables, each divided by its number of degrees of freedom.
Hence,

$$F(v_1, v_2) = \frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2}$$

Monalisa Sarma
IIT KHARAGPUR

Now, *F* distribution first: *F* distribution, I think you remember when I have just mentioned about it in my first class or second class of sampling distribution, what is *F* distribution basically, we use when we want to compare 2 different populations, when you want to compare the variances of 2 different populations, comparing the mean of 2 populations that we have used *Z* distribution even we can use *T* distribution as well.

When we want to compare 2 different means, see, when we want to infer something about the population mean, we can use either *Z* distribution or we can use *T* distribution, again, the same 2 distribution we use when we want to compare 2 different populations. So, but in case of variance, when we want to infer about the population variance, we use chisquare distribution. And when we want to infer about when we want to basically compare 1 different population on variance, then we do not use chisquare distribution, we use *F* distribution.

And one more thing which I forgot to mention in my last class, chisquare distribution, I think I have mentioned that it is very much sensitivity to normality assumption, that means my parent population has to be normal. Same is the case with *T* distribution that also the parent population has to be normal. So, *F* distribution also same, let us parent population has to be normal population. So, now *F* distribution has one more application that we will be seeing while we will be discussing ANOVA.

Now, what is first let us come to see what is *F* distribution; what is the PDF of *F* distribution? So, PDF of *F* distribution as similar to chisquare and *T* distribution, there can be different representations. One of the representations which we; will be using, which we need us,

basically for our sampling distribution. So now, first we need to know the PDF of the F distribution. So, when we talk about the PDF of the F distribution, there can be again different representation.

So, we will be using the representation which will be easier for us to use in order to find out the sampling distribution fine. So, this is the PDF for the F distribution. So, what this is a PDF for the F Distribution? See, we have 2 chisquare distribution first in the numerator gives 1 chisquare distribution whose degrees of freedom is v_1 divided by the degrees of freedom and in the denominator, we have another chisquare distribution with degrees of freedom v_2 divided by the degrees of freedom. So basically, we will be using this.

(Refer Slide Time: 03:57)

Now, we know what is chisquare distribution we have seen, what is the statistics associated with chisquare this is $n - 1 S^2$ by σ^2 what is the degrees of freedom for this? Degrees of freedom for it is $n - 1$, so we will be using this, this use substitute this and F value. So, if we substitute this chisquare $n - 1 S^2 / \sigma^2$ would you say $n - 1$ I am writing here say, if I use $n - 1 S^2 / \sigma^2$, because these are 2 different chisquare distribution, is not it?

So, $n - 1$ what is that this is my chisquare and I have here again what to say divided by the degrees of freedom, so degrees of freedom is $n - 1$. So, here also similarly S^2 divided by σ^2 again divided by the degrees of freedom that is $n - 1$. So, my degrees of freedom, degrees of freedom gets cut and if I get simplified this is my F distribution. So, what is the PDF of F distribution is S^2 / σ^2 that means the concerning the first population.

Because we are talking about comparing 2 different population so, this is the, S_1 is the standard deviation of the sample of the first population, σ_1 is the standard deviation of the first population, S_2 is the standard deviation of the sample of the second population, σ_2 is the standard deviation of the second population.

(Refer Slide Time: 05:36)

Characteristics of the F distribution

- The F -distribution is defined only for nonnegative values.
- The F -distribution is not symmetric.
- A different table is needed for each combination of degrees of freedom.
- The choice of which variance estimate to place in the numerator is somewhat arbitrary; hence the table of probabilities of the F -distribution always assumes that the larger variance estimate is in the numerator.

So, again F distribution since F distribution is like a division of 2 chisquare values, is not it numerator we have a chisquare denominator we have a chisquare so, definitely it will also have only nonnegative values, chisquare we already we have seen it always has nonnegative, it cannot have negative values because it is just 2 term. So, here f also will have only nonnegative values.

And here in F distribution there are 2 degrees of freedom and numerator we have 1 degrees of freedom that is the v_1 degrees of freedom denominator we have another degrees of freedom that is v_2 degrees of freedom. Basically, the numerator is the sample size that we have taken from the first population and denominator is the sample size that we have taken for the second population. So, here see there are 2 degrees of freedom 10, 30 6, 10 these are the degrees of freedom, the 2 parameters are the F distribution.

So, F distribution is defined only for non terminal and F distribution is also not symmetric. Chisquare distribution is not symmetric we have seen similarly F distribution is also not symmetric. And we need a different table for each combination of degrees of freedom. For each combination of the degrees of freedom, we need a different table. So that is why it is

very difficult to have a table for different probability values. So, in most of the standard textbook, you will find this of course, this lookup table for F calculating the F value.

Also, we have the lookup table like we have for binomial distribution for chisquare for T distribution at all. Similarly, for here, also, we have the lookup table, but we have to for very limited probability values, most of the standard textbooks have F distribution values, probability values for the 0.05 and 0.01 maybe this range. Maybe 05, I can say this is the probability, this whole is a 05 probability and 01 maybe this one this portion, only 2 probability value we have.

But there is 1 theorem, if we know this value, we can find out this value as well. How do we find that out? First, we will have to one more important point thus we have 2 variance estimate 1 in numerator, 1 in denominator; now which one will put in a numerator, which one will put in the denominator, this choice of which variance estimate to be placed in a numerator is somewhat arbitrary.

Hence the table of probabilities of the F distribution always assumed that a larger variance estimate in the numerator, the table in the standard textbook the table that we have always it puts with a larger variance in the numerator. But this is arbitrary, like you can put anything. If you can calculate the value, you can calculate yourself as well.

(Refer Slide Time: 08:23)

The F-Distribution

Writing $f_{\alpha}(v_1, v_2)$ for f_{α} with v_1 and v_2 degrees of freedom, we obtain

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_{\alpha}(v_2, v_1)}$$

So, as I told you, you just know the value of f of alpha that means this portion. Similarly, if I want, there is a way if I can find out this portion as well, how do I find out this is there is a

theorem for that this is theorem, f of $1 - \alpha$ if I know this value corresponding to area of 0.05. This is an area of probability means what this is nothing but this area, is not it? If I know the value of f a corresponding with a probability of 05, I can also know corresponding to the value of $1 - 0.05$ that is 0.95, I can also know the value of f of 0.95.

But maybe that means this whole area, how do I find out this is the formula? But see, there is a difference f of $1 - \alpha$ v 1 v 2 $1 / f$ of alpha v 2 v 1 , whatever I am using numerator that gets change when I am doing $1 - \alpha$ numerator becomes denominator, denominator becomes numerator, we will be solving some problems then it will be things will be more clear.

(Refer Slide Time: 09:27)

The F-Distribution

Writing $f_{\alpha}(v_1, v_2)$ for f_{α} with v_1 and v_2 degrees of freedom, we obtain

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_{\alpha}(v_2, v_1)}$$

Thus, the f-value with 6 and 10 degrees of freedom, leaving an area of 0.95 to the right, is

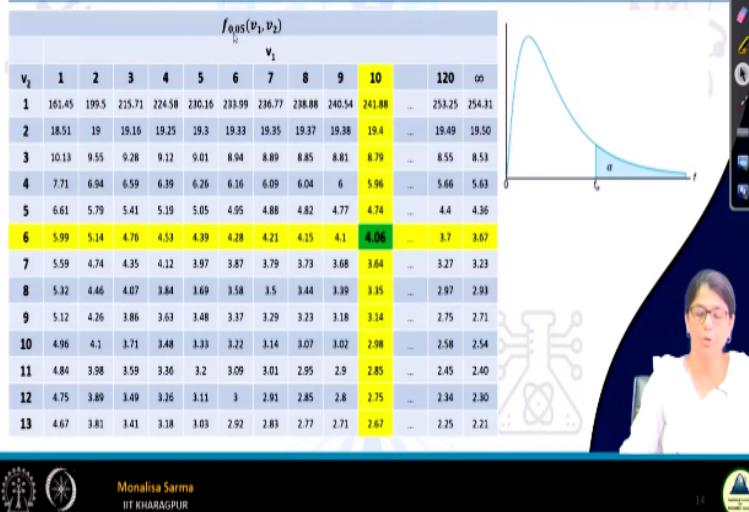
$$f_{0.95}(6,10) = \frac{1}{f_{0.05}(10,6)} = \frac{1}{4.06} = 0.246$$

Monalisa Sarma
IIT KHARAGPUR

So, if thus the f value which 6 and 10 degrees of freedom having an area of 0.95. So, I suppose we are interested in the area of 0.95 total, so to find an area of 0.95 what we will be doing, so 0.95 where we do not have in the F table, it is not given in the F table. So, what we will do I am interested in this I will calculate this $f_{0.05}(10, 6)$ from the table I can get this value is not it? This is 0.05 values I will get in a table on the table I can get this value and 1 by of that I will give me this value.

(Refer Slide Time: 10:07)

The F-Distribution



So, here we have seen F of 0.05 this is 10 this is 6 numerator is v_1 denominator is v_2 I got this is 4.06. So, then my F of 0.95 6 10 will be $1 / 4.06$ that is 0.246 this value 0.246.

(Refer Slide Time: 10:34)

The F-Distribution: Example – 1

Problem

Consider the following measurements of the heat-producing capacity of the coal produced by two mines (in millions of calories per ton):

Mine 1: 8260 8130 8350 8070 8340 ✓ 9
Mine 2: 7950 7890 7900 8140 7920 7840 ✓ 5

Can it be concluded that the two population variances are equal?



So, how we will do from will to take a simple example, as I told you F distribution we use to compare 2 different variance of 2 different population this is the example, consider the following mechanism of the heat producing capacity of the coal produced by 2 mines that is in millions of calories per ton. So, this is one first mine, this is a second mine how many sample size this is 1 2 3 4 5 from the first population, we have taken a sample size of 5.

And from the second population 1 2 3 4 5 6 from the second population, we have taken a sample size of 6. So, degrees of freedom here the degrees of freedom is 4, here the degrees of freedom is 5 can it be concluded that the 2 population variances are equal? So, see what it is asking can it be concluded that the 2 population variances are equal that means, we are

estimating we are considering that the 2 population variances are equal, if the 2 population variances are equal than this data what we got from this data we will be getting a variance.

So, whatever this variance, the difference of these 2 variances whether is it, it has a higher probability basically that we need to find out, is not it? If we compare the variance of these 2 population and if we find out considering both the population variance is the same, if we find out the probability is very less that means, we can conclude that the 2 population variances are not equal, if we find a population is quite high, we can say yes, the population variance is equal like whatever we have done for all the other same type of things.

(Refer Slide Time: 12:14)

The F-Distribution: Example – 1

Solution

From previous discussion, we know, $f = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{\sigma_1^2/s_1^2}{\sigma_2^2/s_2^2}$

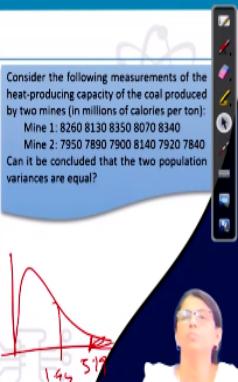
Considering the two population variances are equal, therefore, $\frac{\sigma_1^2}{\sigma_2^2} = 1$

Here, $s_1^2 = 15750$ and $s_2^2 = 10920$ which gives $f = 1.44$

Now, from Table $f_{0.05}(4,5) = ?$

Consider the following measurements of the heat-producing capacity of the coal produced by two mines (in millions of calories per ton):
 Mine 1: 8260 8130 8350 8070 8340
 Mine 2: 7950 7890 7900 8140 7920 7840

Can it be concluded that the two population variances are equal?



A photograph of a woman, Monalisa Sarma, speaking into a microphone.

Monalisa Sarma
IIT KHARAGPUR

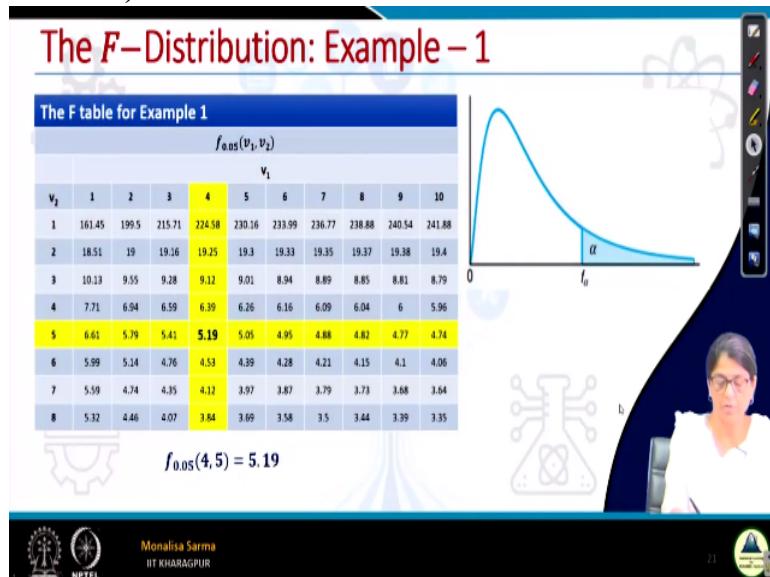
So, from the previous discussion, we know this is the f value and we are considering that the both the population variances are equal, population variance is equal means, $\sigma_1^2 / \sigma_2^2 = 1$. So, from the sample given data we can calculate the variance so, you can calculate and see by yourself just S_1^2 I can founded this S_2^2 founded this, then putting this in the numerator putting this in the denominator, then I got $f = 1.44$.

If when I put this upper one because higher value in a table, it is assumed that it will keep the higher value in that numerator. If I keep in higher this in the numerator, what are the degrees of freedom for here, it is 4. So, that means, I need to find out f of 1.44 for degrees of freedom 4 and 5, is not it? So, I am interested in finding out for a value of f 1.44 for degrees of freedom 4 and 5, but the f value also.

It is as I told you it is only given for some very limited value most of the books you will find that this value is given only for 0.05 and 0.01. So, just seeing the value 1.45 only we can very well make out that it is very near to 0. So, probability of that will be definitely a bigger probability, but still let us see what we have because we have only 0.5 that is 5% probability and 1% probability.

First let us check for 5% probability, what is the f value corresponding to the 5% probability first you see the figure this is the figure. So, from the figure you will be able to make it say this is if I consider this is the figure. So, this is for 5% what is the value?

(Refer Slide Time: 14:08)



Let me find out what is the value for this say for 5% for degrees of freedom 4 and 5 it is 5.19 for that means this is 5.19 and my value calculated value f is 1.44. So, value is 1.44 is definitely to the left of this may be somewhere in this 1.44 and definitely this means this area is greater than 0.05 that greater than 5% when this probability much greater than 5% then it is obvious that whatever this is predicting that the both of the population parents are equal that is a reasonable estimate.

When we get a very less probability much lesser than 5%, much lesser than 1%, then we can say, there is no probability. But we are getting definitely more than 5% is not it? But we do not have the F distribution value to find out what is the exact probability for 1.44 but we can understand when this value 5.19 will be somewhere here and 1 will be somewhere here definitely this area will be much more.

(Refer Slide Time: 15:12)

The F-Distribution: Example – 1

Solution

From previous discussion, we know, $f = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{\sigma_1^2 s_1^2}{\sigma_2^2 s_2^2}$

Considering the two population variances are equal, therefore, $\frac{\sigma_1^2}{\sigma_2^2} = 1$

Here, $s_1^2 = 15750$ and $s_2^2 = 10920$ which gives $f = 1.44$

Since, from Table $f_{0.05}(4,5) = 5.19$, the probability of $f > 1.44$ is much bigger than 0.05, which means the two variances can be considered as equal.

Consider the following measurements of the heat-producing capacity of the coal produced by two mines (in millions of calories per ton):
Mine 1: 8260 8130 8350 8070 8340
Mine 2: 7950 7890 9000 8140 7920 7840
Can it be concluded that the two population variances are equal?

Monalisa Sarma
IIT KHARAGPUR

So, since from table the probability of $f 1.44$ is much bigger than 0.05 which means the 2 variances can be considered as equal. So, now, we will be discussing the next topic that is the sampling distribution of the sample proportion this is the last topic on the sampling distribution.

(Refer Slide Time: 15:38)

Sampling Distribution of the Sample Proportion

- In a random sample of 200 parts from a manufacturing process, 18 had major manufacturing defect.
- Here, \hat{p} the sample proportion
- $$\hat{p} = \frac{18}{200}$$
- \hat{p} represents the proportion of the individuals or objects in the sample that has a certain characteristics.
- In statistical inference scenario, we use the sample proportion \hat{p} to estimate the population proportion p

Monalisa Sarma
IIT KHARAGPUR

So, first to start before starting sampling distribution and sample proportion let us take an analogy are very common feature which we get to see during election time exit poll you know what is an exit poll and exit poll what happens? We from the exit poll they try to predict which party will come to power or maybe which candidate will win the election. So, what is then so, in a locality suppose a particular political party affiliation or suppose there are 3 political affiliation party x, y and z 3 different parties.

So, in the election if we want to know which party will win let us make it simple a 2 party x and y I am just making it simple. So, which party will win the election? So, what is the exit poll what we do is that do we interview from we know exit poll means do we interview the people after when they both cast their vote and come out then we sort of take the information from them and accordingly based on that we give the information.

But is it that we try to find out from all the people, no we do not try to it is not possible many term lakhs and lakhs of people come and cast your vote we do not try to find out the information from all those lakhs and lakhs of people, but just we take a small sample of those lakhs and lakhs of people from the whole population we take a small people from a subset of this people, but it has to be unbiased, it should not be biased, unbiased.

In the sense maybe we have taken the feedback maybe some from different age category, different profession, different locality, that that may be an unbiased estimator, we try to take out the thing what to say there which political party they have cast a vote and based on whatever result we get, based on that we try to infer about the election result that means, based on that, we try to infer what the whole population might have voted, is not it? That is the exit poll.

So, now what is that is means is a big population from the population we want to infer something about the population we cannot do that. So, when we take a sample of the population that means we take a proportion of that, from that we try to infer about the population that is for that we need to do again this proportion is a random variable there can be different values if we take different samples there can be different values.

So, random variables means there will be different frequency of observation meaning it will also have a probability distribution that probability distribution is sampling distribution of sample proportion. With that diagram, let us now see. So, in a random sample of 500 parts from a manufacturing process 18 has major manufacturing defect. 2 from 200 parts 18 had major manufacturing defects it is that is a sample again, that is not a population. So, let me tell sample proportion, which how much proportion is defective?

How much proportions of the people have caused their 4 for political party x? That proportion I am telling it as a p cap so my p cap is $18 / 200$. So, what this p cap represents? p

p_{cap} represents the portion of the individuals or objects in the sample that has a certain characteristic here, what is the characteristic that p_{cap} has, that is the portion the defective portion of the whole sample as it poll the portion maybe the portion of the population who have voted for political party x or political party y that is p_{cap} .

In statistical inference scenario, we use the sample proportion p_{cap} to estimate the population proportion p . We now from this sample, we have to estimate the population proportion from the exit poll we have to make an estimate of the whole population. The whole population may have voted for which political party which person is basically may have voted for which political party, so that is from this p_{cap} we have to find out estimate of the population proportion let me take population proportion.

Let me take that I have written it is p . So, from p_{cap} I have to estimate for p . Now for doing that, I need a sampling distribution of p_{cap} , sampling distribution p_{cap} means I need to know p_{cap} represent what sort of distribution and accordingly whatever distribution I need to know the characteristics of the distribution, if it is a normal distribution then I will be needing to know the mean and the variance of the distribution.

Similarly, whatever distribution is this I need to know the characteristics of the distribution then only once I have the characteristics of the distribution then only I can use the distribution to evaluate the probability of a particular occurrence. So, now, I have to first find out what is the p_{cap} we will have what distribution then once we know that people what distribution then accordingly what are the different parameters of this distribution, we need to find out the values of those parameters.

(Refer Slide Time: 20:47)

Characteristics of the sampling distribution of \hat{P}

- Let us assume we are sampling from an infinite population, or rather we are sampling a small fraction from a large population.
- We can view the sample proportion as \hat{P} , where $\hat{P} = \frac{X}{n}$
where
*X is a random variable that represents the no of individuals in the sample with the characteristic of interest,
and n represents the no of individuals in the sample.*

Monalisa Sarma
IIT KHARAGPUR

So, in that angle first we will see, you know, what we will assume? Let us assume that we are sampling from an infinite population or rather we are sampling a very small fraction of the last population. So, we can view the sample proportion \hat{P} as X / n what is X ? X is a random variable that represents a number of individual in the sample with the characteristics of interest they are what is X and my example X is the number of defective components.

The characteristics of interest means here are characteristics of interest is defective component, characteristic of interest is people who have voted for X political party X and n represents the number of individual in the sample total sample what we have taken that is so, \hat{P} is X / n .

(Refer Slide Time: 21:43)

Sampling Distribution of the Sample Proportion

Characteristics of the sampling distribution of \hat{P}

- Let us assume we are sampling from an infinite population, or rather we are sampling a small fraction from a large population.
- We can view the sample proportion as \hat{P} , where $\hat{P} = \frac{X}{n}$
- Here, X is a discrete random variable, and the value it can take is $0, 1, 2, \dots, n$; i.e. $(n+1)$ possible values.
- X can be thought as a binomial random variable with parameter n and p
- Recall that the binomial random variable X has a
 - mean = np
 - variance = $np(1-p)$
 - It is approximately normally distributed for large sample size

Monalisa Sarma
IIT KHARAGPUR

So, now X what is X here? That is very important now, X is very much a discrete random variable it X cannot take any values in an interval, it will only take some discrete values is

not it? So, X is definitely not a continuous random variable it is a discrete random variable and find it is a description or we will not watch sort of district random variable and what the value it can take 0, 1, 2 up to n what value 0 X can take?

X can take value 0 no defective component that means p cap is $0 / n$, 1 defective component in a sample $1 / n$, 2 defective components in a simple $2 / n$ and n defective component in a simple n / n . So, it can take total $n + 1$ values, is not it? Now, this X cannot we see that this X is very much a binomial random variable is not it? Because what does X indicates either the presence of the characteristics or the absence of the characteristic, in binomial random variable X what is that X ? X is either what is X ?

X is the total number of failure or the total number of heat, total number of miss, whatever it is success or failure heat miss, whatever it is like here similarly, X is the thing whether it contains the desired characteristics what is X here? Characteristics of interest. So, what is my characteristic of interest in our case in our example, if needed a defective and what value it can take? It can take 0, 1, 2 like the probability of 0 success, probability 1 success, probability 2 success.

Similarly, probability of 0 defective, probability of 1 defective, probability of 2 defective. So, X can take this total number of defective in n trials and n is the number of trials in total n trials how many defectives can be there, is not it? And it is probability of each trial getting a defective in each trail is independent there is no any dependency in it yes or no? So, we can very well say that X is a discrete of course it is this thus the X is a binomial random variable.

When it is a binomial random variable, then we need to know binomial there are 2 parameters n and p what is n sample size, p is the probability of success or the probability failure whatever it is, so, we know n we know p , X can be thought as a binomial random variable parameter n and p . Now, this is X now what is p cap? p cap is X / n let us go there later. Now binomial random variable X as parameter n and p what is the mean of this?

Mean is np we know mean of binomial random variable is np variances npq let us $np \times 1 - p$ and moreover, binomial interval it is approximately normally distributed for large sample size. If we take a larger sample size it is approximately normally distributed we have seen this well we have discussed binomial distribution know when we have discussed normal

distribution then we have mentioned it even binomial distribution also when the sample size is larger number of trials is very large.

Instead of considering binomial distribution we consider normal because it is almost same it can be approximated. So, if sample size is large it can be approximated as normally distribution, we know what is mean we know what is variance?

(Refer Slide Time: 25:09)

Sampling Distribution of the Sample Proportion

Characteristics of the sampling distribution of \hat{p}

- Now to derive the characteristics of the sampling distribution of the sample proportion \hat{p} :
- The mean of the sampling distribution $= E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n} \times np = p$
- And so, we can say on an average the sample proportion equals the population proportion.
- The variance of the sampling distribution of \hat{p}

$$= \sigma_{\hat{p}}^2 = Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$
- The standard deviation $= \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Monalisa Sarma
IIT KHARAGPUR

Now, we need to derive the characteristics of p cap, we understood what is X ? X we can consider it a binomial random variable, we know the different mean and variance parameter as well. And we know that X can be remember if n is large X can be approximated by normal distribution. Now, what is p cap see p cap is what? p cap is X / n , p cap is nothing but X / n . Now, what values p cap can take? p cap can take $0 / n, 1 / n, 2 / n, 3 / n$ there is also again discrete value.

Again p cap also cannot take any continuous as well a p cap also can take this discrete the values are $0 / n, 1 / n, 2 / n, 3 / n$ and n / n that is a valid p cap can take similar to the values that extremity and the probability that p cap will take $0 / n$ is same as the probability that X will take value 0 probability that p cap will take $1 / n$ value is the same as the probability that X will take value 1.

So, say p cap is also when X is a binary random variable n is a fixed with a constant value, then definitely p cap also we can consider that it is a binomial random variable. So, if p cap is a binomial random variable now that means, we need to know the characteristics of n and p

that means the mean and the variance of the p cap. So, mean of the sampling distribution E of p is what for this p cap? p cap is X / n we have initially when we have learned a mean of variable random variable.

So, if there is a constant number comes out mean of X / n is nothing but $1 / n E$ to the power X and E to the power X we have already seen it as np so, mean of p cap is p see, mean of p cap is p that means, we can say on an average the sample proportion equals the population proportion so, on an average the sample proportion equals the population proportion. Now, the variance of the sampling distribution of p cap is where of p cap nothing but we have X / n .

X remember where of X / n when there is a constant this is $1 / n^2$ of X. So, where of X is what npq then we got is the variance. Here also since p cap is again and we can consider is a normal random variable if the size of N is large, we can p cap as a binomial random variable, and it is the size of n is large we can consider it as a normal random variable.

(Refer Slide Time: 28:09)

Sampling Distribution of the Sample Proportion

Characteristics of the sampling distribution of \hat{p}

- ④ The sampling distribution of \hat{p} is approximately normal if the sample size is large
- ④ So, finally we can say, for large sample size, the sampling distribution of \hat{p} is approximately normal, represented as

$$N \left(p, \frac{p(1-p)}{n} \right)$$

The slide features a background image of a scientist in a lab coat and a computer monitor icon. The footer includes the NPTEL logo, the name 'Monalisa Sarma', and the text 'IIT KHARAGPUR'.

So, therefore, and we can write so, sampling distribution of p cap is approximately normal if the sample size is large and this is how we can represent, the sampling distribution of p cap is approximately normal distribution we need mean and variance. So, this is mean, this is variance so, we have the sampling, we have the distribution, we have the mean we are the variance now, we can calculate a probability.

(Refer Slide Time: 28:40)

Sampling Distribution: Example – 2

Problem

It is believed that 20% of voters in a certain city favor a tax increase for improved schools. If these percentage is correct, what is the probability that in a sample of 250 voters 60 or more will favor the tax increase?



Monalisa Sarma
IIT KHARAGPUR

40



When we will do 1 example before complete this topic. So, it is believed that 20% of the voters in a certain city favour a tax increase for improved school it is believed that is estimated it is predicted 20% of the voters in a certain city favour a tax increase. So, that is the population proportion is given what is the proportion is 20%, if this percentage is correct, what is the probability that in a sample of 250 voters 60 or more will favour the tax increase. If this is correct, what is the probability of this happening? So, my p cap is what 60 / 250 I have to find out probability that of p cap greater than 60 / 250.

(Refer Slide Time: 29:21)

Sampling Distribution: Example – 2

Solution

The sampling distribution of proportion is approximately normal with mean $\mu_p = 0.2$

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.2 \times 0.8}{250}} = 0.025$$

$$\begin{aligned} \therefore P\left(\hat{p} > \frac{60}{250}\right) &= P\left(Z > \frac{\frac{60}{250} - 0.2}{0.025}\right) \\ &= 1 - P(Z < 1.6) \\ &= 1 - 0.9452 \\ &= 0.05 \end{aligned}$$

It is believed that 20% of voters in a certain city favor a tax increase for improved schools. If these percentage is correct, what is the probability that in a sample of 250 voters 60 or more will favor the tax increase?



Monalisa Sarma
IIT KHARAGPUR

41



For that first I found out first I know this is a normal distribution first I need to know the mean I need to know the standard deviation what will be my mean? Mean is on an average we have seen here on and every sample proportion equals the population proportion, is not it? So, my mean is 20%. So, my mean of p cap is 20% that is the population proportion. My

variance this is the formula for variance and standard deviation is the formula for standard deviation.

Then what I need to find out p cap is greater than this is the probability I need to find out what is the probability given, what is the probability that in a sample of 250 voter 60 or more will favour a tax increase? So, p cap greater than $60 / 250$, normal distribution that means we will have to convert it to Z, because we do not have the normal table lookup table here, the Z lookup table, so we have converted it to Z probability of Z. So, what is this, Z is what? X bar - μ / σ .

So, what is X bar here X bar is nothing but p cap my X bar value is nothing but a p cap here. So, this is p cap -0.2, this is the σ value and what I got is 1.6. So, probability of Z greater than 1.6 is $1 - \text{probability of } Z \text{ less than } 1.6$. 1.6 and this from the table, I am not showing it I have to show it many places from the Z table I can find out Z probability corresponding to Z less than 1.6 this probability. So, this is the probability. If that is correct probability then a sample of 250 voters 60 or more will favour the tax increase that probability is 5% probability.

Now, it is up to you whether you will accept that what it is believed that 20% of voters is correct or not if you find this probability is very small, then we will say that no whatever is believed it is not correct, if you find that there is this 5% this may be some variation, then we can consider that.

(Refer Slide Time: 31:29)

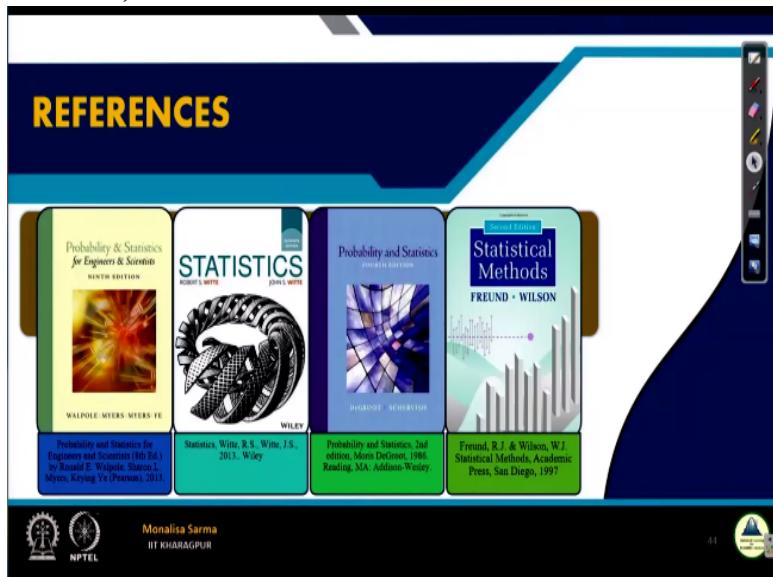
The slide has a dark blue header with the word "CONCLUSION" in yellow capital letters. Below the header is a teal bar. The main content area is white. At the top left, it says "In this lecture we learned about". Below that is a bulleted list:

- ⌚ F –distribution
- ⌚ Sampling distributions of sample proportions

On the right side of the slide, there is a video feed of a person with glasses and a white shirt, sitting at a desk and gesturing with their hands. The video feed is framed by a black border. In the bottom left corner of the slide, there are two logos: NPTEL and IIT Kharagpur. In the bottom right corner, there is a small circular logo with the text "NPTEL" and "IIT KHARAGPUR".

So, with this I conclude this topic on sampling distribution. In this lecture, we have learned distribution and sampling distribution of sample proportion. So, here I conclude the portion on sampling distribution. So, sampling distribution, we have learned Z distribution, sampling distribution of mean using both Z distribution and T distribution, sampling distribution of variance using chisquare distribution, sampling distribution for what is a combination of 2 population variance, the product we have used F distribution, then again we have seen sampling distribution of proportion.

(Refer Slide Time: 32:10)



These are my references and thank you guys.

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

Lecture - 21
Tutorial on Sampling Distributions

Hello everyone. So, today we are going to end the topic on sampling distribution like last lecture, I already told that we have covered almost everything of sampling distribution, I should not say everything I have covered, but then whatever is necessary at this point. So, I have covered those topics and today I will end those classes basically with a tutorial that is a tutorial on sampling distribution. And after that, I will be starting a new topic.

(Refer Slide Time: 00:54)

The slide has a dark blue header bar with the title 'Concepts Covered' in white. Below the title, there is a list of objectives:

- ④ Solving objective type questions
- ④ To test the level of understanding from Lecture 16-20

- ④ Problems to ponder
- ④ To build problem solving aptitude

On the right side of the slide, there is a video feed of a woman with glasses and a white shirt, who appears to be the professor. At the bottom of the slide, there is a footer bar with the IIT Kharagpur logo and the name 'Monalisa Sarma'.

So, as usual in a tutorial, I always try to keep few objective questions, which will basically help you to take a quick recap of whatever we have learned. So, and then we will be doing few problems as well.

(Refer Slide Time: 01:12)

Problem-6.1

T 6.1: State true or false

- a) The t – distribution is used as the sampling distribution of the mean if the sample is small and the population variance is known. [False]
- b) The standard error of the mean increases as the sample size increases. [False]
- c) The sampling distribution is used to describe the variability of sample statistics. [True]
- d) The sample mean is a reliable estimate of the population mean for populations with larger variances. [False]



So, coming to the objective type questions, it is simple I do not have much objective question in this tutorial was there some simple true or false statement? So, first question here is the t distribution is known as the sampling distribution of the mean if the sample is small, and the population variance is known, so, what it is given? It is given that, so, let me take the pen, so, it is given that sampling distribution of the mean.

So, for sampling distribution of a mean we use 2 distributions if you can remember that one is t distribution and one is z distribution, when we use t distribution and when we use z distribution? We use z distribution when it is when the population variance is known, and we have to estimate about the mean of the population and population variance is known and we also have an estimated value of the population mean then we use z distribution on the contrary when the population variance is not known, which is actually more practical.

Knowing the population variance is not very much possible they actually so, when the population standard deviation whether population variance is not known, then we use t distribution. So, this statement is false. Second statement the standard error of the mean decreases as the sample size increases. So, while in a sampling and what is the central limit theorem do you remember when we have done central limit theorem? So, what was that central limit theorem what does it say?

It says the mean of the sampling distribution is the mean of the population and the variance of the sampling distribution is given by σ^2 / \sqrt{n} right this was the variance of the sampling distribution of mean. So, what does this variance indicate? Variance indicates if this is the variance sorry it is variance is σ^2 / n not \sqrt{n} . So, if I this is whole square is variance and if I removed a square and this is the standard deviation, what the standard deviation means?

Here in this case, standard deviation is basically the standard error of the mean, means how much the mean varies among themselves if you take mean of different samples, how much the mean of different samples varies among themselves that is, we call it a standard error among the means. So, the standard error of the mean increases the sample size increases. So, what happens this is the standard error. So, if the sample size increases that means, if n increases then what happens the denominator is higher than the result will be lower.

So, the standard of the error of the mean it will decrease when the sample size will increase. So, this is again a false statement. Third, the sampling distribution is used to describe the variability of sample statistics. That is, of course, that is the reason why we use the sampling distribution. So, represent the variability of the sample statistics different sample will have if we take from the same population if we take different sample all the sample will have the same value for the statistic it is quite unlikely.

So, the sampling distribution is used to describe the variability of the sample statistics this is a correct statement. Then, the sample mean is a reliable estimate of the population mean for population with larger variances. Same similar to question number b, the sample mean is a reliable estimate of the population mean for population with larger variance. So, for population with again same thing is what will be the standard error? The standard error will be σ / \sqrt{n} .

So, if this factor is more, definitely the standard error will be more if the standard error is more than on every sample mean is equals to the population mean that is not a very much a reliable estimate. So, sample mean is a reliable estimate of the population mean for population has larger variance that statement is false. So, that is all from the objective side. Then we will be solving few problems.

(Refer Slide Time: 05:41)

Problem-6.2

T 6.2: A production company making machined auto parts, checks the consistency of its dimensions by sampling 15 auto parts and found the standard deviation of those parts, $s = 0.0125 \text{ mm}$. If the allowable tolerance of these parts is specified so that the standard deviation may not be larger than 0.01 mm, we would like to know the probability of obtaining that value of S (or larger) if the population standard deviation is 0.01 mm.

F.Y. $n = 15$

Monalisa Sarma
IIT KHARAGPUR

So, first problem you will see a production company making machine auto parts, checks the consistency of its dimensions by sampling 15 auto parts and found a standard deviation of those parts. So, standard deviation of those part is given, how many parts we have taken a sample sizes that means my $n = 15$, sample size I have taken 15 and of this 15 the standard deviation is given when it is as that means it is talking about the sample.

If it is population, I would have written as σ . So, then the standard deviation of the sample is given this value 0.0125mm if the allowable tolerance of this part is specified, so, that the standard deviation may not be larger than 0.01mm. So, what we want is that our requirement is that the standard division should not be larger than 0.01mm we would like to know the probability of obtaining that value of S or larger if the population standard division is 0.01mm.

So, given if the population standard deviation is 0.01mm, if this is the population standard deviation, what is the probability that if we take a sample from the sample standard deviation will be this value that is what we need to find out. If you go through the question, you will understand it is basically this asking that if the population standard deviation is this what is the probability from a sample we will get this much standard division. So, it is talking about variance.

So, that means, we have to infer about the population variance, because this is something what we got from the sample this cannot be wrong, we have collected the sample we have calculated it and we have got this value. So, this value cannot be wrong, but we are not sure about this value. This value may be wrong this value may be correct. We do not know about this, we have just estimated this. Now, from this value, we will find out if that means what we will assume that this is correct, if this is correct, what is the probability of getting this.

If the probability of getting this in the probability of getting this is very less that means our assumption is wrong means we are trying to prove it the other way basically. So, for this remembers what we have done, this is the inference of a population variance that means, we will be using chi square distribution. So, what was the statistic? $n - 1 s^2 / \sigma^2$.

(Refer Slide Time: 08:08)

Problem-6.2 : Solution

The statistic to be compared to the χ^2 distribution has the value

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{14 \times 0.00015625}{0.0001} = 21.875$$

T 6.2: A production company making machined auto parts, checks the consistency of its dimensions by sampling 15 auto parts and found the standard deviation of those parts, $s = 0.0125$ mm. If the allowable tolerance of these parts is specified so that the standard deviation may not be larger than 0.01 mm, we would like to know the probability of obtaining that value of S (or larger) if the population standard deviation is 0.01 mm.

v	α								
	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01
1	0	0	0	0	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.1	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.3	0.48	0.71	1.06	7.78	9.49	11.14	13.28
11	2.6	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.4	5.23	6.3	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.64	6.57	7.79	21.06	23.68	26.12	29.14
15	4.6	5.23	6.28	7.26	8.55	22.31	25	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.3	28.85	32
17	5.7	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41

Monalisa Sarmas
IIT Kharagpur

So, the statistics will compute is chi square $n - 1 s^2 / \sigma^2$. So, this is $n - 1 14 s^2 / \sigma^2$ value is given σ^2 value so, we got is 21.875. Now, we will find out what is the probability of getting this value 21.875 remember chi square always the value is towards the unlike this normal distribution what we get? We get cumulative distributed value, but here we get is the area towards the right.

So, if suppose, this is my 21.875 this value this portion corresponds to this 21.875 then what is this area basically, and what is the degrees of freedom in chi square distribution parameters is degrees of freedom. So, what is my degree of freedom? Degree of freedom is 14. So,

corresponding to 14 I am trying to find out for what probability I will be getting this value probability means this area here this areas are given for degrees of freedom from 14 for what probability I will be getting this area?

But that area is not available in the table what is available here? You see here I am getting a value of 21.06 and after that, it is 23.68 that value is not there. That means from here we can find it out if this is 21.06 that means this one let me first proceed and also it properly. So, if this value is suppose this value is 21.06 and this value is suppose 23.68. So, my value will be somewhere here, is not it? Somewhere in between this from 21 to 23.

So, that means, corresponding to 23 what is my area what is my probability is 0.05 corresponding to 21.06 my property is 0.1 that means, my probability will be when between 0.1 and 0.05. So, that is the probability it is smaller than 0.01 but greater than 0.05 that is my probability.

(Refer Slide Time: 10:25)

Problem–6.2 : Solution

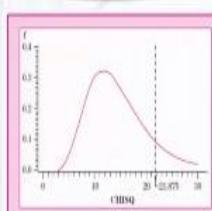
The statistic to be compared to the χ^2 distribution has the value

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{14 \times 0.00015625}{0.0001} = 21.875$$

The desired probability is the area to the right of that value.

ν	α								
	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01
1	0	0	0	0	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.1	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.3	0.48	0.71	1.06	7.78	9.49	11.14	13.28
11	2.6	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.4	5.23	6.3	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.08	26.11	29.14
15	4.6	5.23	6.26	7.26	8.55	22.31	25	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.3	28.85	32
17	5.7	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41

T 6.2: A production company making machined auto parts, checks the consistency of its dimensions by sampling 15 auto parts and found the standard deviation of those parts, $s = 0.0125$ mm. If the allowable tolerance of these parts is specified so that the standard deviation may not be larger than 0.01 mm, we would like to know the probability of obtaining that value of s (or larger) if the population standard deviation is 0.01 mm.



Monalisa Serma
IT BHARATPUR

(Refer Slide Time: 10:28)

Problem-6.3

T 6.3: Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute.



So, next question travelling between 2 campus of a university in the city via shuttle bus takes on average 28 minutes with a standard deviation of 5 minutes, this is something it is this estimated or we can say from our previous knowledge from a past knowledge we can say travelling between 2 campus it takes one and a average 28 minutes with a standard deviation of 5 minutes. In a given week, the bus transported passengers in 40 times that means my n is 40 here. What is the probability with that average transport time was more than 30 minutes.

In a week, the bus transported passenger, 40 times bus as shuttle it to and from and then average transport time was more than 30 minutes. As in the mean time is measured to the nearest minute. This say assume the mean time is measured to the nearest minute, that means measure to the nearest minute means when I have means specified is more than 30 minutes means if any of my journey, if I got is 30.1 then I will write it 30 only 30.2 I will still at 30, 30.3 I will still write it 30 till 30.5 I will write it 30.

Above 30.5 I will write it 31 minutes. So, it is nearest to the nearest minute means still 30.5 I will write it 30 minutes. So, essentially, when it is asking what is the probability that every transport time has more than 30 minutes. So, essentially I need to find out what is the average transport time what is the probability that the average transport time is more than 30.5? Because 30.5 also we will write it as 30 minutes because we are measuring it to the nearest minute.

Now what we have to find out here? Let us move this to the nearest minute the surrounding and follows the different and simple issue. Now coming to the main question what we need to find out here, but it is given so an average is 28 minutes standard deviation is 5 minutes. So, it is we have to infer about the mean 28 is the mean and 5 is the standard deviation. So, we have taken a sample, sample is doing to and from 40 times and from there we got it on an average 30 minutes.

(Refer Slide Time: 12:50)

Problem-6.3 : Solution

Given $\mu = 28$ and $\sigma = 5$.
Need to find $P(\bar{X} > 30)$ for $n = 40$.

As the time is measured on a continuous scale to the nearest minute, an $(\bar{x} > 30)$ is equivalent to $(\bar{x} \geq 30.5)$.

$$\begin{aligned} P(\bar{x} > 30) &= P\left(\frac{\bar{x} - 28}{\frac{5}{\sqrt{40}}} \geq \frac{30.5 - 28}{\frac{5}{\sqrt{40}}}\right) \\ &= P(Z \geq 3.16) \\ &= 1 - P(Z < 3.16) \end{aligned}$$

T 6.3: Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute.

Monalisa Sarma
IIT Kharagpur

So, what we have to find? μ is 28 given σ is 5 given means this is estimated this is if this is known only why at all we are doing all this is not it? These values are not known this or we are just what to say predicting it or estimating it. I do not want to use the word hypothesize. Basically we are hypothesizing that term hypothesis this value, I will be using this word after this lecture on what basically. So, this is we are just let me use the term estimate. Now we are just estimating or we are just guessing it of that estimation.

Now from the sample whatever we got, I am repeating this again and again from the sample what we get that is the actual value. So, from the sample basically, we need to find out assuming this is correct, whatever we got from the sample, what is the probability of that? If that probability is very less that means our assumption is wrong, same thing in all the questions. So now, since here, we have to find out the mean, we have to infer about the population mean, and our standard deviation of the population is given.

So, it is a straightforward question, we will just have to use z distribution. So, probability of this is \bar{x} greater than 30, mean is \bar{x} greater than 30.5. When I am telling actually greater than 30, that means actually, I am actually implying \bar{x} greater than 30.5, because I am nearing it to the nearest minute. I am measuring it to the nearest minute. So, what is this? What is the formula for \bar{z} for sorry, so z is equals to basically $\bar{x} - \mu / \sigma \sqrt{n}$.

So, my \bar{x} is 30.5, μ is 28. What is my $\sigma \sqrt{n}$? $5 / \sqrt{40}$ the root of a 40. So, this is what probability of Z greater than equals to 3.16. z distribution always I get it from $-\infty$ to that value. Whereas for chi distribution, t distribution we get the opposite side from that value, value to the infinite. So here, z equals to, so I will find it $1 - P(Z < 3.16)$ at least, probability that Z is less than 3.16. So this value we can get it from the table this I have not shown here the table I have shown it many times.

(Refer Slide Time: 15:08)

Problem-6.3 : Solution

$$P(\bar{x} > 30) = 1 - P(Z < 3.16)$$

From the Z-score table, we get $P(Z < 3.16) = 0.9992$

$$\therefore P(\bar{x} > 30) = 1 - 0.9992$$

$$= 0.0008$$

So there is a slim chance that the average time of one bus trip will exceed 30 minutes.

T 6.3: Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute.

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5	0.51996	0.52098	0.51997	0.51998	0.51999	0.52092	0.52098	0.52108	0.52109
0.1	0.53195	0.54348	0.54476	0.55172	0.55567	0.55962	0.56459	0.56781	0.57142	0.57539
0.2	0.57926	0.58517	0.58706	0.59055	0.59483	0.59871	0.60259	0.60642	0.61025	0.61409
0.3	0.61791	0.62172	0.62552	0.62941	0.63307	0.63681	0.64058	0.64431	0.64803	0.65171
0.4	0.65542	0.65916	0.66294	0.66674	0.67051	0.67434	0.67714	0.68093	0.68473	0.68851
0.5	0.69285	0.69657	0.69935	0.70213	0.70491	0.70769	0.71047	0.71325	0.71602	0.71879
0.6	0.73025	0.73395	0.73664	0.73934	0.74203	0.74472	0.74741	0.75010	0.75278	0.75546
0.7	0.76765	0.77135	0.77404	0.77674	0.77943	0.78212	0.78481	0.78751	0.79018	0.79278
0.8	0.80494	0.81752	0.81918	0.82676	0.82945	0.83213	0.83481	0.83749	0.84017	0.84285
0.9	0.84223	0.84819	0.85285	0.85833	0.86380	0.86827	0.87264	0.87701	0.88138	0.88565
1.0	0.87953	0.88600	0.89247	0.89800	0.90352	0.90893	0.91435	0.91976	0.92517	0.92954
1.1	0.91682	0.92366	0.92991	0.93613	0.94236	0.94859	0.95482	0.96104	0.96725	0.97346
1.2	0.95411	0.96104	0.96796	0.97488	0.98181	0.98873	0.99565	0.99946	0.99948	0.99949

Monalisa Sarma
IIT Kharagpur

So, from the value if we can calculate it I have here I have given a table I forgot it. So, my value compared to 3.16 see how do we remember 3.1 is this then 6 is this so 3.16 this is the value this is the probability the Z less than 3.16, how is the z distribution remember the figure if I draw it this way, so, this is 0 the side is minus this side is plus so, it is 3.16 maybe somewhere in here 3.16. So, this whole area this whole area is this is 0.9992 and what I want is greater than this, I do not want less than this, I want greater than this that means, I want this portion. So, what is this portion? 1 minus of this will be given this portion.

(Refer Slide Time: 16:01)

Problem-6.4

T 6.4: A normal population with unknown variance has a mean of 20. Is one likely to obtain a random sample of size 9 from this population with a mean of 24 and a standard deviation of 4.1? If not, what conclusion would you draw?



So, next question in normal population with unknown variants has a mean of 20. Is one likely to obtain a random sample of size 9 from this population with a mean of 24 and a standard deviation of 4.1. If not, what conclusion would you draw? So, see here it is given unknown variance has a mean of 20. If this is the case, it is a population is normal. Always remember I am repeating this again chi square population, a chi square distribution and t distribution and f distribution we can use only if the parent population is normal.

So, it is mentioned here the normal population with unknown variance, variance is not known. That means somehow we can we are guessing the value mean is 20. If says this is the case, is one likely to obtain a random sample of size 9 with a mean of 24 standard deviation of 4.1. If not, what conclusion would you draw? Same we will use here, we will use t distribution because variance is not known and we have to infer about the population mean is not it? Because it is asking is it likely to obtain a random so, for with a mean of 24.

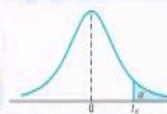
(Refer Slide Time: 17:17)

Problem-6.4 : Solution

$$t = \frac{(24-20)}{\left(\frac{4.1}{\sqrt{9}}\right)} = 2.927$$

$t_{0.01} = 2.896$, with 8 degrees of freedom

Conclusion NO; $\mu > 20$



T 6.4: A normal population with unknown variance has a mean of 20. Is one likely to obtain a random sample of size 9 from this population with a mean of 24 and a standard deviation of 4.1? If not, what conclusion would you draw?

v	0.5	0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	0	1	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0	0.816	1.061	1.386	1.886	2.92	4.303	6.965	9.925	22.327	31.599
3	0	0.765	0.978	1.25	1.630	2.353	3.182	4.541	5.841	10.215	12.924
4	0	0.741	0.941	1.19	1.533	2.132	2.776	3.747	4.604	7.173	8.61
5	0	0.727	0.92	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0	0.718	0.906	1.134	1.44	1.943	2.447	3.143	3.707	5.205	5.959
7	0	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0	0.706	0.889	1.108	1.397	1.86	2.306	2.806	3.355	4.501	5.041
9	0	0.703	0.883	1.1	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	0	0.7	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0	0.695	0.873	1.084	1.356	1.782	2.179	2.681	3.055	3.93	4.318
13	0	0.694	0.87	1.079	1.35	1.771	2.16	2.65	3.012	3.852	4.221



Monalisa Sarma
IIT Kharagpur

IIT Kharagpur

23



So, we are using t distribution. So, using the value for t distribution, what is the thing representation, I mean the formula for t distribution what the t statistics what this gives $t = \bar{x} - \mu / s / \sqrt{n}$. So, what is my s here? s is 4.1 \sqrt{n} is what is the size is 9? So, $\sqrt{9}$ is 3 so, 4.1 / 3 24 - 20 this is the value 2.927 so, 2.927 with 8 degrees of freedom because it is sample size is 9, 8 degrees of freedom 2.927, 8 is not there in a table, so, 2.896 closer value is there. So, it will be above this.

So, the probability of this will be lesser than one person less than 0.01, it will be between 0.01 and 0.005 less than 0.01. So, that means it is a very less probability. So, that means the whatever we have estimated that is this our estimation this mean of 20. So this our estimation is not correct. So, conclusion no μ is greater than 20 μ has to be greater than 20 it cannot be less than equal to 20.

(Refer Slide Time: 18:41)

Problem-6.5

T 6.5: A certain type of thread is manufactured with a mean tensile strength of 78.3 kilograms and a standard deviation of 5.6 kilograms. How is the variance of the sample mean changed when the sample size is

- a) increased from 64 to 196
- b) decreased from 784 to 49?



So, next question is certain type of thread is manufactured with a mean tensile strength of 78.3 kilograms and a standard deviation of 5.6 kilograms, mean is given this is the mean and this is the standard deviation how is the variance of the sample mean changed when the sample size is increased from 64 to 196 or decrease from this I will just one second one will also be the same. So, it is a very simple question what it is given a certain type of thread is manufacture mean tensile strength is 78.3 and a standard deviation is 5.6 kilogram

So, that means, for this for sampling distribution of mean if we draw a sampling distribution of mean what will have? Mean of the sampling distribution of mean will be this 78.3 and what will be this variance? Variance will be 5.6 divided by the sample size, is not it? So, what it is given? So, how is the variance of the sample mean when a sample sizes so, sample size initially sample size is 64 then we have seen the sample size to 196. If we do that, how the variance will change.

Remember we have already done in the objective type my second question is the same thing actually, when my sample size is increased, whatever as my variance decrease my sample sizes decrease my variance increase. So, here initially sample size is this then so, see what happens.

(Refer Slide Time: 20:09)

Problem-6.5 : Solution

a) For $n = 64$, $\sigma_{\bar{x}} = \frac{5.6}{8} = 0.7$, whereas, for $n = 196$, $\sigma_{\bar{x}} = \frac{5.6}{14} = 0.4$.

So the variance of the sample mean is reduced from 0.49 to 0.16, when the sample size is increased from 64 to 196.

T 6.5: A certain type of thread is manufactured with a mean tensile strength of 78.3 kilograms and a standard deviation of 5.6 kilograms. How is the variance of the sample mean changed when the sample size is

- a) increased from 64 to 196?
- b) decreased from 784 to 49?

b) For $n = 784$, $\sigma_{\bar{x}} = \frac{5.6}{28} = 0.2$, whereas, for $n = 49$, $\sigma_{\bar{x}} = \frac{5.6}{7} = 0.8$.

So the variance of the sample mean is increased from 0.04 to 0.64, when the sample size is decreased from 784 to 49.



Monalisa Sarma
IIT Kharagpur

IIT Kharagpur

30

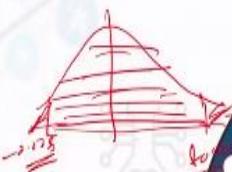


So, this is how we find out the standard deviation that is the standard error just the standard deviation of the population divided by the what is a sample size root of our sample size that is root of 64 is 8 I got 8 is 0.7 whereas, if $n = 196$ if I got variance I got 0.4. See, when the sample size increased my variance our standard error decreased that is it obviously, but we have seen in objective type question similarly, for the next question decrease from 784 to 49. So, variance will increase. So, I am not discuss this it is a solution is given there you can see.

(Refer Slide Time: 20:50)

Problem-6.6

T 6.6: A manufacturing firm claims that the batteries used in their electronic games will last an average of 30 hours. To maintain this average, 16 batteries are tested each month. If the computed t-value falls between $-t_{0.025}$ and $t_{0.025}$, the firm is satisfied with its claim. What conclusion should the firm draw from a sample that has a mean of $\bar{x} = 27.5$ hours and a standard deviation of $s = 5$ hours? Assume the distribution of battery lives to be approximately normal.



Monalisa Sarma
IIT Kharagpur

IIT Kharagpur

31



So, next question a manufacturing firm claims that the batteries used in the electronic games will last an average of 30 hours. So, this is a claim. So, they are claiming that their mean lasting time that the mean time that the game will last is 30 hours to maintain this average 16 that is tested

each month. So, whether this average is maintained at a manufacturing firm it is claiming there, but before it goes to the market, it wants to check again and again. So, what is so, it tries to check it taking 16 batteries each month.

So, how it does? If the computed t value falls between this t value is also similar to t distribution is also similar to normal distribution just to the fatter tail is not it? So, the computed t value falls between t value of 0.025 means 0.025 maybe says this is t value of 0.025 maybe this portion and t value of -0.025 maybe this portion. So, if my computed value falls in this range, that means falling in this range means, from the sample what I am getting assuming this is true from a sample what I am getting is be significant probability.

Probability is quite good, it is not a very less probability, because if it falls in this region, it falls in this region means very less probability. So, the manufacturer or the firm is satisfied if it falls within this range that means, if it is falls within this 95% of this area, it is satisfied with this claim. So, what we need to find out from the question, only it is very clear, that means we need to find out a t value first from the sample. And this our sample t values would fall within this range.

So, we will have to find out what is the t value corresponding to this area, we will have to find out what is the t value corresponding to this area, what is the t value corresponding to this area? How do we find out? This we will find out we will be able to find out from the table using what is the degrees of freedom 15 degrees of freedom. For 15 degrees of freedom what is the t value corresponding to 0.025 as I told you t distribution as symmetry and whatever we get it if we get value we have x then this will be $-x$ is not it?

(Refer Slide Time: 23:17)

Problem-6.6 : Solution

From table, we get $t_{0.025} = 2.131$ for $v = 15$ degrees of freedom.

Since the value $t = \frac{27.5-30}{\frac{5}{\sqrt{16}}} = -2.00$ falls between -2.131 and 2.131 , the claim is valid.

T 6.6: A manufacturing firm claims that the batteries used in their electronic games will last an average of 30 hours. To maintain this average, 16 batteries are tested each month. If the computed t-value falls between $-t_{0.025}$ and $t_{0.025}$, the firm is satisfied with its claim. What conclusion should the firm draw from a sample that has a mean of $\bar{x} = 27.5$ hours and a standard deviation of $s = 5$ hours? Assume the distribution of battery lives to be approximately normal.

V	0.5	0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	0	1	1.376	1.963	3.078	6.314	12.71	31.82	61.66	316.31	698.62
2	0	0.816	1.011	1.388	1.886	2.92	4.201	6.935	9.925	22.327	31.599
3	0	0.798	0.978	1.25	1.938	2.331	3.182	4.541	5.841	10.215	12.924
4	0	0.765	0.892	1.080	1.476	1.943	2.447	3.078	3.736	6.964	9.121
13	0	0.094	0.87	1.079	1.35	1.771	2.18	2.85	3.013	3.052	4.311
14	0	0.092	0.88	1.079	1.345	1.781	2.145	2.828	2.977	3.767	4.14
15	0	0.091	0.816	1.07	1.341	1.751	2.133	2.602	2.847	3.733	4.073
16	0	0.09	0.815	1.071	1.337	1.746	2.12	2.588	2.921	3.696	4.015
17	0	0.089	0.813	1.069	1.333	1.74	2.11	2.567	2.898	3.646	3.935
18	0	0.088	0.812	1.067	1.33	1.734	2.101	2.552	2.878	3.61	3.921



Monalisa Sarma
IIT KHARAGPUR

IIT Kharagpur

33



So, for 15 degrees of freedom t value of 0.025 is 2.131 that means from the sample statistics from the sample sorry from the sample the sample statistics that I will calculate if my calculated value falls within plus minus 2.131 then the manufacturing firm is satisfied. So, what is my t value t value formula you know how to calculate the t value it found is -2. So, it falls between the - 2.31 and 2.31. So, claim is valid.

(Refer Slide Time: 23:58)

Problem-6.7

T 6.7: Pull-strength tests on 10 soldered leads for a semiconductor device yield the following results, in pounds of force required to rupture the bond:

19.8 12.7 13.2 16.9 10.6 18.8 11.1 14.3 17.0 12.5 ✓

Another set of 8 leads was tested after encapsulation to determine whether the pull strength had been increased by encapsulation of the device, with the following results:

24.9 22.8 23.6 22.1 20.4 21.6 21.8 22.5 ✓

Comment on the evidence available concerning equality of the two population variances.



Monalisa Sarma
IIT KHARAGPUR

IIT Kharagpur

34



So, we have seen question on the sampling distribution of mean, sampling distribution of mean we have seen using both z distribution and t distribution then we have seen the sampling distribution of variance that means to infer about the population variance we have seen some problems on that we have also seen problems on how the standard error varies with the as regard

to the sample size. Now, what is remaining whatever we have learned of sampling distribution what is remaining?

This f distribution is remaining, f distribution and sampling distribution of proportion, sampling distribution of proportion. I do not think I have a problem here already we have discussed when we discuss 1 or 2 problems while we have discussed sampling distribution of proportion. So, now we will see f distribution. When we use f distribution, remember when we want to compare the variance of 2 different population as I had mentioned you earlier it is usually we used in the food and beverage industry.

Where food and beverage industry where there are many players for the same type of product like cold drinks, there are many products for fruit juice for say mango juice, mango juice there are different players in the market. So, this is one of for this sort of products, usually this variance and comparing the variance of 2 different populations or different chemical factories comparing the variance of 2 different products, it is very important.

So, f distribution you will see a different example here, pull strength test on 10 soldered leads for a semiconductor device is the following results in pounds of force required to capture the bond, how much strength we have to give basically to rupture the bond. So, these are the values given for sample size is 10 here for f distribution, it is not required the sample size for both the population has to be same it is not required. So, as we have already seen that n_1 and n_2 we have used 2 different terms remember.

So, here another set of 8 leads was tested after encapsulation, we have encapsulated the thing with some material maybe and then to determine whether the pull strength had been increased by encapsulation of the device, we have encapsulate the soldier lead by some sort of material after that, do we need to what to say put more stress it is expected at we need to put more stress to rupture the bond. So, increased with the following results we got the following results. Comment on the evidence available concerning equality of the 2 population variances.

But the manufacturer whatever they are claiming that both the population variances are same both the population variances are same. So, we have to comment on this whether the population variances same are not how we will come in? We will find out f value if the f value from the because these are the statistics. This is one sample, this is one sample from this 2 sample we will be able to calculate the f value and if the calculate the f value has a significant probability. Then equal the population variances are in the 2 population variances are equal, we will agree to that claim.

(Refer Slide Time: 27:23)

Problem-6.7 : Solution

$s_1^2 = 10.441 \text{ and } s_2^2 = 1.846 \text{ which gives } f = 5.66$

$F = \frac{s_1^2}{s_2^2} = \frac{\sigma_1^2}{\sigma_2^2} = 1$

T 6.7: Pull-strength tests on 10 soldered leads for a semiconductor device yield the following results, in pounds of force required to rupture the bond:
19.8 12.7 13.2 16.9 10.6 18.8 11.1 14.3 17.0 12.5
Another set of 8 leads was tested after encapsulation to determine whether the pull strength had been increased by encapsulation of the device, with the following results:
24.9 22.8 23.6 22.1 20.4 21.6 21.8 22.5
Comment on the evidence available concerning equality of the two population variances.

NPTEL Monalisa Sarma IIT Kharagpur

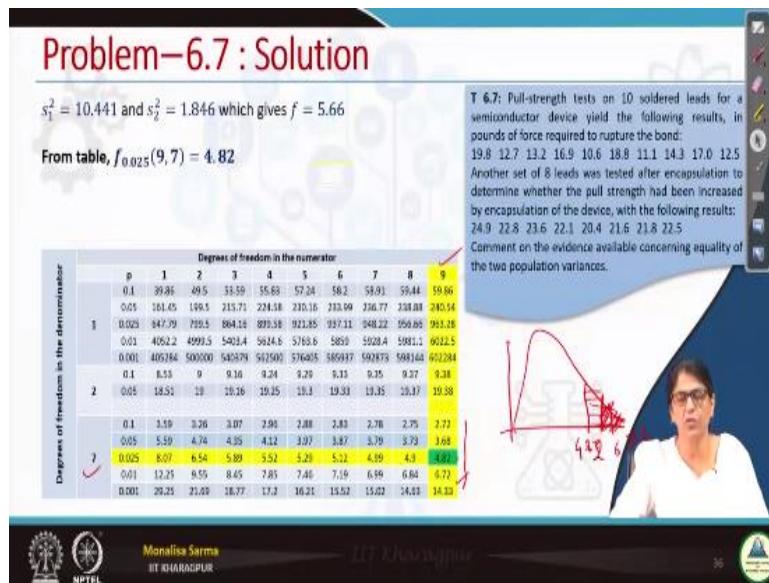
So, what we got so, from this set of data you will calculate the S^2 that is the standard variance I have not shown you how to you know how to calculate the variance with the same method calculating variance or we know that from class 9 onwards. So, calculated the variance from this value, calculate out the variance from this value, so one is this $S_1^2 S_2^2$, then we will find out the f value. What it is mentioned?

We have to assume that the 2 population variances are equal if the 2 population variances are equal. So, what will be the f statistics value will be f statistic was $S_1^2 \sigma_2^2 / S_2^2 \sigma_1^2$ is not it? So, now $\sigma_1^2 \sigma_2^2$ becomes one in both are equal. So, if that is one, then σ_1^2 / σ_2^2 this is equals to 1 because if both the population variance are assumed to be equal, under this situation, what is my f? f turns out to be S_1^2 / S_2^2 .

Now, whether I should which one I will consider as S 1 which will consider S 2 as a general convention. The higher value I will consider S 1 is a general convention based on that also the tables are given in the standard textbooks. So, S_1^2 / S_2^2 , which indeed I got this is my f value, $f = 5.66$. So, now, this is my f value what is the 2 degrees of freedom? One is the numerator degrees of freedom and the denominator degrees of freedom, what is the numerator degree of freedom, so, that is the total 10 size sample size is 10.

So, that is 9 and this is 8, that is 7. So, my degree of freedom is 9 and 7. For this f value, and this is my degrees of freedom, what is the probability corresponding to this that we will see, but again mind it for f value has very limited probability in that standard books. So, sometimes we will have to either we have to interpolate or we will have to find a value which like closely in this and then we can basically give a interval like we have done in t distribution or x chi square distribution in previous one for example.

(Refer Slide Time: 29:37)



So, here what it is given, we got the value of 5.66 so, and degrees of freedom is 9 and 7. So, this is the numerator degrees of freedom here this is 9, this is 7, we in the table we have only this 5 probabilities given 0.1, 0.05, 0.025, 0.01, 0.001 and that means only these are the values which are given here which value closely resembles is 5 we have our value is 5.66. So, we do not have this value in the table what we have is 4.82 we have and 4.82 corresponding to area 0.025.

And we have 6.72 corresponding to area 0.01 that means, our value will lie within this range. So, if f distribution is something like that, so, this is one value this is another one value this is one value that is 4.82 and this is 6.72 it will so, 4.82 corresponding to 4.82 what is the area it is 0.025 that means, this area to the right and what is the area corresponding to 6.72 is this portion. So, my value is in between 4.82 and 6.72. So, my value may be somewhere in this portion. So, my area will be this.

So, it will be in between 0.025 and 0.01 means it will be greater than 2% but lesser than 1% I mean sorry it will be lesser than 2% but greater than 1%.

(Refer Slide Time: 31:32)

Problem-6.7 : Solution

$s_1^2 = 10.441$ and $s_2^2 = 1.846$ which gives $f = 5.66$									
From table, $f_{0.025}(9.7) = 4.82$ and $f_{0.01}(9.7) = 6.72$, the probability of $P(F > 5.66)$ should be between 0.01 and 0.025, which is quite small. Hence the variances may not be equal.									
T 6.7: Pull-strength tests on 10 soldered leads for a semiconductor device yield the following results, in pounds of force required to rupture the bond: 19.8 12.7 13.2 16.9 10.6 18.8 11.1 14.3 17.0 12.5 Another set of 8 leads was tested after encapsulation to determine whether the pull strength had been increased by encapsulation of the device, with the following results: 24.9 22.8 23.6 22.1 20.4 21.6 21.8 22.5 Comment on the evidence available concerning equality of the two population variances.									
Degrees of freedom in the numerator	9	8	7	6	5	4	3	2	1
Degrees of freedom in the denominator	9	8	7	6	5	4	3	2	1
1	0.1	38.66	49.5	53.59	55.81	57.24	58.2	58.91	59.66
0.05	161.05	194.5	213.71	224.51	226.16	228.99	236.77	238.16	240.34
0.025	647.79	793.5	864.18	899.9	921.85	937.11	948.22	958.88	961.21
0.01	403.12	499.5	542.4	562.6	576.6	585.9	592.4	598.11	602.25
0.001	4032.84	5000.0	5403.79	5625.00	5766.05	5859.57	5938.73	5981.44	6027.94
0.1	8.53	9	9.18	9.28	9.29	9.33	9.35	9.37	9.38
2	0.05	18.31	19	19.16	19.25	19.3	19.33	19.35	19.37
3	0.1	1.58	1.26	1.07	2.36	2.08	2.03	2.78	2.75
0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.8	4.62
0.01	12.15	9.35	8.45	7.85	7.46	7.19	6.99	6.84	6.72
0.001	16.75	21.07	18.77	17.2	16.21	15.12	15.02	14.83	14.33

Monalisa Sarma
 IIT Kharagpur

38

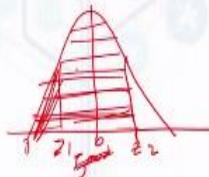
So, probability of that f is greater than 5.66 so should be between this value. So, which is quite small this is a very less probability and so, it is less than 2%, less than 2% is a very less probability. So, hence the variation variance whatever it is assumed that the 2 population variances are equal that may not be true. In probability and statistics we cannot tell deterministic it is not true, this is correct, this is false, this is right nothing we can tell deterministic because it is a science of uncertainty probability. So, everyone see the statements where variance may not be equal.

(Refer Slide Time: 32:16)

Problem-6.8

T 6.8: The breaking strength X of a certain rivet used in a machine engine has a mean 5000 psi and standard deviation 400 psi. A random sample of 36 rivets is taken. Consider the distribution of \bar{X} , the sample mean breaking strength.

- What is the probability that the sample mean falls between 4800 psi and 5200 psi?
- What sample size would be necessary in order to have $P(4900 < \bar{X} < 5100) = 0.99$?



Monalisa Sarma
IIT Kharagpur

There is the last question. The breaking strength X of a certain rivet used in a machine engine has a mean of 5000 psi and a standard deviation of 400 psi random sample of 36 rivets is taken consider the distribution of \bar{X} the sample mean breaking strength, what is the probability that sample mean falls between 4800 psi and 5200 psi here it is one problem on error in the equation. If you can notice then it is very good if you till now if you did not notice I am mentioning it to you.

Because this is always it is a general convention random variable we always write it using capital letters. And a value that a random variable takes we write it in small letters. So, when we talk about the random variable has a particular distribution that means we are talking about a random variable we are not talking about a value then always it will be what it will be capital letter. So, consider a distribution it is not small it is \bar{X} . So, what is the probability that sample means falls between so it is asking that a sample between falls within this.

So, say what it is given let us let us just draw it when we draw things becomes very easy, \bar{X} has a mean of this 500 psi this is the mean value. So, this is something like that and a standard deviation of 400 psi. So, consider the distribution of \bar{X} sample mean distribution of \bar{X} has a sample mean breaking straight what is the probability that sample means fall between 400 psi and 5200 psi? So, for 1400 will be somewhere maybe here and 52 will be somewhere maybe here I need this probability.

So, how do I find out? First this is something which we this type of problem we have done when we have discussed normal distribution. This is basically a question of that only not a question of sampling distribution basically. So, how do I find this value first is that this is in normal distribution, I will have to convert it to z distribution that means, which has mean 0 and standard division 1, that means each value I will have to convert it to z value.

And then from that z value I will be able to whatever some I will get some value say this is z 1, this is z 2 and this mean will be 0 basically, when it is converting it to z distribution and I have to find out this area how do I find this area? This area will be this whole area minus this portion because normal distribution I get cumulative from $-\infty$ to this point, is not it? So, I got this whole area and from here I got this area. So, this minus this area, I will get this area.

So, it is this repetition of that type of question which we have done many questions on while discussing normal distribution.

(Refer Slide Time: 35:18)

Problem-6.8 : Solution

a) Using approximate normal distribution (by CLT),

$$P(4800 < \bar{X} < 5200) = P\left(\frac{4800 - 5000}{\frac{400}{\sqrt{36}}} < Z < \frac{5200 - 5000}{\frac{400}{\sqrt{36}}}\right) \\ = P(-3 < Z < 3) = 0.9974$$

b) To find a z such that $P(-z < Z < z) = 0.99$, we have $P(Z < z) = 0.995$, which results in $z = 2.575$.

So, I will not be discussing in details here you can just see for calculated the Z value, so, this P of Z value falls between this. So, you can from the table you will be able to do this. So, I am not discussing that next also what sample n would be necessary in order to have probability of this is 0.99 here it is given that again, if you solve it yourself, it is given probability 4900. So, this is

4900 because mean is 5% 49 will be definitely this side and 5100 it is given that probability of this is that means this portion is given 0.99.

If this portion is 0.99 that means this to put together is 1% means this will be 0.05, this will be 0.05 is not it? So first, we will have to convert it to z. So that means this portion what will be the area of this portion? This portion will be 0.995 is not it? This portion will be 0.995 and what it will be sorry and this portion will be 0.05. So, subtracting from this portion to this portion corresponding to z value of that and we will get the n value.

(Refer Slide Time: 36:39)

Problem-6.8 : Solution

a) Using approximate normal distribution (by CLT),

$$P(4800 < \bar{X} < 5200) = P\left(\frac{4800 - 5000}{400/\sqrt{36}} < Z < \frac{5200 - 5000}{400/\sqrt{36}}\right) = P(-3 < Z < 3) = 0.9974$$

b) To find a z such that $P(-z < Z < z) = 0.99$, we have $P(Z < z) = 0.995$, which results in $z = 2.575$. Hence by solving $2.575 = \frac{5100 - 5000}{400/\sqrt{n}}$, we have $n \geq 107$. Note that the value n can be affected by the z values picked (2.57 or 2.58).

T 6.8: The breaking strength X of a certain rivet used in a machine engine has a mean 5000 psi and standard deviation 400 psi. A random sample of 36 rivets is taken. Consider the distribution of \bar{X} , the sample mean breaking strength.

- a) What is the probability that the sample mean falls between 4800 psi and 5200 psi?
- b) What sample n would be necessary in order to have $P(4900 < \bar{X} < 5100) = 0.99$?

So, this all this I am not repeating it because we have done this sort of questions.

(Refer Slide Time: 36:46)



So, these are the references and thank you guys.

Statistical Learning for Reliability Analysis

Dr. Monalisa Sarma

**Subir Chowdhury of Quality and Reliability
Indian Institute of Technology – Kharagpur**

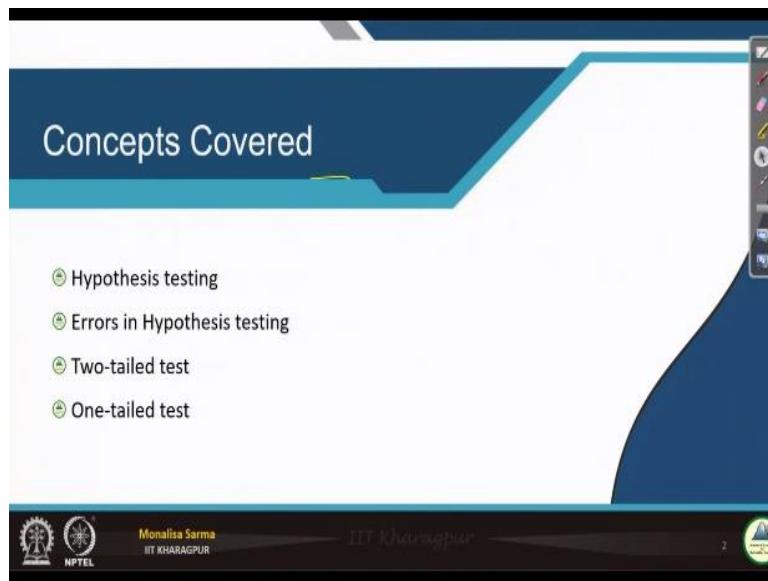
Lecture – 22 Statistical Inference (Part 1)

Hello everyone warm greetings. So, today we will start a new topic a very interesting topic, I am sure you all will enjoy, learning this topic, this is statistical inference. This is what basically what we are talking from the first class our main aim of statistical learning, statistical method is for statistical inference. In fact, sampling distribution, what we have learned? Sampling distribution is the basically the backbone of statistical inference.

Like, what I want to say is that like in computer all of you even if you know from computer science background, you know how a computer works. So, if we do programming, if we write a program, the main part that means your CPU, the CPU does all the job, but for the CPU to do all the jobs, there are some other parts also which have some activities is not it? So, like that is what if I tell take the whole thing that is the statistical inferences like for solving executing the program, the different steps.

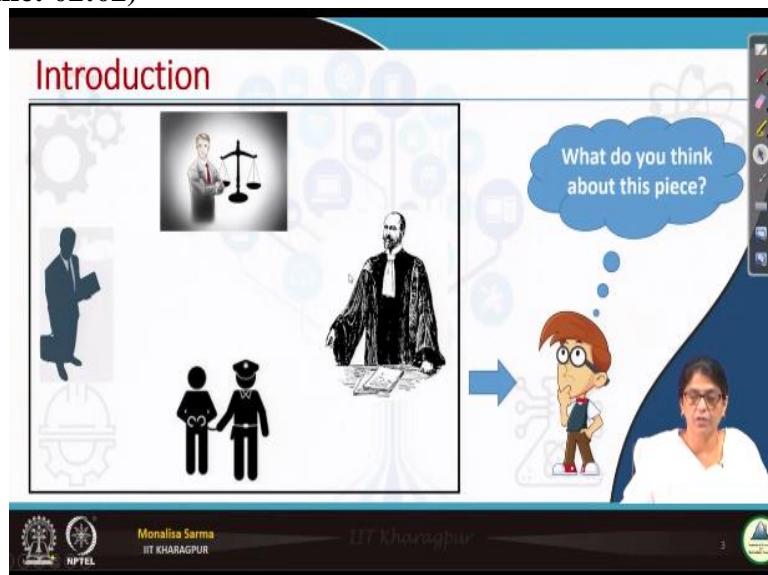
And the main part is the CPU as I told you, so main part in statistical inference is a sampling distribution. Sampling Distribution is not a standalone something that basically we have learned sampling distributions so that we can do statistical inference. So that is what we will be learning in this topic. So, statistical inference is a long chapter. So, we will be doing in many parts so this is the part 1.

(Refer Slide Time: 01:51)



So, here in this today's lecture, we will be learning what is hypothesis testing, what do we mean by errors in hypothesis testing, then what is a 2-tailed test and what is a 1-tailed test?

(Refer Slide Time: 02:02)



So, now what do we think about this space? If you see here, from this picture, what you can see? So, it is like so here, what we see? Here we can see this is a public prosecutor, maybe this may be a public prosecutor, this may be a defence lawyer and this is a suspected criminal and this is maybe the police resolving this criminal and this is the judge is not it? So, from this piece, what you can infer?

I will come to this piece, again, this piece of painting, whatever you say, I will come to this again, after a few slides. Then I am sure all of you will be able to understand why I have in the beginning I have kept this picture here.

(Refer Slide Time: 02:51)

The primary objective of statistical analysis is to use data from a sample to make inferences about the population from which the sample was drawn.

μ, σ

Sample

The mean and variance of students in the entire country?

Mean and variance of GATE scores of all students of IIT Kharagpur

This lecture aims to learn the basic procedures for making such inferences.

Monalisa Sarma
IIT Kharagpur

IIT Kharagpur

So first, so now introduce it first is the primary objective of statistical analysis is to use data from sample to make inference about population from the sample that was drawn, we have repeatedly I am telling this again and again, the primary what is my primary objective, we will use data from the sample to infer about this population is not it? So, this is one such example is from sample we will get some item sample mean or sample variance from there we will try to infer about maybe this is a so this is a population.

This is my sample from the sample I am inferring another population. Similarly, like suppose I know the mean and variance of the GATE score of all the students of IIT Kharagpur, say from this sample this I am taking it a sample mean score of the students of IIT Kharagpur. From this, I am trying to predict something about the mean and variances of the students of the entire country get results. I am not have all the things in exams. So, this is what we call statistical inference.

So, this lecture aims to learn the basic procedures for making such inference. So, how do we make some inference? There are some procedures, there is certain steps which you follow to make this inference. Of course, sampling distribution is the main part that is the backbone, but where we use sampling distribution, how do we use? What are the steps that we follow? That we will be discussing here.

(Refer Slide Time: 04:22)

Basic Approaches

Approach 1: Hypothesis testing

- We conduct test on hypothesis.
- We hypothesize that one (or more) parameter(s) has (have) some specific value(s) or relationship.
- Make our decision about the parameter(s) based on one (or more) sample statistic(s)
- The reliability of decision is expressed as the probability that the decision is incorrect.

Approach 2: Confidence interval measurement

- We estimate one (or more) parameter(s) using sample statistics.
- This estimation usually done in the form of an interval.
- The reliability of this inference is expressed as the level of confidence we have in the interval.

Monalisa Sarma
IIT KHARAGPUR

So, to the basic approach, when we talk of doing this statistical inference, so there are 2 approaches, the first approach is called hypothesis testing. So, we conduct tests on hypothesis that means what we conduct based on hypothesis means what we hypothesize that one or more parameters have some specific value or relationships. So, initially, when I was discussing sampling distribution, I was always using the word estimated or from past results or whatever it is in some way, I have also mentioned that.

I do not want to use the term hypothesis here because we will be talking about that time later. see from here on, I will not be using estimated on there from here on I will be using the word hypothesize that so, what we do we hypothesize that one or more parameters have some specific value like when we told when we in previous question, the mean of the population is this mean of the variance of the population is this that means, population mean and population variance.

How do we know it is such from the huge thing, from the huge thing how we do know the mean and the population? Definitely, we have hypothesized that value now what do we mean by hypothesis? First let me come to what do you mean by I will come to this slide again.

(Refer Slide Time: 05:36)

So, let me take a quick see what is hypothesis? If you just Google it and just Google it you will see hypothesis you will see different definition. So, what one definition is says a hypothesis is an educated prediction that can be tested is a prediction. Educated prediction means you are predicting not just like that you are predicting, you are predicting the best in some knowledge. Again hypothesis is a proposed explanation for a phenomenon.

Again dictionary what Oxford dictionary tells a hypothesis is used to define the relationship between 2 variables that of course, is not very much for our use. Then what does Walpole? Walpole is a standard book on statistical methods what does it say a supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation on the basis of limited investigation, on the basis of whatever evidence past experience whatever we have.

We give some value, we predict will specify certain value and that is basically a starting point for further investigation, this value can be anything it is a value or relation between 2 variables whatever it is, so but why do we so that we can start further investigation for in starting investigation, it is better if we have some value from there, we will go up. That is what Walpole says so, that is hypothesis.

So, now coming back to this so, the first hypothesis testing what is that we hypothesize that one or more parameters have certain values or relationship. So, we then what happens? First, we hypothesize that then we make our decision about the parameters based on the sample

statistics. That is what we have seen in sampling distribution? Sampling distribution, we did not use that what hypothesis, but here I am using it.

So, we hypothesis some value then we collect the sample from the sample we get the data from whatever data we collect, based on that we give our decision whether whatever value we have hypothesis is correct or not. Then now the reliability of the decision whatever decision we give that it is not the problem, it may not be this, the mean may not be this whatever is estimated or variance may not be this whatever it is specified the decision.

The reliability of the decision is expressed as the probability that a decision is incorrect. What is the probability that a decision what we have given what is the probability that is incorrect that also we specify, we just do not tell that this is we do not agree with this hypothesis parameter, the hypothesis parameter may be wrong or hypothesis parameter is right. We just cannot we just only do not say that we also explicit in terms of reliability, what is this reliability? This reliability is the probability that the decision is incorrect.

The reliability of the decision our we explicit and the probability that a decision is incorrect. There is one more approach this is the first approach for statistical inference and this approach we call it hypothesis testing second approach confidence interval measurement confidence, what we do we basically estimate the value of the parameters based on the sample statistics means like in hypothesis testing, we hypothesize the value and then based on that we do the further investigation.

But in confidence interval we do not hypothesis any value we just try to estimate the value of the parameter based on the sample values. So, this we will not be discussing here this I will go in details when we will be discussing and I think after 3, 4 lectures I will come to this confidence interval measurement. So, let us not discuss about this now so, this we have already seen.

(Refer Slide Time: 09:30)

Statistical Hypothesis

- If the hypothesis is stated in terms of population parameters (such as mean and variance), the hypothesis is called statistical hypothesis.
- Data from a sample (which may be an experiment) are used to test the validity of the hypothesis.
- A procedure that enables us to agree (or disagree) with the statistical hypothesis is called a **test of the hypothesis**.

Example

- To determine whether a teaching procedure enhances student performance.
- A product in the market is of standard quality.
- Whether a particular medicine is effective to cure a disease.

Monalisa Sarma
IIT Kharagpur

Now, we know what is hypothesis we have known now? Now what is statistical hypothesis because this hypothesis, maybe we may hypothesis about anything. So, what is statistical hypothesis? If the hypothesis is stated in terms of population parameter such as mean and variance, the hypothesis is called statistical hypothesis. So, that means we will be talking about statistical hypothesis only. So, data from the sample are used to test the validity of the hypothesis.

Validity of the hypothesis means we will talk about the reliability of a decision how will give a reliability of a decision? That is the probability this may be incorrect? A procedure that enables us to agree or disagree with a statistical hypothesis is called a test of hypothesis that we will see what is the test of hypothesis? So, some example of statistical hypothesis like someone maybe we want to determine whether teaching procedure enhances student performance.

So, whether new teaching procedure has come to market whether the teaching procedure will enhance the student performance what, how we will do maybe we will take a conductive test before us introducing the testing procedure and after introducing the teaching procedure, maybe after some time, we will again take a test and we will evaluate the score and what is the score before this introduction of this new procedure.

After the introduction of this new procedure does base on the mean score of this we will be able to find out whether the teaching procedure is good or not. So, that is one way of there is one example of statistical hypothesis. Another example maybe product is the market is of

standard quality, then when I talk of standard quality the maybe the time it last may any is suppose the LED bulbs, LED bulbs how put is maybe the average working average life of the bulb.

So, the standard quality means it has to it is average has to be a particular value say x . So, what we take a sample from there we find out whether it satisfies that or not. So, first we hypothesize a valid and from we take a sample and whether it is correct or not that is again another example of statistical hypothesis. One more example whether a particular medicine is effective to cure a disease these are some of the examples, there is many such examples.

(Refer Slide Time: 11:57)

The main purpose of statistical hypothesis testing is to choose between two competing hypotheses.

Example

One hypothesis might claim that wages of men and women are equal, while the alternative might claim that men make more than women.

- ④ Hypothesis testing start by making a set of two statements about the parameter(s) in question.
- ④ The hypothesis actually to be tested is usually given the symbol H_0 and is commonly referred as the null hypothesis.
- ④ The other hypothesis, which is assumed to be true when null hypothesis is false, is referred as the alternate hypothesis and is often symbolized by H_1 .
- ④ The two hypotheses are **exclusive** and **exhaustive**.

NPTEL Monalisa Sarma IIT KHARAGPUR IIT Kharagpur

Now, when we talk about hypothesis testing, basically, we have 2 different hypotheses. So that is what the main purpose of statistical hypothesis is to choose between 2 competing hypotheses, 2 hypotheses and both are competing hypotheses, we have to choose between 2 we will see. One example 1 hypothesis might claim that wages of men and women are equal that is 1 hypothesis. Another hypothesis, maybe men make more money than women. So, it has 2 competing hypotheses. So, we will have to choose between 2.

How will you choose? We will choose based on the sample data. So, hypothesis testing how do we start? We start by making a set of 2 statements about the parameters in questions. So, hypothesis testing firstly, that we always start by making 2 statements about the parameters in question by what I mean maybe, if we want to infer about the mean. So, we will make 2 statements about the mean of the population. If we want to infer about the variance, we will make 2 statements about a variance of the populations.

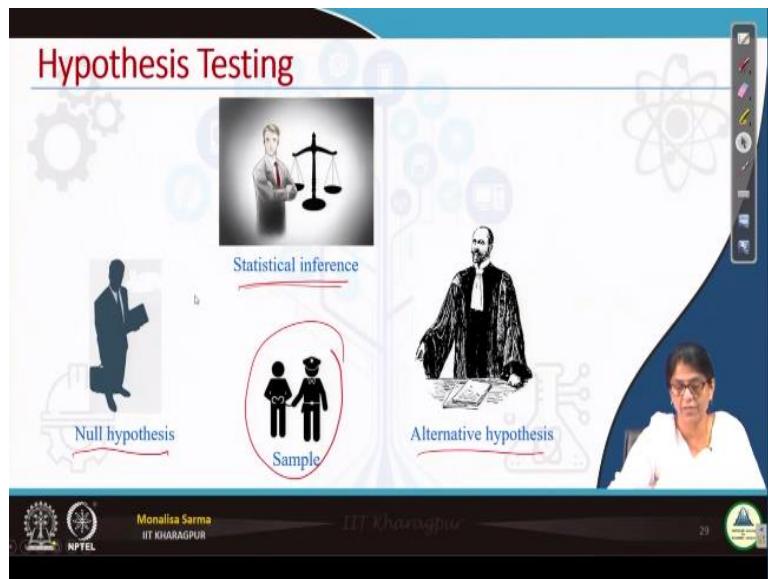
So, the hypothesis actually to be tested easily given by a symbol H_0 and is commonly known referred as null hypothesis. So, there are 2 hypotheses one is called null hypothesis the null hypothesis is usually the status quo, usually that whatever it is already what is existing. So that is specified as null hypothesis. And we want to test it whether what we are claiming or, what we want to be correct whether that is true or not, that we specified null hypothesis and the alternate is what we want to test it.

Whether, I should not have what we want to test it whether what may mean this, which is assumed to be true when null hypothesis is false. So, what actually to be tested is given in the null hypothesis that means what we actually want it to be true, we want something to be true, whatever we are claiming we want that claim to be true. There are reasons I will come to there so that is called null hypothesis.

And another one which actually we want to test. We do not want that to be true, that might be true, we want to test it so that we call it as alternative hypothesis. So, null hypothesis is the status quo that means maintaining the same status. So, null hypothesis symbolizes at H_0 and alternate hypothesis symbolize at H_1 . So, these 2 hypotheses are exclusive and exhaustive, exclusive and exhaustive I am sure all of you know the meaning of what this means or exclusive means both the hypothesis cannot be true.

Both the hypothesis cannot be true and there are only these 2 hypotheses is possible that is why it is exhaustive and both cannot be true only one has to be true. One of the hypotheses has to be true.

(Refer Slide Time: 14:56)



Now, the same picture that we have seen in the first slide. Now we can understand so this may be the null hypothesis which we want to calculate the suspect. We do not want to be convicted. So, there is a null hypothesis alternative hypothesis or we can say the other way around, then this is the based on this maybe this is the sample or some witness or whatever it is maybe the suspects of what to say whatever the suspect has to say, and based on that what we infer. So, now you can understand the significance of this picture.

(Refer Slide Time: 15:41)

The Hypotheses

Problem

To measure the engineering aptitudes of graduates, Ministry of Human Resource Development (MHRD) conducts GATE examination for a total marks of 500 every year. Based on an informal analysis the mean marks for GATE 2020 is hypothesized to be 220.

In this context, statistical hypothesis testing is to determine the mean mark of all GATE-2016 examinee. What are the two hypothesis in this context?

Solution

The two hypotheses in this context are:

$$H_0: \mu = 220$$

$$H_1: \mu < 220$$

So, one simple example like on what I have already given in the first slide, a second slide maybe. So, to measure the engineering aptitude of graduate Ministry of Human Resource Development MHRD conducts GATE examination for total marks of 500 every year. Based on the informal analysis, the main marks for GATE is hypothesized to be 220 from an analysis were done and it is founded mean marks to be around 220.

So, in this context, statistical hypothesis testing is to determine the mean marks of all GATE examinee. What are the 2 hypotheses? In this context what maybe there are 2 hypothesis? The 2 hypotheses first is the null hypothesis definitely, μ that is the mean means μ we are talking about a population. So, we write as μ . So, $\mu = 220$ this is my null hypothesis. Now, my alternate hypothesis maybe if we want to test it, whether it is less than 220 or we want to test it whether it is greater than 220.

Accordingly, I will specify the alternate hypothesis. Now, suppose I want to test it I have a doubt though I have assumed it 220 but I have a doubt it is it may be less than 220 then my alternate hypothesis 220 less than 220.

(Refer Slide Time: 16:54)

Note:

- ④ As null hypothesis, we could choose $H_0: \mu \leq 220$ or $H_0: \mu \geq 220$
- ④ It is customary to always have the null hypothesis with an equal sign.
- ④ As an alternative hypothesis there are many options available with us.
 - ④ Examples
 - ④ $H_1: \mu > 220$
 - ④ $H_1: \mu < 220$
 - ④ $H_1: \mu \neq 220$
 - ④ The two hypothesis should be chosen in such a way that they are **exclusive** and **exhaustive**.
 - ④ One or other must be true, but they cannot both be true.

Monalisa Sarma
IIT Kharagpur

So, now and null hypothesis always it is customary to always add a null hypothesis with an equal sign. So, the null hypothesis we have specified to take it is always necessary the null hypothesis should be with an equal sign, why I will not explain this again this point will be discussing I think the next lecture or maybe next to next lecture I will discuss because dates need some other clarification. So, but you remember till them null hypothesis always it has to be with an equal sign.

So, whenever, I say this question $\mu = 220$, $\mu < 220$ that means, $\mu = 220$ because, I told us 2 hypotheses are exhaustive. So, this does not mean exhaustive there. What about a greater than 220 that means, this $\mu = 220$ it is basically it is telling that μ is greater or equal to 220. So, we could choose less equals to or greater equals to whatever it is or simply equals to if it is simply equals to then alternate hypothesis will be not equal to.

So, it has to be exhausted it is as an alternate hypothesis there are many options available greater or less than if not equal to, but null hypothesis is always equal to should be the other along with equal along with less than or greater than or equal to has to be there we will see where one or other must be true but they cannot both be true very important.

(Refer Slide Time: 18:28)

The Hypotheses

Definition: One-tailed test

A statistical test in which the alternative hypothesis specifies that the population parameter lies entirely above or below the value specified in H_0 is called a one-sided (or one-tailed) test.

Example.

$$H_0: \mu = 100$$
$$H_1: \mu > 100$$

Definition: Two-tailed test

An alternative hypothesis that specifies that the parameter can lie on either sides of the value specified by H_0 is called a two-sided (or two-tailed) test.

Example.

$$H_0: \mu = 100$$
$$H_1: \mu \neq 100$$

So, now, what is one-tailed test? One-tailed test is a statistical test in which the alternate hypothesis specified that the population parameter lies entirely above or below the values specified in H_0 whatever we have specified in H_0 it specifies the population parameter lies either above that or below that then we call it is one-tailed test. The last example what we have seen that a get value GATE score is less than 220 that is a one-tailed test if I would have written not equals to 220 that is a two-tailed test. So, this is an one-tailed test example.

So, two-tailed test an alternative hypothesis that specifies the parameter can lie on either sides of the value specified by H_0 whatever we have specified H_0 that it can lie on both sides then it is called 2 sided that is two-tailed test. So, this is an example of two-tailed test $\mu = 100$ $\mu \neq 100$ this is an example of two-tailed test. We will see one-tailed, two-tailed what is it we will see then get ((19:30)) using the what 2 numbers tailed mean here? Tailed definitely I think you could understand I am talking about a distribution tailed.

(Refer Slide Time: 19:37)

Note:

In fact, a 1-tailed test such as:

$$H_0: \mu = 100$$

$$H_1: \mu > 100$$

is same as

$$H_0: \mu \leq 100$$

$$H_1: \mu > 100$$

Monalisa Sarma
IIT Kharagpur

So, in fact a 1-tailed test such as this as same as I told you it has to be exhaustive. So, if I am writing a greater than means equal that means I am actually meaning less than equals to, but I am not very bothered whether it is less than or what, I just want to test whether it is greater than 100 or not. I am assuming this 100 and I want to test whether it is greater than 100 or not means less than 100 I am not giving any importance to that. So, when I am talking equals to, that means I am specifying actually, I am meaning less than equals to 100 only.

(Refer Slide Time: 20:13)

Errors in Hypothesis Testing

In hypothesis testing, there are two types of errors.

Type I error:	A type I error occurs when we incorrectly reject H_0 (i.e. we reject the null hypothesis, when H_0 is true).
Type II error:	A type II error occurs when we incorrectly fail to reject H_0 (i.e. we accept H_0 when it is not true).

		Observation	
Decision	H_0 is true ✓	H_0 is false ✗	
H_0 is accepted ✗	Decision is correct	Type II error ✗	
H_0 is rejected	Type I error ✗	Decision is correct	

Monalisa Sarma
IIT Kharagpur

Again, in hypothesis testing, there is one unescapable part there is an error, error in our decision whatever decision we take in hypothesis testing as a tool, we will be there 2 steps, 1 is the null hypothesis another is the alternative hypothesis. So, we take a sample, from the sample we get the data from the gate data, we tell whether null hypothesis is true or the alternate hypothesis is true means we basically give a decision on based on these 2 hypotheses.

Now, this decision, always there is a probability that there may be some error in this decision. So, doing an error in then on escapable part in this statistical inferences, this is something we cannot escape there may not be any error, but still the probability of error will always be there. So, basically, there are 2 different types of error what are the 2 errors first is called a type 1 error, see what is type 1 error? So, we will be using this type 1 error type 2 error for coming few classes. So, please remember this again.

So, type 1 error occurs when we incorrectly reject H_0 that means, our null hypothesis whatever we have specified in the null hypothesis, that like in an example, for GATE score, we are specified null hypothesis is $\mu = 220$ the actually if you see population μ is 220. But whatever sample we got from the sample, we found that $\text{mean} \neq 220$ probability of $\text{mean} = \mu$ very very less from the sample what we got, that is why we have rejected the hypothesis.

But actually if we say if there was some mechanism to find out I mean, actually it is mean is 220 only. So, that is what so then that means, we have made it type 1 error. So, type 1 error occurs when we incorrectly reject H_0 that is we reject the null hypothesis when H_0 is true, that is type 1 error. Now, what is type 2 error? Type 2 error occurs when we incorrectly fail to reject H_0 actually H_0 is false, but we fail to reject H_0 like to get example, actually $\text{mean} \neq 220$ what is an alternate hypothesis?

Alternate hypothesis was greater or less than 220. So, actually if you see the actual result actually mean is actually less than 220. But on a sample what we got sample indicates that the mean = 220 that means, it type 2 error occurs that means, we incorrectly because we take the decision based on a sample and sample statistics or variable for different samples we may get different, different sample statistics is value, as well as what matters most of the sample is there biased sample or unbiased sample.

Sometimes we may incorrectly take a biased sample as well and sometimes what because the selection what we have selection? We have selected even random for us to make a select the sample unbiased, our selection has to be random, but however while picking taking randomly picking different subjects also it is it may so, happen that our sample is not very diverse sample so, we may come up with a wrong result is not it?

So, that is what so, a type 2 error occurs when we incorrectly fail to reject H_0 actually H_0 is false, but the simple result indicators no H_0 is correct so, that is type 2 error. Some examples will make it very clear. So, what is this is an observation you see. So, these are a decision H_0 is accepted H_0 is rejected. There are 2 cases H_0 is accepted 1 if H_0 is true and we are accepting the decision that means this is a correct, when H_0 is false and we are accepting H_0 then it is a type 2 error. We are rejecting H_0 whereas H_0 is true and this is a type 1 error H_0 is rejected is false and decision is correct.

(Refer Slide Time: 24:17)

So, now this type 1 error and type 2 error. So, we have some representation for type 1 and type 2 error we usually indicate type 1 error by α type 2 error by β . So, here type 1 how to calculate this type 1 error. Now question is how to calculate the type 1 error? Type 1 error α denotes the probability of making a type 1 error. So, what is α ? Alpha is nothing but probability of rejecting H_0 given H_0 is true see we have talked learn conditional problem is not it? What is conditional probability, what is the probability of A given B?

So, what is the probability of rejecting H_0 given H_0 is true that is α . Similarly, what is β ? Beta is the type 2 error probability of accepting H_0 given H_0 is false is H_0 is false and we accepting H_0 what is that probability of β . So, we will see with some example the formation of different types of hypotheses formation of α β and region. So, α and β are not independent of each other as one increases the other decreases if α increases β decreases, β increase α decreases.

When the sample size increases both decreases, because sampling error is reduced that we have seen. So, in general we focus on type 1 error but type 2 error is also important, particularly when sample size is small anyway, in general we focus on type 1 error why we focus on type 1 error, this thing also I will come later. So, there are 2 points in this today's class there are 2 points which I mentioned that I will be discussing in some other class because now you do not know all the technical details to understand this.

So, 2 things why we focus on type 1 error and secondly, why it is what to say null hypothesis has equality sign needed for null hypothesis in this 2 we will be discussing later.

(Refer Slide Time: 26:15)

Case Study 1: Formation of Hypotheses

There are two identically appearing boxes of chocolates. Box A contains 60 red and 40 black chocolates whereas box B contains 40 red and 60 black chocolates. There is no label on the either box. One box is placed on the table. We are to test the hypothesis that "Box B is on the table".

Let us express the population parameter as
 p = the number of red chocolates in Box B.

The hypotheses of the problem can be stated as:

$H_0: p = 0.4$	// Box B is on the table
$H_1: p = 0.6$	// Box A is on the table

Monalisa Sarma
IIT Kharagpur

So, now, with some very simple example not very technical type of example, we will see how we form the hypothesis. So, we will take 2 examples basically. So, first example, as the simply a box of chocolates there are 2 boxes, box A and box B. Box A contains 60 red and 40 black chocolates, the 60 red chocolate and 40 black chocolate, box B contains 40 red and 60 black chocolate, if our interest in identifying the boxes based on the number of red chocolate it has.

So, probability the box number of red chocolate boxes has what is that probability is 0.6. But for Box B it is 0.4. So, now there is no level in both the boxes one box is placed on the table, we are to test the hypothesis that box be on the table box B on the table that we are to test that hypothesis whether this is correct or not. So, it is a very disturbing non technical example. We will take a technical example also. So, let us start with this.

So, now, for what will be my null hypothesis? Null hypothesis is probability that box B is on the table we will consider the main deciding factor is the red chocolate. So, p is the number of red chocolate in the box. So, my null hypothesis is $p = 0.4$ and if I write it informally, informally, I can write our null hypothesis is the box B is on the table instead of writing that I am writing as $\pi = 0.4$ taking that direct chocolate as the deciding factor for the boxes. So, my alternate hypothesis is $p = 0.6$.

Now, I have to find out whether my null hypothesis is true or not I will test null hypothesis the one which we will test which you want to be correct here. This is a whether boxes in A is on the table B is on the table. It does not make any difference, but anyway, but in actually in real example, we want the null hypothesis to be true and we are testing on null hypothesis. So, we will take the sample from the sample we will take this we will find out which hypothesis is true.

(Refer Slide Time: 28:30)

Case Study 1: Box of Chocolates

Problem

There are two identically appearing boxes of chocolates. Box A contains 60 red and 40 black chocolates whereas box B contains 40 red and 60 black chocolates. There is no label on either box. One box is placed on the table. We are to test the hypothesis that "Box B is on the table".

Hints:

To test the hypothesis an experiment is planned, which is as follows:

1. Draw at random five chocolates from the box.
2. We replace each chocolates before selecting a new one.

Note:

Since each draw is independent to each other, we can assume the sample distribution follows binomial probability distribution.

NPTEL
Monalisa Sarma
IIT Kharagpur
IIT Kharagpur

So, first thing is that to find out which hypothesis is true from the sample what we have to find out we always first will have to decide a rejection region what do we mean by rejection region? Rejection region means the probability for what probability we will tell that this is not true. Remember we while doing sampling distribution we are specified when we get very less probability we tell that whatever we have estimated whatever we have guessed or whatever it is conjectured, it is not true when we get a very less probability.

Now, what is this less probability? This less probability meaning we will specify this as a rejection region. So, if I take this sampling distribution of mean and this is a normal

distribution, my rejection region will be this very definitely this is a very less probability is not it? Probability is nothing but the area of this under this, is not it? So, this is very less, but this is my rejection region. So, now for this simple example, similarly, I will have to find my rejection region.

Rejection region means from the sample I will have to calculate a test statistics from the sample I have to calculate the test statistics. And I will have to find out the probability of that test statistics. So, here so what I will do, what I am interested in proving? I am interested in proving that box B is on the table, is not it? This is box B is on the table I want to put that box B is on the table that means box with just less than number of red chocolates.

So, what I will do? I will do my experiment how I have designed an experiment, draw at random 5 chocolates from the box we replace each chocolates before selecting a new one. Selected one put it back; selected one put it back that means this is nothing but a binomial distribution simple. This choice independent since each draw independent to each other we can assume simple distance using we can assume that a sample distribution follows binomial probability distribution.

(Refer Slide Time: 30:35)

Case Study 1: Calculating α

Let us express the population parameter as
 p = the number of red chocolates in Box B.

The hypotheses of the problem can be stated as:
 $H_0: p = 0.4$ // Box B is on the table
 $H_1: p = 0.6$ // Box A is on the table

There are two identically appearing boxes of chocolates. Box A contains 60 red and 40 black chocolates whereas box B contains 40 red and 60 black chocolates. There is no label on either box. One box is placed on the table. We are to test the hypothesis that "Box B is on the table".

Calculating α :

A video feed shows a lecturer wearing glasses and a white shirt, sitting at a desk with a chalkboard in the background. The chalkboard has some handwritten notes, including a circle with a diagonal line through it and some text.

NPTEL

Monalisa Sarma
IIT Kharagpur

IIT Kharagpur

60

Now, we need to calculate α . So, how do we calculate α ? For calculating α that means probability of type 1 error. Type 1 error means what is type 1 error remember that probability that given that H_0 is true and we are rejecting H_0 . So, when that will happen, that will happen when we get some when we will reject H_0 we will get reject H_0 when we will get some very less probability of happening is not it?

So, this portion it may value have falls in this area then I will reject it. So, this is nothing my α if the sample results fall in this region, maybe it actually it is correct, but from the sample I got this result, because I have just taken from one sample is not it? Maybe you would have to add many more sample I would have seen this it does not fall here it falls in this region. So, this is my α this region basically. So, I have to calculate α .

So, for calculating α first I will have to find out which is my rejection region at one point α means there has to be some point corresponding to this then only this portion I call it as α what is the point specific today that means, my rejection region starts from here. So, this is the starting of my rejection region, this point what is this point? This is my rejection region. So, for this simple example, what rejection region we may consider.

(Refer Slide Time: 32:17)

Case Study 1: Calculating α

Let us express the population parameter as
 p = the number of red chocolates in Box B.

The hypotheses of the problem can be stated as:

$H_0: p = 0.4$	// Box B is on the table
$H_1: p = 0.6$	// Box A is on the table

There are two identically appearing boxes of chocolates. Box A contains 60 red and 40 black chocolates whereas box B contains 40 red and 60 black chocolates. There is no label on the either box. One box is placed on the table. We are to test the hypothesis that "Box B is on the table".

Calculating α :

- In this example, the null hypothesis (H_0) specifies that the probability of drawing a red chocolate is 0.4.
- This means that, lower proportion of red chocolates in observations (i.e., sample) favors the null hypothesis.
- In other words, drawing all red chocolates provides sufficient evidence to reject the null hypothesis.

Monalisa Sarma
IIT KHARAGPUR

So, in this example, null hypothesis specifies that a probability of drawing a red chocolate is 0.4. This means that the lower proportion of red chocolates in observation favours the null hypothesis. So, let us take the rejection region maybe if we since it blocked box B in the table that means very less red chocolates are there. So, if you let us take our rejection region we may assume here if we draw all the written that if we are drawing how many chocolates? 5 chocolates.

So, you are doing 5 when you draw 5 chocolates, all the 5 chocolates if we get red, so, that is a rejection region if all the 5 chocolates we get red, then we can the conclusion that box B is not on the table box A is on the table because we get so many red chocolates we are picking

but we got all the chocolates red that may be a rejection region see here. Now we can understand the concept of type 1 error say I have decided my rejection region as if I get all my chocolates red then I will decide that is not blocked box B but box A is on the table.

Because boxes more number of red chocolate. But see there may be the case actually blocked box B is on the table. But when I am picking it randomly I pick all the red chocolates, is not it? That thing is there so that is the probability of error that is the type 1 error I have committed. So, now, what is the probability of type 1 error that means what is the α that we need to calculate?

(Refer Slide Time: 33:44)

Case Study 1: Calculating α

- The probability of making a Type I error is the probability of getting five red chocolates in a sample of five from Box B. That is,
$$\alpha = P(X = 5 \text{ when } p = 0.4)$$
- Using the binomial distribution
$$= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \text{ where } n = 5, x = 5 \\ = (0.4)^5 = 0.01024$$
- Thus, the probability of rejecting a true null hypothesis is ≈ 0.01 . That is, there is approximately 1 in 100 chance that the box B will be mislabeled as box A.

Monalisa Sarma
IIT Kharagpur

So, what is α ? α is probability of that $X = 5$ when $p = 0.4$, I have specified my rejection region, rejection region is all the 5 chocolates are there. So, now I can calculate what is α that is the type 1 error actually hypothesis is null hypothesis is true, but we sample is also that it is not true that means we have rejected a null hypothesis. So, α is probability that x is = 5 when $p = 0.4$ this is my α type 1 error.

So, using simple we can use binomial distribution to find out what is this probability. So, if $X = 5$ so from 5 draw 5 chocolates, but it will be using this formula I will be getting this. So, this is my value simple binomial distribution, I will not go into the details of this. I am sure you know that we have done this sort of thing binomial distribution, so this is my value of α . So, there is the probability of rejecting a true null hypothesis.

If that is my rejection region, true null hypothesis is 1% that is there is approximately 1 in 100 chance that a box B will be mislabelled as box A there is very less sense if this is the thing if I consider if this is my rejection reason I consider that there is very less chance that it is actually though it is box B is on the table, I say that no it is not box B box A is on the table.

(Refer Slide Time: 35:19)

Case Study 2: Machine Testing

Problem

A medicine production company packages medicine in a tube of 8 ml. In maintaining the control of the amount of medicine in tubes, they use a machine. To monitor this control a sample of 16 tubes is taken from the production line at random time interval and their contents are measured precisely. The mean amount of medicine in these 16 tubes will be used to test the hypothesis that the machine is indeed working properly.

Monalisa Sarma
IIT Kharagpur

So now, that is how we find out type 1 error that is 1 example where we have seen how we can form the hypothesis, how the type 1 error is calculated. Now, you may ask why type 2 error we did not see we will come later, for this question type 2 error calculation will be very easy that I think I have in this lecture only, but actually calculation of type 2 error is very difficult we will come since next I think next to next lecture.

Now, let us take a bit technical problem earlier problem was a very simple toy game so, think toy example. So, a bit technical problem example. So, what is a medicine production company? We are not doing statistical inference till now we are just going step by step just reaffirming the hypothesis and try to find out the probability of type 1 error? That is α a medicine production company packages medicine in a tube of 8 ml.

So, in maintaining the control of the amount of medicine in tubes they use a machine to monitor this control a sample of 16 tube is taken from the production line at random time interval and the contents are measured precisely what we are done in medicine production company packages medicine in a capsule of say 8 ml. So, that means, it is expecting that the liquid there has to be 8 ml it is expecting that that is the expected value. So, whether this is correct or not.

So, what is the some to monitor this it has taken a sample of 16 and from each sample it is it tries to measure the content. And the contents or measure precisely the mean amount of medicine in this 16 tubes.

(Refer Slide Time: 37:02)

Case Study 2: Formation of Hypotheses

Consider the two hypotheses are

The null hypothesis is

$$H_0: \mu = 8$$

The alternative hypothesis is

$$H_1: \mu \neq 8$$

Assume that given a sample of size 16 and standard deviation is 0.2 and the population follows **normal distribution**.

Monalisa Sarma
IIT KHARAGPUR

It is not given here the data are not given. So, the mean amount of medicine in the 16 tube will be used to test the hypothesis that the medicine machine is indeed working properly. So, we have to test the hypothesis that whether it is working properly or not. So, that we will know from the sample if we test the values. So, here how what will be our null hypothesis because we want to be true, we wanted 8 ml should be true to our $\mu = 8$ that is why null hypothesis.

What is my alternate hypothesis? It is not equals to 8 because if it is less than 8 then also it is because it is a medicine that means it will affect the general population and general people and if it is greater than 8 and also it will affect the population. If we talk of Institute of medicine if we talk away fruit drinks, so, if it is some say 100 ml and if it is less than 100 ml then it is cheating to the customers if it is more than 100 ml customers will be happy, but the company manufacturing company will be at a loss. So, both ways not needed.

So, it is we will try for not equality. So, $\mu \neq 8$ is the alternate hypothesis. This is how we form the hypothesis. Assume that given a sample size of 16 and a standard deviation of 0.2 and the population follow normal distribution.

(Refer Slide Time: 38:35)

Case Study 2: Calculating α

We can decide the rejection region as follows.

Suppose, the null hypothesis is to be rejected if the mean value is less than 7.9 or greater than 8.1. If \bar{X} is the sample mean, then the probability of Type I error is

$$\alpha = P(\bar{X} < 7.9 \text{ or } \bar{X} > 8.1, \text{when } \mu = 8)$$

Given σ , the standard deviation of the sample is 0.2 and that the distribution follows normal distribution.

Thus,

$$P(\bar{X} < 7.9) = P\left[Z = \frac{7.9-8}{0.2/\sqrt{16}}\right] = P[Z < -2.0] = 0.0228$$

and

$$P(\bar{X} > 8.1) = P\left[Z > \frac{8.1-8}{0.2/\sqrt{16}}\right] = P[Z > 2.0] = 0.0228$$

Hence, $\alpha = 0.0228 + 0.0228 = 0.0456$

So, now we can decide the rejection region. Now, what is precisely rejection region like earlier case we have taken the rejection region if we get the 5 chocolates now here a rejection region. So, mean we want this 8, but we will accept if it is within 8.1 then within 7.9. So, within 7.9 to 8.1 we will accept it, if it is 8.1 also still it is acceptable if it is 7.9 that also it is acceptable it will not affect a patient 0.1 difference 0.1 ml it will not affect. So, it is 7.9 and 8.1 is acceptable.

So, what is the rejection region? Rejection region is greater than 8.1 and the rejection region is less than 7.9 this is my rejection region. So, what will be my thing type 1 error when my type 1 error will happen my type 1 error will happen that means, actually the medicine is producing correctly, machine is producing correctly only on an average it is around 8 lies within this range only, I will not tell it exactly A.

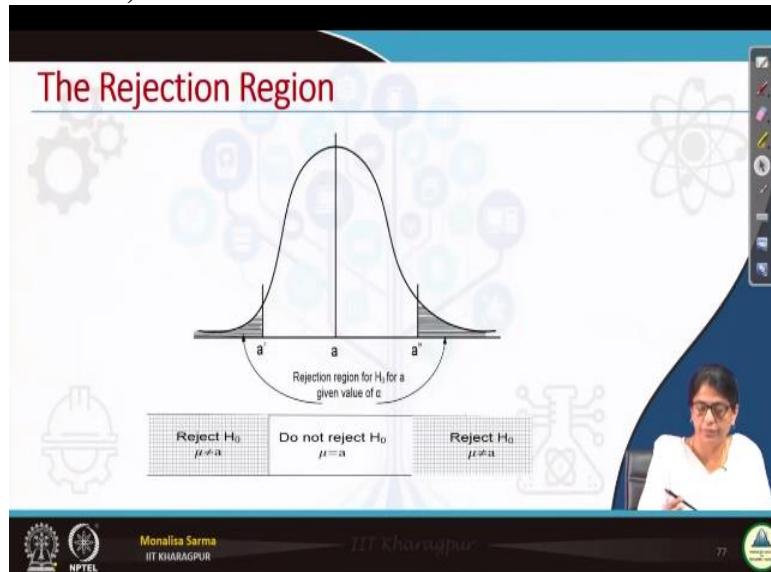
But it is on an average within this range on the but somehow we have picked some sample there may be some outliers, somehow we have picked some sample and from the sample we found that it is not in this interval but it is either in this interval or in this interval. So, what happens when we found that it is in this or this interval in a lesser or greater interval, then whatever maybe we will scrap the machine totally we will buy in total new machine investing crores and crores of rupees.

Whereas, actually it was the machine was working properly only. So, that is the type 1 error. So, what is type 1 error? Type 1 error is if we have decided that 7.9 and 8.1 is the rejection region. So, my type 1 error is the probability that my X bar is less than 7.9 and greater than

8.1 when μ is actually 8 when the sample mean is actually 8, but what is the probability that I got X bar is 7.9 and less than 7.9 or greater than 8.1 that is my α .

So, how do I calculate X bar getting 7.9 or X bar greater than 8.1 it is nothing but again come find out z distribution from that region then I can get the value from the z table. So, given the standard deviation of the population is 0.2. So, X bar less than 7.9 we know all this we have done lots of this problem, is not it? So, I found probability that is that less than -2 this is the probability.

Similarly, for 8.1 I found this that means if this is 8.1, this is 7.9 that means, this value and this value is this and this, so, this 2 together. So, my α is addition of this and this, this is my α
(Refer Slide Time: 41:44)



So, this is the rejection region. When I am talking about 2-tailed tests that means I have a 2 rejection region. This is 2-tailed test when I am specifying alternate hypothesis I am specifying is not equal, then I have 2 rejection region this is one reason this is another one region.

(Refer Slide Time: 42:02)

Two-Tailed Test

For two-tailed hypothesis test, hypotheses take the form

$$H_0: \mu = \mu_{H_0}$$

$$H_1: \mu \neq \mu_{H_0}$$

In other words, to reject a null hypothesis, sample mean $\mu > \mu_{H_0}$ or $\mu < \mu_{H_0}$ for a given α .

Thus, in a two-tailed test, there are two rejection regions (also known as critical region), one on each tail of the sampling distribution curve.

Acceptance and rejection regions in case of a two-tailed test with a rejection region of 5%.

Monalisa Sarma
IIT Kharagpur

This is same thing I do not have to repeat it here again. So, here you see if my α is 5% α that is type 1 error is 5% and it is a 2-tailed test 5% means 0.025 will be this side 0.025 will be this side, that is in a 2-tailed test there are 2 rejection region and this rejection region is also known as critical region. So, here it is shown if it is a α is 5%. So, it is 0.025 and 0.025.

(Refer Slide Time: 42:45)

One-Tailed Test

A one-tailed test would be used when we are to test, say, whether the population mean is either lower or higher than the hypothesis test value.

Symbolically,

$$H_0: \mu = \mu_{H_0}$$

$$H_1: \mu < \mu_{H_0} \quad [\text{or } \mu > \mu_{H_0}]$$

Wherein there is one rejection region only on the left-tail (or right-tail).

Monalisa Sarma
IIT Kharagpur

And 1-tailed test to be used when we have to take as I have already specified where we use 1-tailed test when whether the population mean is either low or higher than the hypothesis test, well, then we use 1-tailed test, then 1-tailed test means we will have only 1 rejection region. If I am talking of less than this, then my rejection region is this. And if I am talking I am greater than my rejection region is this alternate hypothesis.

If alternate hypothesis if I tell it is less than that then my rejection region will be left if I tell it the alternate hypothesis is greater than my rejection region will be right. Now in this 1-tailed

test if I specify my $\alpha = 5\%$, so in 2-tailed it when my α is 5%, both side 0.025, 0.025. If in 1-tailed test α is 5% means it is only 1-tailed 1 side. So, 5% is only this side this is 5%, this is 5% we will do problems on this.

(Refer Slide Time: 43:39)

Example: Calculating β

- The Type II error occurs if we fail to reject the null hypothesis when it is not true.
- For the current illustration, such a situation occurs, if Box A is on the table but we did not get the five red chocolates required to reject the hypothesis that Box B is on the table.
- The probability of Type II error is then the probability of getting four or fewer red chocolates in a sample of five from Box A.
- That is,
$$\beta = P(\bar{X} \leq 4 \quad \text{when } p = 0.6)$$
- Using the probability rule:
$$P(X \leq 4) + P(X = 5) = 1$$

That is, $P(X \leq 4) = 1 - P(X = 5)$

Now, $P(X = 5) = (0.6)^5$

Hence, $\beta = 1 - (0.6)^5 = 1 - 0.07776 = 0.92224$

- That is, the probability of making Type II error is over 92%.
- This means that, if Box A is on the table, the probability that we will be unable to detect it is 0.92.

And type 2 error occurs when you fail to reject a null hypothesis when it is not true. So, for the case study 1 box A is on the table, but we did not get the 5 red chocolates required to reject a hypothesis that box B is on the table. And case study 1 what we have seen, actually it was box A was in a table, but we to reject that we needed 5 red chocolates, we did not get the 5 red chocolates. So, we have told that so what we what hypothesis the null hypothesis is correct. That is what we have found out.

But actually it is wrong that is what that is the type 2 error. Type 2 error is when we failed a false null hypothesis. So, for this example, we will be able to calculate β for the other example it is difficult. So, we will solve for this example what is what is β ? β is probability of X when X is less than equals to 4 when p is 0.6. Actually, box A is on the table. So, p is 0.6 and when p is 0.6. I am getting X less than or equal to 4.

What is this probability that is my β ? So, I found this is my β a very high β that is the problem making type 2 error is over 92%. Now, how can we reduce this β because this is very high β , this problem the type 2 error of 92% is quite high. So, how can we reduce this?

(Refer Slide Time: 45:12)

Relation between α and β

- How to decrease type II error?
By making rejection easier
- Suppose we decide to reject H_0 if either four or five of the chocolates are red.

$$\alpha = P(Y \geq 4 \text{ when } p = 0.4) = 0.087$$

$$\beta = P(Y < 4 \text{ when } p = 0.6) = 0.663.$$

- By changing the rejection region, β is decreased but α is increased.
- This will always true if the sample size is unchanged.
- Neither error can have a probability of 0

In fact, the only way to ensure that $\alpha = 0$ is to never reject a hypothesis, while to ensure that $\beta = 0$ the hypothesis should always be rejected, regardless of any sample results.

NPTEL Monalisa Sarma IIT Kharagpur 301

How to decrease type 2 error by making rejection easier, how we can reduce the type 2 error if I made a rejection is easier. That means if we decide to reject H_0 if 4 or 5 of the chocolates are red, earlier what is the rejection region we have considered if 5 added then we will reject. Now, let us make rejection easier that means what we if 4 or 5 of the chocolates are red, then we will reject then what happens?

Then α is probability of Y it is actually X , X is greater than equal to 4 when $p = 0.4$ this is now my α has become 0.08 earlier my α is 0.01 remember, now my α has increased my β will decrease now for this my β has become 0.663. So, when β increase α decreased, so, by changing the rejection region β is decreased α is increased this is always true and the sample sizes unchanged, however increase the sample size then we will get a better result.

Neither error can have a probability of 0. In fact, the only way to ensure α is 0 is to never reject a hypothesis. If we never reject the hypothesis, then α will be 0 probability of committing a type 1 error will never be there, we are not rejecting it only while to ensure that $\beta = 0$ the hypothesis should always be rejected regardless on the sample result this is a hardly deserved a possible this is not at all possibilities, we cannot do that.

And what is the point of conducting the test of hypothesis? So, the neither can have a probability of 0 it may be that there may be no error, but there is always a chance of error.

(Refer Slide Time: 46:51)

CONCLUSION

- ④ In this lecture we learned the theory of statistical Hypothesis testing that includes the knowledge of –
 - ④ Formulation of null and alternate Hypothesis ✓
 - ④ Type I and Type II errors in Hypothesis testing ✓✓
 - ④ Probabilities of Type I and Type II errors ✓✓
 - ④ And lastly, the relationship between these errors ✓
- ④ Above-mentioned concepts were illustrated with few examples for clear understanding.
- ④ In the next lecture, we will cover some more theoretical aspects of statistical inference.



Monalisa Sarma
IIT KHARAGPUR

IIT Kharagpur

104

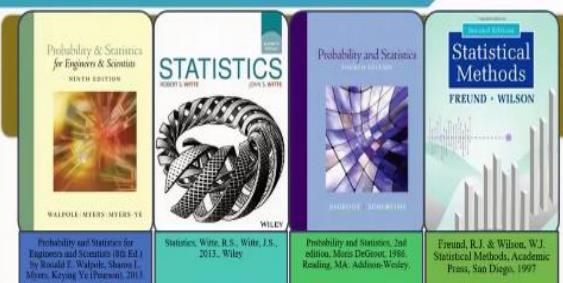


So, here we have calculated β , but actually calculation of β is not easy this is for this toy example we could do that, but calculation of β is difficult. So, we will be learning in I think the next to next lecture what is this. So, now to conclude in this lecture, we have learned about formulation of null and alternative hypothesis, how we formulate the hypothesis, what is type 1 error and type 2 error, what is the probability of type 1 and type 2 error what is α , what is β and lastly the relationship between these error.

We have also illustrated this with some examples. And in the next lecture we will cover some more theoretical aspects of statistical inferences.

(Refer Slide Time: 47:34)

REFERENCES



Monalisa Sarma
IIT KHARAGPUR

IIT Kharagpur

105



With that these are the references and thank you guys.

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

Lecture - 23
Statistical Inference (Part 2)

Welcome back guys so, today in continuation of our earlier lecture on statistical inference, today is the second lecture on statistical inference.

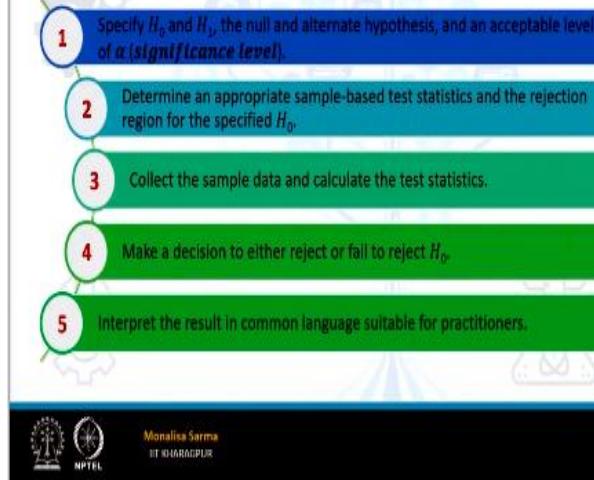
(Refer Slide Time: 00:35)



So, in this lecture, we will learn about hypothesis testing procedure. And we will see some test cases I mean some case studies for hypothesis testing and we will also see what is p value in the context of hypothesis testing.

(Refer Slide Time: 00:48)

Hypothesis Testing Procedures



Monalisa Sarma
IIT KHARAGPUR



So, like in the last class, if you remember, we have discussed what do we mean by statistical inference. And what are the different types of error that can happen in a statistical inference how to form the hypothesis, what does rejection region mean rejection region and critical region what are the same thing actually, what does this mean those we have discussed in my last lecture. So, today will directly come to the hypothesis testing procedure.

So, there are many steps in hypothesis testing procedure many means, there are total 5 steps. So, what is the first step? First step is that we have to specify H_0 and H_1 what is H_0 and H_1 I have already mentioned in my last class, where we specify H_0 and H_1 the null and alternate hypothesis and an acceptable level of α what is α ? Remember, α is called the probability of type 1 error. What is type 1 error?

Type 1 error is that we are rejecting a true null hypothesis when a null hypothesis is true and we are rejecting it that is the called type 1 error. So, α is the probability of that and this α is also called significance level. So, since this α is called significance level in hypothesis testing procedure, we use the significance level. So, this hypothesis testing is also called significance testing.

So now, then the after specifying H_0 , H_1 null and alternate hypothesis then what we need to do, we need to find out an appropriate sample based test statistics. We will have to find from the sample will have to calculate the test statistic and the rejection region for the specified H_0 . After

calculating the test statistics in a second step, what we will do? We will find out what is an appropriate test statistics.

Like when I talk about appropriate test statistics means, if I want to do sampling distribute if I were to infer about the population mean, then my test statistics maybe Z and my test statistic maybe T value also depending on the whether, I know the standard deviation of the population or not, then again if I want to infer something about the population variance, then my sample test statistics will be chi square value. If I want to compare 2 variants, then my test statistics will be F value.

So, accordingly based on the problem, we will find out by sample based test statistics and the rejection region. How do I find out a rejection region? Rejection region is I find the rejection region based on the value of α that is the significance level that we have already seen in the last class. Then, we will collect the sample data calculate the test statistics. Next step after calculating the test statistics.

Since we know what is the rejection region then based on a test statistic we will be able to say whether it falls in a rejection region or it does not falls in the rejection region. If it falls so, based on that we will make a decision to either reject or fail to reject H_0 . Then finally, the last step is interpret the result in common language suitable for practitioners. After once I have reject the null hypothesis or we do not reject a null hypothesis.

Then after that, we will interpret the result in common language that is depending on the application. Now, from point number 4 when we make a decision to either reject or fail to reject, here I want to bring to your notice 1 important point is that one thing. If my null hypothesis is rejected, then definitely null hypothesis is rejected means the alternate hypothesis is accepted there is no other way out.

But if my null hypothesis is not rejected, that does not only mean that I accept my null hypothesis, it may mean I accept my null hypothesis, or I can also say failure to reject the null hypothesis unable to reject the null hypothesis. Both are accepting the null hypothesis or unable

to reject the null hypothesis both are not different and both are not same. We will see in some examples how in what way it is different you may think both are the same thing.

We are not rejecting the null hypothesis means we are accepting no that is not both are 2 different perspectives. We will see with examples.

(Refer Slide Time: 05:08)

Case Study 1: Coffee Sale

Problem

A coffee vendor nearby Kharagpur railway station has been having average sales of 500 cups per day. Because of the development of a bus stand nearby, it expects to increase its sales. During the first 12 days, after the inauguration of the bus stand, the daily sales were as under:

550 570 490 615 505 580 570 460 600 580 530 526

Monalisa Sarma
IIT KHARAGPUR

So, now this hypothesis testing procedure, we will see with some very simple example. First is a very simple example see, what it is given a coffee vendor nearby Kharagpur railway station since I am from Kharagpur I have mentioned it is Kharagpur railway station it can be any railway station. So, having an average sale of 500 cups per day this pen is really irritant do you know. So, as an average sale of 500 cups per day that is a hypothesized value.

Now, we will be using the hypothesized value. Because it has been this coffee shop has been running for ages it is not possible for a person to daily find out the average and come up with a service based on an as we know hypothesis, maybe an educated guess or based on some solid evidence. So, it can be hypothesized that sale is around 500 cups per day that means mean value is 500.

I can say mean of the population that is μ is 500 because of the development of a bus stand nearby it expects to increase it sells, obviously, bus stand is there so, many people will come. So, during the first 12 days after the inauguration of the bus then that daily sales were as under so,

after the inauguration of bus stand, I noticed that sales amount of sales per day for 12 days and this is the data.

(Refer Slide Time: 06:50)

Case Study 1: Coffee Sale

Problem

A coffee vendor nearby Kharagpur railway station has been having average sales of 500 cups per day. Because of the development of a bus stand nearby, it expects to increase its sales. During the first 12 days, after the inauguration of the bus stand, the daily sales were as under:

550 570 490 615 505 580 570 460 600 580 530 526

Question: On the basis of this sample information, can we conclude that the sales of coffee have increased? Consider 5% significance level.

Monalisa Sarma
IT KHARAGPUR



So, on the basis of this sample information, can we conclude that the sales of coffee have increased? We have where it has once and here we can consider 5% significance level. That means my significance level α is 0.05. So, now here null hypothesis is maintaining the status quo. Alternative hypothesis is what we want to test. Null hypothesis is with maintaining the status quo that means $\mu = 500$. That will make H₀ alternate hypothesis what we want to test we want to test whether the coffee sales has increased more than 500 earlier mean was 500.

(Refer Slide Time: 07:27)

Case Study 1: Step 1

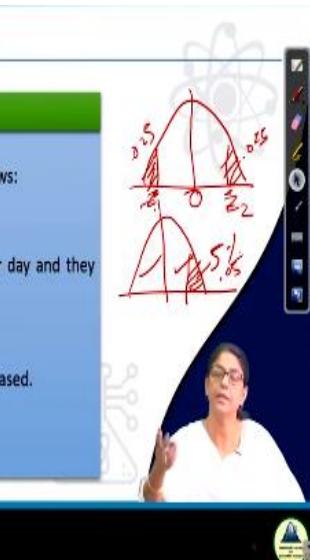
Step 1: Specification of hypothesis and acceptable level of α

Let us consider the hypotheses for the given problem as follows:

$H_0: \mu = 500$ cups per day
The null hypothesis that sales average 500 cups per day and they have not increased.

$H_1: \mu > 500$
The alternative hypothesis is that the sales have increased.

Monalisa Sarma
IT KHARAGPUR



So, specification of first step specification of hypothesis and acceptable level of α , so, this is my first step, $\mu = 500$ that is H_0 and this is H_1 , sorry this is not H_0 this is H_1 alternate hypothesis H_1 . Alternate hypothesis μ is greater than 500. The alternate hypothesis is that sales have increased means I want alternate hypothesis something what I want to test. So, and significant level of acceptable level of α is 5%. Why it is called an acceptable level of α ? Significance level α what does that mean actually?

In a hypothesis test α is the probability which is the maximum probability which is acceptable maximum probability of rejecting a true null hypothesis maximum accepting probability of rejecting a true null hypothesis. That means, α we call it a significance level that is like we can this much we can accept this much significance level this much type 1 error we can accept.

The type 1 definitely in all hypothesis as I mentioned before yesterday only in hypothesis testing is making an error is an inescapable part there may not be error, but then we cannot say that there will never be error, it is an inescapable part of null hypothesis testing. So, when it is an inescapable part let us accept the fact. So, what we accept and we accept give specifying α . we will accept the maximum acceptable level of rejecting a true null hypothesis this is α . I will maximum I will accept 5% that does that is the meaning of significance level.

So, here it is 5% now, again one more thing to say here, when as α is 5%. Now, I yesterday we had discussed what is 2 tailed tests and what is single tailed test. So, if it is a 2 tailed test, then rejection region will be in the both side rejection reason you say or critical reason you say it will be in the both side. If it is a 2 tailed test. So, when I specify 5% that means this 2 is put together is 5% that means, what is this? This is 0.025 this is 0.025.

When it is 2 tail and when it is single tail it will be one way either this way or this way, whatever way it is depending now, it is I want to prove that μ greater than 500. So, definitely my rejection region will be this side. So, it is 5% so, 5% is total this area will be 5%, that is 0.05. Because this side we are not considering. So, why we need α based on this α we will find out the corresponding Z value. What is the Z value cost?

Because Z value give me this value. But is this is the axis is the y axis what is the value corresponding to this point that is the Z value that is the starting of my critical region try to remember all this, this is just starting on my critical region or rejection region. So, this I am finding it in terms of Z distribution. So, this is 0 this side it will be minus this side it will be plus. So, once I found out my critical region Z let me tell this Z 1, let me tell this Z 2, this are my 2 critical region.

So, this is when I convert it to Z distribution, but however, in a given problem it is not in Z distribution it will I will be getting a sample and say any random variable say x . So, corresponding to this Z, what is my rejection region that also I can again find out there is nothing corresponding to Z 1 value what is the \bar{x} value? Correspondingly what is Z 2 value what is the \bar{x} value? That is my starting of the critical region we will see in some examples. So, now, given the; acceptance of 5% level of significance.

(Refer Slide Time: 11:35)

Step 2: Define a sample-based test statistics and the rejection region for specified H_0

Degree of freedom = $n - 1 = 12 - 1 = 11$

- As H_1 is one-tailed, we shall determine the rejection region (applying one-tailed in the right tail because H_1 is more) at 5% level of significance.
- Since the sample size is small and the population standard deviation is not known, we shall use t -distribution. The test statistic is t -value.
- Using table of t -distribution for 11 degrees of freedom and with 5% level of significance:

R: $t > 1.796$

Monalisa Sarma
IT BHARATPUR

So, now, step 2 is, define a sample based test statistics and the rejection region. So, what will be the test statistics here see here what it is given? The mean of the population is given but the standard division where the population is not known, when the standard deviation of the population is not known, and we have to infer about the population mean then definitely we will be using on my test statistics will be t value that means, we will be using t distribution.

So, with the degrees of freedom I have 1 parameter that is the degrees of freedom. Degrees of freedom will be sample size minus 1. So, it is 11 degrees of freedom and at 5% level of significance I need to find out what is the t value.

(Refer Slide Time: 12:24)

Case Study 1: Step 2

Using table of $t - distribution$ for 11 degrees of freedom and with 5% level of significance: $R: t > 1.796$

α	0.5	0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	0	1	1.376	1.961	3.078	6.314	12.71	31.82	63.66	318.31	616.62
2	0	0.816	1.061	1.386	1.866	2.92	4.103	6.965	9.925	22.327	31.599
3	0	0.765	0.978	1.25	1.638	2.353	3.882	4.541	5.841	10.215	12.924
4	0	0.741	0.941	1.19	1.533	2.131	2.775	3.747	4.604	7.173	8.61
5	0	0.727	0.92	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0	0.718	0.906	1.134	1.44	1.941	2.447	3.143	3.707	5.208	5.959
7	0	0.711	0.896	1.119	1.415	1.895	2.365	2.994	3.499	4.785	5.608
8	0	0.706	0.889	1.108	1.397	1.86	2.306	2.896	3.355	4.501	5.041
9	0	0.703	0.883	1.1	1.383	1.833	2.262	2.821	3.25	4.237	4.781
10	0	0.7	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0	0.697	0.876	1.088	1.363	1.783	2.201	2.718	3.108	4.025	4.637
12	0	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.025	3.93	4.318
13	0	0.694	0.87	1.079	1.35	1.771	2.16	2.65	3.012	3.852	4.221

Monalisa Sarma
IIT Kharagpur

So, this if you see the table t table here, see the t, this is 11 and this is corresponding to because 5% 0.05 it is 1 tail. So, I will take total 5% in one side only. So, it is 11 degrees of freedom 11 point by 5% significance level. So, this is my rejection region 1.796 that means, t distribution is 1.796 so, my rejection rates are 1.796. This is my rejection region t distribution fatter tails, so, I have drawn it this way. So, my rejection region is somewhere here. If my test statistics value falls in this region, then I reject the hypothesis.

(Refer Slide Time: 13:05)

Case Study 1: Step 3

Step 3: Collect the sample data and calculate the test statistics

Given the sample as

550 570 490 615 505 580 570 460 580 530 526

The test statistics t is

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

To find \bar{X} and s , we make the following computations.

$$\bar{X} = \frac{\sum X_i}{n} = \frac{6576}{12} = 548$$

Monalisa Sarma
IIT Kharagpur

So, given the sample is given from the sample I need to find out this value x bar value I need to find out the standard deviation that is S. So, given from a set of data how to calculate x bar? That is the mean how to calculate the standard deviation that you know it I am just skipping it.

(Refer Slide Time: 13:23)

Case Study 1: Step 3

Sample #	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	550	2	4
2	570	22	484
3	490	-58	3364
4	615	67	4489
5	505	-43	1849
6	580	32	1024
7	570	22	484
8	460	-88	7744
9	600	52	2704
10	580	32	1024
11	530	-18	324
12	526	-22	484
$n = 12$		$\sum X_i = 6576$	$\sum (X_i - \bar{X})^2 = 23978$

Now,

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

$$= \sqrt{\frac{23978}{12-1}}$$

$$= 46.68$$

Hence,

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

$$= \frac{48}{\frac{46.68}{\sqrt{12}}}$$

$$= \frac{48}{13.49}$$

$$= 3.558$$



So, I got the mean value I got standard deviation value then I put it into a t value and then I got my t is 3.558. Where is my what to say my what was my rejection region starting up rejection region my value was remember 1.796 my rejection region value was 1.7 this value is 1.796 and this value t value I got something 3 point something so, definitely it is in this area is not it? 3 point something would be maybe somewhere maybe. So, that means it lies in the rejection region.

(Refer Slide Time: 14:03)

Case Study 1: Step 4

Step 4: Make a decision to either reject or fail to reject H_0

The observed value of $t = 3.558$ which is in the rejection region and thus H_0 is rejected at 5% level of significance.



So, the observed value of $t = 3.5$, which is in the rejection region and H_0 that is null hypothesis is rejected at less than 5% level of significance.

(Refer Slide Time: 14:16)

Case Study 1: Step 5

Step 5: Final comment and interpret the result

We can conclude that the sample data indicate that coffee sales have increased.

Monalisa Sarma
IIT ROORKEE

Now, the final comment based on the application we can give the form it is come in simple language, we can conclude that the sample data indicate that a cup of coffee sales have increased because we are rejecting the null hypothesis means we are accepting the alternate hypothesis. What is the alternate hypothesis μ is greater than 500. So, that means we are accepting the fact that coffee sales have increased.

Now here I also want to bring to your notice is that some we call a result statistically significant when we call a result as a statistically significant. Please, this is very important. Please try to remember when a statistically significant is basically when we reject a null hypothesis how we reject the null hypothesis because the value that we get from the sample it is very much it is significantly different from the statistically different from the null hypothesis value.

That is why only we are rejecting the null hypothesis is not it? That why we are rejecting the null hypothesis? Because in a null hypothesis we have assumed a certain value of the mean assuming the value of the mean from the sample whatever data we got, if that data correlates with this mean, then this is that means the mean is correct, is not it? But when we reject when we get such a value, that it falls probability of which is very, very less that we specified whatever probability is acceptable to us and accordingly we specify the rejection region.

So, if the value sample test statistics values fall in the rejection region, that means, what sample value is statistically significant compared to the null hypothesis value. So, such results are called statistically significant, that means, when we say a result is statistically significant that means, definitely we are rejecting the null hypothesis. So, statistically significant is important term please try to understand this.

(Refer Slide Time: 16:14)

Case Study 2: Machine Testing

A medicine production company packages medicine in a tube of 8 ml. In maintaining the control of the amount of medicine in tubes, they use a machine. To monitor this control a sample of 16 tubes is taken from the production line at random time interval and their contents are measured precisely. The mean amount of medicine in these 16 tubes is calculated as 7.89, the standard deviation of the population is estimated to be equal to $\sigma=0.2$. Use hypothesis testing to infer if the machine is indeed working properly.



Monalisa Sarma
IIT KHARAGPUR

So, we will see one more example. So, say this example thing we have written I have explained it while explaining the hypothesis testing, how to form the; what to say hypothesis. So, here however, there are some other information are given compared to the previous one. Previously, and I am just quickly going because we have already discussed this medicine production company packages medicine in a tube of 8 ml that is my mean is 8 ml in maintaining the control of the amount of medicine in tubes, they use a machine.

To monitor this control a sample of 16 tubes is taken from the production line at random time interval and the contents are measured precisely. We have already done that that was we have already formulated the hypothesis. What is now what is given the mean amount of medicine of the 16 tubes is calculated as we have taken a sample from the sample the mean amount sample of 16, mean amount is calculated 7.89.

And the standard deviation a population not the sample, but the population is estimated to be 0.02 use hypothesis testing to infer it a machine is indeed working properly. So, see here there are 2 things so, in some problem the significance level may not be mentioned. So, there are 2 ways one is if a significance level is not mentioned then we will report a P value what is p value I will come later.

Now, another test for another way, if the significance level is not mentioned, in general 5% significance level is considered. If it is not mentioned, or if it is if you are asked to get a P value, then it is a different thing then you do not have to consider a significant level. But if you are not asked to give the p value, what is p value we will later again. So, then you will have to assume is 5% significance level, here the significance level is not given.

So, let us take the significance level of 5%. Now here how to formulate what to say hypothesis first is definitely $\mu = 8$ ml status quo that means we know that mean is 8 ml, we are hypothesizing that it is 8 ml. And what we want to test that it is not equal 8 ml not neither less or greater, we are not interested in less or greater, we have what we want to test that it is not equal to 8 ml. So, that is my alternate hypothesis, alternate hypothesis is $\mu \neq 8$.

(Refer Slide Time: 18:40)

Case Study 2: Step 1

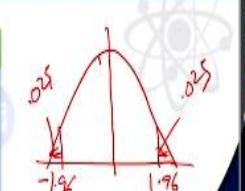
Step 1: Specification of hypothesis and acceptable level of α

The hypotheses are given in terms of the population mean of medicine per tube.

The null hypothesis is
 $H_0: \mu = 8$

The alternative hypothesis is
 $H_1: \mu \neq 8$

We assume α , the significance level in our hypothesis testing ≈ 0.05 .




NPTEL

Monalisa Sarma
IIT KHARAGPUR

So, this is my that is how I found a hypothesis $\mu = 8$ H 1 $\mu \neq 8$, then some significant level 5%. We have assumed, so, since it is 5%, it is 2 tailed means one side it will be 0.025, aside will 0.025 and as I said it will be 0.025. So, my when I talk of my rejection region. So, 5% means say

this is 2 things 2 sides this and this both put together to 5 person so that means this will be 0.025, this will be 0.025 and then I will have from the table I will be able to find out the Z value corresponding to this 0.025.

So, Z value corresponding to this is I can remember actually it is minus 1.96. And this is plus 1.96. So, for the thing to be added to and null hypothesis to be accepted my values should be within minus 1.96 to plus 1.96. You can see the table.

(Refer Slide Time: 19:49)

Case Study 2: Step 2

Step 2: Sample-based test statistics and the rejection region for specified H_0

Rejection region: Given $\alpha = 0.05$, which gives $|Z| > 1.96$ (obtained from standard normal calculation for $n(Z:0,1)$).

NPTEL

Monalisa Sarma
IIT KHARAGPUR

So, which sample based test statistics we will be using here. See here what it is given population mean is given population standard deviation is given, we have to infer about the population mean then definitely we will be using their distribution.

(Refer Slide Time: 20:03)

Case Study 2: Step 3

Step 3: Collect the sample data and calculate the test statistics

Sample results: $n = 16$, $\bar{x} = 7.89$

Estimated standard deviation, $\sigma = 0.2$

With the sample, the test statistic is

$$Z = \frac{7.89 - 8}{\frac{0.2}{\sqrt{16}}} = -2.20$$

Hence, $|Z| = 2.20$

$|Z| = 1.96$

Monalisa Sarma
IIT KHARAGPUR

And collect from the data we will come to find out the Z value. So, that will what we got minus 2.20. So, minus 2.20 means, it is greater than the rejection region is not it? What is my rejection region? Rejection region is $Z = 1.96$ from minus 1.96 to plus 1.96 and my calculated Z value is greater than that and so, it is in the rejection region.

(Refer Slide Time: 20:30)

Case Study 2: Step 4

Step 4: Make a decision to either reject or fail to reject H_0

Monalisa Sarma
IIT KHARAGPUR

So, you see this figure this is my rejection region and I got my value here in this region it may be positive or negative whatever it is I got in this region so, we reject H_0 .

(Refer Slide Time: 20:45)

Case Study 2: Step 5

Step 5: Final comment and interpretation of the result

We conclude $\mu \neq 8$ and recommend that the machine be adjusted.

NPTEL
Monalisa Sarma
IIT KHARAGPUR

So, we conclude that μ is not equal to 8 and recommend that a machine be adjusted that means, there is some problem in the machine and we recommend that that machine is adjusted.

(Refer Slide Time: 20:56)

Case Study 2: Alternative Test

Suppose that in our initial setup of hypothesis test, if we choose $\alpha = 0.01$ instead of 0.05, then the test can be summarized as;

1. $H_0: \mu = 8, H_1: \mu \neq 8 \quad \alpha = 0.01$
2. Reject H_0 if $Z > 2.576$
3. Sample result $n=16, \sigma = 0.2, \bar{X}=7.89, Z = \frac{7.89-8}{0.2/\sqrt{16}} = -2.20, |Z| = 2.20$
4. $|Z| < 2.20$, we fail to reject $H_0 = 8$
5. We do not recommend that the machine be readjusted.

NPTEL
Monalisa Sarma
IIT KHARAGPUR

Now, why I have taken this example actually, to prove a point, I basically wanted to come gradually to p values that you have to this example again though we have discussed this example in my first lecture, see here specifying significance level for the same problem, same problem everything data remains everything same. Now, let me take a significance level of $\alpha = 0.01$ instead of 0.05. So, if I take a significance level of $\alpha = 0.01$.

So, what happens, my reject H_0 my Z value will be 2.576, corresponding to 0.01. So, what it will be my both sides it will be 0.005. This side is 0.005, this side it will be 0.005. If you see the

Z table, you will get corresponding value to that is 2.576. So, from the test statistics what value I got from the test statistics I got value 2.2. So, now, that means, it is not rejected failed I failed to reject the status the data statistics the sample data failed to reject the null hypothesis I got value 2.0 and this is 2.57 says it is less than that.

So, similarly the same problem suppose, as I previously when I was discussing sampling distribution I told them if you take different sample data statistics from the sample will different is a very slight chance that statistics of many 2 samples are same. So, I have taken a sample from the sample I got \bar{X} bar is 7.89. In this example, I have that is what I have assumed this from \bar{X} bar I got 7.89 suppose I took a different sample and from there suppose in the \bar{X} bar I got \bar{X} bar = 7.91.

Let me happen I took a sample 1 sample from which I got 1.89. Another sample I got from which I got \bar{X} bar = 7.91. Now, for the 7.91 only, if I take 5% significance level only no need of taking 1%. If I take 5% significance level only still what will happen my I will fail to reject H_0 null hypothesis. See such a slight change from 7.89 and I got 7.91, for 7.89 I had to reject the hypothesis for 7.9. I could not reject the hypothesis.

If I change the significance level 5% significance level I have rejected the hypothesis for 1% significance level we fail to reject the hypothesis then what about if I take 2% significance level if I take 3% significance level. So, that is why usually this sort of problem is not carried out manually the sort of problem is the problem is done automatically in a computer. So, in the computer when we feed the algorithm basically for this particular this is the critical reason.

Then from what sample whatever data it gets, and what the computer checks this data is less than this then do not reject if it is greater than this reject it is just reject or no reject there is no other option here. But see for slight changes of data we get different results. So, that is why there is one more approach instead of directly telling it is reject or not reject one more approach of mentioning the statistical results the same significance test results or hypothesis test results is by reporting the p value.

(Refer Slide Time: 24:31)

p Values

- The p value is the probability of committing a type-I error if the actual sample value of the statistic is used as the boundary of the rejection region.
- It is also interpreted as an indicator of the weight of evidence against the null hypothesis.

Why p value reporting is desirable?

- The significance level need not be specified by the statistical analyst.
- If the person who makes the decision and the statistical analyst are two different person, p value reporting is more preferred.
- The analyst provides the p value and the decision maker determines the significance level based on the costs of making the type I error.



Monalisa Sarma
IIT KHARAGPUR



So, instead of directly telling that to reject the hypothesis or do not reject the hypothesis, we do not what we do is that we report the p value when first of all when this p value is used mainly suppose the person who takes the decision is someone else and the person who does the statistical test is someone else. So, suppose I am doing the statistical test. So, I got all the data and I will just report the data to the decision maker.

The decision maker will see the data and will try to find out whether, we need to based on this data whether I need to reject the hypothesis or I need to accept it that is totally based on the decision maker because he has total knowledge about the system, he will be able to take proper decision whether to reject or not to reject based on the, how the data has come. So, as a statistician, I am just doing all the data, I am just collecting the data calculating all the values and giving you so now in that case.

So, when I am not directly specifying, reject or not reject what I am doing is I am specifying the p value. So, what is p value? p value is that basically if u see the definition p value is the probability of committing a type 1 error is the actual sample value of the statistic is used as the boundary of the rejection region, it is the definition of p value. Let us not go into those definitions just simple p value basically what we from the statistics whatever value you got like here this is the from the statistics you got the value 2.20.

From the sample this is a sample statistic these statistics we got from the sample. So, what we will do is that from the table, we will find out what is the probability of this occurrence that is nothing but the p value of this set of data, what is the probability of this 2.20 that is nothing but the p value. So, in fact, now, here this is the p value is the probability of committing a type 1 error in the actual sample value of a statistic is used as a boundary region or the rejection region.

Why is it telling that why is defining p value in such a way when suppose, in significance level is not mentioned. So, whatever p value I get based on that if I reject the null hypothesis that means, that for the p value corresponding Z value corresponding for that p value that is my starting region that is my rejection region starting of the rejection region. So, you do not have to remember this definition just if you understand what is p value, p value is the probability of getting that t statistic value.

So, it is also interpret it as an indicator of weight of evidence against the null hypothesis why it is indicator of weight of evidence? Because we have assumed something null hypothesis is what we have assumed what we have hypothesized and from the sample whatever we got now, as you mean this what we got what is the probability of weight that is the way it is, is not it? It is the indicator of the weight of evidence against the null hypothesis.

So, why p value reporting is desirable? The significance level need not be specified by the statistical analyst. It the person who makes the decision and the statistical analysts are 2 different person p value reporting is more preferred. The analysts provide the p value and the decision maker determines the significance level based on the cost of making the type 1 error. If the cost type 1 error already I have specified I have mentioned some what is this type significance level?

That is the type 1 error significance levels is the maximum acceptable significance level or maximum probability of type 1 error that will be accepted maximum acceptable type 1 error acceptable means what? Acceptable probability of rejecting a true null hypothesis. So, when so, it is the analyst a decision maker will decide how much acceptable well how much acceptable error I can take 1% error 2% error, it totally depends upon the application if the application is a very critical application, then my error what do I mean by critical applications?

How do I select a null hypothesis? How do I select a significance level? I will be discussing in my next lecture. So, basically, so, there are different factors based on the application. So, it is so, the decision maker will based on these different factors will take the decision whether this is the p value of the statistical analysis given me this is the p value. Now, based on I know what sort of what is what application is this?

What is the cause of the cost of an error? So, based on that I can decide whether I will accept the null hypothesis or I will not. Whether I will reject the null hypothesis or I will not reject a null hypothesis.

(Refer Slide Time: 29:29)

CONCLUSION

- ④ In this lecture we learn in details about
- ④ Hypothesis testing procedure
- ④ Hypothesis testing – case studies
- ④ p values
- ④ In next lecture we will learn more about statistical inferences

NPTEL
Monalisa Sarma
IIT KHARAGPUR

So, in this lecture we learn in detail about the hypothesis testing procedure. We have seen few case studies we have also understood the concept of p values. And in the next lecture, we will learn some more about statistical inferences.

(Refer Slide Time: 29:42)



So, these are the references and thank you guys.

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

Lecture - 24
Statistical Inference (Part 3)

(Refer Slide Time: 00:33)

The slide has a dark blue header bar with the title 'Concepts Covered'. Below the header, there is a list of three items:

- ➊ Type-I error, Type-II error: Which is more important?
- ➋ Plotting of OC curve
- ➌ Power curve

In the bottom right corner of the slide area, there is a video feed of a woman, identified as Prof. Monalisa Sarma, speaking. The video feed is framed by a white border. At the very bottom of the slide, there are two small logos: NPTEL on the left and IIT KHARAGPUR on the right.

Hello, everyone so in continuation of our earlier lecture on statistical inferences. Today, we will be learning few more topics. So, in my earlier lecture, we have learned what is type 1 error, what is type 2 error. Now we will see which is more important, why we should, mainly we will see we focus more and more on type 1 error. That is why we have that significance level what I have mentioned significant level is mainly with respect to the type 1 error.

I did not mention about type 2, error, why we focus more on type 1 error, why it is considered more important? We will be discussing that and then how to plot the OC curve, what is an OC curve? How to plot that? And then we will discuss what is an power curve.

(Refer Slide Time: 01:08)

Type-I Error

In hypothesis testing we focus more on Type I error.

- It is the hypothesis that requires no action to be taken, no money to be spent, or in general nothing to be changed.
- It is usually costlier to incorrectly reject the status quo than it is to do the reverse.

Monalisa Sarma
IIT KHARAGPUR

So, in hypothesis testing, we focus more on type 1 error, why? That you have noticed we focus more on type 1 error why? Because see, when the first step is, we specify a H_0 and a significant level what is the significant level? Significant level is the maximum acceptable probability of rejecting a true null hypothesis, is not it? So, that is α that is what type 1 I did not mention anything about type 2 that means what type 2 is not important, we will see that. So, what is basically the null hypothesis?

Null hypothesis is the hypothesis that requires no action to be taken no money to be spent, or in general nothing to be changed. Already I have mentioned about when I talked about what is and how to form the hypothesis. So that is my null hypothesis. Null hypothesis is the maintaining the status quo like where we do not want to do any action. No, we do not have to do any more.

We do not have to spend any money we do not have to change anything so that is my null hypothesis. So say, like in my example, that machine that fills medicines in the 8ml bottle, so that there is a machine which fills exactly 8ml medicine in this bottle that is my null hypothesis. Remember, I have specified my null hypothesis $\mu = 8$, then what I wanted to prove? I wanted to prove μ is not equals to 8. So, see why we have focused here on type 1 error.

Now, suppose I have rejected a true null hypothesis. What happens if I have rejected a true null hypothesis suppose that the medicines that was actually producing 8ml actually filling up 8ml medicines only in the bottle, but then the sample results but I got we already saw. So

when I discussed about the p value remember, if you forgot, then please, I suggest you to go back to those slides and see again, when I sometimes what have been a slight change for a slight change, we can have different results is not it?

So, maybe it is for such sample we got such a result for with my value is falling the rejection region, my value is falling in the rejection measured region means I am rejecting the null hypothesis. But what happens, but actually maybe my null hypothesis is actually true, that was a stray sample which I took and I got that value. So, what happens are I have rejected the null hypothesis.

When I rejected the null hypothesis, that means the machine is not working properly, I will have to change the whole machine maybe I have to invest crores of money for changing the machine or whatever spending lots of money, lots of time, energy, everything, which actually it is not necessary. So, it is easily costlier to incorrectly reject the status quo than it is to do the reverse. So that is the main thing why we focus on type 1 error.

(Refer Slide Time: 04:21)

The slide is titled "Type-I Error". It features a cartoon character at the top left. Two buttons are present: "How to choose the value for Type I error probability" and "Let us understand this with few examples.". A video player window on the right shows a woman speaking. Below the video, there is a blue box containing text for "Example 1". The text reads: "Assume that six-year-old children should average about 10 kg in weight to be considered normal. Considering that a sample of children from a low-income neighborhood is to be tested for subnormal weight. Design your test for the above." Logos for NPTEL and IIT Kharagpur are at the bottom left, and a small circular logo is at the bottom right.

So now, knowing this, so hypothesis testing, our first step is we specify the hypothesis null and alternate. And we specify the significance level, significance level is totally depends on the null hypothesis means significance level corresponds to the acceptable probability of the rejection of the true null hypothesis. So, now the question is, how do we understood why we had to focus on type 1 error.

So, now for a given application when we have to do hypothetical testing, so we will have to frame our hypothesis in such way, so that our null hypothesis is that which maintains the status quo. So, and when it is that which maintains the status quo and then our significance level should be such that we have to select the significance level such that, so, that we can even if there is some error there will not be much harm.

So, there as I told you depending on the situation the significance level is selected. So, when I also have mentioned when I talk a p value the decision maker will decide based on different factors what should be the rejection region. So, depending on the how costly it will be if I reject a true null hypothesis? How difficult it will be if I reject a true null hypothesis? Based on that level of difficulty based on that level of expenses.

I can decide my type 1 error probability that is my value of α suppose it is a very critical application. So, if I cannot manage to reject the null hypothesis for no reason if I cannot manage to reject a true null hypothesis very critical it is, because if I reject the true null unnecessarily I will spend crores of money then I will keep my type 1 error probability very, very less, maybe say 0.001.

So, however, maybe if some sort of some error I can manage, I do not need to bother much about the even if it is a type 1 error that will do not make much of a difference then I will keep my type 1 error probability a bit higher. Now, thing is that why should I keep it higher? Because already in my first lecture on statistical inferences, I have already seen which I will again talk more details on it was I have already seen when I am increasing α my β decreases that means, if I increase my type 1 error, my type 2 error decreases.

If I decrease my type 1 error, if I take my type 1 error very, very less value, my type 2 error will become very, very high, one increases the other decreases. So, for situation where the type 1 error is very, very critical, I will have a very less type 1 error probability, but at the same time my type 2 error will increase. So, since because of this relation, so, wherever the situation that I can bear a bit greater type 1 error probability then I will take a better and greater type 1 error probability so, that I get my type 2 error also reduces.

So, we will understand this with few examples. So, first consider this case assume that a 6 year old children would average about 10kg in weight and that is considered normal. So,

what are the; suppose consider that a sample of children from a low income neighbourhood is to be tested for subnormal weight suppose in a slum area like this children are malnourished. So, government is trying to bring out some nutrition program it just in the first day.

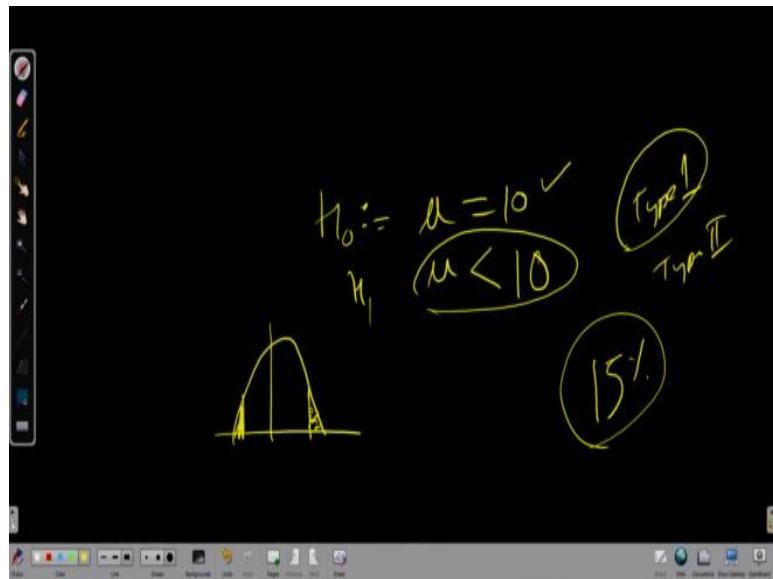
They will first find out whether the children are really malnourished if they are malnourished some program will be some meal some nourishing program government will start. If they find that the children are okay the children weight of 6 year old children on an average weight is 10kg then government need not start that program. So, now see here what will be my normal? What will be my null hypothesis?

What will be my alternate hypothesis what will be my significance level? So, here my null hypothesis definitely μ is 10kg that is I want to maintain definitely I will always want to maintain the status quo that will I will $\mu = 10\text{kg}$ but what I want to prove that μ is less than 10kg because if it is less than 10kg, then I will have to start some nutritional program for the children of that slum area. So, what I want to prove is less than 10kg.

Now here if I consider what to say $\mu = 10\text{kg}$ and I sorry let us not take this this way. For this example, what I want to prove is that μ greater than 10kg why? I want to prove μ greater than 10kg if I want to prove me greater than 10kg, then I will not have to start this nutritional program why it actually totally depending on the situation that different hypotheses are formed.

So, here maybe if my intention is if not required, I will not start a nutritional program in that case I will make the type alternate hypothesis is μ greater than 10 but see the significance here. So, that means when my null hypothesis is 10 if I reject a null hypothesis if it is true and I still reject the null hypothesis let me first write it.

(Refer Slide Time: 10:21)



So, this is that it is colon $\mu = 10$. Then let us take that will be more easier, where μ is less than 10 this will fit more as an example this we are taking just an example to bring to the point how we select the significance level. So, here if I take $\mu = 10$. So, if my null hypothesis is rejected, then what happens this hypothesis will be accepted H_1 will be accepted and the nutritional program will start.

So, if what happens if I do not if by mistake the null hypothesis is actually true. So, here, but null hypothesis is actually true, but I have rejected a true null hypothesis under what situation I have accepted it rejected a true null hypothesis that means, if it falls in my rejection region, null hypothesis is actually true, but the sample what I got from the sample I found that no null hypothesis and the value is falling the critical reason.

So, I will reject the null hypothesis then what happens a nutritional program will start an unnecessary expense to the government. But then you see the other way around what happens if the null hypothesis is not true actually, but I still accept it that means a type 2 error I got such a sample from the sample I found that null hypothesis is true only. But actually the null hypothesis is not true.

Actually the sample what I got that was a some outlier I got and I found that, the weight of the children is more equal or greater than 10. So, that means I do not have to reject the null hypothesis then what happens the nutritional program did not start now which will have more effect. So, if in the slum area if there is children even if they do not need it, if the nutrition program is started.

Government will spend some money that is all but this children definitely it will be good for the children only because as it is their poor children they will get to eat. So, what so, here if my type 1 error is not very serious, just that government has to spend some money, but if I commit type 2 error, what is type 2 error? Type 2 error means my null hypothesis is not true, but still I accept it, then it may be dangerous, the children are malnourished, still the government did not start the program.

So, it will have a dangerous effect on the children. So, here my type 2 error is very important understood the point. So, in such case hypothesis testing we always specify the type one error acceptable type one error. So, this is the case where type 1 error is not very critical. So, what I will do, I will specify a greater type 1 error so that I get a lesser type 2. So, maybe it is the in standard type 1 error is usually 5% or 1%, maybe I take some type 1 error of say 15%.

Poor children even if their weight is fine if the null hypothesis if and by mistake also we have rejected the null hypothesis, let the nutritional program start. So, I am ready to accept it 15% error probability understood so, now, this is 1 example.

(Refer Slide Time: 13:57)

The slide is titled "Type-I Error". It contains two examples:

- Example 1:** Assume that six-year-old children should average about 10 kg in weight to be considered normal. Considering that a sample of children from a low-income neighborhood is to be tested for subnormal weight. Design your test for the above.
- Example 2:** A drug company wants to test a new drug for (1) the toxicity (side effects) and (2) the effectiveness. Design the test for both the cases.

Handwritten notes on the slide include:

- $H_0: \text{Drug is toxic}$
- $H_1: \text{It is not toxic}$
- $\alpha = 0.05$

At the bottom left, there are logos for NPTEL and Monalisa Sarma, IIT KHARAGPUR. At the bottom right, there is a small circular logo.

The next say the next example, a drug company wants to test a new drug for the toxicity it wants to test the new drug for toxicity see here. So, here what will be the null hypothesis will be a drug is toxic alternate hypothesis, because what I want to prove that it is not toxic. Here again, the as I told you, this formation of this hypothesis is totally depends on who is trying to do here the drug company itself is trying to do.

So, drug companies what to say its objective will be trying to prove that it is not toxic. But if some third party a third party is trying to prove third party will try to prove that it is toxic. So, then accordingly my hypothesis will be different. It totally depends on who forms the hypothesis and which is more important while forming that hypothesis. Alternate hypothesis is something which we always want to test that is the alternate hypothesis. So, my H_1 is it is not toxic.

So, now here the drug is toxic. So, here suppose here what happens if I by mistake if I reject the null hypothesis then what happened given the drug is toxic I have reject the null hypothesis then what happens the drug is then it is proved that drug is not toxic see the effect it will have the drug is toxic and you got your specifying the drug is not toxic. So, it will have a very, very dangerous effect.

So, in this case my type 1 error is very, very significant in such case maybe I will select my type 1 error will be say 0.0001 let β the more we will see even if the β is more it will not have much effect we will see what but here let β be more but we cannot manage a higher type 1 error because a higher type 1 error means it will result in a toxic drug turned as non toxic.

(Refer Slide Time: 16:15)

Type-I error

Can we completely ignore Type II error?

If the probability of making a type II error is very large, then the test may not be useful.
Because of the trade-off between α and β , we may find that we may need to increase α in order to have a reasonable value for β .

Monalisa Sarma
IIT KHARAGPUR

So, now, the question is can we completely ignore type 2 error? Now, in the second question see, if we have selected a very, very, very small type 1 error that means our type 2 error will become very big. So, that means is it that we can completely ignore type 2 error? No, we

cannot ignore, but there is a twist to it, we will see how that is what if the probability of making a type 2 error is very less than this may not be useful also.

So, we will see, because the trade off between α and β , we may find that we may need to increase α in order to have a reasonable values of β . So, that totally depends on the application what α will keep, which is more important type 1 or type 2.

(Refer Slide Time: 16:54)

Type-II error

- ➊ Calculating β is not always straightforward.
- ➋ Consider an alternative hypothesis, $H_1: \mu <> 8$, encompasses all values of μ not equal to 8.
- ➌ Hence there is a sampling distribution of the test statistic for each unique value of μ , each producing a different value for β . Therefore β must be evaluated for all values of μ contained in the alternative hypothesis, that is, all values of μ not equal to 8.
- ➍ However, for practical purposes it is sufficient to calculate β for a few representative values of μ .
- ➎ Use these values to plot a function representing β for all values of μ not equal to 8.
- ➏ A graph of β versus μ is called an "operating characteristic curve" or simply an OC curve.

Monalisa Sarma
IIT KHARAGPUR

Now, calculating β remember in the first lecture, I told 2 things that we will be discussing one is how to calculate β though for the toy example that is box of chocolate I calculated β but this is a toy example that is not applicable for any that is just to understand the concept, but calculating β is not a very easy job it is not straightforward. That is so that is the reason I did not show, how to calculate β over there.

Now, you have enough knowledge. So, now we can see how to calculate β . Secondly, I also mentioned one thing and that lecture itself the null hypothesis always should be with equal sign that why I will come here thirdly, why we were to say focus more on type 1 error that we have seen here. Now, how to calculate β and why null hypothesis we need an equal sign that we will see here.

So, see any sort of hypothesis testing my backbone is the; my main CPU I should say my CPU is the sampling distribution everything I put data I give to the simple CPU and the CPU gives me the results, is not it? So, every data I give to the sampling distribution from the sampling distribution, whatever result I get based on the result only I give the decision. So,

for sampling, when I have the sampling distribution, what happened my sampling distribution means that is one distribution which is normally distributed.

And which has a mean and its mean is what its mean is equals to the population mean and it has is variance is population mean divided by population variance divided by sample size that is the variance, forget about the variance. Now, the sampling distribution, it mean is equals to the population mean. So, if my hypothesis if always I try to find out the data based on my null hypothesis what we have a hypothesis that is $\mu = 8$.

So, when I find out the sampling distribution in my sampling distribution, so, my mean is 8. So, if I do not have an equality sign, then I cannot have 1 sampling distribution, I will have to have many distribution with differs each values of μ if I write μ not equals to suppose that medicine example that is mean is equals to 8 mean not equals 8, it means not equals to 8 means I will have to have distribution for all different all possible values of 8 is not it? All possible values of 8 means 7.9.

For 7.9 that means, I will have to have a distribution μ 7.9 and variance whatever it is, then whatever suppose 8.1, I will have to have a value for distribution for 8.1. So that way, I will have to find out the α values for all different values of μ . That was the reason why to simplify the process. My null hypothesis always need to have an equal sign so that we can have 1 sampling distribution which mean is equals to the population mean.

Now, when we are trying to find out the type 2 error for finding out the type 2 parameters, we have to find out the rejection regions or basically type 2 error means, we are not rejecting a true null hypothesis means, it is false but it is falling in the acceptance region. So, definitely we need to have a sampling distribution. So, when we have a sampling distribution we need to have the mean of the distribution.

So, when the value is not equals to 8 what will be the value? So, that means, we will have different distribution for different values of μ that is why calculating β is not always straightforward that was a toy example that box of chocolates so, we could do it very easily. So, here so, hence the sampling distribution hence, there is a sampling distribution of that for each unique value of μ for each unique value of μ there will be 1 sampling distribution each producing a different value of β .

Therefore, β must be evaluated for all values of μ containing the alternate hypothesis that is all values of μ not equals to 8 we will have to take μ not equal to 8 there can be any value in final values. So, for all values of μ not equals to 8 we will have to find out β . So, however, for practical purpose we will see it is sufficient to calculate β for a few represented values of μ if we take some few representative values of μ .

Then it is sufficient to find out the value of β then this represent the values of β for this values of μ different values of μ what β we get, we can plot this and what we call this this plotting is nothing but it is called operating characteristic curve. So, we will see what is an operating characteristic curve.

(Refer Slide Time: 21:48)

Example: Plotting of OC Curve

Problem

A medicine production company packages medicine in a tube of 8 ml. In maintaining the control of the amount of medicine in tubes, they use a machine. To monitor this control a sample of 16 tubes is taken from the production line at random time interval and their contents are measured precisely. The mean amount of medicine in these 16 tubes will be used to test the hypothesis that the machine is indeed working properly. Assume that we know that σ , the standard deviation of the population of volume, is 0.2 and that the distribution of volume is approximately normal.

Monalisa Sarma
IIT KHARAGPUR

So, we have again taken the same example to plot the operating characteristic curve.

(Refer Slide Time: 21:55)

Solution: Plotting of OC Curve

Solution: Hypothesis formulation and acceptable α

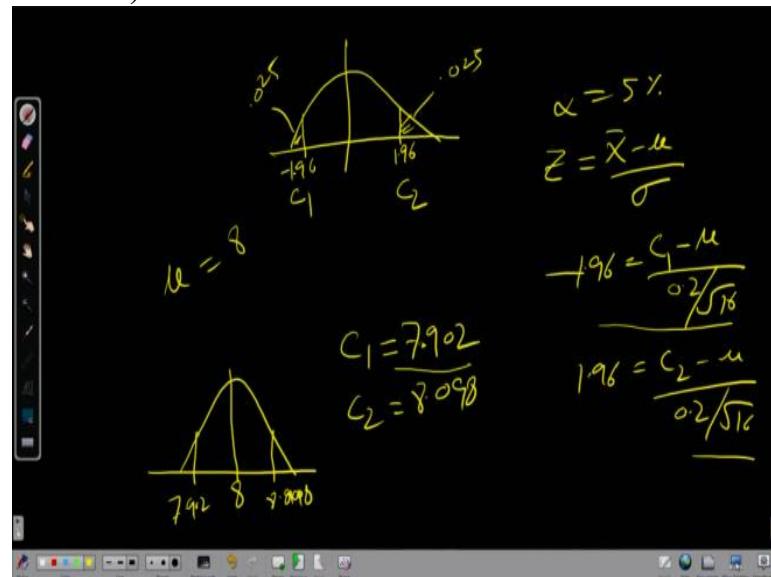
- ④ Specification of hypothesis and acceptable level of α
- ④ The hypotheses are given in terms of the population mean of medicine per tube.
- ④ The null hypothesis is $H_0: \mu = 8$
- ④ The alternative hypothesis is $H_1: \mu \neq 8$
- ④ We assume α , the significance level in our hypothesis testing ≈ 0.05 .

A medicine production company packages medicine in a tube of 8 ml. In maintaining the control of the amount of medicine in tubes, they use a machine. To monitor this control a sample of 16 tubes is taken from the production line at random time interval and their contents are measured precisely. The mean amount of medicine in these 16 tubes will be used to test the hypothesis that the machine is indeed working properly. Assume that we know that σ , the standard deviation of the population of volume, is 0.2 and that the distribution of volume is approximately normal.

NPTEL
Monalisa Sarma
IIT KHARAGPUR
37

So, here you see what was our hypothesis, hypothesis is $\mu = 8$ alternate hypothesis is μ not equals to 8 significance level we have considered 5% significance level only.

(Refer Slide Time: 22:07)



So, here when I have found out the significance level, the see very carefully z is equals sorry significance level $\alpha = 5\%$, when $\alpha = 5\%$ means, what if this is my what to say distribution that means, this bellow this area this area is 0.525 and this area is 0.025. Then I found out what is the z value corresponding to this z value corresponding to this. So, I think you remember it was 0.96 and - 1.96 this is in terms of z, is not it?

But I need the value in terms of to find out the rejection region I may also need to value in terms of the actual parameter value random variable what does my random variable my random variable was x based on that I found out x bar, is not it? So, this z I got value

converting from x to z how are we what is the z value that is equals to $\bar{x} - \mu / \sigma$ is not it? This is how we calculate the z value.

Now, this let me tell this as this counterpart in the as a random variable. So, this is c_1 this is c_2 then what is the z value $- 1.96 = c_1 - \mu$ is 8 and what was σ ? Σ was in this example is I think 0.2 divided by 16. We have taken a sample of size 16 and standard deviation is 0.2. So, similarly again $1.96 = c_2 - \mu$ $0.2 / \sqrt{16}$. So, we calculate the 3 we are from here we calculate the value of c_1 one from here we calculate the value of c_2 .

So, c_1 we get if I can simplify it c_1 will get I think 7.902 I remember correctly and c_2 will we will be getting some 8.0 90 so, my rejection region is $\mu = 8$, so, my rejection region will start from 7.902 to 8.098. That means, here I will draw I am drawing the graph again. So, if this is the thing, this is my rejection region 7.902 and this is 8.098. So, this is my rejection region here this is 8 that means acceptance region $\mu = 8$.

Remember why what is this acceptance is I am repeating it again $\mu = 8$. So, $\mu = 8$ means the under what situation I will accept it μ is 8 but still a bit of variance I can consider that so, that was my logic remember bit of variance I can consider. So, how much variance I can consider maximum and this side it is 7.9 and this side it is 8.1 maximum this much variants I can consider, is not it?

That is what so, basically that is I have just told 7.9 or 8.1, but when I specify the significance level from the significance level I could find out what is my this critical region or start of the rejection region.

(Refer Slide Time: 25:26)

Solution: Plotting of OC Curve

To construct the OC curve we first select a few values of μ
— let, $\mu = 7.80, 7.90, 7.95, 8.05, 8.10$, and 8.20

Next, we calculate the probability of a type II error at these values.

$$\begin{aligned}\beta &= P(7.902 \leq \bar{X} \leq 8.098 \text{ when } \mu = 7.95) \\ &= P\left(\left|\frac{7.902 - 7.95}{0.05}\right| \leq Z \leq \left|\frac{8.098 - 7.95}{0.05}\right|\right) \\ &= P(-0.96 \leq Z \leq 2.96) = 0.8300\end{aligned}$$

$$\begin{aligned}\beta &= P(7.902 \leq \bar{X} \leq 8.098 \text{ when } \mu = 8.05) \\ &= P\left(\left|\frac{7.902 - 8.05}{0.05}\right| \leq Z \leq \left|\frac{8.098 - 8.05}{0.05}\right|\right) \\ &= P(-2.96 \leq Z \leq 0.96) \\ &= 0.8300\end{aligned}$$

Note that, for $\alpha = 0.05$, the rejection region is $\bar{X} < 7.902$ or $\bar{X} > 8.098$.

Monalisa Sarma
IIT KHARAGPUR

So, now here so, now, we will have to find out the value of β for the different values of μ , because μ is not equal 8. So, different value we have taken some representative value. So, what we have taken we have taken see here 7.80. And this side we have taken 8.20 we have taken equal what to say symmetric value both the side. So, we have taken 7.90 have taken 8.1, 7.95, 8.05.

So, now, we have to calculate the type 2 probability, type 2 probability means, actually, my null hypothesis is false, but I am accepting it that means it should move on when a null hypothesis is false, but my value is falling in the acceptance region that means μ value is falling between 7.9 to 8.1, but actually it is false. So, that is my type 2 error. So, β is a probability that my X bar value falls within this range when μ is actually my mean with 7.95 μ is not a 8.

But when my μ is 7.995 my X bar is falling within this range that is my β . So, what is this corresponding to this X bar value, corresponding to this X bar value what is that value I can find out so, this is the Z value. So, what is the probability of β is 0.8300 from the Z table I can find out and I can remember how do we find out this and again repeating it we got it from here this portion then this portion.

So, then this portion minus this portion will give me this portion. So, this we found out for musical 7.95 my β value is 0.8300. Similarly, I found out for $\mu = 8.05$ what is my β value, I got the same thing, same β value, it is symmetric. So, considering this is a 7.902 and 8.098 is

the rejection region that is what I have just now shown why this is the reduction in corresponding to minus 1.96 and + 1.9645% significance level we got this as the x value.

(Refer Slide Time: 27:35)

Solution: Plotting of OC Curve

The probability of a type II error when $\mu = 7.90$, which is the same as that for $\mu = 8.10$.

$$\beta = P(7.902 \leq \bar{Y} \leq 8.098 \text{ when } \mu = 7.90) = P(0.04 \leq Z \leq 3.96) = 0.4840$$

Similarly for $\mu = 7.80$ and $\mu = 8.20$,

$$\beta = 0.0207$$

So, we get,

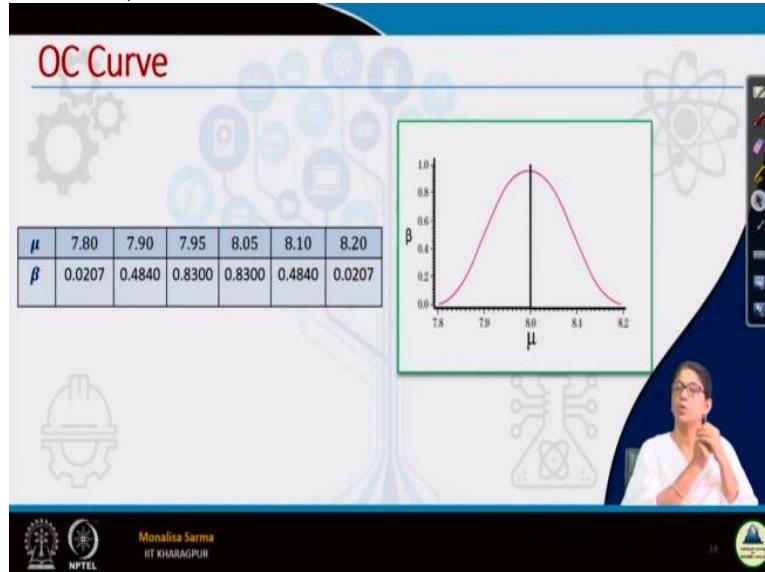
μ	7.80	7.90	7.95	8.05	8.10	8.20
β	0.0207	0.4840	0.8300	0.8300	0.4840	0.0207

Plotting these points (β, μ) we get the OC curve.

Monalisa Sarma
IIT KHARAGPUR

So, now, we will consider for $\mu = 7.10$ and 8.10 we will see that if we calculate we will get this value 0.4840. Similarly, if we calculate for 7.80 and 8.20 we will get this this β value. So, for this represented values of μ different values of μ this is the these are the values of β we got. Now, we can plot this value.

(Refer Slide Time: 28:03)



So, if we plot this value, this is the this curve is called operating characteristic curve operating characteristic curve means, where we are trying to plot β for different μ what is β for different μ , where μ is not illustrated, you see here the beauty of this curve you see, see when my μ actual μ that means, population mean when my mean is very close to the hypothesized value, then what happens then my type 2 error is very high.

It in fact β in fact, a process $1 - \alpha$, whatever α , we specified β process $1 - \alpha$, when my actual value approaches goes very near to the hypothesized value. So, see what happens when my actual value is very near to the hypothesized value, then my type 2 error is very high meaning what means I am accepting a false null hypothesis. But you see when my value is not very different, but very near to the hypothesis value.

If I am even if I am accepting a false null hypothesis, then also it will not have a very serious effect that is one thing. Second point is that already I told you that when my null hypothesis is not rejected, that does not mean that I am accepting the alternate hypothesis that does not mean that when a null hypothesis is not rejected, does not mean that I am accepting the null hypothesis. It may mean 2 things.

It may mean I am accepting the null hypothesis. It may mean I am I am not able to reject the null hypothesis. So, if I am accepting the null hypothesis, that means I am complicit I am okay fine, my null hypothesis I have accepted, but I am not being able to reject the null hypothesis means, again, I will do in future again, I will do the experiment and again, I will see whether really, my machine is really working fine or not.

I will again check put on my check whether my null hypothesis is accepted or rejected, I will maybe I will continue this experiment for 2 times 3 times 4 times. And if it is really, if my population value is actually a really away from 8, the time will come when I will see that actually it is falling in a critical region and a null hypothesis will be rejected. So that is what when my even if I have a very high type 2 probability.

Since my value is very close to μ , it will not have a serious effect that is the first point. Second point on if the μ is really different after the 1, 2 experiment, it will definitely come to the conclusion that null hypothesis actually rejected when will when I will take one sample again after sometimes I will take another sample for some sample maybe bias all sample will not be bias.

So, I may gradually see that my null hypothesis is rejected that is the first point now, when my actual value comes very much away from the null hypothesis, then type 2 error is very less means that times accepting a false null hypothesis is very less and that is good that is

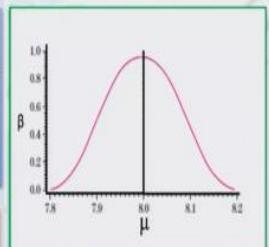
required, is not it? That means when I require my type 2 error to be very less, I am getting that when it is very much away from the hypothesis below my type 2 error is very less.

(Refer Slide Time: 31:46)

OC Curve

The OC curve shows that the probability of making the type II error is larger when the difference between the true value of the mean is close to the null hypothesis value, but decreases as that difference becomes greater.

In other words, the higher probabilities of failing to reject the null hypothesis occur when the null hypothesis is "almost" true, in which case the type II error may not have serious consequences.



Monalisa Sarma
IIT KHARAGPUR

So, the OC curve shows that the probability of making the type 2 error is larger when the difference between the true value of the mean is close to the null hypothesis value, but decreases as the difference becomes greater. In other words, the higher probabilities of failing to reject the null hypothesis occur when the null hypothesis is almost true, in which case the type 2 error may not have serious consequences.

(Refer Slide Time: 32:18)

Power

As a practical matter we are usually more interested in the probability of not making a type II error, that is, the probability of correctly rejecting the null hypothesis when it is false.

The power of a test is the probability of correctly rejecting the null hypothesis when it is false.

The power of a test is $(1 - \beta)$ and depends on the true value of the parameter μ .

The graph of power versus all values of μ is called a power curve.

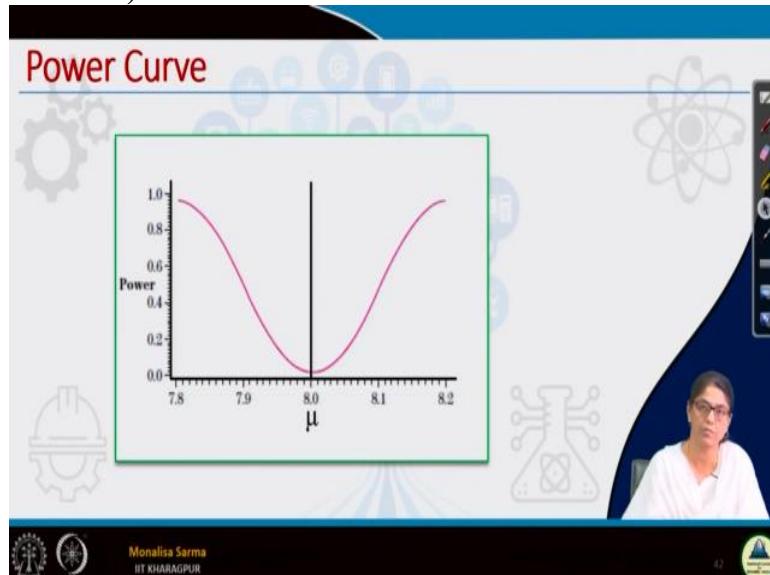
Monalisa Sarma
IIT KHARAGPUR

Now there is one more concept that is the last concept in this lecture, we shall be discussing that is power. Power is nothing but it is β is type 2 error, power is nothing but $1 - \beta$. So, basically as a practical method we are easily more interested in the probability of not making

a type 2 error, practically if we see we will be more interested in making not making type 2 error. That is the probability of correctly rejecting the null hypothesis when it is false.

Not making the type 2 error, not making the error means when the null hypothesis is false we are correctly rejecting it, instead of accepting it, that is my power. The power of the test is the probability of correctly rejecting the null hypothesis when it is false, we define as a power of a test. So, power of test is nothing but $1 - \beta$, β is accepting a false hypothesis so $1 - \beta$ is rejecting a false hypothesis. The graph of power versus all values of μ is called a power curve. This whatever OC curve we got is just a power curve is just an opposite of that.

(Refer Slide Time: 33:22)



So, this is the power curve OC curve is β versus μ and what is power curve? $1 - \beta$ versus μ . So, my power is very less when my actual value is very closer to the null hypothesis value. When my actual value goes far away from my null hypothesis value my power is very high that is desirable.

(Refer Slide Time: 33:51)

Features of Power Curve

- 1** The power of the test increases and approaches unity as the true mean gets further from the null hypothesis value.
- 2** As the true value of the population parameter approaches that of the null hypothesis, the power approaches α .
- 3** Decreasing α decreases the power.
- 4** Increasing the sample size increases the power.

Monalisa Sarma
IIT KHARAGPUR

So, the power of the test increases and approaches unity as the true mean gets further from the null hypothesis value. As the true value of the population parameter approaches that of the null hypothesis, the power approaches α . When the true value of the population it approaches near the null hypothesis value my power becomes very less, it approaches α that is it approaches to the value of type 1 error.

Decreasing α decreases the power why? When it decreases α β increases when β increases $1 - \beta$ decreases is not it? So, decreasing α decreases the power. However, increasing the sample size increases the power.

(Refer Slide Time: 34:33)

CONCLUSION

- ④ In this lecture we learned about-
 - ④ Relative importance of Type-I error and Type-II error
 - ④ Plotting OC curve and its significance
 - ④ The power of a test and how it varies with respect to α
- ④ In the next lecture, we will cover a tutorial.

Monalisa Sarma
IIT KHARAGPUR

So, to conclude this in this lecture today we have learnt couple of concepts we have learned relative importance of type 1 error type 2, error how we plot OC curve and what is its significance; the power of a test and how it varies with respect to α . In the next lecture we

will cover a tutorial encompassing all 3 lectures what we have done in statistical inference. We will move to statistical inference we shall be discussing.

Before you forget all the things what we have discussed let us see the tutorial first and then we will be going to do lectures other lectures other topics on statistical inference.

(Refer Slide Time: 35:07)



So, these are my references and thank you guys.

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

Lecture - 25
Tutorial on Statistical Inference

Hello guys, so in continuation of our discussion on statistical inference, today we will be doing a tutorial on whatever we have learned. It is not that statistical inference is complete, yet, there are lots of other things which we will have to learn. But, before going to those topics, till now, what we have covered, let us do a quick tutorial on that.

(Refer Slide Time: 00:47)

Concepts Covered

- ⌚ Solving objective type questions
⌚ To test the level of understanding from Lecture 22-24
- ⌚ Problems to ponder
⌚ To build problem solving aptitude

NPTEL Monalisa Sarma
IIT KHARAGPUR

So, in this, we will be covering we will test the level of understanding from lecture 22 to 24 that is the last 3 lectures, and then we will also solve few problems. So, first we will see the objective type questions.

(Refer Slide Time: 01:01)

Question-7.1

T 7.1: State whether the given statements are True or False:

- a) In a hypothesis test, the p value is 0.043. This means that the null hypothesis would be rejected at $\alpha = 0.05$.



Monalisa Sarma
IIT Kharagpur

So, it is again a true false type of question let us given in a hypothesis test the p value is 0.043 that means that a null hypothesis would be rejected at $\alpha = 0.05$. So, like, what will be the answer to this question, so, if I draw the figure so, if this is my what is say sampling distribution, So, α is 0.05 means assuming both tail α is 0.05 means it will be both side it will be how much the loop 0.025, if it is both side. So, and this means that a null hypothesis will be rejected at α equal to 0.05.

This is since it is not specified whether it is 2 tails or single tail it is this written α is 0.05. So, you can just consider it as a single tail if it is a single tailed then so, it is 0.05 will be this portion is 0.05 then in that case 0.043 will fall in this region. So, in that case the null hypothesis will be rejected. However, if it would have been a 2 tailed then it would not have been rejected.

(Refer Slide Time: 02:29)

Question-7.1

T 7.1: State whether the given statements are True or False:

- a) In a hypothesis test, the p value is 0.043. This means that the null hypothesis would be rejected at $\alpha = 0.05$. [True]
- b) If the null hypothesis is rejected by a one-tailed hypothesis test, then it will also be rejected by a two-tailed test.



Monalisa Sarma
IIT Kharagpur

So, this is true. So, if the null hypothesis is rejected by a 1 tailed hypothesis then it will also be rejected by 2 tailed test. What it is given if the null hypothesis is rejected by a 1 tailed hypothesis test, if it is rejected by 1 tail means like just the last example what we have seen if it is rejected by 1 tail, when suppose we have taken say 0.05 is my α . So, if it is rejected by 1 tail 1 tail means my critical region is this portion is 0.05. So, if it is rejected that means it is less than 0.05. That is why it is rejected then it will also be rejected by a 2 tailed test.

So, that we cannot say it may be rejected it may not be rejected because, if it is 2 tail then what happens? 2 tail means this will be 0.025, this will be 0.025. So, that means suppose my test statistics value suppose I got 0.04. So, if it is a 1 tailed hypothesis test then 0.04 definitely it would have been rejected, is not it? But then in a 2 tailed test 0.04 will not be rejected because for in a 2 tailed test the rejection region starts only from 0.025.

So, if the null hypothesis is rejected by a 1 tailed hypothesis test then it will also be rejected by 2 tailed test no it is false, it may be rejected it may not be rejected.

(Refer Slide Time: 04:06)

Question–7.1

T 7.1: State whether the given statements are True or False:

- a) In a hypothesis test, the p value is 0.043. This means that the null hypothesis would be rejected at $\alpha = 0.05$. [True]
- b) If the null hypothesis is rejected by a one-tailed hypothesis test, then it will also be rejected by a two-tailed test. [False]
- c) If a null hypothesis is rejected at the 0.01 level of significance, it will also be rejected at the 0.05 level of significance. [True]
- d) If the test statistic falls in the rejection region, the null hypothesis has been proven to be true. [False]

Monalisa Sarma
IIT KHARAGPUR

So, it will always be rejected by 2 tail test that is definitely false. So, if a null hypothesis is rejected at a 0.01 level of significance, it will also be rejected at the 0.05 level of significance. So, what it is given if it is rejected at 0.01 level of significance 0.01 it is rejected at 0.01 means my value is less than 0.01, t statistics whatever t statistics whatever it is the sample statistics value, I got is less than 0.01. That is why it got rejected.

Now, it will also be rejected at 0.05 something which is less than 0.01 will also be less than 0.05. So, if it is rejected at 0.01, level of significance it will also be rejected at the 0.05 level of significance that is true. So, if the test statistics falls in a rejection region, the null hypothesis has been proven to be true. So, if it is if the test statistic is false in a rejection region that what happens we reject the null hypothesis. So, null hypothesis program, this program proven to be true is false.

(Refer Slide Time: 05:20)

Question-7.1

T 7.1: State whether the given statements are True or False:

- e) The risk of a type II error is directly controlled in a hypothesis test by establishing a specific significance level. [False]
- f) If the null hypothesis is true, increasing only the sample size will increase the probability of rejecting the null hypothesis.



$$Z = \frac{Z - \mu}{\sigma/\sqrt{n}}$$



Monalisa Sarma
IIT KHARAGPUR



Next question the risk of a type 2 error is directly control in hypothesis test by establishing a specific significance level. Since specific significance level means so, means α by having a specific significance level α , the risks of type 2 error is not directly controlled it is indirectly controlled, because when we specify α we control we directly control the type 1 error. But that does not mean by controlling α of course, when we increase α my β decreases.

When we what to say decrease α my β increases that is of course true, but then if we control α it directly controls the type 1 error because type 2 error because we have seen when we tried to find out the operating characteristics curve, we have noticed that it is not α does not directly dictate the risk of the type 2 error risk of the type 2 error is mostly it depends on how much it is away from the true value of the population parameter to the true value of the parameter how much it is away from the hypothesis value. So, it is false.

If the null hypothesis is true increasing only the sample size will increase the probability of rejecting the null hypothesis. So, how we find out the z value if you are interested in my statistic is z , z is $\bar{x} - \mu / \sigma / \sqrt{n}$. So, my status the rejection region these are very much conceptual question if you can understand a concept you will be able to answer all the questions like so, rejection reason is always this means, when my z value is rejection region my z value is always more.

So, when my n is sample size n is increased what happened, my z value is increased. Value of z increase means same size going into the critical region becomes more and more. So, the null hypothesis is true increasing only the sample size will increase the probability of rejecting the null hypothesis is true.

(Refer Slide Time: 07:34)

Question-7.1

T 7.1: State whether the given statements are True or False:

- e) The risk of a type II error is directly controlled in a hypothesis test by establishing a specific significance level. [False]
- f) If the null hypothesis is true, increasing only the sample size will increase the probability of rejecting the null hypothesis. [True]
- g) If the null hypothesis is false, increasing the level of significance (α) for a specified sample size will increase the probability of rejecting the null hypothesis. [True]

Monalisa Sarma
IIT Kharagpur

If the null hypothesis is false, increasing the level of significance for a specified sample size will increase the probability of rejecting the null hypothesis. If the null hypothesis is false given that the null hypothesis is false, but we do is that increasing the level of significance means we are increasing α , we are going to increasing α that my critical results become more. So, what happens that will increase the probability of rejecting the null hypothesis, this is true.

(Refer Slide Time: 08:03)



So, this objective question what I suggest is that immediately do not see the answer just from the question you first try to once I have already discussed now, next when you will be seeing this video again definitely it is not once you will see the video in maybe that for your exam time again, you will revise it right. So, first try to you answer yourself because, these are very much conceptual questions. If the concepts are cleared and you will be able to solve all the problems and your quiz answers for this objective will also be correct.

(Refer Slide Time: 08:34)

A screenshot of a NPTEL slide titled "Problem-7.2". The slide contains a math problem: "T 7.2: An aptitude test has been used to test the ability of fourth graders to reason quantitatively. The test is constructed so that the scores are normally distributed with a mean of 50 and standard deviation of 10. It is suspected that, with increasing exposure to computer-assisted learning, the test has become obsolete. That is, it is suspected that the mean score is no longer 50, although σ remains the same. Test the suspicion with sample size of 500, whose mean is 51.07." Handwritten notes on the slide show the formula $n = 500$, $\mu = 50$, and $\sigma = 10$. The slide also features a video feed of a woman on the right and various icons related to reliability analysis in the background. At the bottom left, there is the NPTEL logo and the name "Monalisa Sarma IIT KHARAGPUR".

Next the problem we try to solve so, an aptitude test has been used to test the ability of 4th graders to reason quantitatively. The test is constructed so, that the scores are normally distributed with a mean of 50 and a standard deviation of 10. The test is constructed, so that the

test is constructed in such a way so that the test scores are normally distributed and what we got, we got a mean equal to 50 and standard deviation equal to 10 standard division is σ not σ^2 .

Test is conducted in such a way so, that we get this value that means, this is what we want 50 and 10 it is suspected that with increasing exposure to computer assisted learning, nowadays there is a computer assisted learning the test has become obsolete. Whatever the tests were hardly it was there it is the suspected that the test has become obsolete. That means it has become obsolete means that means it is suspected that the mean score is no longer 50.

Although the standard division remains the same, now, a new computer assisted learning has been introduced with this the tests what was the earlier the same test it says they are suspecting that the test is no longer viable it is suspect that the mean score is no longer 50. So, it is either less than 50 or greater than 50 it is not telling anything whether you need to test for less than or it need to test for greater it is this telling that it is no longer 50.

So, although the standard deviation remains the same, it is this suspicion with a sample size of 100 whose mean is. So, we have to test the suspicion with samples we have taken a sample size of 50 and mean of the sample that is 51.07 that is the mean. Now, first thing this will have to frame the hypothesis. So, how will frame the hypothesis always I told you will discuss your hypothesis while framing the hypothesis, we have to frame the hypothesis in such a way and where type 1 error is more significant.

Because we are in a hypothesis testing procedure we specify the significance level that is α what is α ? α is the type 1 error. So, while framing the hypothesis always we should frame in such a way that the type 1 error is more significant. Because we are giving more importance to type 1 error, that is why we are specifying α if we would have given more importance to be a type 2 error with a specified β .

But in hypothesis testing, we are giving more importance to α more importance to type 1 error that is why we are specifying the error probability of type 1 that is α . So, now while framing this hypothesis, we should keep this in mind. First of all, there are many things while framing the

hypothesis we should keep in this in mind first is that. Second, we understood we will have to frame it in such a way so, that type 1 error becomes more significant. And secondly, 1 important thing the type 1, the null hypothesis should always be with a equality sign.

Because we have seen if the null hypothesis is not equal, as without equal sign, if null hypothesis we have defined some parameter greater than some parameter or less than some parameter, then it is very difficult to find out α . Because in that case, we will have to find out critical region from the α what we do? From the α we find the critical region. So in that case, we will have to find the critical region for different values of the parameter.

When we specify a hypothesis null hypothesis if you specify the parameter is greater than a particular value, so greater means it can take any value greater than that particular value, say x greater than x . So, we will have to take any value greater than x . So, for all says x , we will have to find out α . So, that is a very tedious process we have seen while doing it for β . So, to keep it simple, always null hypothesis is specified with an equality sign so, the second thing.

Third thing now, while specifying the significance level, it should always be specified in such a way based on the cost of the type 1 error. If the cost of the type 1 error is huge, and the cost of type 1 error is huge then we will give a very less α means we do not want to reject a true null hypothesis in rejecting a true null hypothesis should be very, very remote that means α should be very, very less the cost of the rejecting a true null hypothesis is very high, my value of α will be very, very low.

So, keeping all these things in mind, we will have to frame the hypothesis. Now, here you can see the significance level is not mentioned, but we can understand this is a first we need to understand what you will form the null hypothesis? So, here you see the null hypothesis we will be maintaining the status quo maintaining means if we have to say something, we need to either invest lots of lots of lots and lots of money, resource, time, many such things many such parameters.

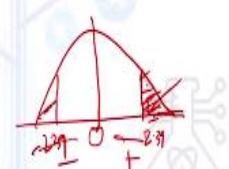
So, null hypothesis always we try to maintain the status quo. So, here the status quo is that the test is no longer obsolete, that means when the test is no longer and obsolete, when you will find that μ is what is expected what is the hypothesis value μ equal to 50 that specifying that μ we are getting with the computer assisted learning also we are getting the same μ that means test is no longer what to say obsolete. So, that is my null hypothesis. So, $\mu = 50$ is my null hypothesis.

(Refer Slide Time: 14:21)

Problem-7.2 : Solution

$H_0: \mu = 50$
 $H_1: \mu \neq 50$

Test is done on a sample of size 500.

$$Z = \frac{\bar{X} - 50}{10/\sqrt{500}} = \frac{51.07 - 50}{10/\sqrt{500}} = 2.39$$


T 7.2 : An aptitude test has been used to test the ability of fourth graders to reason quantitatively. The test is constructed so that the scores are normally distributed with a mean of 50 and standard deviation of 10. It is suspected that, with increasing exposure to computer-assisted learning, the test has become obsolete. That is, it is suspected that the mean score is no longer 50, although σ remains the same. Test the suspicion with sample size of 500, whose mean is 51.07

Monalisa Sarma
IIT Kharagpur

And now next what I need to check that is the alternate hypothesis what I need to check? I need to check whether μ is not equal to 50. So, I have to check μ is not equal to 50. So, for that, I will take a sample the size of sample size is given 500 and mean of the sample is given 51.07 and I know standard deviation of the population is given 10. So, the standard deviation of the population would not have been given then I would have gone for t distribution.

Now I will be taking the standard deviation of the sample. Now, since the standard deviation of the population is given, I will consider z distribution. Now, secondly, there are 2 things first significance level is not mentioned as I told you when a significance level is not mentioned and while discussing this at all, if it is not mentioned you can do 2 things. First is you do not have to write accept reject just specify the p value, then the decision maker will decide whether to accept the null hypothesis or to reject the null hypothesis.

Just specify the p value that is one option. Second option is that you select your own significance level. So, that significance level usually it is considered 0.05 as a standard significance level is considered, but for this example, where if the true null hypothesis is rejected unnecessarily the whole test has to be constructed it is there because constructing a test is not an easy task while constructing a test lots of men brain has to be means you really have to work hard on it.

It may not be lots of money, but then lots and lots of effort. So, here we do not want to undo something which is working fine. So, here we may consider a very less significance level. So, now, since we will consider z distribution, so, what will be the value of the z, z we know $x \bar{ } - \mu \sigma / \sqrt{n}$. So, this is $x \bar{ } - \mu$, $\mu = 50$ so, computing we got value 2.39. So, now, we can find out the p value of this since its significant level is not mentioned. So, now see while finding out the p value, there are 2 things again you have to remember.

If it is a single tail what z value you get? For z value you get the corresponding probability of that is the p value if it is a single tail, the corresponding probability of the z value is the p value remember this z table we have seen it is not it? In z table what we get? In z table we get the z probability corresponding to different z values. Now, here z value is given. So, probability corresponding to z value maybe it is this since it is what to say standard normal distribution definitely it is zero and this side to be minus this side it will be plus this side it will be plus.

This said minus so, z we got -2.39 means some somewhere here maybe 2.39. So, whatever area is corresponding, the probability of 2.39 from the table we can get and said if it is a single tail. If it is a double tail, then what happens, z corresponding area we get on for 2.39 we will have to check for -2.39 as well. If it is a 2 tail single tail only this probability this or this, whatever it is, it is the same thing symmetric z distribution is totally symmetric. If it is double tail, we will have to find out my p value will be corresponding to the probability of 2 point + - 2.39 so, this plus this.

(Refer Slide Time: 18:05)

Problem-7.2 : Solution

$H_0 : \mu = 50$
 $H_1 : \mu \neq 50$

Test is done on a sample of size 500.

$$Z = \frac{\bar{X} - 50}{10/\sqrt{500}} = \frac{51.07 - 50}{10/\sqrt{500}} = 2.39$$

$$P(|Z| > 2.39) = 0.0084 + 0.0084 = 0.0168$$

Monalisa Sarma
IIT KHARAGPUR

So, it is see 2.39, -2.39 this is the area this is the area from the z table you will see this area corresponds to 0.0084. So, this is my z value that is my p value. So, now see p value is 0.01, it is a very less actually if you consider it is very less. So, if we since constructing a new test, so, it is very expensive. So, we may consider significance level very less.

(Refer Slide Time: 18:40)

Problem-7.2 : Solution

$H_0 : \mu = 50$
 $H_1 : \mu \neq 50$

Test is done on a sample of size 500.

$$Z = \frac{\bar{X} - 50}{10/\sqrt{500}} = \frac{51.07 - 50}{10/\sqrt{500}} = 2.39$$

$$P(|Z| > 2.39) = 0.0084 + 0.0084 = 0.0168$$

T 7.2 : An aptitude test has been used to test the ability of fourth graders to reason quantitatively. The test is constructed so that the scores are normally distributed with a mean of 50 and standard deviation of 10. It is suspected that, with increasing exposure to computer-assisted learning, the test has become obsolete. That is, it is suspected that the mean score is no longer 50, although σ remains the same. Test the suspicion with sample size of 500, whose mean is 51.07.

As construction of a new test suite is quite expensive, the level of significance should be less than 0.01. In such case, this will not be rejected. However, the p-value is sufficiently small & needs to be investigated further.

Monalisa Sarma
IIT KHARAGPUR

If we consider a significant levels of 0.01 in such case what happened? We have considered a significance level of 0.01 and we get the value 0.0168 it is slightly greater than point 0.168. It means if this is 0.01 my value is slightly more than this. So, since it is more than the significance level, then it is not rejected. If it is more than the significance level means it does not fall in the critical region then we do not reject a null hypothesis.

But however, as I told you remember x when we reject the null hypothesis there is only one solution one on what to say one option that means we accept an alternate hypothesis am repeating again when we reject the null hypothesis there is only one option that is we accept the alternate hypothesis. But when we cannot reject the null hypothesis when the null hypothesis is not rejected, there are 2 option either is that we accept the null hypothesis or we just consider that we could not reject the null hypothesis.

We could not reject enable to reject the null hypothesis, these are 2 different things, unable to accept the null hypothesis means we are not complacent we may still carry out the experiment again, again and again and we will try to find out whether actually it is true. Actually, my null hypothesis is true or not. If my null hypothesis is not true, the sample which I considered maybe an out layer sample, then at one time, it will really show that, that null hypothesis has to be rejected.

So, now in this case, I got a very less significance value, P value. Sorry, it is not significance P value I got a very less P value. So, I need to be it is written it needs to be investigated further, that means I will carry out the test again. I am not satisfied, because I got very less value might be that if I can carry out the test might be that it gets rejected or might be that this value, maybe quite, this value becomes more maybe that instead of 0.1, and maybe I got 0.06.

Maybe my sample was not good that is why I got 0.06 means, I am comfortable, I can accept the or maybe 0.10, so then I can accept the hypothesis, I can tell that I am accepting the null hypothesis. So, here you see properly as the construction of the new test suite is quite expensive, the level of significance should be less than 0.01, which I have already mentioned, in such case, this will not be rejected, because this is more than this. However, the P value is sufficiently small and needs to be investigated further.

(Refer Slide Time: 21:40)

Problem-7.3

T 7.3: An apple buyer is willing to pay a premium price for a load of apples if they have, as claimed, an average diameter of more than 2.5 in. The buyer wants to test the claim of sufficiently large apples, so he takes a random sample of 12 apples from the load and measures their diameters. The results are given in table below:

2.9	2.8	2.7	3.2
2.1	3.1	3.0	2.3
2.4	2.8	2.4	3.4



Monalisa Sarma
IIT KHARAGPUR

24



Second question see here, an apple buyer is willing to pay a premium price for a load of apples if they have as claimed an average diameter of more than 2.5. An apple buyer, someone who is buying one to buy apple lots and lots of apple he wants to he will pay a premium price means you will pay more price, if the size of the average diameter of the apple is more than 2.5. He will pay more price, the average diameter means for a big size apple, he will pay more.

The seller is telling you these apples are very big apples, I will give you bigger apples, but he is not satisfied with that. So, wants to test the claim. So, the buyer wants to test the claim of sufficiently large apples. So, he takes a random sample of 12 apples from the load and measures the diameters these are the diameters of the 12 apples which you have taken. So, now, thing is that you have to find out whether the buyer will pay the premium price or not.

Now, in this case, what will be my null hypothesis? My null hypothesis as I told you, one is maintaining status quo of course, and I should make this null hypothesis in such a way so that my type 1 error is more significant. So, here what is the null hypothesis type 1 error is most significant means my null hypothesis will be my apple is less than or equal to 2.5 that means I will tell it is my size of apple is 2.5 inch. M is 2.5 inch.

(Refer Slide Time: 23:17)

Problem-7.3 : Solution

$$H_0: \mu = 2.5$$

$$H_1: \mu > 2.5$$

Let us consider that the buyer is willing to take a 10% chance of unnecessary paying the premium price. That is, $\alpha = 0.10$

As this is a One-tail test, the rejection region is ?? (for $df = 11$)

T 7.3: An apple buyer is willing to pay a premium price for a load of apples if they have, as claimed, an average diameter of more than 2.5 in. The buyer wants to test the claim of sufficiently large apples, so he takes a random sample of 12 apples from the load and measures their diameters.



And what I want to prove that it is greater than 2.5. Now, send here also the significance level is not mentioned. Let us consider that a buyer is willing to take a 10% chance of unnecessary paying a premium price reeks of type 1 error means even if it is not large, he was fine if it even if it is not large he will take care of error that is error he will take a chance that chances 10%. 10% means α is equal to 0.10 and this question α is not mentioned.

So, based on the case the buyer maybe he can consider α is 0.10 when α is not mentioned always you can take for safety sake you can 0.05 or always specify the P value and then maybe give the decision based on 0.05. If you yourself can consider that what to say that application is very sensitive then consider 0.01 if you considered application is okay we can take care. Here say if apple is a bit small picking you will pay a bit extra or bit more.

So, then here we have to take a very high what to say significance level we do not have to take a very low significance level. Then we can take point α equal to 0.10 why we are doing this? Say here by being α is equal to 0.10 when we are increasing α my type 2 is also decreasing. Type 2 thing is also decreasing is not it? The type 2 error is also decreasing. What is type 2 error? Type 2 error means I am accepting a false null hypothesis.

That is also I do not want a false null hypothesis means the apples are not actually big, but I am accepting it, if I make my α very small, my type 2 error will also be increased. So, in this case, I

do not want to increase the type 2 error also. So, I will keep in such a way that my α is also not very big. Anyway, you do not have to worry about this it is application specific whichever application person who wants to do this experiment and they will specify the α you just need to know the logic behind it, the science behind it.

And as well as how to do the techniques basically, as this is a 1 tailed test, because we are finding μ greater than 2.5. So, it is a 1 tail test. So, my significance level remains $\alpha = 0.10$ only it will not be divided by 2 and here what to say the average diameter that you see the standard deviation is not given when the standard deviation is not given it is just specified as claimed an average diameter of more than 2.5 inch.

Standard deviation is not given that means we will be considering t distribution now degrees of freedom how much degree of the freedom is 11. For degree of freedom 11 $\alpha = 0.10$ what is my rejection region.

(Refer Slide Time: 26:16)

Problem-7.3 : Solution

H ₀ :	$\mu = 2.5$
H ₁ :	$\mu > 2.5$

Let us consider that the buyer is willing to take a 10% chance of unnecessary paying the premium price. That is, $\alpha = 0.10$

As this is a One-tail test, the rejection region is ?? (for df = 11)

V	0.5	0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	0	1	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0	0.816	1.061	1.386	1.886	2.92	4.303	6.965	9.925	22.327	31.599
3	0	0.765	0.978	1.25	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0	0.741	0.941	1.19	1.533	2.132	2.776	3.747	4.604	7.173	8.61
5	0	0.727	0.92	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0	0.718	0.906	1.134	1.44	1.943	2.447	3.143	3.707	5.208	5.959
7	0	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0	0.706	0.889	1.108	1.397	1.86	2.308	2.896	3.355	4.501	5.041
9	0	0.703	0.883	1.1	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	0	0.7	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.93	4.318
13	0	0.694	0.87	1.079	1.35	1.771	2.16	2.65	3.012	3.852	4.221

Monalisa Sarma
IIT KHARAGPUR

So, $\alpha = 0.10$, degree of the freedom is 11, 1.363 this is my rejection region. So, if from my sample, if I get value which is greater than this, then I will reject the null hypothesis that means, I can consider that means my apple or bigger size and I can pay that premium price. But if it is less than 1.363, that means the null hypothesis is I could not reject the null hypothesis. So, I will not pay the premium price.

(Refer Slide Time: 26:46)

Problem-7.3 : Solution

$H_0: \mu = 2.5$
 $H_1: \mu > 2.5$

Let us consider that the buyer is willing to take a 10% chance of unnecessary paying the premium price. That is, $\alpha = 0.10$

As this is a One-tail test, the rejection region is $t > 1.3634$ (for $df = 11$)

From the sample,

$$\bar{X} = 2.758, S^2 = 0.1554$$
$$t = \frac{2.758 - 2.5}{\sqrt{0.1554/12}} = 2.267$$

2.9	2.8	2.7	3.2
2.1	3.1	3.0	2.3
2.4	2.8	2.4	3.4

T 7.3: An apple buyer is willing to pay a premium price for a load of apples if they have, as claimed, an average diameter of more than 2.5 in. The buyer wants to test the claim of sufficiently large apples, so he takes a random sample of 12 apples from the load and measures their diameters. The results are given in table below:

Monalisa Sarma
IIT KHARAGPUR



So, in this case, what happens so, the rejection region is greater than $t 1.363$. So, from the sample this is the sample you can calculate X bar you can calculate S^2 you can calculate by yourself S^2 is nothing but the variance you can calculate X bar is the mean. So, from the t value I got 2.267 so, 2.267 is better than this that means my null hypothesis is rejected. So, now hypothesis is rejected means alternate hypothesis accepted.

(Refer Slide Time: 27:16)

Problem-7.3 : Solution

$H_0: \mu = 2.5$
 $H_1: \mu > 2.5$

Let us consider that the buyer is willing to take a 10% chance of unnecessary paying the premium price. That is, $\alpha = 0.10$

As this is a One-tail test, the rejection region is $t > 1.3634$ (for $df = 11$)

From the sample, $\bar{X} = 2.758, S^2 = 0.1554$

$$t = \frac{2.758 - 2.5}{\sqrt{0.1554/12}} = 2.267$$

So, the null hypothesis is rejected.

T 7.3: An apple buyer is willing to pay a premium price for a load of apples if they have, as claimed, an average diameter of more than 2.5 in. The buyer wants to test the claim of sufficiently large apples, so he takes a random sample of 12 apples from the load and measures their diameters. The results are given in table below:

Monalisa Sarma
IIT KHARAGPUR



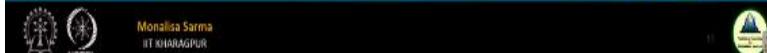
So, the buyer can pay the premium price for the apple.

(Refer Slide Time: 27:23)

Problem-7.4

T 7.4: An NRI wants to take a property for rent for business use either in Bandra or Dadar. Anticipating Bandra will be more expensive, he is considering setting up office space in Dadar, essentially to reduce the office rental costs. In the 2021 issue of the Dadar real-estate report, the mean cost of leasing office space for all downtown buildings in Dadar was quoted as being \$12.61 per square foot with a standard deviation of \$4.50. To compare costs with those in Bandra, the businessman sampled 36 office buildings in Bandra and found a mean leasing cost of \$13.55 per square foot.

Does this mean that leasing office space in Bandra is really higher? Should the businessman consider setting up at Dadar to save money on rent (assuming other factors equal)?



So, one last question this question is a bit different in till now what we have tried to see, we have hypothesis something about the population and we are taking a sample from the sample we are trying to infer about the population whatever hypothesis value is correct or not. This is a bit different. We are just first go through the question then you will understand. See an NRI wants to take a property for rent for business use either in Bandra or Dadar, so this is Bandra, this is Dadar.

Anticipating Bandra will be more expensive he is considering setting up office space in Dadar, essentially to reduce the office rental costs so obviously he wants to reduce the office rental in the 2021 issue of Dadar real estate report, the mean cost of leasing office space for all downtown building in Dadar was quoted as being the mean costs is quoted as been 12.61 dollar with standard deviation of 4.50.

In a real estate report head is quoted for sorry, it is quoted for Dadar this is Dadar and this is Bandra. So, you want to set up in Dadar anticipating that it is less cost. So, from Dadar real estate report, he got this value of 12.61 per ² foot and this is the standard division is 4.50 for Bandra, he does not have such a real estate report is not available. So, what he does to compare costs with those involved the businessman sample 36 office building in Bandra he himself sample 36 office buildings belongings he inquired basically.

He picked up different office buildings and inquired and found a mean leasing cost is 13.55 he found 13.55. See here for Dadar mean is 12.61 here it is 13.55. But in Dadar it is real estate report means it is as a whole it is given because means this is about the whole population of Dadar. And this whole Bandra has just taken some sample. Now from this sample basically he has to find out does this mean that leasing office space in Bandra is really higher.

So, from this sample, basically he has to find out whatever it is reflecting this 13.55, does that mean leasing office space in Bandra is really higher. So, that means from this sample he just he had to find out this is he got it from the sample of Bandra. So, from this he has to infer about the population of Bandra so, population of Bandra, if you find if it is very higher than he will lease space in Dadar.

(Refer Slide Time: 30:25)

Problem–7.4 : Solution

In this case,

$$H_0 : \mu = \$12.61$$

$$H_1 : \mu > \$12.61$$

$$Z = \frac{13.55 - 12.61}{4.50/\sqrt{6}} = 1.25$$

So,

$$P = P(Z > 1.25) = 0.1056$$

T 7.4: An NRI wants to take a property for rent for business use either in Bandra or Dadar. Anticipating Bandra will be more expensive, he is considering setting up office space in Dadar, essentially to reduce the office rental costs. In the 2021 issue of the Dadar real-estate report, the mean cost of leasing office space for all downtown buildings in Dadar was quoted as being \$12.61 per square foot with a standard deviation of \$4.50. To compare costs with those in Bandra, the businessman sampled 36 office buildings in Bandra and found a mean leasing cost of \$13.55 per square foot. Does this mean that leasing office space in Bandra is really higher? Should the businessman consider setting up at Dadar to save money on rent (assuming other factors equal)?

Important points

- For a given level of significance, power is larger in a One-tail test than a two-tail test, when the value of μ is in the range of alternative hypothesis.
- The power is essentially zero on the other side.

NPTEL Monalisa Sarma IIT KHARAGPUR

So, what will be the hypothesis, definitely he wants to test because from the population, what he has for Dadar, he can assume that the same thing for Bandra. So, we can assume that μ is 12.61 and what he wants to test it is greater than 12.61 because he got 13.55 in the sample if you would have got less than 12.61 then he get tested it for less. So, he wants to test whether it is greater than this. So, now, since the standard deviation is given, he has to inquire about the mean definitely z distribution he will use. So, z is 1.25.

So, if the significance level is not mentioned we can just specify the P value. So, P value this is again this is one tail just greater than one tail, what is the probability, corresponding to z is value is 1.25 see here we are not adding in the previous question we have added this because z was 2 tails in the single tail the probability corresponding to 1.25 is this. So, this is the P value. Now, depending on the significance level, you will take the decision or depending on what how much error he can be.

So, one important point for a given level of significance power is larger in 1 tailed test, then a 2 tailed test, when the value of μ is in the range of alternative hypotheses, why we have learned power is not it? What is power? Power is $1 - \beta$ what is power when the null hypothesis is false, we are truly we are rejecting a false null hypothesis that is power, is not it? Rejecting a false null hypothesis is power that is $1 - \beta$.

So, given for 1 tailed test for the same level of significance, for 1 tailed test what happens α is higher because for 2 tailed tests α becomes $\alpha / 2$ significance level becomes $\alpha / 2$. So, when the 1 tailed test significance of level of α remains α , so when significant level is α , that way, my β gets reduced, is not it? So, when my β gets reduced. So, what happens is $1 - \beta$ becomes more so for a given power is larger in a 1 tailed test then a 2 tailed test when the value of μ is in the range of alternative hypothesis.

When the value of μ is actually in the range of alternative hypothesis but if the value of μ is in the here see value of μ in less than it is equal to means is less than equal to when value of μ is less than that than actually the power is essentially 0. But we are not worried about that also, because we want to if it is greater than, so we need to find out the power on that. We do not want to find the power on it is less because it may be equal it may be less we are fine with that.

(Refer Slide Time: 33:12)



So that is all so these are the references as I mentioned before, thank you guys.

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

Lecture – 26
Statistical Inference (Part - 4)

Hello guys. So, we have done one tutorial in my last lecture. So, now, we will start with our next continuing with our discussion on statistical inference, as I told you, there are lots of things to learn in statistical inferences.

(Refer Slide Time: 00:40)

Concepts Covered

- ④ Confidence interval estimation
- ④ Relationship between hypothesis testing and confidence interval
- ④ Error of estimation

NPTEL
Monalisa Sarma
IIT KHARAGPUR

So, in today's lecture, what we will see? What is confidence interval estimation then what is the relationship between hypothesis testing and confidence interval and also we will see what is an error of estimation.

(Refer Slide Time: 00:52)

Quick Recap

Statistical inference is carried out with two different approaches. The objectives, however, of both the approaches are similar:

Objectives of Statistical Inference

- Hypothesis Testing
- Estimation

Monalisa Sarma
IIT KHARAGPUR

Now, confidence interval estimation I think you remember when we initially started to talk on the statistical inferences, then I have already mentioned statistical inference we can carry out between 2 different approaches one approach is hypothetical testing and another approach is confidence interval estimation. We have discussed hypothesis testing for last 3 lectures and in today's lecture we will discuss the other one that is the confidence interval estimation.

(Refer Slide Time: 01:24)

Quick recap

Up-to this point, we learnt about:

Statistical Hypothesis

Null and Alternate Hypothesis

Rejection region/ Critical region

Type 1 and Type 2 error

Choosing between α and β

Ideas Covered

Monalisa Sarma
IIT KHARAGPUR

So, before going to confidence interval because this is we are studying with a different approach let us just have a quick recap of what we have learned till now. So, we have of course, learn what is a statistical hypothesis if the hypothesis is done with the parameters of this population and we call it a statistical hypothesis I think you will remember what is null hypothesis, what is alternate

hypothesis, null hypothesis is always we try to keep that hypothesis as a null which maintains the status quo we always try to give that hypothesis is null hypothesis.

Which has type one error may be quite significant accordingly to; based on the cost of the type one error we will consider the significance level. And moreover the alternate hypothesis is something which we want to test it that we may keep it as an alternate hypothesis, then what is rejection region or critical region that we have seen based on the significance level whatever significance level is mentioned based on the significance level we find out the rejection region.

Like from the table we can find out the z value corresponding to the probability that is the α , α is nothing but probability of type 2 error, probability is nothing but the area under the curve is not it? So, when the z curve so, then the corresponding z value corresponding to α is nothing but the rejection region like if I take if I remember the action which we have taken the medicine company that it should fill in the 8ml tube if it is 8ml tube so, initially what I was discussing how can we pick the rejection reason.

Suppose, find $\mu = 8$ that is okay but then there can be some changes we can accept. So, that suggests we have decided it to be it if it is from 7.92 8.1 we can accept that. So, what is the 7.9 and 8.1 this is the starting of the rejection region or the starting of the critical region and the probability so, the area corresponding to 7.9, 8.1 it is nothing but α . So, what in a hypothesis testing rejection region is not given α is given from α we find out the rejection region.

Then what is type 1 error, type 2 error if remember type 1 error is true null hypothesis we are what to say rejecting a true null hypothesis that is type 1 error, type 2 error is we are accepting a false null hypothesis then choosing between α and β that also we have seen how to choose α , how to choose β that we have seen some example based on that example we have seen how to do that.

(Refer Slide Time: 04:01)

Up-to this point, we learnt about:

- Five steps of HT
- Significance level of a HT
- Statistically significant result
- Why do we focus on type 1 error
- What is p – value
- Operating characteristic curve
- Power of a test

Ideas Covered

Monalisa Sarma
IIT KHARAGPUR

Then on different ideas what we have covered is the 5 steps of hypothesis testing has 5 steps. Remember then what is the significance level of a HT hypothesis testing that is α , statistically significant results do you remember what is statistically significant results when we call a result a statistically significant essentially a results that is statistically significant when we reject the null hypothesis why it is say? Means the value what we get from the sample and whatever value we have hypothesis this both the values are statistically different.

This result is very different whatever we have hypothesis based on the sample we got the result, this result is very much different from whatever we hypothesis so, we call this result as a significant result. So, that is why it is called statistically significant results. So, easily statistically significant result when we get statistically significant results; that means we reject a hypothesis or we get some value like it. If you remember we have taken an example of constructing a test.

So, in the last tutorial we have seen in constructing a test, so, we got the value very close very, very less that means, but then we still we could not reject the null hypothesis that value also we can consider it as a statistically significant result. When the value is very, very different from whatever we have hypothesis value of below what is indicates? This indicates this probability of this occurrence if this is true what is the property of disappearance, is not it. So, now, why do we focus on type 1 error I have already mentioned.

Then what is p value? p value reporting is necessary is necessary why because sometimes it is very difficult to just give a reject a hypothesis or do not reject a hypothesis given just a yes no result it becomes too much because what I have explained it in some examples for similar type of results with slight change we see sometimes it gets accepted sometimes it gets rejected for very slight changes even for difference slight change of significance level also may get a different result.

And moreover the person who is carrying out the hypothesis testing that is the statistician and the person who is the actual decision maker maybe the 2 different ways in person. So, we should leave the decision to the decision maker whether we will give the result now, the decision maker will take a decision whether to reject or not to reject. So, this is call p value reporting basically based on what is the calculated test statistic whatever value we get the probability of test statistic is the p value.

Operating characteristic curve it is nothing but β versus μ . So, in power of test is? Power of a test is $1 - \beta$ versus μ power of test means rejecting a false null hypothesis that is the power of a test.

(Refer Slide Time: 06:51)

Confidence interval

Definition: Confidence Interval

A confidence interval consists of a range of values together with a percentage that specifies how confident we are that the parameter lies in the interval.

Point Estimate vs Confidence Interval

- Estimation of parameters with intervals uses the sampling distribution of the point estimate.
- A point estimate appears to be precise, but the precision is illusory.
- A confidence interval is not as precise as a point estimate, but it has the advantage of having a known reliability.**

Monalisa Surma
IIT Kharagpur

So, now coming to the confidence interval so, before looking at a slide first let us take an example suppose a retail chain retail chain wants to open a big retail what to say store basically in one area where the so, before opening such a counter he needs to know the earning capacity of

the people in that locality. So, if because why he why it is necessary, because if he can find out what is the earning capacity of the people accordingly based on that he will keep the things in his counter in his outlet.

So, if I mean the spending capacity not on it the spending capacity is more definite he can keep some high end products in the spending capacity is very, very less. So, high end product will just be left into outlet and it will not be sold. So, he will have to go for some less expensive products specialty some expensive product. Now it is a totally new area. So, how to find out the spending capacity of people it is a huge area with diverse populations and totally know what to say pass knowledge from which you can have a hypothesis value.

In such a case there are many such examples again the another one example which I can say is that so some what to say in a chemical industry the amount of what to say byproduct it gives until and unless you do the experiment, you really cannot say what will be the amount of mean of those byproducts. So, in this when you cannot basically when you cannot hypothesis about the parameter of a population, then the only option left in front of you is there then you will have to take a sample.

From the sample you will have to find out the value of the statistics and based on the statistics we will infer about the population like for here is the person who wants to open a retail outlet here. So, he will since he has no idea about the spending capacity of the people he will select he will take a non bias sample he will take a set of people and from the set of people he will try to find out what is the spending capacity of the people from with that spending capacity of the people from the sample then he can infer about the whole population.

So, basically here is hypothesis value is not there. So, he has to estimate based on the sample. Now, this estimation there are 2 different ways the first one is whatever he got from the sample suppose he has taken his interview around I should not say interview he has he tried to find out from around 20 person what is the spending capacity by spending capacity means how much a person arms in annually. So, from there suppose we found out on an average person in that

industry 20 around among these 20 people on an average a person are not around 8 lakhs per annum.

So, 8 lakhs per annum is the mean salary of this 20 people what he has interviewed. Now, based on this, if he directly tells what you said the mean of earning mean of the spending capacity of this people or the earning of these people in this area is 8 lakhs per annum. So, that is basically that is called point estimation we have just we found out the mean and we are just telling this mean only we are telling that it is the population mean if we take this mean of the sample as the population mean.

If we infer the mean of the population mean of the sample as the population mean this estimate is called point estimate of course, and secondly, this estimate the probability that it is correct is very very less because, as I told you before this samples this statistics vary if I took a particular sample of 20 people I found as many as 8 lakhs per annum, I took a different sample maybe I found it 7 lakhs per annum I took in different sample maybe I got some other value. So, if I take a point estimation reliability of this estimate is very less.

So, what we can do is that instead of just giving a point estimate we can just hedge a bit what we do we will give an interval that means the earning capacity earning of this people in this locality is around from 5 lakhs to 10 lakhs we have given an interval directly without specifying the earning of this people is 8 lakhs per annum, we are given telling us a 5 lakhs to 10 lakhs. So, we have given an interval so, we are estimating within an interval in this is called interval estimation.

Now comes the question of what is confidence? Confidence is the how much reliability of this estimation that is called as confidence interval estimation. Now, how we can give the statement about the reliability of that this is again very much dependent on the significance level what a significance level remember significance level is the probability of type 1 error here also if the significance level is α that means my confidence coefficient is 95%. So, accordingly I will find out the confidence interval. So, let us see in the slide.

So, first say what we have a confidence interval consists of a range of values together with a percentage that specifies how confident we are that the parameter lies in the interval. Interval consists of a range of values together with a percentage that specifies how confident we are. So, it is 5 lakhs to 10 lakhs is a range now along with that we have to specify a percentage that we will have to give a percentage that specifies how confident we are that the parameter lies in that interval, that is the confidence interval estimation.

So, estimation of parameters with interval it also again like hypothesis testing hypothesis testing my main backbone is the sampling distribution my CPU is the sampling distribution is not it? here also for confidence interval estimation also my CPU was the sampling distribution estimation parameters with interval uses the sampling distribution of the point estimates we use sampling distribution based on the point estimation whatever mean I got from the sample I have collected one sample from the sample whatever I got mean.

Mean or variance or anything whatever we want to infer. So, this I will find the sampling distribution of that particular point estimate like how we do in hypothesis testing similarly. A point estimate appears to be precise, but the precision is illusory. The point estimate if I tell the mean income of the population of this area is 8 lakhs that is a very precise statement, rather than if I say may mean income of this area is from 5 lakhs to 10 lakhs this is not very precise statement, I am giving a huge range.

But if I am specifying just one value that is a very precise statement, but this precision is illusory as I told you why. A confidence interval is not as precise as point estimation, it is not as precise, but it has the advantage of having a known reliability that is the confidence level.

(Refer Slide Time: 14:12)

Interval estimate of mean μ

Mathematical Formulation

We can write,

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

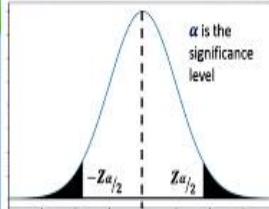
Again, $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

$$\text{So, } P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

$$\text{Or, } P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Therefore, our interval estimate of mean μ is,

$$(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) \text{ to } (\bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}})$$



 Monalisa Sarma
IIT KHARAGPUR



So, we will see how it is. So, what we do is that? We have seen till now, this concept of significant level that critical region or the rejection region when I specify α as my significance level, so for two tailed, when I am not specifying then I will always consider 2k. So, it is α is a significance level. So, this value $z \alpha / 2$ corresponding to value $z \alpha / 2$, this is my rejection region, is not it? This is my rejection region. So, if my value falls within this region, then I accept it if my value falls within this region.

Now, if this area put together as α , what is this area? This area is $1 - \alpha$, this area and this area total is α then this area is $1 - \alpha$ is not it? So, probability meant that my z will lie within this range is what is $1 - \alpha$. So, now what is z ? z is a $\bar{X} - \mu \sigma / \sqrt{n}$. So, I will substitute z with that now, I will do a bit of simplifying what I get I will be getting this sort of an expression μ this value is less than μ , μ less than is this value.

And the quality of this is $1 - \alpha$ where α is the significance level. So, this is my interval the my μ value will lie between these to this my μ will lie between this to this, this is my interval this is my interval assist estimation and what is the confidence of this is $1 - \alpha$ probability that my μ value will lie between this is $1 - \alpha$. So, there are interval estimate a mean μ is this where \bar{X} is the mean of the sample.

(Refer Slide Time: 16:16)

Interval estimate of mean μ

Formula of Confidence Interval

- Interval estimate of mean μ is expressed as, $(\bar{x} - Z_{\alpha/2} \sigma / \sqrt{n})$ to $(\bar{x} + Z_{\alpha/2} \sigma / \sqrt{n})$, if \bar{x} is the mean of a random sample of size n from a population with known variance σ^2 .
- This interval estimate is called a confidence interval.

- The lower and upper boundary values of the interval are known as confidence limits.
- The probability used to construct the interval is called the level of confidence or confidence coefficient.
- The confidence coefficient is often given as a percentage; for example, a 95% confidence interval.



Monalisa Sarma

IIT KHARAGPUR



This interval estimate is called a confidence interval the lower and upper boundary values of the interval are known as confidence limit the lower confidence limit upper confidence limit and here this is my lower confidence limit, and this is my upper confidence limit and what is the confidence the percentage the probability is called a confidence coefficient. So, what is my confidence coefficient here $1 - \alpha$ so, if α is 5 a confidence coefficient is 95%.

The probability used to construct the level is called the level of confidence or the confidence coefficient the probability used to construct the interval I call it as level of confidence or the confidence coefficient. So, if α is my significance level, what is my confidence coefficient or level of confidence it is 95%.

(Refer Slide Time: 17:20)

Revisiting... case study 2

Case Study 2

A medicine production company packages medicine in a tube of 8 ml. In maintaining the control of the volume of medicine in tubes, they use a machine. To monitor this control a sample of 16 tubes is taken from the production line at random time interval and their contents are measured precisely. The mean volume of medicine in these 16 tubes is 7.89 ml. Calculate the confidence interval assuming a significance level of 5%. Assume $\sigma = 0.2$.



Monalisa Sarma
IIT KHARAGPUR



Now we will see with the help of an example, we have seen this example again that is why I have given a 3 digit the case study 2 we are considered when we have just tried to learn what is hypothesis testing? So, same example and bringing it out here. Here we have done using this we have used this example to find out hypothesis testing. Now, we will use this example to find the confidence intervals. Now, there are 2 things first is that when we use confidence interval first there are 2 cases.

Once we use confidence interval when we cannot hypothesis a value then we try to estimate the value of the parameter within an interval with a known confidence coefficient. When we cannot hypothesis if we can hypothesis the value then we can do the hypothesis testing and then we can find out whether it is accepted or rejected. If we cannot do that, we can find out the confidence interval and we can tell that our parameter value lies within this range. This is one reason why we do confidence interval estimation.

There is another one reason another one reason is that like when we do an experiment we have done an hypothesis testing like any other example what we have taken suppose we have taken $\mu = 8$ is null hypothesis $\mu \neq 8$. So, it is the alternate hypothesis. Now, suppose on the hypothesis testing whatever we have done and we found that null hypothesis is rejected that means $\mu \neq 8$. Now $\mu \neq 8$ state find that but μ is how much μ is not equals to it is okay.

But μ is how much that how we will get we can calculate it in using this confidence interval of course, we will not go to specific values of μ but we can tell my μ falls within this interval. So, these are the 2 reasons why we do confidence interval estimation. So, now this question what it is given we know all this but still going through medicine production, company packaged medicine in a tube of 8ml. That means I am I required 8ml in maintaining the control of the volume of medicine in tubes they use a machine.

To monitor this control a sample of 16 tubes is taken from the production line at random time interval and the contents are measured precisely. That is required is 8ml is required and whether it is really filling 8ml what we have done we have taken a; what to say random sample random sample of 16 tubes. And we have measured it precisely but we found the mean volume of the medicine in this 16 tubes is 7.89 mm. Calculate the confidence interval assuming a significance level of 5% as the mean $\sigma = 0.2$.

Now, it is asking us to calculate the confidence interval only, we are not asked to do the; what to say hypothesis testing. Remember this problem when we have done hypothesis testing for same data, we had rejected the hypothesis null hypothesis, we have the null hypothesis was $\mu = 8$. Alternative hypothesis is μ not equals to 8 and we have rejected the null hypothesis we found that μ not equal to for the same question now, let us see what we get in a confidence interval.

(Refer Slide Time: 20:24)

Case study

Computing Confidence Interval for Mean in Case Study 2

In this problem, $\bar{x} = 7.89$, $\sigma = 0.2$, $n = 16$, and $Z_{\alpha/2} = 1.96$.

So the confidence interval is given by,

$$(7.89 - 1.96 \times \frac{0.2}{\sqrt{16}}) \text{ to } (7.89 + 1.96 \times \frac{0.2}{\sqrt{16}})$$

$$\text{Or, } (7.89 \pm 1.96 \times 0.05)$$

$$\text{Or, } (7.89 \pm 0.098) \text{ Or, } 7.792 \text{ to } 7.988$$

A medicine production company packages medicine in a tube of 8 ml. In maintaining the control of the volume of medicine in tubes, they use a machine. To monitor this control a sample of 16 tubes is taken from the production line at random time interval and their contents are measured precisely. The mean volume of medicine in these 16 tubes is 7.89 ml. Calculate the confidence interval assuming a significance level of 5%. Assume $\sigma = 0.2$.

Hence, we say that we are 95% confident that the true mean volume of medicine is between 7.792 to 7.988 ml per tubes.



Monalisa Sarma
IIT KHARAGPUR



So, from this what it is given from the question it is \bar{x} bar is given σ of the population is given n is 16. So, and significance level is 5% it is two tailed means $\alpha / 2$. So, $z \alpha / 2$ its value is 1.96 you can find it from z table it is 1.96. So, confidence interval will be this is a confidence interval this is the formula $X - z \alpha / 2 \sigma \sqrt{n}$ X bar X bar $- z \alpha / 2 \sigma \sqrt{n}$ means this is the relative value of z corresponding to $\alpha / 2$ it means if α is 0.05 value of z corresponding to 0.025.

That is that $\alpha / 2$ this is also a value of z corresponding to 0.025 means this area is 0.025. So, what is this value of z ? z that I am specifying as $z \alpha / 2$. So, confidence interval I am putting all the values here and I got a confidence interval this 0.7 this is my confidence interval my μ lies in this range. Now, see here in hypothesis testing, I have rejected this null hypothesis if you can remember if you cannot remember you can go to the lecture again I think it is in the second lecture on statistical inference.

So, we have rejected this null hypothesis that μ is not equal to now, what is confidence interval value what it gives? confidence interval value also gave us that μ lies in this range 7.79 to 7.988 it means our hypothesis testing result is so, true means we have rejected that it is not equals to 8, but actually it is true you see in from the confidence interval also we found our μ lies in 7.79 to 7.98 8 is not included in this range.

Hence, we can say that we are 95% confidence because a significant level is 5% that the true mean volume or dimension is between this.

(Refer Slide Time: 22:23)

Reframing case study 2: case study 2a

Case Study 2a

We are interested in a quality control problem of a medicine production company, in which we want to test the hypothesis with a significance level of 5%, that the mean volume of medicines being put in the tubes was the required 8 ml. Table below lists the data from a sample of 16 tubes. Find out if the hypothesis is true. Also calculate the 0.95 confidence interval on the mean volume of medicine per tube for the given sample.

8.08	7.71	7.89	7.72
8.00	7.90	7.77	7.81
8.33	7.67	7.79	7.79
7.94	7.84	8.17	7.87

Monalisa Sarma
IIT KHARAGPUR

The same question whatever we have seen, but we have slightly changed up here whatever what we are doing is here our standard deviation in the first question our standard deviation of the mean was given here to see where I have mentioned here assume $\sigma = 0.2$, the standard deviation or mean is not given, then what we can do we cannot use the z distribution then we will have to use t distribution we will have to estimate the s^2 from the sample and then we will have to use t distribution.

So, same question just a standard deviation is not given that means we will from the sample will estimate s^2 and use t distribution.

(Refer Slide Time: 23:05)

Case study 2a

Solution: Hypothesis Testing for Case Study 2a

Here, the population standard deviation is not known, hence we will use t-distribution.

$$\alpha = 0.05$$

The t-value for the two-tailed rejection region for 15 degrees of freedom is:

$$|t| > 2.1314$$

We are interested in a quality control problem of a medicine production company, in which we want to test the hypothesis with a significance level of 5%, that the mean volume of medicines being put in the tubes was the required 8 ml. Table below lists the data from a sample of 16 tubes. Find out if the hypothesis is true. Also calculate the 0.95 confidence interval on the mean volume of medicine per tube for the given sample.

Monalisa Sarma
IIT KHARAGPUR

So, first we are solving it by hypothesis simple hypothesis testing how we have this we know already I am not going into details. So, $\alpha = 0.05$ and t value means degrees of freedom we have taken 16 samples so, degrees of freedom is 15 degrees of freedom 15 for $\alpha = 0.05$ that means α by 0.025 we found value of t should be absolute value of t should be greater than 2.13 means if my flatted tail t has flatted. So, the rejection region is this side is 2.1314 this is -2.1314 this is my rejection region.

(Refer Slide Time: 24:03)

Case study 2a

Solution: Hypothesis Testing for Case Study 2a

Here, the population standard deviation is not known, hence we will use t-distribution.

$$\alpha = 0.05$$

The t-value for the two-tailed rejection region for 15 degrees of freedom is:

$$|t| > 2.1314$$

From the sample, $\bar{x} = 7.8925$, $s^2 = 0.03174$

So, the test statistic value $t = \frac{(\bar{x} - 8)}{\sqrt{s^2/16}} = -2.4136$

$|t|$ exceeds the critical value of 2.1314 => Reject the hypothesis.

We are interested in a quality control problem of a medicine production company, in which we want to test the hypothesis with a significance level of 5%, that the mean volume of medicines being put in the tubes was the required 8 ml. Table below lists the data from a sample of 16 tubes. Find out if the hypothesis is true. Also calculate the 0.95 confidence interval on the mean volume of medicine per tube for the given sample.

Monalisa Sarma
IIT KHARAGPUR

So, from the sample I have calculated \bar{x} bar I have calculated s^2 from the sample data what is given and I found my calculated value is -2.4136. So, it is greater than rejection region so, that

means my null hypothesis is rejected. Obviously, we got it for other case is also the null hypothesis is rejected.

(Refer Slide Time: 24:26)

The slide is titled "Case study 2a". It has a green header bar with the text "How to Compute μ ?". The main content area is orange and contains the following points:

- So, we have seen that the hypothesis $\mu_0: \mu = 8$ is rejected.
- This implies, $\mu_1: \mu \neq 8$ is true.
- Now the question is how much is μ ?
- This can be found out by confidence interval estimation.

On the right side of the slide, there is a video feed of a woman speaking. The video interface shows a progress bar at approximately 14% and a volume icon. The bottom of the slide features the IIT Kharagpur logo and the name "Monalisa Sarma IIT KHARAGPUR".

Now, if we want to so, we have seen that a hypothesis $\mu = 8$ is rejected this implies that μ is not equals to 8 is true. Now, the question is how much is μ ? This can be found by confidence interval estimation.

(Refer Slide Time: 24:43)

The slide is titled "Case study 2a". It has a green header bar with the text "Estimation of μ for t - distributions". The main content area is blue and contains the following points:

- Confidence intervals on μ for t - distribution are constructed in the same manner as those in normal distribution except that σ is replaced with s .
- The general formula of the $(1 - \alpha)$ confidence interval on μ is given by:

$$\bar{x} \pm t_{\alpha/2} \sqrt{\frac{s^2}{n}}$$

On the right side of the slide, there is a video feed of a woman speaking. The video interface shows a progress bar at approximately 14% and a volume icon. The bottom of the slide features the IIT Kharagpur logo and the name "Monalisa Sarma IIT KHARAGPUR".

Now, how do we find out the confidence interval estimation here? Because here we do not know the σ so, we cannot use z distribution. So, in that case we will have to use t distribution. So, t distribution also confidence interval are constructed in the same manner as a z distribution but

instead of $z \cdot \alpha / 2$ we find out to take $t \cdot \alpha / 2$. So, same formula so, instead of σ / \sqrt{n} , we have taken s / \sqrt{n} or $\sqrt{S^2 / n}$ is the same thing is not it?

So, instead of $z \cdot \alpha / 2$ I have taken $t \cdot \alpha / 2$ everything remains same and instead of σ I have used s .

(Refer Slide Time: 25:20)

The slide title is "Case study 2a". A green box at the top left contains the text "Estimation of μ for Case Study 2a". Below it, a blue box contains a question: "The 0.95 confidence interval on the mean weights of peanuts is". To the right of the question is the formula $7.8925 \pm 2.1314 \times 0.04453 = 7.8925 \pm 0.0949 = 7.793$ to 7.987 , with the result 7.793 to 7.987 underlined in red. The background features a watermark of a woman speaking and various engineering-related icons like gears and a circuit board. At the bottom, there are logos for IIT Kharagpur and NPTEL, and the name "Monalisa Sarma" is mentioned.

So, that is how I found a confidence interval in this range. So, it agrees with the hypothesis testing.

(Refer Slide Time: 25:30)

The slide title is "Hypothesis testing and confidence intervals: how are they related?". A green box at the top left contains the text "Relationship Between Hypothesis Testing and Confidence Intervals". Below it, a blue box contains two points: "A confidence interval on μ gives all acceptable values for that parameter with confidence $(1 - \alpha)$ " and "The probability of being incorrect in making this statement is α ". To the right, a red box contains a statement: "This implies a hypothesis test for $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$ will be rejected at a significance level of α , if μ_0 is not in the $(1 - \alpha)$ confidence interval for μ ". Below this, another red box contains the statement: "In other words, any value of μ inside the $(1 - \alpha)$ confidence interval will not be rejected by an α -level significance test." The background features a watermark of a woman speaking and various engineering-related icons. At the bottom, there are logos for IIT Kharagpur and NPTEL, and the name "Monalisa Sarma" is mentioned.

So, the relationship between the hypothesis testing and confidence interval, the confidence interval on μ so see a confidence interval and μ gives all acceptable values for that parameter we

did with confidence $1 - \alpha$ with $1 - \alpha$ confidence I can see that I am 95% confidence my μ value lies in this range if my significance level is 5 that my μ value given a significance level of 5 I can say with 95% confidence that my μ will rise in this range say this is lower bound and this is upper bound.

The probability of being incorrect in making this statement is α . So, this implies a hypothesis for a taste $\mu = \mu$ naught against μ is not equal to μ naught will be rejected at a significance of α if μ is not in a $1 - \alpha$ confidence interval is not it? What does this mean a hypothesis test it will reject the null hypothesis that μ is known, if μ is not in the when it is at the high null hypothesis if it finds that μ naught is not in the confidence interval like here 8 was not in that interval in this example, but we got see 8 is not in this interval.

So, the null hypothesis rejected is not it? This imply in other words, any value of μ inside the $1 - \alpha$ confidence interval will be rejected by α level significance it will not be rejected. In other words, any value of μ inside this confidence level will not be rejected by α level significance tests. So, hypothesis testing and confidence interval it is basically the 2 sides of the same point we can say if hypothesis testing it is rejected that means, it will not be in the confidence level.

If the value is not in the confidence level definitely it will be reflected in the hypothesis test as well. So, it is like 2 sides of the same coin.

(Refer Slide Time: 27:30)

Now, here it is one more concept in confidence interval estimation that is the margin of error. So, \bar{x} is the point estimation. So, from \bar{x} till how much we can go how much we can drill, as I told you we can hedge a bit. So, how much we can hedge from \bar{x} how much we can hedge that is my error of estimation. So, a maximum error of estimation is called a margin of error is an indicator of the precision of an estimate if point estimate is highly precise, is not it?

So, it is an precision of an estimate and is defined as one half of the width of the confidence level. So, it can tell that means some value this lower interval less than equals to μ less than equals to upper interval is not it? So, what is this μ ? I got basically this interval one half of this interval I call it other margin of error that is $E = \bar{x} + - E$ because μ will lies in this range what is $\bar{x} + E$. This is the upper interval $\bar{x} - E$ that is the lower interval.

(Refer Slide Time: 28:49)

Error of estimation

Definition: Margin of Error

The maximum error of estimation, also called the margin of error, is an indicator of the precision of an estimate and is defined as one-half the width of a confidence interval.

- The formula for the confidence limits on μ can be written as $\bar{x} \pm E$, where

$$E = \frac{z_{\alpha/2} \sigma}{\sqrt{n}}$$

$$\bar{x} - \frac{z_{\alpha/2} \sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{z_{\alpha/2} \sigma}{\sqrt{n}}$$



Monalisa Sarma
IIT KHARAGPUR



So, this is what is E ? E is this value $z_{\alpha/2} \sigma / \sqrt{n}$ because what we got $\bar{x} - z_{\alpha/2} \sigma / \sqrt{n}$ this is less than equals to μ , μ is less than equals to $\bar{x} + z_{\alpha/2} \sigma / \sqrt{n}$ this was our interval is not it? So, this is our lower interval this is our upper interval, this portion we call it as a error of estimation $z_{\alpha/2} \sigma / \sqrt{n}$ this is called the error of estimation. So, from this expression E is equals to $z_{\alpha/2} \sigma / \sqrt{n}$ and from this expression we can come to some observation.

(Refer Slide Time: 29:37)

Error of estimation

Definition: Margin of Error

The maximum error of estimation, also called the margin of error, is an indicator of the precision of an estimate and is defined as one-half the width of a confidence interval.

- The formula for the confidence limits on μ can be written as $\bar{x} \pm E$, where
- $E = \frac{z_{\alpha/2} \sigma}{\sqrt{n}}$
- The quantity E can also be described as the farthest that μ may be from \bar{x} and still be in the confidence interval.
- This bound on the error of estimation, E , is associated with a confidence coefficients.



Monalisa Sarma
IIT KHARAGPUR



So, first what is E the quantity E can also be described as the farthest that μ maybe from \bar{x} and still be in the confidence interval. What does it E describe far does that μ can be from \bar{x} bar but it is in the confidence interval. This bound on the error of estimation is associated with a

confidence coefficient definitely it is associated with a $z \alpha / 2$ is what it is associated to a confidence coefficient of α .

(Refer Slide Time: 30:06)

Maximum error of estimation

Relationships among E , α , n , and σ

- If the confidence coefficient is increased (or decreased) and the sample size remains constant, the maximum error of estimation will increase (the confidence interval will be wider).
$$E = z_{\alpha/2} \sigma / \sqrt{n}$$
- In other words, the more confidence we require, the less precise a statement we can make, and vice versa.
- If the sample size is increased and the confidence coefficient remains constant, the maximum error of estimation will be decreased (the confidence interval narrower).
- In other words, by increasing the sample size we can increase precision without loss of confidence, or vice versa.
- Decreasing σ has the same effect as increasing the sample size. This may seem a useless statement, but it turns out that proper experimental design can often reduce the standard deviation.

Confidence significance level of α confidence coefficient is $1 - \alpha$. So, in this if the confidence coefficient is increased that means, if α is decreased confidence coefficient means $1 / \alpha$ is increase means α is decrease, if α is decreased, then what happens if α is decreased my value of z will be more is not it? In this this famous concept, so, if my α is decrease initially suppose my α is this much α was this much now I am reducing this portion that means.

What my z value is becoming more desert plus side as well as my saddle is becoming less it is going more decide. So, my Z value is increasing absolute value of Z is increasing when my absolute value of z is increasing that means, what my $z \alpha / 2$ value is also increasing. So, this value is increasing that means, what this value E will be more in fact, if you can see it this way, when my rejection region become less my acceptance region becomes more when α is decrease my rejection region becomes very, very less.

When my rejection region become less accepting region becomes more, is not it? So, what happens when my confidence coefficient is increased my precision decreases I am telling you value lies between 2 to 4 one statement another statement I am telling my value lies from 1 to 10 which is more precise 2 to 4 is more precise here 1 to 10 means have been reduced interval. So,

my precision has gone down. So, similarly, when α is decreased, my confidence interval has increased that means, what my precision has decreased.

In other words, the more confidence we require, the less precise a statement we make our confidence has increased whatever statement precision has decreased if the sample size is decreased now, if we increase the sample size what happens if we increase the sample size here from the expression on only you can see if you increase the sample size my error estimation becomes less for the same of confidence coefficient.

So, it might increase the sample size without sending the confidence coefficient I can increase the precision because E value become less so, the sample size is increased and the confidence coefficient remain constant the maximum error estimation will be decreased the confidence interval becomes narrower. In other words, by increasing the sample size we can increase precision without loss of confidence or vice versa. And decreasing σ also has the same effect but decreasing σ is that something which we do not have any control is not it?

That is why it is in dismissing a useless statement, but when we how can we decrease σ how can we decrease the standard deviation or variance of a population if we have a proper experimental design while doing the well manufacturing the things are while doing the things only if I have my σ is reduced by good experimental design will have a very less σ . So, that is why it is decreasing σ it is not in the hands of a statistician it is totally on the hands on the people who are actually producing this stuff.

(Refer Slide Time: 33:31)

Maximum error of estimation

Sample Size

- Given values for σ and α and a specified maximum E , we can determine the required sample size for the desired precision.

$$E = \frac{z_{\alpha/2} \sigma}{\sqrt{n}}$$

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2}$$

Monalisa Sarma
IIT KHARAGPUR

So, from this expression, this is the expression for you now, if we are interested in finding out what is the sample size required for a particular precision and particular significance level given the precision that is E and given the significance level what is the sample size required just from this expression we can find out what is the value of E given values for σ and α and a specified maximum E we can determine the required sample size for the desired precision.

(Refer Slide Time: 34:04)

CONCLUSION

- In this lecture we learned the idea of confidence interval estimation that includes the knowledge of –
 - Interval estimate of mean, for a population with and without known standard deviation
 - Relationship between hypothesis testing and confidence interval
 - Precision of an estimate, how it depends on significance level, sample size and standard deviation of population
- In the next lecture, we will cover interval estimate of population variance and proportion.

Monalisa Sarma
IIT KHARAGPUR

So, in this lecture what we learned we learned about idea of confidence interval in estimation that includes the knowledge of interval estimate of mean of a poor population with and without known standard of deviation, if we know the standard deviation, if we do not know the standard division how we do then the relationship between the hypothesis testing and confidence interval

we have seen that on the precision of an estimate what do you mean by precision of an estimate? How it depends on a significant level?

Precision of estimates how it depends on a significant level? It has significance level is decreases my precision decreases, is not it? I am sorry, my significance level is decreased my precision also decrease how sample size if my sample sizes increase my precision increased it my standard deviation is decreased my precision increased and the next lecture will also cover interval estimate or population variance and proportion here we have only seen interval estimation of population mean.

(Refer Slide Time: 35:02)



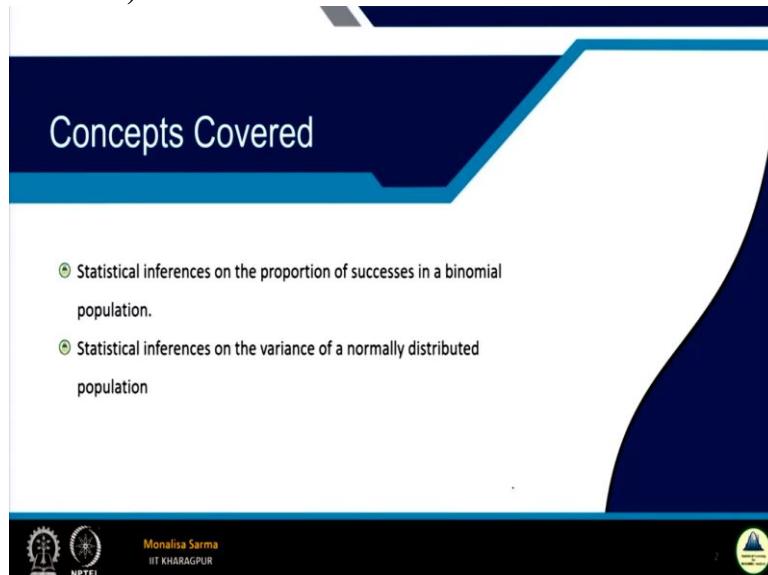
So, these are the reference. Thank you guys. Thank you.

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

Lecture - 27
Statistical Inference (Part - 5)

Hello, everyone so in continuation of our earlier discussion on statistical inferences.

(Refer Slide Time: 00:34)



Today again we will be learning few more topics on that. Today what we will see is statistical inferences on the proportion of success in a binomial population. So, we have seen statistical inferences on mean, we have also seen how to find out a confidence interval as well. So, now we will be seeing statistical inferences on the proportion of success. And when we talk about proportion, I think you guys can remember.

So, when we want to infer something about the proportion of our populations, their populations and then we could find out that it is basically it has the characteristics of a binomial populations random variable basically, it has the characteristics of a binomial distribution and that is why we can consider the population as a binomial population. And moreover, the sampling distribution also for a sampling distribution of the proportion we have seen that we can use normal distribution if the sample size is larger.

So, that as I told you again and again sampling distribution is basically the backbone of any inferences. So, when we will be seeing the proportion of success statistical inferences on a proportion of success, we will actually consider the sampling distribution of the proportion.

Similarly, we will also see the statistical inferences on the variance of a normally distributed population.

When we have to infer about the variance rather I should say when we consider when we use chi square distribution or t distribution, always our population has to be a normal population, if not normal, at least not too much away from a normal population, otherwise, the data that we will get will not be a precise result. So, that chi square distribution, F distribution and t distribution is very much sensitive to normality assumptions that we have seen again and again. So, now, first inference on a proportion, we will start with an example.

(Refer Slide Time: 02:32)

Example 1: Using Hypothesis Testing

Problem

An advertisement claims that more than 60% of doctors prefer a particular brand of pain killer. An agency established to monitor truth in advertising conducts a survey consisting of a random sample of 120 doctors. Of the 120 questioned, 82 indicated a preference for the particular brand. Is the advertisement justified?

$\hat{P} = \frac{82}{120}$

NPTEL Monalisa Sarma IIT KHARAGPUR

So, this example is a good way to understand that concept, is not it? So, like what in this example, what we have here? An advertisement claims more than 60% of doctors prefer a particular brand of painkiller more than 60% that is basically it is talking about a population. So, it is talking about the whole population as a whole about the whole lot it is talking that more than 60% of the doctors prefer a particular brand.

That means I can say $p = 0.6$ that this is the claim an agency established to monitor truth and the agency advertisement is claiming that more than 60% of the doctors prefer this particular brand, and the agency third party basically third party to monitor the truth in advertising conducts a survey consisting of a random sample of 120 doctors. So and a third party basically wants to see whether it is really correct.

So, what that says that it takes a random sample of 120 doctors, so out of 120 questions 82 indicated a preference for a particular brand. So, what is my proportion is $82 / 120$ is not it? My p cap is $82 / 120$ proportion of the sample which has preferred the particular brand. So, it is the advertisement justified. Now, the question is with this proportion when assuming that because advertisement is claiming that this is true.

Advertisement is claiming that $p = 0.6$ and from the sample what result I got this proportion of the sample which actually prefers this brand is $82 / 100$. So, this is of course, this is correct, because this we have tested it this is not something which we have tested, this is we have tested it. So assuming this correct is this possible, that is what we will check right. So, is the advertisement justified so, if we get a quite acceptable probability for this, Then we can say yes, acceptance, advertisement is justified.

Where actually I want to bring to your notice is that first, definitely, it is a question of hypothesis testing, we will do hypothesis testing, now how to frame the hypothesis?

(Refer Slide Time: 04:54)

Example 1: Solution

Solution: Hypothesis formation

Given, the proportion of doctors in the population who prefers a particular band = 60%
That is, we can say, $p = 0.6$

The hypothesis are

$H_0: p = 0.6$

$H_1: p > 0.6$

An advertisement claims that more than 60% of doctors prefer a particular brand of pain killer. An agency established to monitor truth in advertising conducts a survey consisting of a random sample of 120 doctors. Of the 120 questioned, 82 indicated a preference for the particular brand. Is the advertisement justified?

Monalisa Sarma
IIT KHARAGPUR

See here, if we frame now here who is doing the research a third party. So, when the third party is conducting the research what the third party wants to prove as I told you the alternate hypothesis is always something what you wants to prove. So, the third party wants to prove whether it is greater. It has doubt in the advertising. So, it wants to prove that whether it is better, so, definitely alternative hypothesis is p greater than 0.6.

And null hypothesis is $p = 0.6$. Why I am discussing this question, because I want to bring to your notice one more thing. Suppose this same example the same thing if instead of the third party even if what to say the particular manufacturer, the manufacturer whoever advertise is the manufacture wants to do this experiment wants to check then how he would have formed a hypothesis.

Because the manufacturer he believes that 60% of the doctor prefers this brand, he believes that. So, for him that is the status quo that is greater than 0.6. So, he wants to check whether it is less than, he does not want that less than to happen, he wants because type 1 is something which we want, what is the null hypothesis something which you want that to happen. That is why we have type 1 error very less amount, the type 1 error we have the significance level that is the significance level is a very less value we keep is not it?

Because we always put that in null hypothesis, which we want it to happen because which it maintains the status quo. So, if the particular manufacturers may have wanted to test this, then for them, they would have framed a hypothesis, they would not have framed this hypothesis instead what do you have different hypothesis? The alternative hypothesis will be p less than 0.6 they would have want to put is it p less than 0.6.

They just for checking purpose they are sure that it is greater. So, for them step this is it is greater that means greater or equal. So, as I told you null hypothesis, we always specify with equality. So, if it is less other, because again one more thing I have mentioned, remember the both the hypotheses are exhaustive, is not it? It should cover all the things so if it is less than definitely will be other one will be greater and equal.

So, has to cover all the things all the values. So, if the firm would have manufacturing, firm would have frame the hypothesis, they would have framed in this way, but now it is a third party they want to prove that it is really greater than 0.6. So, they are framing the hypothesis as p greater than 0.6. So, fine we have this.

(Refer Slide Time: 07:40)

Example 1: Solution

Solution: Hypothesis formation

Given, the proportion of doctors in the population who prefers a particular band = 60%
 That is, we can say, $p = 0.6$
 The hypothesis are

$H_0: p = 0.6$
 $H_1: p > 0.6$

Consider a significance level of 5%

An advertisement claims that more than 60% of doctors prefer a particular brand of pain killer. An agency established to monitor truth in advertising conducts a survey consisting of a random sample of 120 doctors. Of the 120 questioned, 82 indicated a preference for the particular brand. Is the advertisement justified?

Now, how to test it? It is given a significance level of 5%. So, it is a one thing, so, when this is a one tailed will have only one rejection region. So, for greater than it is sender, greater than 0.6 that means, our rejection region will be in the upper tail, when we are talking of this is the normal distribution, this is the upper tail, this portion is the upper tail, this portion is the lower tail.

So when it is, when we are looking for whether it is better, so, our rejection region will be in the upper tail, because our null hypothesis is less than equals to is not it? A null hypothesis if it is greater what is a null hypothesis? The null hypothesis basically less than equals to so, if we get a value which is much greater that means, in the upper tail then we can say that your null hypothesis is rejected.

So, this is the case p greater than 0.6 so, it is given significance level of 5% so, we have to find out the rejection region in this region in the upper tail.

(Refer Slide Time: 08:50)

Example 1: Solution

Solution: Computing Test Statistics and p-Value

So, the test statistics is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.68 - 0.6}{\sqrt{\frac{0.6 \times (1-0.6)}{120}}} = 1.86$$

An advertisement claims that more than 60% of doctors prefer a particular brand of pain killer. An agency established to monitor truth in advertising conducts a survey consisting of a random sample of 120 doctors. Of the 120 questioned, 82 indicated a preference for the particular brand. Is the advertisement justified?



Monalisa Sarma
IIT KHARAGPUR



So, we will definitely be using the z statistics here, z is the z distribution what is the value for this that is \hat{p} bar and p_0 cap - p_0 so this p_0 cap - p_0 then it is the σ / \sqrt{n} , what is the standard deviation from the binomial population that is we got a $p \times 1 / 1 - p$. So, this is $p \times 1 - 2$ we have done it while discussing sampling distribution here the standard deviation is $p \times 1 - p$ that is, so, this is $p - p / 120$.

So, we got 1.86 value so, corresponding to 1.86 there is 2 way of doing it, either you find true corresponding to this 5% significance, what is the critical region that you find if the value what we get from the z if this value is greater than the critical region, then we reject the hypothesis. Another way is that we find the p value of this and critical significance level is given 5% that means our 5% means it is 0.05.

Area is 0.05, so, if it is the area corresponding to this is lesser than 0.05 then we will reject a hypothesis. So, let us do it by computing p value, you can you can do the other one that whatever it is.

(Refer Slide Time: 10:06)

Example 1: Solution

Solution: Computing Test Statistics and p-Value

So, the test statistics is

$$z = \frac{\frac{82}{100} - 0.6}{\sqrt{\frac{0.6 \times (1 - 0.6)}{120}}} = 1.86$$


The p-value of this statistics is

$$p = P(z > 1.86) = 0.0314; \text{ therefore reject } H_0$$

An advertisement claims that more than 60% of doctors prefer a particular brand of pain killer. An agency established to monitor truth in advertising conducts a survey consisting of a random sample of 120 doctors. Of the 120 questioned, 82 indicated a preference for the particular brand. Is the advertisement justified?

Monalisa Sarma
IIT KHARAGPUR



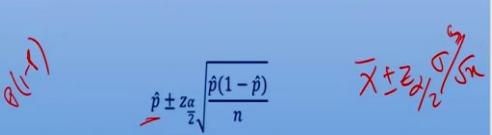
So, p value for this is we will get it from the table it is one tail. So, you do not have to plus it twice. So, it is 0.03 and 0.03 less than 0.05. So, if this portion is 0.05 so, 0.03 will be somewhere set this portion. So, if fall in the critical region is so, therefore, reject H_0 .

(Refer Slide Time: 10:40)

Inferences on a proportion

Using Estimation Approach:

A $(1 - \alpha)$ confidence interval on p based on a sample size of n with y successes is given by:

$$\hat{p} \pm \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\hat{p}(1 - \hat{p})}$$


Note

Since, there is no hypothesized value of p , the sample proportion \hat{p} is substituted for p in the formula for the variance.

Monalisa Sarma
IIT KHARAGPUR



Now, this is using hypothesis testing, we have also learned the other one approach that is the confidence interval estimation. So, confidence interval estimation I have already discussed under what situation we use? The 2 situation first one is, number one is that when we do not have any hypothesis value we cannot guess we cannot make a guess what will be the hypothesis value, in that case, we just have to estimate how we estimate?

We take a sample from the sample we calculate a particular statistics whichever we are interested and based on that statistics that statistic we call it as a point estimate based on the statistics we form the sampling distribution of that statistic, and then we find out the

confidence interval that is one way, another way is that not another way that is the one reason why we go for confidence interval another reason is it supposing the hypothesis is they have rejected.

So, like we have done an example $\mu = 8$ alternate hypothesis is $\mu_{\text{naught}} = 8$ so, if $\mu_{\text{naught}} \neq 8$ is rejected that means, μ is not equals to 8. Now, $\mu \neq 8$ it means what is the value of μ ? So, in that case also we can find the confidence intervals to know actually μ falls in what range. So, now, similarly for proportion also we can find a confidence interval.

So, saying whatever formula we use to find out the confidence interval for inferences and mean same formula to find out what to say confidence interval estimation for a mean what is the formula if you remember it was $\bar{x} + - z_{\alpha/2} \sigma / \sqrt{n}$ is not it? $\bar{x} + - z_{\alpha/2} \sigma / \sqrt{n}$. So, here instead of \bar{x} , \bar{x} is the mean of the sample.

So, we have proportion that is the p_{cap} and we have σ what is the σ distance placing the see note it, the σ is the population standard deviation remember σ is the population standard deviation. So, and when we use when the population standard deviation is not known to us, then we use t distribution, in case of t distribution instead of σ we had s is not it? s that is the standard deviation of the sample.

Here similarly, here when we consider the standard deviation, so, what is the standard deviation of the population? Standard deviation of the population is equals to $p \times 1 - p$, but if we do not know the hypothesise value what is the p of the population then what we will use like for t distribution we use s similarly here if the p is not known to us, we the sample proportion \bar{p} is p_{cap} is substituted for p in the formula for the variance.

We substitute p_{cap} when we consider, when we calculate the estimation, it is mostly because we do not have the value of p we do not have the proportion value and proportion of the population. So, we substitute p with p_{cap} so, my standard deviation is $p_{\text{cap}} \times 1 - p_{\text{cap}}$.

(Refer Slide Time: 13:50)

Example 2

Problem

A pre election poll using a random sample of 150 voters indicated that 84 favored candidate Smith. Construct a 0.99 confidence interval on the true proportion of voters favoring Smith. Comment on whether Smith can predict with 0.99 confidence that she will win the election.

$\hat{p} = \frac{84}{150}$

$\alpha = .01$

$z = .995$

Monalisa Sarma
IIT KHARAGPUR

So, you see this example here, if pre-election polls using a random sample of 150 voters indicated that 84 favoured candidates Smith. So that means my \hat{p} cap this is not p cap that is not talking about the whole population, it has taken a random sample of 150. So, what is that \hat{p} cap? \hat{p} cap is $84 / 150$. Construct a 0.99 confidence interval we have to construct a 0.99 confidence interval. In the 0.99 confidence interval what is the significance level that means? α , α is 0.01.

So, if it is a two tailed if you have to check for two tailed 0.01 means it will be both side is α will be is equal to 0.005 and then if it is two tailed, but if it is single tailed we will just consider $\alpha = 0.01$, now see this situation is what two tailed or single tailed whatever what it is asking, construct a 0.99 confidence interval on the true proportion of the voters favouring Smith. So, it is not talking about a single population one tailed or two tailed it is not talking about anything.

And we do not know the value of p as well we just whatever information we have just from the sample like the example what I have given, so, the person wants to open a retail outlet and he does not know the spending capacity of the people, so, he has this taken a sample and try to find out what is the annual income of the people, so, he does not have any idea so, he just took out the value from the sample.

So, similarly, we just have the sample value, we do not have any idea what maybe the population proportion we do not know p . So, p , we found out \hat{p} cap here. So, comment on whether Smith can predict with 0.99 competence that she will win the election.

(Refer Slide Time: 15:49)

Example 2: Solution

Solution

Given, $\hat{p} = \frac{84}{150} = 0.56$, $\alpha = 0.01$

Therefore, the confidence interval is

$$0.56 \pm 2.576 \sqrt{\frac{0.56(1 - 0.56)}{150}} = 0.456 \text{ to } 0.664$$

A pre election poll using a random sample of 150 voters indicated that 84 favored candidate Smith. Construct a 0.99 confidence interval on the true proportion of voters favoring Smith. Comment on whether Smith can predict with 0.99 confidence that she will win the election.



Monalisa Sarma
IIT KHARAGPUR

So, now here we have p cap is this $\alpha = 0.01$. So now, what is the confidence interval here the confidence interval same value whatever we used here this is the confidence interval p cap $\pm z_{\alpha/2}$ p cap $\times \sqrt{1 - p \text{ cap}} / n$. So, just putting the value this is z of $\alpha/2$ values corresponding to that is 2.576 you can see it from the z table. So, we got this is the value we got.

This is the confidence interval from 0.45 to 0.66 that means what is the last question it is asking comment on whether Smith can predict with 0.99 confidence that you can win the election for winning the election at least more than 50% so, both. So, here the person who are voting is we have values below 0.5 also of course, we have values above 0.5 but we have below values be below 0.5 as well.

(Refer Slide Time: 16:52)

Example 2: Solution

Solution

Given, $\hat{p} = \frac{84}{150} = 0.56$, $\alpha = 0.01$

Therefore, the confidence interval is

$$0.56 \pm 2.576 \sqrt{\frac{0.56(1 - 0.56)}{150}} = 0.456 \text{ to } 0.664$$

Since the interval contains values below 50% also, this implies Smith can not predict with 0.99 confidence that she will win the election.

A pre election poll using a random sample of 150 voters indicated that 84 favored candidate Smith. Construct a 0.99 confidence interval on the true proportion of voters favoring Smith. Comment on whether Smith can predict with 0.99 confidence that she will win the election.



Monalisa Sarma
IIT KHARAGPUR

So, since the interval contains values below 50% also this implies Smith cannot predict with 0.99 confidence that she will win the election. So, next is; we will see how we infer on the variance of one population.

(Refer Slide Time: 17:12)

Inferences on the variance of one population

Comparing the Variance of a Population with a Prescribed Value

To test the null hypothesis that the variance of a population is a prescribed value, say σ_0^2 , the hypotheses are

$$H_0: \sigma^2 = \sigma_0^2, \quad \checkmark$$

$$H_1: \sigma^2 \neq \sigma_0^2$$

Monalisa Sarma
IIT KHARAGPUR

All this we are considering for one population. So, variance for one population similarly, the way we do it for mean we have done it for proportion. Similarly, this will be our null hypothesis this is an alternate hypothesis.

(Refer Slide Time: 17:27)

Inferences on the variance of one population

Comparing the Variance of a Population with a Prescribed Value

To test the null hypothesis that the variance of a population is a prescribed value, say σ_0^2 , the hypotheses are

$$H_0: \sigma^2 = \sigma_0^2,$$

$$H_1: \sigma^2 \neq \sigma_0^2$$

The statistic used to test the null hypothesis is

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum(y-\bar{y})^2}{\sigma^2} = \frac{SS}{\sigma^2}$$

Monalisa Sarma
IIT KHARAGPUR

And for proportion I am in for variance that is the statistics we use remember for what to say to find out the sampling distribution of a variance we have used the statistics $n - 1 s^2 / \sigma^2$ this is the value and this $n - 1 s^2 / \sigma^2$ it it has a chi square distribution remember we have seen, it as a chi square distribution with degrees of freedom $n - 1$ is not it?

So, this is has a chi square distribution and what is this $n - 1 s^2$ you can also tell is a sum of SS is called sum of squares basically, I have shown you why it is called sum of squares, because if we find out the formula for s^2 what you get remember the formula for variance the whole portion is basically and then in it is $1 / n - 1$ summation of $x_i - \bar{x}$ bar 2 this is the formula for a square. So, if I bring $n - 1$ here $n - 1 s^2$ what remains is this summation of $x_i - \bar{x}$ bar 2 .

So, this is a summation of square that is why when you say s^2 is also called summation of squares. So, there is nothing so, heard about that. So, you can use either this or you can use this. So, basically $n - 1 s^2 / \sigma^2$ is suite.

(Refer Slide Time: 18:45)

Inferences on the variance of one population

Comparing the Variance of a Population with a Prescribed Value

- To test the null hypothesis that the variance of a population is a prescribed value, say σ_0^2 , the hypotheses are

$$H_0: \sigma^2 = \sigma_0^2,$$

$$H_1: \sigma^2 \neq \sigma_0^2$$
- The statistic used to test the null hypothesis is

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum(y-\bar{y})^2}{\sigma^2} = \frac{SS}{\sigma^2}$$
- If the null hypothesis is true, this statistic has χ^2 distribution with $(n - 1)$ degrees of freedom.

Monisha Sarma
IIT KHARAGPUR

The statistics chi square is the null hypothesis is true this statistics $n - 1 / s^2 / \sigma^2$ this statistic says chi square distribution with $n - 1$ degrees of freedom, that we have already seen. So, now we will have to find a for if we want to infer on the variance we will have to find out the value of the statistics and then we will have to see whether this has a chi square distribution basically it meaning it has a chi square distribution means it has to have a significant probability not if it says very less probability.

We have seen it is 95% of the value falls within what range I am, if you remember have discussed that. So, if it is so, chi square distribution chi square distribution we have something of this chart. So, 95% of the value basically satisfies this, satisfy the chi square distribution of the for the hypothesised σ value. If our value falls in this region why will a value fall in this region?

If we assume variance to be very small which is very unlikely very small of course, it can happen it is not that is but it is very unlikely, if we assume a variance to be very small we will get in this range. Again if we assume our variance to be very big, very high then it will fall in this range that is also quite unlikely a very high variance, it is a very unstable process. So, that is also quite unlikely.

So, if our value falls within this 95% range, then we can say that it is it satisfies the chi square distribution. So, now here, so what is the rejection region corresponding to a particular significance level.

(Refer Slide Time: 20:29)

Inferences on the variance of one population

The rejection region is:

$\text{reject } H_0 \text{ if: } \left(\frac{SS}{\sigma_0^2} \right) > \chi_{\alpha/2}^2,$
 $\text{or if: } \left(\frac{SS}{\sigma_0^2} \right) < \chi_{1-\alpha/2}^2$

Remember how to find out a rejection region. So, it has like let me write it here only so, it is $n - 1 s^2 / \sigma^2$ basically from the confidence interval it will be easier to write it this way $\chi^2 1 - \alpha / 2$, this is less than $\chi^2 \alpha / 2$ this probability is equal to $1 - \alpha$ is not it? My this value $n - 1 s^2 / \sigma^2$ it should fall between this $\chi^2 \alpha / 2$ or $\chi^2 1 - \alpha / 2$.

So, that means, what is my rejection region if my this statistics, if this statistics is if it is greater than $\chi^2 / 2$, greater than $\chi^2 \alpha / 2$ means this region then we reject it or this value if it is less than this value $\chi^2 1 - \alpha / 2$ is value, $\chi^2 \alpha / 2$ is this portion. So, if it falls in this region or region then we will reject the hypothesis. So, reject H_0 if this is true or if this is true.

(Refer Slide Time: 21:46)

The rejection region is:

$$\text{reject } H_0 \text{ if: } \left(\frac{SS}{\sigma_0^2} \right) > \chi_{\alpha/2}^2$$

$$\text{or if: } \left(\frac{SS}{\sigma_0^2} \right) < \chi_{(1-\alpha/2)}^2$$

Important Point

- Hypothesis tests on variances are often one-tailed because variability is used as a measure of consistency, which is indicated by small variance.
- Thus, an alternative hypothesis of a larger variance implies an unstable or inconsistent process.

Monalisa Sarma
IIT KHARAGPUR

Coming back so now, if you see here hypothesis tests on variance are often one tailed because variability is a measure of consistency hypothesis testing when we try to find out infer on the variance usually when we try to infer on the variance we never try to inference if variance is great what to say smaller than this value, because we need smaller variance what does smaller variance indicates? Smaller variance indicates consistency.

So, any product, any equipment, any material whatever it is we want variance to be as less as possible. So, smaller variance is a positive thing for us. So, definitely we will not try to check for usually we do not try to check for a smaller variance what we try to check for whether it is greater variable. So, mostly in hypothesis testing for inference and variance it usually for us it is usually that means we test for high and greater variances that way it is single tailed.

See alternative hypothesis of a larger variance implies an unstable or inconsistent process. So, in unstable process, no point carrying out hypothesis testing or confidence interval estimation.

(Refer Slide Time: 22:58)

Confidence interval of variance

The Lower and Upper Limit of Confidence Interval

- As the χ^2 distribution is not symmetric, the confidence interval is not symmetric about S^2
- Now, to calculate the upper and lower confidence interval,

$$P\left[\chi_{(1-\alpha/2)}^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_{\alpha/2}^2\right] = 1 - \alpha$$

$$\chi_{(1-\alpha/2)}^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_{\alpha/2}^2$$

$$\Rightarrow \frac{\chi_{(1-\alpha/2)}^2}{(n-1)S^2} < \frac{1}{\sigma^2} < \frac{\chi_{\alpha/2}^2}{(n-1)S^2}$$

$$\Rightarrow \frac{(n-1)S^2}{\chi_{\alpha/2}^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_{(1-\alpha/2)}^2}$$

So, The lower limit of the

confidence interval: $L = \frac{SS}{\chi_{\alpha/2}^2}$,

and, the upper limit of the

confidence interval $U = \frac{SS}{\chi_{(1-\alpha/2)}^2}$



Monalisa Sarma
IIT KHARAGPUR

22



So, now, what is the confidence interval we have seen the using hypothesis testing how we do this now, we will see how the confidence interval for variance. So, this is what just now what I have mentioned if the confidence coefficient is $1 - \alpha$ if the significance level is α , my confidence coefficient is $1 - \alpha$, if it is 5% confidence coefficient is 95%. So, probability that it falls within this range is $1 - \alpha$.

But my this statistics falls is the less than the X and $\chi^2 \alpha / 2$ and greater than $\chi^2 1 - \alpha / 2$. So, if I simplify a bit this whole thing if I simplify a bit how I have simplified just divided both sides by $n - 1 S^2$ just simple simplification I will get this value. So, my σ^2 would fall within this range what range $n - 1 S^2 / \chi^2 \alpha / 2$ $n - 1 S^2 / \chi^2 1 - \alpha / 2$.

So, this is the confidence interval, confidence interval for confidence coefficient $1 - \alpha$. So, the lower interval lower limit which is a lower limit, lower limit is $n - 1 S^2 / \chi^2 \alpha / 2$ or I can tell $SS / \chi^2 \alpha / 2$ this is my lower limit what is my upper limit is assessed by $\chi^2 1 - \alpha / 2$ this is my upper limit.

(Refer Slide Time: 24:25)

So, we will see a problem for that in processing grain in the beverage industry the person is extract recovered is measured, a particular beverage industry introduce a new source of grain and the percentage extract on 11 separate days is as follows. This is the sample what we have taken regarding the sample as a random sample from a normal population variance of course the population has to be normal.

Calculate the 90% confidence interval for the population variance we have to calculate the 90% confidence interval, here we do not know what is the population the variance is not given. We do not know what maybe the population variance or population standard deviation, we just may have taken a sample from the sample we got this data and from this data we will can calculate the 90% confidence interval.

So, basically from this sample we will try to find out the value of S^2 once you find out the value of S^2 that is the variance from the sample then $n - 1$ that is here total 11 days $n - 1$ is 10 $n - 1 S^2 / \sigma^2$. So, this is what to say this is the statistics we have to find out. So, what happened, what does the confidence interval say? See here to find out this, I do not have the value of σ^2 , I do not know what is the value of σ^2 here, I know only S^2 .

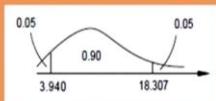
So, I cannot use sampling distribution based I cannot hypothesise the value because I do not have any idea. So, definitely I cannot use hypothesis testing here what I will do I will just go and find out the confidence interval. To find out the confidence interval what was this? This is the formula remember, this is the upper limit lower limit this is the upper limit.

(Refer Slide Time: 26:16)

Example 3

Solution

Given, $n = 11$, $\bar{x} = 94.045$, $s = 1.34117$



In processing grain in the beverage industry, the percentage extract recovered is measured. A particular beverage industry introduces a new source of grain and the percentage extract on eleven separate days is as follows: 95.2, 93.1, 93.5, 95.9, 94.0, 92.0, 94.4, 93.2, 95.5, 92.3, 95.4. Regarding the sample as a random sample from a normal population, calculate a 90% confidence interval for the population variance.

90% confidence interval for variance is given by,

$$\text{Lower confidence interval } L = \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2} = \frac{10 \times 1.34117^2}{18.307} = 0.98$$

$$\text{Upper confidence interval } U = \frac{10 \times 1.34117^2}{3.940} = 4.57$$



Monalisa Sarma

IIT KHARAGPUR

27



So, I found out \bar{x} bar I found that s than what is the $\chi^2 \alpha / 2$ $\chi^2 \alpha / 2$ is this value that we will find it from the χ^2 table, what is α ? α is given is 10%. So, $\alpha / 2$ will be 5% that is 0.05. But this chi square value corresponding to 0.05 with 11, it is 10 degrees of freedom, 10 degrees of freedom we will get it from the chi square table. And similarly for this is $1 - \alpha / 2$ χ^2 of $1 - \alpha / 2$ for 10 degrees of freedom we will get this value.

So, now, when once we know this value, we can find out the interval low interval, upper interval and lower interval using the formula.

(Refer Slide Time: 27:06)

CONCLUSION

- ④ In this lecture we learned the estimation of confidence intervals for population proportion and population variance.
- ④ The concepts were illustrated with few examples for clear understanding.
- ④ In the next lecture, we will cover a quick tutorial before starting discussion on inferences for two populations



Monalisa Sarma

IIT KHARAGPUR

28



So, that is all in this class. So, what we have learned? We have learned the estimation of confidence interval for proportion and population variance. So, I have tried to illustrate the concept with some examples. And in the next lecture, we will cover a quick tutorial before starting the discussion on the inferences for 2 populations.

(Refer Slide Time: 27:27)

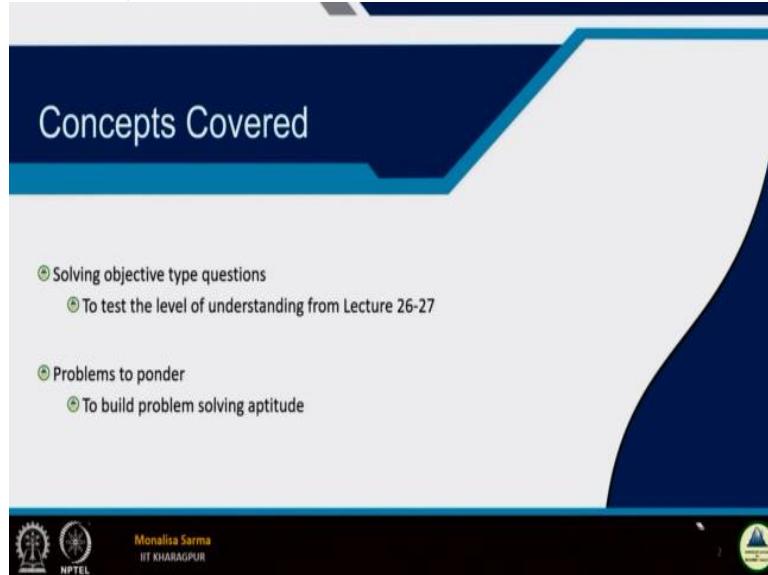


So, these are the reference. Thank you guys. Thank you.

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

Lecture - 28
Tutorial on Confidence Interval

(Refer Slide Time: 00:30)



The slide has a dark blue header bar with the text "Concepts Covered". Below the header, there is a list of bullet points:

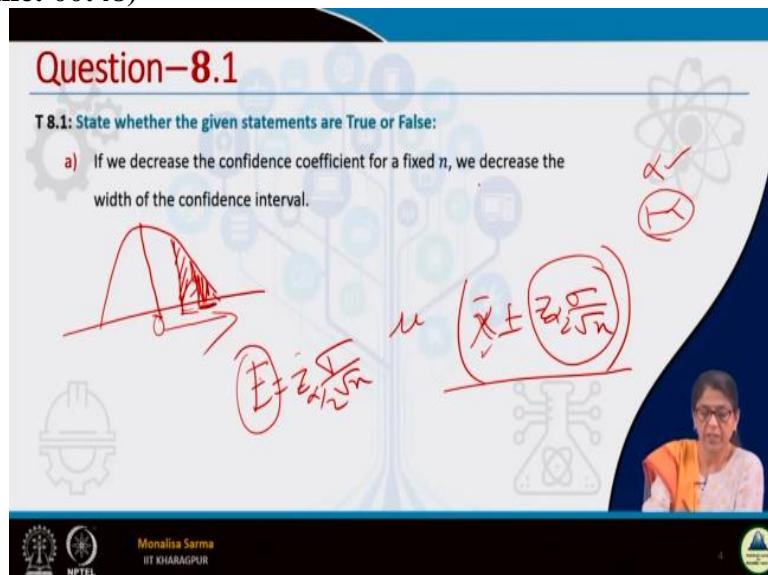
- ④ Solving objective type questions
- ④ To test the level of understanding from Lecture 26-27

- ④ Problems to ponder
- ④ To build problem solving aptitude

At the bottom of the slide, there are logos for NPTEL and IIT KHARAGPUR, along with the name "Monalisa Sarma" and "IIT KHARAGPUR".

Hello guys, so we will do a tutorial today based on the topics what we have learned. So, last 2 lectures basically lecture 26 and 27. So, we are doing a couple of tutorials like because this things we will understand it more better using more and more problem solving. So that is the reason why I am giving so many tutorials.

(Refer Slide Time: 00:48)



The slide is titled "Question-8.1" in red. Below the title, it says "T.8.1: State whether the given statements are True or False:". There is a question listed: "a) If we decrease the confidence coefficient for a fixed n , we decrease the width of the confidence interval." To the right of the text, there is a hand-drawn diagram of a normal distribution curve with a mean μ and a confidence interval centered around it. The width of the interval is labeled $2\sigma/\sqrt{n}$. A red checkmark is placed next to the statement, indicating it is true. The slide also features a video feed of the professor at the bottom right and various engineering-related icons in the background.

So, now, first we will start with objective type question which basically will help us to keep making a quick recap of whatever we studied. So, first is if we decrease the confidence

coefficient for a fixed n we have decreased the confidence coefficient. We have decreased the confidence coefficient meaning what? Means we have increased α we have decreased the confidence coefficient that means, we have increased α we have decreased $1 - \alpha$ this $1 - \alpha$ is divisible.

So, we have increased α if we decrease the confidence coefficient for a fixed n we decrease the width of the confidence interval, what is the width of the confidence interval is basically the error of estimation remember my what to say μ^2 value will lie within what range $x_{\bar{}} + z_{\alpha/2} \sigma / \sqrt{n}$. So, my μ^2 will lie within this range I sorry my μ not μ^2 my μ will lie it within this range.

Similarly, if we talk of variance proportion whatever it is, accordingly it is value will get changed, so, this is my total range. So, this is my total width, my width is basically E , what is this? This value is E , so, I can write $E = z_{\alpha/2} \sigma / \sqrt{n}$. So, here when I am increasing α , you see the diagram here from the table, increasing α means what? I am increasing the critical region, increase it; suppose if my α was here.

But basically here, when say 0.1, now I am increasing to 0.5, I am increasing this area, α I am increasing means from here I am bringing to here before it was this portion, now increasing means I added this portion as well. So, when I am increasing α , what is happening my z value is decreasing because this is 0 from 0, I am going this way it is increasing z value is increasing, when I am increasing α my z value is decreasing.

So, in this expression, when my z value is decreasing, what is happening? z value is decreasing my means my E will be smaller, my error of estimation will be smaller when my error of estimation will be smaller that means my width will become less, is not it? So, we decrease the width of the confidence interval that is true.

(Refer Slide Time: 03:47)

Question–8.1

T 8.1: State whether the given statements are True or False:

- a) If we decrease the confidence coefficient for a fixed n , we decrease the width of the confidence interval. [True]
- b) If a 95% confidence interval on μ was from 50.5 to 60.6, we would reject the null hypothesis that $\mu = 60$ at the 0.05 level of significance. [False]



Monalisa Sarma
IIT KHARAGPUR



So, next question, if a 95% confidence interval on μ was from 50.5 to 60.6, this is a confidence interval given we would reject the null hypothesis that $\mu = 60$ at 0.05 level of significance so, this is my confidence interval. So, μ very well lies within this region, because this ranges to 50.5 to 60.6, 60 lies in this region. So, any value that lie in this region, we will not reject the null hypothesis that is what we have seen.

If the value lies within the confidence interval, then the null hypothesis is not rejected. So, this value is lying within this region. So, we will not reject so this is false.

(Refer Slide Time: 04:31)

Question–8.1

T 8.1: State whether the given statements are True or False:

- a) If we decrease the confidence coefficient for a fixed n , we decrease the width of the confidence interval. [True]
- b) If a 95% confidence interval on μ was from 50.5 to 60.6, we would reject the null hypothesis that $\mu = 60$ at the 0.05 level of significance. [False]
- c) If the sample size is increased and the confidence coefficient remains constant, the width of the confidence interval will decrease. [True]



Monalisa Sarma
IIT KHARAGPUR



So, if the sample size is increased, and the confidence coefficient remains constant, the width of the confidence interval will decrease. Again same E what is my E? $E = z \alpha / 2 \sigma / \sqrt{n}$, \sqrt{n} is in the denominator. If denominator is bigger what happens our value becomes smaller. So,

what happens if the sample size is increased? If I increase the sample size, my E will decrease. E means the width will decrease.

Then and the confidence coefficient remains constant when α remains same, but I am the increase in the sample size, the width of the confidence interval will decrease, the width will decrease. That is one of the advantages, if I increase the sample size, then what happens? My precision is also increasing. At the same time, I am not losing on the confidence also. Usually, what happens when I change the confidence coefficients.

Change the confidence coefficients means, when I reduce the confidence coefficient my precision is increasing, but my confidence is going down is not it? But if I increase the sample size, still my precision is increased, but my confidence does not change my confidence still remains the same. So that is the advantage of increasing having a bigger sample size so, this is true.

(Refer Slide Time: 05:52)

Question–8.1

T 8.1: State whether the given statements are True or False:

- a) If we decrease the confidence coefficient for a fixed n , we decrease the width of the confidence interval. [True]
- b) If a 95% confidence interval on μ was from 50.5 to 60.6, we would reject the null hypothesis that $\mu = 60$ at the 0.05 level of significance. [False]
- c) If the sample size is increased and the confidence coefficient remains constant, the width of the confidence interval will decrease. [True]
- d) The variance of a binomial proportion is npq [or $np(1 - p)$]. [False]

Monalisa Sarma
IIT KHARAGPUR

The variance of a binomial proportion is npq or $np(1 - p)$, is it true? Just now, we have seen it I mean sorry, not just now, I mean, in the last lecture, we have seen, so, the variance of a binomial proportion is $p \times 1 - p$, $p \times q$, that is the variance of a binomial proportion so, this is false.

(Refer Slide Time: 06:17)

Question–8.1

T 8.1: State whether the given statements are True or False:

- e) The sampling distribution of a proportion is approximated by the χ^2 distribution. [**False**]

NPTEL
Monalisa Sarma
IIT KHARAGPUR

13

Again, there are some more questions, the sampling distribution of a proportion is approximated by chi square distribution as a true we have seen it in the last lecture, the sampling distribution of proportion is not approximated by chi square distribution it is approximated by the normal distribution considering this as the binomial population and considering a bigger sample size it is approximated by normal distribution it is chi square distribution.

(Refer Slide Time: 06:44)

Question–8.1

T 8.1: State whether the given statements are True or False:

- e) The sampling distribution of a proportion is approximated by the χ^2 distribution. [**False**]
- f) The t test can be applied with absolutely no assumptions about the distribution of the population. [**False**]

NPTEL
Monalisa Sarma
IIT KHARAGPUR

13

The t test can be applied with absolutely no assumption about the distribution of the population completely false t test is very much sensitivity to normality assumption of the population.

(Refer Slide Time: 06:59)

Question-8.1

T 8.1: State whether the given statements are True or False:

- e) The sampling distribution of a proportion is approximated by the χ^2 distribution. [False]
- f) The *t* test can be applied with absolutely no assumptions about the distribution of the population. [False]
- g) The degrees of freedom for the *t* test do not necessarily depend on the sample size used in computing the mean. [**True**]



Monalisa Sarma
IIT KHARAGPUR



The degree of freedom for a *t* test do not necessarily depend on the sample size used in computing the mean this thing I did not take this in the class so, we will explain it here usually what happens whatever we have seen in the *t* test or chi square it has one parameter that is the degrees of freedom and degrees of freedom how do we take? It is the sample size is not it?

Sample size - 1 is out in degrees of freedom for *t* test chi square test *f* square test or *f* test for all these days we have seen the they have just 1 parameter that parameter is the degrees of freedom what is the degrees of freedom? Degrees of freedom is the sample size - 1 but this question, but you see there are some situation like let me take an example suppose I am interested suppose a factory is producing stone chips.

You know this small shown stone chips so, a factory is producing that and I want to know the mean width of the stone chips. So, factory is producing huge number of stone chips and I am interested in me not be stones is because I need a particular size of stones for building something some I want to build something somehow some whatever some hotel whatever it is. So, now, how do I take for that?

Definitely, I should to know the mean of the population, I will have to take a sample. So, suppose taking a small sample will definitely make no sense in such cases. We will take a bigger sample suppose we took 100 a sample of say around 100 stones, definitely we will not count 1, 2 on an estimate we have taken around 100 stones and now will have to take the mean of this 100 stones how we will have to take the mean of the each.

And every how will have to weigh it what is the weight and then we will have to find out the mean no that we will not do that we are not so jobless. So, what we will do is that we will take this whole 100 samples in a weighing machine we will just weigh the weight of this 100 stones and divided by 100 that becomes my mean weight of the stones, that is my mean weight of the stones now, what is the variance of the stones?

So, variance definitely I cannot just weigh it together and can find out the variance and it is also not possible for me to find out the variance of this 100 stones my sample size after 100. So, it is not possible for me to find a variance of these 100 stone also from each stone I will subtract from the weight of each stone I will subtract it from the mean whatever mean I got that is also not a very interesting job.

So, what I will do, I will take suppose I took 10 stones, and from for the 10 stones, I have sincerely calculated a variance. For each stone variance you know how to calculate the I will have to calculate from the mean I will subtract this value of weight of each stones. So, I have taken a from this sample I have taken a sub sample of say 10 and from there I found out the variance. So, in this case, then my degrees of freedom will not be $100 - 99$. Later my degrees of freedom will be $10 - 1$.

So that is very much correct. And in this sort of situations, where it is really not feasible to take the find out the variance of whole sample, this sort of thing this process is also taken up. So, the degrees of freedom for t tests do not necessarily see the word depend on the sample size using computing the mean yes, it is very true it depends on the sample size use for computing the variance so, this statement is true.

(Refer Slide Time: 10:31)

Problem-8.2

T 8.2: The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per milliliter. Find the 95% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3 gram per milliliter.

$$\begin{aligned} n &= 36 \\ \bar{x} &= 2.6 \end{aligned}$$



Monalisa Sarma
IIT KHARAGPUR

19



Now, we will see some problems. So, last 2 lectures whatever what we have learned, we try to find out what to say confidence interval basically confidence interval and how it changes with the sample size, what is the error of estimation and all those things we have learned and how to infer about the population how to infer about the variance, population proportion how to infer the variance of population we have learned in the last 2 lectures we have learned all those. So, based on that we; will be seeing some problems.

The first question you see the average zinc concentration recovered from a sample measurements taken in 36 different location in the river. So, 36 different locations that means, my n is 36 in a river is found to be 2.6 grams per millilitre. So, every is that means is this μ or x bar? This is x bar because this is we are talking about the sample. So, what is my x bar? x bar is 2.6.

Find a 95% confidence interval for the zinc concentration in the river we have to find a confidence interval assuming the population standard deviation is 0.3 good it is they have given the population standard deviation they have assumed the population standard deviations 0.3. And we have to find the confidence interval we can very well find out a confidence interval because we will be using $z_{\alpha/2}$ instead of $t_{\alpha/2}$.

(Refer Slide Time: 11:58)

Problem-8.2 : Solution

Given, estimate of μ , that is, $\bar{X} = 2.6$

Confidence co-efficient = 95%

\therefore The significance level, $\alpha = 5\%$

\therefore The confidence interval,

$$\left(2.6 - (1.96) \left(\frac{0.3}{\sqrt{36}} \right) \right) < \mu < \left(2.6 + (1.96) \left(\frac{0.3}{\sqrt{36}} \right) \right)$$

$$\Rightarrow 2.50 < \mu < 2.70$$

\Rightarrow 95% Confidence Interval

T 8.2 : The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per milliliter. Find the 95% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3 gram per milliliter.



Monalisa Sarma
IIT KHARAGPUR

13



So, this is given significance level so, this is how we calculate the confidence interval is not it? Confidence interval how we calculate $\bar{x} - z \alpha / 2 \sigma / \sqrt{n}$ this is less than equals to μ this is less than or equal to $\bar{x} + z \alpha / 2$ what to say σ / \sqrt{n} is not it? So, that is all we have all the value what is the 5% means? 0.025 corresponding to 0.025 z table we will find our value is 1.96.

You can check it that is 1.96 this is \bar{x} value this is the variance σ and we have taken from 36 different locations so, \sqrt{n} is 36. So, this is the confidence interval.

(Refer Slide Time: 13:10)

Problem-8.3

T 8.3: Suppose that a population mean is to be estimated from a sample of size 25 from a normal population with $\sigma = 5.0$. Find the maximum error of estimation with confidence coefficients 0.95.
What changes if n is increased to 100 while the confidence coefficient remains at 0.95?



Monalisa Sarma
IIT KHARAGPUR

13



Similar type of question suppose that the population mean is to be estimated from a sample size of 25 from a normal population which $\sigma = 5.0$ and previous problem also average problem also population mean is not we do not have any hypothesized value. So, we just have to estimate from the sample find maximum error estimation with confidence coefficient 0.95.

So, here we have to find out maximum error estimation basically we have to find out a width what changes if n is increased to 100 the confidence coefficient remains 0.95.

We have already seen when we increase the n what happens our width decreases when we decrease our confidence coefficient our width decreases when we increase our significance level our width decreases we have seen that.

(Refer Slide Time: 14:07)

Problem-8.3 : Solution

The maximum error of estimation, considering 0.95 confidence co-efficient is

$$E = 1.96 \times \frac{5}{\sqrt{25}} = 1.96$$

If $n = 100$, keeping the other factor same

$$E = 1.96 \times \frac{5}{\sqrt{100}} = 0.98$$

T 8.3 : Suppose that a population mean is to be estimated from a sample of size 25 from a normal population with $\sigma = 5.0$. Find the maximum error of estimation with confidence coefficients 0.95. What changes if n is increased to 100 while the confidence coefficient remains at 0.95?

Monalisa Sarma
IIT KHARAGPUR

So, this is the formula for E what is E? E is nothing but $z \alpha / 2 \times \sigma / \sqrt{n}$. So, this is $z \alpha / 2$ is 0.95 is a significance coefficient that means 5% is a significance level $\alpha / 2$ will be 0.025. So, 0.025% to 0.025 values 1.96. So, this is the E value considering the sample size of 25. Now, if we take a sample size of 100 see here the width was 1.96. Now, the width has become 0.98 our width has become reduced means our precision has increased.

Remember the concept of precision I am telling my x value lie between 2 to 5. I am telling my x value lies between 2 to 5 or if I tell my x value lies from 1 to 10 which is more precise my first statement is more precise. I am telling my x value is lying from 2 to 5. And other when I am telling my x value lies from 1 to 10, I am a bit imprecise precision is going down. So, here when the width decreases, precision increases.

(Refer Slide Time: 15:16)

Problem-8.4

T 8.4: An apple buyer is willing to pay a premium price for large size apples. The buyer wants to test if the apples are sufficiently large with a confidence coefficient of 10%, so he takes a random sample of 12 apples from the load and measures their diameters. The results are given in table below. Calculate the confidence interval.

2.9	2.8	2.7	3.2
2.1	3.1	3.0	2.3
2.4	2.8	2.4	3.4

Check

We investigated a "similar" problem in the last tutorial (Tutorial VII) !!!



Monalisa Sarma
IIT KHARAGPUR



An apple buyer is willing to pay a premium, this question we have solved it, remember this question we have solved it, but we will see there is a difference in this the question what we have solved and what we will be doing now, an apple buyer is willing to pay a premium price for larger size apple there in that question, our size of apple was hypothesized size was given the buyer wants to buy an apple which is better than some 2.5 diameter.

Then only, the buyer will buy the apple in the premium price that was the question. Now, he just he does not know the size of the apple he just know that it is the size is bigger than he will pay the premium price. So, he just wants to know what is the size of apples for that, since he has he has no idea so he is just finds out the confidence interval same question, but the hypothesis value is not given the buyer wants to test if the apples are sufficiently large with a confidence coefficient of 90% that means a significance level 10%.

So, he takes a random sample of 12 apples from the load and measure the damages, the results are given below in this table. So, from here what you will find out from this table, so, he is what he wants to find out the diameter of the apple. So, these are the diameter of the apples are given. So, basically he will find out \bar{x} he will find out S^2 is not it? So, and we investigate a similar problem in the last tutorial, you can check it.

(Refer Slide Time: 16:48)

Problem-8.4

Notice the difference in the problem statement

T 7.3: An apple buyer is willing to pay a premium price for a load of apples if they have, as claimed, an average diameter of more than 2.5 in. The buyer wants to test the claim of sufficiently large apples, so he takes a random sample of 12 apples from the load and measures their diameters. The results are given in table below. Consider a significance level of 5%.

T 8.4: An apple buyer is willing to pay a premium price for large size apples. The buyer wants to test if the apples are sufficiently large with a confidence coefficient of 90%, so he takes a random sample of 12 apples from the load and measures their diameters. The results are given in table below. Calculate the confidence interval.



Monalisa Sarma
IIT KHARAGPUR

29



So, that what was the difference? You will see the difference in the 2 questions here it was given if they have as claimed and average diameter more than 2.5 inch the buyer wants to then the buyer will be paying a premium price, but now nothing of that sort is given an earlier the significance level of 5% is given now, it is given a confidence coefficient of 90% when the significance level of 5% and what is the confidence coefficient confidence coefficient was 10% I mean sorry, confidence coefficient is 95%.

(Refer Slide Time: 17:18)

Problem-8.4 : Solution

The rejection region is $t > 1.3634$ (for $df = 11$)

From the sample,

$$\bar{X} = 2.758, S^2 = 0.1554$$

The one-sided lower 0.90 confidence interval for the mean apple size is

$$\begin{aligned} \bar{X} - t_{\alpha/2} \frac{S}{\sqrt{n}} \\ 2.758 - 1.3634 \frac{0.1554}{\sqrt{12}} \\ = 2.758 - 0.155 = 2.603 \end{aligned}$$

T 8.4 : An apple buyer is willing to pay a premium price for large size apples. The buyer wants to test if the apples are sufficiently large with a confidence coefficient of 10%, so he takes a random sample of 12 apples from the load and measures their diameters. The results are given in the table. Calculate the confidence interval.



Monalisa Sarma
IIT KHARAGPUR

30



So, here from this region see buyer wants to find out if he is not interested in finding out he is basically is interested in finding out if the size of the apple is greater if the size of the apples and more is not it? So, what he will do is that in this case I say try to understand very carefully. So, he wants to find out the; if the size of the apple is more than he will pay the premium price. So, in this case, he will definitely not want to find out what is the size of the apples lies in what he wants bigger apple more the bigger more the better.

So, in that case, he will just find out what is the lower confidence interval because the apple can be how small from the sample apple can be how small. So, more the bigger more the better. So, he do not want to consider the lower upper confidence interval he will just find out the lower confidence interval. And if this question would have been a hypothesis testing; then remember we have here also we have checked for greater.

And condition μ is was to 2.5μ greater than 2.5 when μ was greater than 2 my alternate hypothesis was greater than 2.5 then our rejection region was in the upper tail see the difference our rejection region was in the upper tail when we want to check it for greater at the same time for the same we want to look for greater but when we find a confidence interval for when our interest in finding out a greater.

So, when we look for a confidence interval confidence interval will always go for lower confidence interval why? Because we want to see how less it can go how small it can go greater means it can be any value we do not want to interval for that. So, t value for 1 degree of freedom since we are interested in finding out greater is 1 portion. So, we will not consider is the $\alpha / 2$ it will be just t of α t of α means t of 0.10 .

Sorry t of 0.10 for 11 degrees of freedom in the table if you will see you will get this value in the last problem in the last tutorial I have sold the t table here we could see this value so this is my t α . So, from the sample I got this is my x bar sample this is my from the sample I got x bar value I got a square value. So, now, I just have to find out a lower confidence interval. What is the lower confidence interval?

This is my lower confidence and x bar - $t \alpha$ this is x bar this is $t \alpha x$ bar - $t \alpha$, then what is that $S^2 / n \sqrt{S^2 / n}$. If I would have interested in finding out the upper confidence, then it will be x bar plus, if I am interested in both finding out the lower and upper, then my α will be $\alpha / 2$, is not it? So, since I am interested here, in finding out the greater, I do not want the lower regions, so it will be αx bar - $t \alpha \sqrt{S^2 / n}$.

And so putting this value I got this is my value, my lower bound, that my apple can be as small as 2.60 , the size of the apple will start from 2.60 to any value that I have not checked, I am not interested because more the bigger, it is better. So, this is my lower value. So, this is

larger than the required size of 2.5. And if you see if you remember, if you do not remember I request you to please go back to the tutorial there in that for the same problem same all the data were same there.

We have checked for hypothesis $\mu = 2.5$ μ greater than 2.5 and we have rejected the null hypothesis $\mu = 2.5$. So, here the confidence interval also says so, $\mu = 2.5$ it is definitely it would have been rejected because my lower bound is only 2.6. So, $\mu = 2.5$ will definitely be rejected. So, this is the larger than required value of 2.5. Again, agreeing with the results of the hypothesis tests that we have done in the last tutorial.

(Refer Slide Time: 21:35)

Problem-8.5

Reframing the Scenario Presented in Case Study 2

Case Study 2

A medicine production company packages medicine in a tube of 8 ml. In maintaining the control of the volume of medicine in tubes, they use a machine. To monitor this control a sample of 16 tubes is taken from the production line at random time interval and their contents are measured precisely.

8.08	7.71	7.89	7.72
8.00	7.90	7.77	7.81
8.33	7.67	7.79	7.79
7.94	7.84	8.17	7.87

Monalisa Sarma
IIT KHARAGPUR

So, again, I am coming back to the case study to which I have used in my first lecture of statistical inference, and here I am reframing it a bit how I am reframing it here like so, in medicine production company packages medicine a tube of 8ml in maintaining the control of the volume of medicine in tubes they use a machine to monitor this control a sample of 16 tube is taken from the production line at random time interval and their contents are measured besides these are the volume of the different volume of the samples that we have taken.

(Refer Slide Time: 22:15)

Problem-8.5

The Scenario Presented in Case Study 2				Reframing the Scenario Presented in Case Study 2	
<p>A medicine production company packages medicine in a tube of 8 ml. In maintaining the control of the volume of medicine in tubes, they use a machine. To monitor this control a sample of 16 tubes is taken from the production line at random time interval and their contents are measured precisely.</p>				<p>T 8.5: Suppose the volume of medicine in at least 95% of the tube is required to be within 0.2 ml of the mean. Test if the filling process is in control using the given data. Also estimate the confidence interval of σ^2.</p> 	
8.08	7.71	7.89	7.72		
8.00	7.90	7.77	7.81		
8.33	7.67	7.79	7.79		
7.94	7.84	8.17	7.87		

Monalisa Sarma
IIT KHARAGPUR

Now, this was my scenario which was presented my previous lecture, I have used this as case study 2 have used in many of my discussion the same example now I am changing a bit here, what is my change I am reframing it suppose the volume of medicine in at least 95% of the tube be within 0.2ml of the mean volume should be within 0.2ml of the mean. So, whatever the mean volume, it should be within 0.2ml of the mean that means, if this is the mean.

So, 0.2ml means 0.1 is this side 0.1 is this side 0.2ml of the mean this is 0.1ml this is 0.1, this side is 0.1 this side is 0.1 that is what we want to check is it the filling processes in control using the given data this is what our we want our variance to be I mean our standard division it is when it is given 0.1 it is not variance, when it is given within 0.2ml of the mean I think you can understand it is not talking about a variance it is talking about the standard deviation is not it? So, how much it is changing from the mean it is changing 0.1.

So, we need to check whether our volume is within 0.2ml of the mean that means maximum we can what we can bear is a standard deviation of 0.1 that we have to check. So, if the standard deviation is 0.1, so, whatever the variance will be 0.1^2 .

(Refer Slide Time: 23:55)

Problem-8.5 : Solution

The hypothesis are:

$$H_0: \sigma^2 = 0.01 \\ H_1: \sigma^2 > 0.01.$$

Consider $\alpha = 0.05$

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}, \quad S^2 = 0.03174, \quad n = 16 \\ \chi^2 = \frac{0.4761}{0.01} = 47.61$$

Rejection region will be in the upper-tail.

Rejection region for 15 degrees of freedom & significance level 0.05, is

$$\chi^2 > 24.996$$

Considering the test statistics the null hypothesis is rejected.

T 8.5 : A medicine production company packages medicine in a tube of 8 ml. In maintaining the control of the volume of medicine in tubes, they use a machine. To monitor this control a sample of 16 tubes is taken from the production line at random time interval and their contents are measured precisely. Suppose the volume of medicine in at least 95% of the tube is required to be within 0.2 ml of the mean. Test if the filling process is in control using the given data. Also estimate the confidence interval of σ^2 .



Monalisa Sarma

IIT KHARAGPUR



So, what will be the hypothesis we have achieved σ^2 as a variance is equals to 0.01 here again now, alternate hypothesis we will check σ^2 not equal to 0.01 or greater than I am claiming that we will not check for not equal we will check for greater than why? Because when it is given within 0.1 of the thing so, anything which is less is covered there and null hypothesis.

So, what we have to check whether it is greater than 0.01 within 0.1 means anything it is less than between that will be here only is not it? Null hypothesis so, what we have to check whether it is greater than 0.01. So, it is again a 1 tailed hypothesis that is where σ^2 we have to check for 0.01 and our α is 0.05. So, what is the statistics for here we will be using because we are in have to infer about the population will be using chi square distribution.

The statistics that is used is $n - 1 S^2 / \sigma^2$ as far as σ^2 from the data that is given we found as this is their square value, our sample size is 16 we got our chi square value 47.619. Now, we will have to first we should have calculated a rejection region according to hypothesis testing this rejection region should have been calculated here and before calculating the statistics rejection region first we calculate a rejection region; then we try to find out a statistic here anyway you understand the thing that is okay.

So, the how rejection region see now when we are interested in finding that it is better than 0.01. So, what will be where we will be our rejection region? It will be in the lower tail or in the upper tail? This is the lower tail this is the upper tail this is $\chi^2 \alpha / 2$ this is $\chi^2 1 - \alpha / 2$ if it is two tailed, if it is single tail, first let me delete this since we are interested in here single tail

is something of this sort, but it is not a normal distribution chi square is very much a skewed distribution.

So, we will be interested in the upper tail or lower tail, this is the upper tail this is the lower tail. So, when we are interested in finding out than greater than 0.01. Our detection our null hypothesis is less than or equal to so, if we get a value greater that means, we are rejecting the null hypothesis is not it? So, it will be our rejection result it will be the upper tail. So, basically we have to find out what is the rejection region taking the significance level of 0.05.

Because it is 1 tailed we will not do $\alpha / 2$, but it will be 0.05 what is the chi square value for the $\alpha = 0.05$ with degrees of freedom how much is the degrees of freedom 15 for 15 degrees of freedom and $\alpha = 0.05$ we will find out what is the chi square value that is the rejection region. So, if rejection region if this value is greater than the rejection value the we reject the null hypothesis. So, what is the rejection region?

Rejection region is 24.996 rejection region for 15 degrees of freedom significance level of 0.5 on the chi square table, if you see the chi square table, this table is available in all standard textbook. Even if you do not consult a standard textbook you just Google in the Google is just write chi square table z table whatever it is you will get the table. So, this rejection region is for value greater into 24.96 what is our value?

Our value is 47.61 so, definitely reject the null hypothesis. So, considering the test statistic the null hypothesis is rejected. Now null hypothesis is rejected that means my σ^2 square value is greater than 0.01. Now greater means how much data let us find it out.

(Refer Slide Time: 28:20)

Problem-8.5 : Solution

Calculating confidence interval of population variance:

Since the hypothesis test is one-tailed, we need to construct a corresponding one-sided interval.

In this case, we want the lower confidence limit.

$$\text{Lower limit} = \frac{(n-1)s^2}{\chi_{\alpha/2}^2} = \frac{0.4761}{24.996} = 0.0190$$

Therefore, the lower confidence limit of the standard deviation = $\sqrt{0.0190} = 0.138$

⇒ We are 95% confident that the true standard deviation is at least 0.138.

⇒ Result of the Confidence Interval agrees with the result of Hypothesis Testing.

T 8.5 : A medicine production company packages medicine in a tube of 8 ml. In maintaining the control of the volume of medicine in tubes, they use a machine. To monitor this control a sample of 16 tubes is taken from the production line at random time interval and their contents are measured precisely. Suppose the volume of medicine in at least 95% of the tube is required to be within 0.2 ml of the mean. Test if the filling process is in control using the given data. Also estimate the confidence interval of σ^2 .



Monalisa Sarma
IIT KHARAGPUR

40



How will you find out? We found that our σ^2 value is greater than 0.01. Now, actually how much value is σ so, it is better understood is greater than 0.01. But how much is that so, we will find the cause confidence interval. So, this is how we can find out from here again here which one confidence interval we will find it is greater than 0.01. And I accepted the hypothesis that it is greater than 0.01.

So, when I want to find out the confidence interval, so what confidence interval I will find I will find a lower confidence interval an upper confidence interval, definitely I will find a lower confidence interval greater means it can be any value, but it will start from what value I am interested in the lower value always remember when we are testing for an alternate hypothesis which is greater than my confidence interval will be always the lower.

And the rejection region will be to the right and my confidence interval will be the lower confidence intervals. If I am checking for something which is less than then my rejection region will be towards the left that means the upper tail and my confidence interval I will check for my upper confidence interval. And if it is not equal to definitely I will look for both lower and upper and my rejection region will be both left and right both the lower tail and the upper tail.

So, here it is greater so my confidence interval will be the upper one or sorry greater means my confidence interval will be the lower confidence limit. So, this is the formula for lower confidence in limit we have seen $n - 1 S^2 / \chi(\alpha/2)$. So, here $\alpha/2$ means we will not consider

$1.5 / 2$ we will consider because it is a single tail always $\alpha / 2$ becomes α . So, it is $\chi(\alpha)$ if α we have seen it is 24.996.

So, this is the value this is my lower limit. So, my variance is at least 0.01 variants in this tubes it at least 0.01. It is it can be much more than done but it is at least 0.01 that means the system really needs some change. So, if this variance is 0.01 say this question has asked for standard deviation, standard deviation should be within 0.2ml. So, I will have to give the accordingly I will have to find out the answer standard deviation only if this is my variance so what is my standard deviation?

It is 0.138 so it should be within 0.2 is not it? So that means my standard deviation should be 0.1 but I got standard deviation as 0.138 greater than that. So, you why the null hypothesis is rejected my lower interval is only greater than 0.1 that is 0.138 so results of confidence interval agree with the results of hypothesis testing.

(Refer Slide Time: 31:29)



So, with this I end this lecture here, these are the references and thank you guys.

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

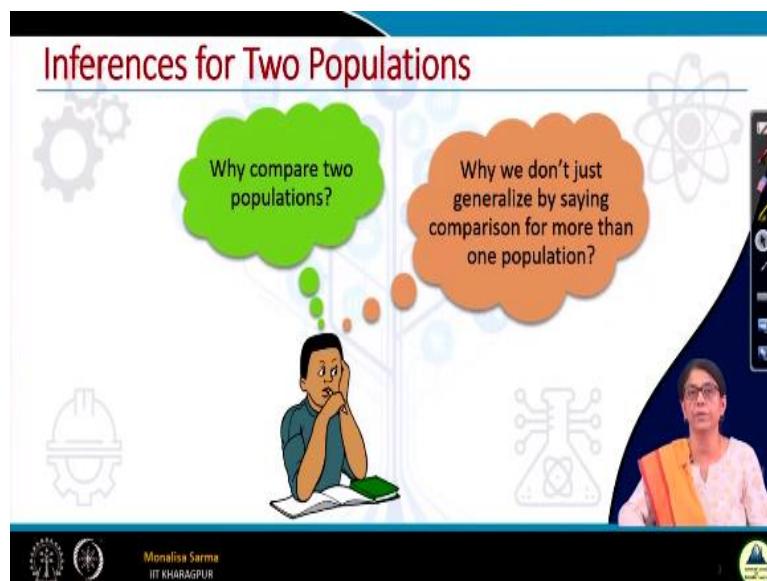
Lecture – 29
Statistical Inference (Part 6)

(Refer Slide Time: 00:29)



Hi guys, welcome again to this talk on statistical inference. So, in this lecture today we will talk on statistical inference for two populations.

(Refer Slide Time: 00:29)



So, when I am talking of inferences for two populations, the first question that may come to anyone's mind is that okay, we have seen to find out how to find out a statistical inferences for one population, that statistical inferences, maybe we may want to try to predict the mean of a population or the variance of a populations. So, we have seen that or maybe the proportion of populations. So, these are the 3 different things we have seen under different conditions.

So, now, next, when we have talked of one population, our next discussion definitely should have been for population inferences for more than one population. So, why suddenly, I thought that we should discuss on inference for two populations that means after that we will be discussed for inference on 3 population then again, after that inference for 4 population is it like that? So, it is definitely not bad. So, the next question that often comes to mind is that, why we do not just generalize by saying comparison for more than one population, is not it?

So, since as I told you, after 2 definitely will not go for 3 or 4, we will go for directly more than 2. So, then why just talk about 2 after one immediately, we can just generalize talking about comprehend for more than one populations.

(Refer Slide Time: 02:00)

The slide has a title 'Inferences for Two Populations' at the top. Below it, a green box states 'Many interesting applications involve only two populations:'. There are four main examples with corresponding images:

- A pink female silhouette and a blue male silhouette with the text 'Any comparisons involving differences between the two genders'.
- Two prescription bottles labeled 'test drug' and 'placebo' with the text 'Comparing a drug with a placebo'.
- Two images of a street, one labeled 'BEFORE' and one labeled 'AFTER' with the text 'Comparing before and after some event etc.'
- A person holding two containers with the text 'Comparing before and after some event etc.'

At the bottom left is the NPTEL logo, and at the bottom right is the name 'Monalisa Sarma IIT KHARAGPUR'.

So, now, the thing is that why we talk of inference as from two populations, there are many reasons the some of the important main reasons are many interesting applications involving only two populations. Like when we try to compare populations, there are many such interesting

applications where there are only 2 different distinct type of populations, like when we say we want to compare between the difference between 2 genders, there are only two populations, then again, we want to compare a drug with a placebo, you know, what is a placebo?

Of course, everyone will know it, I do not have to explain that. So, when to compare a drug with a placebo then we want to compare before and after some events. So, these are the typical case where we need inferences for two population means basically, we want to compare to two populations, like what was the state before some even occurred, says the earthquake what is the state after an earthquake has occurred? So, it is two populations. So, there is no 3 in that it is two populations.

(Refer Slide Time: 03:02)

The slide has a blue header with the title 'Inferences for Two Populations'. Below the title are two thought bubbles: one green bubble says 'Why compare two populations?' and an orange bubble says 'Why we don't just generalize by saying comparison for more than one population?'. A cartoon illustration of a person with their hand to their chin, looking thoughtful, is positioned below the bubbles. To the right, there is a video frame showing a woman speaking. On the left side of the slide, there is a logo for IIT Kharagpur and the name 'Monalisa Sarma'.

Some of the concepts underlying comparing several populations are more easily introduced for the two-population case.

The comparison of two populations results in a single easily understood statistic: such as, the difference between sample means. Such a simple statistic is not available for comparing more than two populations.

So, some of the concepts underlines comparing several populations, which we can very easily understand if we discuss about the two population case, some concept if we directly go to more than one population, it might be a bit difficult to understand. So, it is very easily introduced, if we just first repeat, talk of inferences for two populations. Of course, there are reason behind reasons are that why we compare two populations because as we have seen, there are many such cases where there are only two population only like we have seen the different cases.

Along with that, it is always in before going trading a very complex part it is always easy, it is always better to first know some of the concepts then go to the more complex part. So, that easier

part is first; understand the concept of two populations, then relevant go for more than one population are no more than two populations. The combination of two populations it results in some simple single easily understood static, easily understood statistics.

So, from a comparison of two populations, when we try to compare 2 different populations, we get some easily understood statistics, their statistics like difference of mean of two populations difference or variance of two populations. So, is not it? We while discussing sampling distribution; we have seen difference of means is not it? A difference of means difference of variances also. So, these are some very easily understood concepts. Such a simple statistic is not available for comparing when we try to compare more than two populations.

See when we try to compare more than two populations like we are trying to compare 3 populations, we cannot say difference of 2 or 3 population difference of 3 population they will not be a single value is not it? So, if you try to compare the variance of the 3 population, it will not be single value.

(Refer Slide Time: 04:51)

The slide has a blue header bar with the title "Comparison of Populations". Below the header, there are two blue boxes containing text:

1. The populations are actually different.
2. The populations are a result of an experiment.

The background features a stylized tree with various icons (gears, a lightbulb, a brain, a flask) growing from its branches. On the right side, there is a video frame showing a woman with glasses and a yellow sash, likely the speaker. The bottom of the slide includes the IIT Kharagpur logo and the name "Monalisa Sarma IIT KHARAGPUR".

Again, talking about populations, the populations are also sometimes you see the populations are also very different. Sometimes there populations means type of there are 2 different types of populations that I want to say that some populations are actually different like when you talk of

the population of 2 different genders, population of a drug and a placebo, these are actually different populations.

(Refer Slide Time: 05:16)

The slide has a title 'Comparison of Populations' at the top. Below it, a blue box contains the text '1. The populations are actually different.' An orange box to the right contains the text: 'Example: Male and female students -- a study involving separate populations is in general known as observational study.' To the right of the text are two groups of green stick figures: one group of six female figures in a row, and another group of five male figures in a row. In the bottom right corner of the slide, there is a video frame showing a woman speaking, identified as Monalisa Sarma from IIT Kharagpur. The slide also features a logo for NESTI and the text 'Monalisa Sarma IIT KHARAGPUR'.

Like as I see male and female students is study involving separate population in general is known as observational study this type of population when you study on this type of population and the populations are actually different, and we call this study as observational study and the populations are actually different when we are trying to compare the drug and placebo. So, these populations 2 these populations drug means a particular drug and a placebo. So, these two populations are actually different.

So, when we try to compare we are just are trying to observe the difference between 2. So, it is whatever it is already existing, we are not doing anything to it, we are just trying to compare these 2. And so this type of study is called an observational study.

(Refer Slide Time: 06:02)

Comparison of Populations

2. The populations are a result of an experiment.



- The different populations are usually referred to as "treatments" or "levels of a factor."
- This type of study was referred to as a designed experiment.



Now, sometimes the populations are the results of an experiment. Experiment like you can very well see the example that see the figure what they want to say, like again, I will give you one more example. Suppose say, we want to compare the effect of 2 fertilizer on a land, we want compare the effect of 2 different fertilizers when we want to find out the effect of the 2 different fertilizers on a land. So, what happens if both the land if we take in a different places then we may not get the same effect then what we will do?

We will try to in the same plot of land we maybe we divide the plot of land into 2 parts diagonally maybe and use the what to said use the different fertilizer on a different part and a diagonally upper part we use fertilizer A on a diagonal our lower part we use fertilizer B that way we have created 2 different populations, initially it was the same populations, but we have created 2 different populations, this sort of things have all the populations are add the result of an experiment.

The single homogenous population has been divided into 2 portions, where each has been subjected to some sort of modification, single homogeneous population kind of figure also what you can see in the picture, what you can see? It is a symbol same place where in some place you have a plant at some different type of Lily and another piece of plant a different type of Lily plant. So, simple population, but you are subjecting to some sort of modification. So, this type of populations we have it is because of the result of an experiment.

So, this type of study is referred to as designed experiment, designer experiment is a big topic actually. But anyway, we are not covering this it is totally out of the scope of this course.

(Refer Slide Time: 07:54)



So, now, again, first we saw why we compare two populations, instead of directly going to single population, we have seen it in different regions. Then we also saw what may be the 2 different types of populations. The different types of population, I am not saying to try to telling that we are comparing two populations. So, what are the 2 different populations? No, not that I am just trying to tell there are different types of population that also we saw, there are 2 different types of population.

One study we call it observational study, and another we call it as a designed experiment or experimental study as well. So, now, there are again 2 different ways of collecting data. So, there are 2 different methods for collecting data or designing an experiment for comparing 2 different populations, why can you see the point here different methods, I expected this pain does not work in the first go. Second, it needs a second thing.

So, 2 different methods for collecting data so, all designing an experiment for comparing two populations, where in one case will be collecting data and the first case of population where we are doing observational study in that basically, we will tell that we are collecting data for

different study. And other case, when the same populations, we are modifying some way by to compare the 2 different basically we call it factors that come later. So, that is we are designing an experiment this is we are designing an experiment right in the same plot of land.

What we are doing we are diagonally separating the same plot of land in the upper diagonal part a portion of the land we are using fertilizer A and a lower diagonal portion we are using fertilizer B. So, we are designing an experiment. So, this is one way of collecting data. So that is the second question that designing an experiment for comparing two populations. The first one is called independent samples, the other one is called dependent or paired samples independent samples.

So, independent sample it is very easy to understand you can easily understand like suppose we are trying to compare some characteristics of a boy population and a girl population. So, we have taken some sample from boy we have taken from some sample from the girl populations. So, that is the totally independent sample we are independently we are collecting the samples. Now, the example what I have given let me give one more example suppose what am I want to test the efficacy of 2 different migraine medicines, so, one is the blue pill, another is the red pill.

2 different one, the migraine medicine is a blue color another is a red color blue pill and the red pill. There is a picture also is not it? So, now, there are 2 different ways of comparing these 2 to find out the efficacy of discrimination. One way is that I will take a sample from the populations and I will take a sample means people who suffers from migraine definitely for them only we can use this medicine on them only.

So, among the population of migraine sufferers, we have taken one sample and we have given them blue pill and we have taken another sample and we have given that red pill. This way if I collect the sample, this is one way of collecting the sample. Another way of collecting the sample is that what I have done, so, I have taken from this whole population of migraine sufferers I have taken one sample and I have given randomly I have picked from this suppose I have collected a sample of 2 n people.

And from this randomly I have picked n people and I told them on the first onset of migraine you take blue pill and the second onset, you take the red pill, another and the another and I told them on the first onset take the red pill and the second onset of migraine take the blue pill. Again I am repeating I have taken a sample of size $2n$ from this $2n$ and I randomly picked n people and from for this n people I told you on the first onset of migraine when you get a headache severe they first take the blue pill.

And then when the second time you get a migraine you take the red pill again for the second we have collected $2n$ out of $2n$ we have for n we have given this then the for the remaining n for what we what I told is there on the first onset you take red on the second onset of your headache you take the blue pill then I try to see the effect what I get, these are 2 different way of collecting sample one way what I do from the whole populations, I have collected a sample and I have given the blue pill another sample I have collected I have given the red pill.

This is one way of collecting samples this way is called independent samples. These are independent samples they all are migraine part is a whole population is a migraine suffer people and I have taken independent samples other way is that the same subset of people the same sample but I have given the pill in a different order. So, this way of collecting sample is called dependent or paired sample we will see what is advantage of that we will see visually. Now, let us not talk about that. So, this way is called dependent or paired sample.

Why dependent because the same sample we are using for both the drugs is not it? Same all the person will be taking the blue pill as well as the red pill. But in the other case, only one sample one set of people will be taking the blue pill other set of people will be taking the red pill, so it is totally independent, but here it is dependent. So, this type of collecting data it is called dependent or paired samples.

(Refer Slide Time: 13:52)

Independent Samples

- For two populations we define the difference between the two means as

$$d_0 = \mu_1 - \mu_2$$
- The null hypothesis can be stated as

$$H_0: \mu_1 - \mu_2 = d_0$$
- The alternative hypothesis can be two sided or one sided.
- A sample of size n_1 is randomly selected from the first population and a sample of size n_2 is independently drawn from the second.
- The sampling distribution of the difference between the two sample means $(\bar{x}_1 - \bar{x}_2)$ needs to be considered

Monalisa Sarma
IIT KHARAGPUR

So now, first we will see for independent samples and we will try to infer on the difference between mean, try to compare that mean of 2 different population how we will compare the mean of 2 parameters, we will find the difference definitely that is the only way of comparing 2 means is not it? Whether this one is better or less or equal so, we; are trying to find out the difference of these 2 means. So, for two populations, we define the difference between 2 mean this is how suppose we define the difference between 2 mean.

For the mean of first population this is the mean of second populations and this $\mu_1 - \mu_2 = d_0$. So now, so, my null hypothesis I can state is as $\mu_1 - \mu_2 = d_0$ and my alternate hypothesis if I want those, if suppose if I want to check I want to find out if μ_1 is greater than $\mu_1 - \mu_2$ will be greater than d_0 . If I want to find out μ_2 greater than $\mu_1 - \mu_2$ is less than d_0 , according to the requirement, whichever I want to find out that will be met an alternate hypothesis.

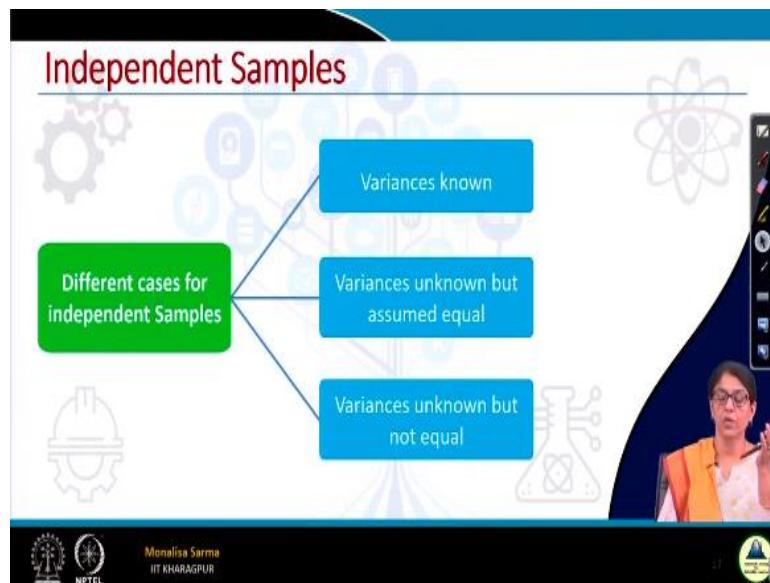
All my condition is that I want to check whether my both the population means are equal then my null hypothesis I can also write it as $\mu_1 - \mu_2 = d_0$ or I can also write as $\mu_1 = \mu_2$, that is my null hypothesis. Because then my alternate hypothesis whatever I want to check whether I want to check if μ_1 is greater than μ_1 is greater than μ_2 will be the null hypothesis if I want to check μ_1 and μ_2 are both are not equal, that is what I want to check.

My alternate hypothesis will be μ_1 not equals to μ_2 . Accordingly, we will frame the hypothesis. So, the alternate hypothesis as I mentioned can be 2 sided or 1 sided. So, this sample of size n_1 is randomly selected from the first population and a sample of size n_2 is independently selected and drawn from the second both are independently strong, 2 different populations, we are randomly we are picking from this and then we are picking from that both are independent of each other.

It is not that here we are picking something that is because of that I am picking something here it has no dependencies there. So, now, remember for finding out the inference always we have to find out the sampling distribution of a statistic, is not it? When we want to infer about a difference of 2 means what will be my sampling distribution of what statistics; value mean sample means so, my 2 sample mean will be $\bar{x}_1 - \bar{x}_2$.

So, I will find out the sampling distribution of $\bar{x}_1 - \bar{x}_2$ that is my test statistics, my test statistics is $\bar{x}_1 - \bar{x}_2$. So, I will find out the sampling distribution of this.

(Refer Slide Time: 16:43)



So, now, there are this as I told you, we are discussing this for independent samples, when the samples are independent, we are trying to compare two populations where the populations are totally different, it is an observational study basically. So, there are different cases of independent samples now independent sample also there are different case. First is variance is

known, when the variance of the population that means variance of the parent populations are known that is one case.

The second case is variance unknown, but we can assume it to be equal variance of the 2 parent population, we do not know the variance of the populations, but okay fine, we can assume it to be equal. Another third cases variance unknown but not equal. So, under this 3 condition, we will see how we can infer the population mean.

(Refer Slide Time: 17:30)

So, to make inference on the difference of 2 means, as I told you, this is my test statistics. So, now, this test statistics $\bar{X}_1 - \bar{X}_2$ bar what it will have $\bar{X}_1 - \bar{X}_2$ bar, it will have a normal distribution. So, when it is a question of normal distribution, I will have to describe 2 parameters, what are the 2 parameters for normal distribution mean and a variance. So, \bar{X}_1 bar - \bar{X}_2 bar will have a normal distribution and what will be the mean of \bar{X}_1 bar mean of that distribution that is the sampling distribution.

Mean of the distribution is $\mu_1 - \mu_2$ is not it? And variance is σ_1^2/n_1 σ_2^2/n_2 is the standard deviation of the first population and σ_2^2/n_2 is the standard deviation of the second population. So, in case of single population what is my variance? Variance of the sampling distribution is σ^2/n is not it? So, this is my σ^2/n side of the population σ^2

by $n/2$. So, this statistic has a normal distribution these are a parameter this is the mean this is the variance.

So, standard deviation is \sqrt{n} of this. So, now, we will like what we do for single population similarly, we can find out the z value corresponding to this. So, how do we find out for a single population how do we find out z value remember $X_{\bar{}} - \mu$ by σ/\sqrt{n} this is my z value is not it? Similarly here it is where $X_{\bar{}}$ is the mean of the sample. So, now here what is the mean of my sample? Mean of my sample is $X_1 - X_2$ I am trying to find out the difference of means, that is why my mean of the sample is $X_1 - X_2$.

And what is my μ ? M is the population mean is not it? So, here what I have hypothesis? I have hypothesis that $X_1 - X_2 = d$ is not it? That means the difference of 2 is the d , d maybe 0 or -1, +1 whatever it is or maybe any value. So, this is what I have hypothesis? I have hypothesis the difference between the two populations is d or if I have hypothesis that can we say that the two populations mean are same. In that case d will be 0 here. So, this is the value of σ/\sqrt{n} .

So, accordingly we will find out the value then we will whatever their significance level is given based on the significance level, we will find out the critical region. So, if the z value falls within the critical region then, we do not reject the null hypothesis if z value falls in the critical region, then we reject the null hypothesis. And even if you are interested in finding out a confidence interval data, so, we can find out a confidence interval same method nothing else.

(Refer Slide Time: 20:30)

Variance Known

To make inference on the difference of two population mean

- The test statistic is $\bar{X}_1 - \bar{X}_2$
- The statistic has a normal distribution with mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
- The statistic $Z = \frac{\bar{X}_1 - \bar{X}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ has the standard normal distribution
- The confidence interval on the difference $\mu_1 - \mu_2$ is

$$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$



Monalisa Sarma
IIT KHARAGPUR



So, the confidence interval is $\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \sigma_{\bar{X}_1 - \bar{X}_2}$. So, this is σ by \sqrt{n} , same thing whatever we have done for single populations, same thing we have done for two populations, same method updating same there we will just use \bar{X} here we are trying to compare 2 different populations. So, it is $\bar{x}_1 - \bar{x}_2$.

(Refer Slide Time: 20:57)

Example

A random sample of size $n_1 = 25$, taken from a normal population with a standard deviation $\sigma_1 = 5.2$, has a mean $x_1 = 81$. A second random sample of size $n_2 = 36$, taken from a different normal population with a standard deviation $\sigma_2 = 3.4$, has a mean $x_2 = 76$. Test the hypothesis that $\mu_1 = \mu_2$ against the alternative, $\mu_1 \neq \mu_2$. Quote a p-value in your conclusion.



Monalisa Sarma
IIT KHARAGPUR



So, this is a small example, a random sample of size $n_1 = 25$ taken from a normal population with a standard deviation $\sigma_1 = 5.2$ sample size is given standard deviation of the population is given and the mean of the sample is given 81. Similarly, the second random sample size is given normal with standard deviation of the second population is given mean of the second population

is given test the hypothesis that $\mu_1 = \mu_2$ against the alternative μ_1 not equals to μ_2 quote a p value in your conclusion. So, we just have to give a p value.

(Refer Slide Time: 21:40)

Example: Solution

The two hypothesis are:

$$H_0: \mu_1 = \mu_2$$

$\mu_1 - \mu_2 = d$

$\mu_1 - \mu_2 = 0$

$\mu_1 = \mu_2$

A random sample of size $n_1 = 25$, taken from a normal population with a standard deviation $\sigma_1 = 5.2$, has a mean $x_1 = 81$. A second random sample of size $n_2 = 36$, taken from a different normal population with a standard deviation $\sigma_2 = 3.4$, has a mean $x_2 = 76$. Test the hypothesis that $\mu_1 = \mu_2$ against the alternative, $\mu_1 \neq \mu_2$. Quote a p-value in your conclusion.

Monalisa Sarma
IIT KHARAGPUR

So, what will be my hypothesis? Hypothesis is William that means, initially what was my hypothesis I have seen $\mu_1 - \mu_2 = d$ here both are test the hypothesis that $\mu_1 = \mu_2$ that means, what this g is nothing but 0 is not it $\mu_1 - \mu_2 = 0$ that means, I can write $\mu_1 = \mu_2$. So, my null hypothesis is $\mu_1 = \mu_2$.

(Refer Slide Time: 22:06)

Example: Solution

The two hypothesis are:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Given, the variances are known.

Therefore, the sample statistic

$$Z = \frac{81 - 76}{\sqrt{\frac{(5.2)^2}{25} + \frac{(3.5)^2}{36}}} = 4.22$$

$Z = \frac{81 - 76}{\sqrt{\frac{(5.2)^2}{25} + \frac{(3.5)^2}{36}}}$

$+ Z \neq Z$

A random sample of size $n_1 = 25$, taken from a normal population with a standard deviation $\sigma_1 = 5.2$, has a mean $x_1 = 81$. A second random sample of size $n_2 = 36$, taken from a different normal population with a standard deviation $\sigma_2 = 3.4$, has a mean $x_2 = 76$. Test the hypothesis that $\mu_1 = \mu_2$ against the alternative, $\mu_1 \neq \mu_2$. Quote a p-value in your conclusion.

Monalisa Sarma
IIT KHARAGPUR

And against the alternative μ_1 is not equals to μ_2 . So, it is a 2 sided it is a 2 tailed hypothesis test given the variances are known, we already it is given. So, therefore, the z statistics so, what

is this $z = \bar{X}_1 - \bar{X}_2$ what is this minus d divided by σ by \sqrt{n} whatever it is σ by \sqrt{n} is this value and d what is d? d is 0 here. So, it is $\bar{X}_1 - \bar{X}_2$ is 81 - 76 this is 81 this is 76. So, we got a z value 4.22.

Now, if you see when we have to give the p value remember when it is 2 tail how we get the p value come from corresponding to the z value we find out the probability suppose probability for z value 4.22 suppose my probability is say x some probability say x whatever maybe, then my p value will be $x + x$ because it is 2 side 2 tail if it is single tail, if the property corresponding to $z = 4.22$ is x then my p value is simply x. So, now from the z table, we will have to find out what is the probability corresponding to z value 4.22.

(Refer Slide Time: 23:25)

Example: Solution

The two hypothesis are:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Given, the variances are known.

Therefore, the sample statistic

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{81 - 76}{\sqrt{\frac{(5.2)^2}{25} + \frac{(3.5)^2}{36}}} = 4.22$$

The p-value corresponding to $Z = 4.22$ is almost 0.



A random sample of size $n_1 = 25$, taken from a normal population with a standard deviation $\sigma_1 = 5.2$, has a mean $\bar{x}_1 = 81$. A second random sample of size $n_2 = 36$, taken from a different normal population with a standard deviation $\sigma_2 = 3.5$, has a mean $\bar{x}_2 = 76$. Test the hypothesis that $\mu_1 = \mu_2$ against the alternative, $\mu_1 \neq \mu_2$. Quote a p-value in your conclusion.



 Monalisa Sarma
IIT KHARAGPUR



Now, corresponding to 4.22 in the table, it will see there is no value for that actually it is so less less less. Let us see, for z value around 3 point something only this probably becomes point 0000 something so, for 4.22 it is almost 0, p value is almost 0, that means p value is almost 0 meaning what? If you can means z value have 4.22 is somewhere at this point 4.22 somewhere you have means it is almost 0 definitely false in a critical region, is not it?

Critical region definitely there has to be some area in the critical region. And almost 0 will be definitely in a critical region so null hypothesis is rejected.

(Refer Slide Time: 24:10)

Example: Solution

The two hypothesis are:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Given, the variances are known.

Therefore, the sample statistic

$$Z = \frac{81 - 76}{\sqrt{\frac{(5.2)^2}{25} + \frac{(3.5)^2}{36}}} = 4.22$$

The p-value corresponding to $Z = 4.22$ is almost 0.

Hence we can conclude, H_0 is rejected.

Infact, from the z-value we can say $\mu_1 > \mu_2$

A random sample of size $n_1 = 25$, taken from a normal population with a standard deviation $\sigma_1 = 5.2$, has a mean $x_1 = 81$. A second random sample of size $n_2 = 36$, taken from a different normal population with a standard deviation $\sigma_2 = 3.4$, has a mean $x_2 = 76$. Test the hypothesis that $\mu_1 = \mu_2$ against the alternative, $\mu_1 \neq \mu_2$. Quote a p-value in your conclusion.



Monalisa Serma
IIT KHARAGPUR

So, in fact on the z value, we can say a null hypothesis is rejected that means and a null hypothesis is rejected alternate hypothesis is accepted μ_1 is not equals to μ_2 in fact the z value only we can say $z = 4.22$ when we are getting $z = 4.22$. So, it is why we are getting for such a big value, that means here when we will get a bit smaller value than 4.22 either when this value is big, or this value is big, is not it?

Now, if I need to get a smaller value here, so I need to take my value sorry, if for to get a smaller value here either this value should be big or these value should be small $81 - 76$ this value should be small then I will be getting a what to say smaller value here. So, corresponding to this when I got this 4.22 what does it indicate that μ_1 is not equals to μ_2 that is of course true, but then it is from here directly we can say this μ_1 is greater than μ_2 .

Because it is this here this value we are getting a very bigger value that is why we are getting this 4.22 from this only we can directly conclude that we do not have to do that it is we just have to tell that okay μ_1 not equals to null alternate hypothesis is accepted that is μ_1 not equals to μ_2 but from the z value you can estimate that μ_1 not equals to μ_2 is that is fine, but, the relation between them is that μ_1 is greater than μ_2 from the z value we can see, if we get a very smallest z value or in the negative side, then we can take tell the reverse basically.

(Refer Slide Time: 26:00)

Variances Unknown

Can we use t- distribution using the two variance estimate s_1^2 and s_2^2 ?

What is the solution?

We need to assume that the two-population variances are equal and find an estimate of that variance.

NPTEL Monalisa Sarma IIT KHARAGPUR

Now, the second case variances are known. So, for single population case when a variance is unknown what we do remember when the variance is unknown, and we try to infer the meaning of a single population, then directly we use t distribution where in t distribution instead of σ we use S is not it? S that is the standard deviation of the sample because we can calculate the sample standard division is not it? So, since we have instead of σ we have S so, we could not use the z distribution rather we use a different distribution.

That is a t distribution which is very similar to normal but has a fatter tail. And what is the parameter t distribution has only one parameter that is the degree of freedom what is the degree of freedom? Degrees of freedom; is the sample size minus 1 that is the degrees of freedom. So, now, here also variance is not known, we are trying to compare 2 different population and a variance is not known. So, if the variance is not known like for a single population case we can very well use the $S 1^2$ instead of $\sigma 1^2 \sigma 2^2$.

We can very well use $S 1^2$ and $S 2^2$. So, can we really use t distribution using the 2 variance estimate $S 1^2 S 2^2$ that is that may be one question in your mind is not it? But, there is one problem to it what is that before coming to the solution what is the problem to it in a t distribution we have seen that we have only one degrees of freedom like an F distribution, remember, we have 2 degrees of freedom like in t distribution there is only one degrees of freedom, but here we are trying to compare 2 different populations 2 different.

So, we are taking 2 different samples the sample sizes may be different maybe same maybe different, but there are 2 different samples. So, that means 2 different samples. So, we will have 2 degrees of freedom, but t has just one degrees of freedom. So, then how can we use t distribution or we cannot use z distribution also then how we will compare the population mean of 2 different populations when the variance is unknown. So, what is the solution basically?

So, what we have to do is that one way that we need to assume that the two population variances are equal, and find an estimate of the variance. What we will do in that case, we will assume that, we are the two population's variances, are equal and from that we will try to find the estimate of that variance. But now, the question is why should we assume that a two population variances are equal? That is something very odd right why should we assume that the two population variances are equal it may not be equal?

Yes, it may not be equal it is true that I will come but usually when we are trying to compare 2 different populations, we will differ definitely never try to compare apples and oranges is not it? When we are trying to compare 2 different populations, these 2 different populations are very much similar that is only we are comparing is not it like apple and potato we will not compare these two populations are very similar. That is why we are comparing when we are comparing 2 similar type of population it is not very unnatural.

If we assume that the population variances are equal. But of course, it may not be equal that is that does not mean that two populations are similar, that means the variance will be equal it is not very, it is not always true that the population variances are equal. That will happen we will see again. Now for the time being let us, assume that we will assume that the two population's variances are equal. And we will find an estimate of that variance.

(Refer Slide Time: 29:25)

Pooled Variance Estimate

The estimate of a common variance from two independent samples is simply the weighted mean of the two individual variance estimates.

The weights being the degrees of freedom for each variance.

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

The pooled variance is now used in the t statistic, which has the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\frac{\bar{x} - \mu}{S/\sqrt{n}}$$



34

So, how will find an estimate of that variance? Very easy, we will just take the weighted mean. The estimate of a common variance from 2 independent samples is simply the weighted mean of the 2 individual variances estimate. Simply we will take the weighted mean variance of one sample whatever is the weight; weight is based on the sample size variance of the other sample it is again weighted based on a sample size divided by the whole sample size.

So, this is let me call it as a S_p^2 what is S_p^2 so $n_1 - 1$ s_1^2 s_1^2 squared is the variance of the first sample s_2^2 squares the variance of the other sample that is $n_2 - 1$ is that multiplying it by the sample size to give the weighted value and we are dividing it by this $n_1 - 1 + n_2 - 1$, this will give me this is called a pooled variance estimate. We are pulling 2 variants together and trying to find out an estimate, we are pulling 2 variances together 2 variants means parents of both a sample and we are trying to find out a common variance.

So, this is called as a pooled variance estimate S_p^2 . So, pooled variance now we can use this pooled variance estimate in the t distribution and is what will be this pooled variance estimate what will be the degrees of freedom for that t distribution only using t distribution we will have to have a degree of freedom so it is 0 degree of freedom will be $n_1 + n_2 - 2$, so, in the same t distribution formula remember $X \bar{ } - \mu$, $X \bar{ } - \mu$ whatever it is μ is the population mean then divided by $s_1 / S \sqrt{n}$, is not it?

So, said same as \bar{x}_1 bar - \bar{x}_2 bar - d_0 , but may be d_0 will be 0 or any value then this is the pooled variance estimator during the t formula what we use \bar{x} bar - μ / S / \sqrt{n} this is the formula for t distribution for single population now, instead of \bar{x} bar - μ instead of μ that means the population mean so, population mean we have what did that is the hypothesis value my hypothesis value is d_0 difference of 2 means is d_0 and S What is my S ? S is this, this whole value.

So, this whole value divided by the various populations sample size the here I am dividing by \sqrt{n} this is a sample size. So, here what is the $1/n_1 + 1/n_2$ it is very similar to the z distribution what we have used see here see here and z distribution $\sigma^2 / n_1 + \sigma^2 / n_2$ remember. So, similarly here we are using S_p^2 bringing it out so, $1/n_1 + 1/n_2$ under that.

So, this is how we will calculate the t value now, we will find out same like previous if it falls in a critical region we reject the null hypothesis if it does not fall in the critical region either we accept the null hypothesis or we will tell that we fail to reject the null hypothesis.

(Refer Slide Time: 32:49)

"Pooled" t test

To compare the mean of two different population with unknown variance but can be considered equal, the corresponding test is called the "pooled t test."

The test statistic used

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$$

It is called pooled t statistic.

Similarly the confidence interval on $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

using values from the t distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

NPTEL

Monalisa Sarma
IIT KHARAGPUR

So, this to compare the mean of 2 different populations is unknown variance but we can be considered equal the corresponding test statistics is called pooled t test. So, this test is not a simple t test, we do not call it a simple we call it a pooled t test, because we are trying to find out the pooled estimate of the 2 variance. It is also sometimes called t test but its actual name is

pooled t test does a different t test paired t test. So, the difference between 2 pair t tests and pool t test paired t tests I will come to that. So, this t test called pooled t test.

And this statistics the t value using this S_p we are calculating the test that is test t using the pooled variance that is S_p the statistics is called pooled t statistics. So, similarly, we can find a confidence interval as well same matter nothing else nothing no difference at all. So, what is the degrees of freedom will be $n_1 + n_2 - 2$. So, this is how we will find out confidence interval.

(Refer Slide Time: 33:52)

Example

An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 81 with a sample standard deviation of 5. Can we conclude at the 0.05 level of significance that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units? Assume the populations to be approximately normal with equal variances.

Monalisa Sarma
IIT KHARAGPUR

$n_1 = 12$
 $n_2 = 10$

So, one simple example the example problem looks bigger with lot many text in it, but it is a very simple example, if you go through it carefully you will find is a very simple example an experiment was performed to compare the abrasive wear of 2 different laminated metrics, we have to compare the abrasive wear of 2 different laminated materials. So, what we have taken we have taken 12 pieces of material 1. So, for sample size measurements $n_1 = 12$ and we test it by exposing each piece to machine measuring wear.

To a machine measuring were we are trying to measure to wear, is not it? Abrasive wear then 10 pieces of material 2 from that means my n_2 was 10 were similarly tested in each case the depth of the wear observed the sample of material one from the sample of material 1 we have taken 10 sample given average wear of 85 units average we have got from all the samples I got a certain

wear to try to find out the mean I got an average of 85 units with a sample standard deviation of 4.

While the sample of material 2 give an average of 81 with a sample standard deviation of 5 mind it the population standard deviation is not given can we conclude that 0.05 level of significance that I have received where a material 1 exceeds that of material 2 by more than 2 units. So, from the sample whatever value you get for selling the sample first sample sizes given mean of the sample is given then standard deviation of the sample is given. So, the sample is given.

Now, from this value can we conclude that with 0.05 level of significance. Significance level is by person with that significance level can I conclude that abrasive wear of material 1 exceeds that of material 2 by more than 2 units, Assume the population to be approximately normal with equal variance. So, we are assuming that the population is approximately normal and has equal variance because the variances are not given. So, we will have to come we are assuming that the variances are equal.

(Refer Slide Time: 36:01)

Example: Solution

Let,
 μ_1 = population means of the abrasive wear for material 1
 μ_2 = population means of the abrasive wear for material 2

The hypothesis,
 $H_0: \mu_1 - \mu_2 = 2$
 $H_1: \mu_1 - \mu_2 > 2$

Given,
 $\alpha = 0.05$
Variance of both the populations are equal

Critical/Rejection region corresponding to $\alpha = 0.05$
 $t > 1.725$ (one tailed)

An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 81 with a sample standard deviation of 5. Can we conclude at the 0.05 level of significance that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units? Assume the populations to be approximately normal with equal variances.

NPTEL
Monalisa Sarma
IIT KHARAGPUR

So, μ_1 is the population mean; these are the different think. So, now, what is us sorry I should have come here only can we that abrasive wear material 1 exceeds that of material 2 by more than 2 units can we conclude that it is it exceeds more than 2 units. If that is the case, then what will be my null hypothesis null hypothesis is this abrasive wear of both the unit different both the

units is 2 and we want to test what we have to test whether it is more than 2 that is greater than 2. So, this is my null hypothesis.

This is my alternate hypothesis with a one tailed test. So, α is equals to 0.5 means we will consider 0.5 only one tail only. So, will that means for confidence interval will not find out α by 2 and even if we have to find out the p value we will not sum it up twice we will just use once. So, α is sorry α is not 0.5 it is 0.05. This is 0.05 level it is given 5 person. It is variance of both the population are equal it is assumed so we will have to find out what is the corresponding full variance estimate.

That means we will have to find out the S_p^2 . Here also does so, based on α is equals to 0.05 If you see the t table and for finger what to say what is the degrees of freedom? Degrees of freedom will be $n_1 + n_2 - 2$ we will see that this is the value computed $n_1 + n_2 - 2$ this is the degrees of freedom. So, what is n_1 ? n_1 is 12, $12 + 10 - 2$. So, it is for degrees of freedom 20 if you see in a table for $\alpha = 0.05$ not 0.5 we will get the value 1.725 that means, if a t statistic value if it is greater than 1.725 then we will reject the null hypothesis.

(Refer Slide Time: 38:04)

Example: Solution

The value of the sample statistics is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad v = 20, \quad S_p = 4.478$$

$$\Rightarrow t = 1.04$$

Null hypothesis is not rejected.

Unable to conclude that the difference of the abrasive wear between the two materials is more than 2 units.

An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 81 with a sample standard deviation of 5. Can we conclude at the 0.05 level of significance that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units? Assume the populations to be approximately normal with equal variances.

Monalisa Sarma
IIT KHARAGPUR

So, S_p we will calculate because standard deviation of both the samples are given we can calculate the S_p . So, from here we will calculate the t value same formula everything same will calculate the t value t value we got 1.04. That means initially what is the rejection region greater

than 1.725. Now, we got 1.04 that means it is not in the rejection region. So, null hypothesis is not rejected we are unable to conclude that a difference of the 2 abrasive wear between the 2 materials is more than 2 units we are unable to conclude that it is more than 2 units.

This is one way of telling another way of telling us that no it is different is that it is equals to 2 unit that means we accepting the null hypothesis as I already mentioned, we are when we are rejecting the null hypothesis it is definitely we accepting the alternate hypothesis, but when we are not rejecting the null hypothesis there can be 2 things one is we are accepting the null hypothesis or we are telling that we are unable to reject the null hypothesis.

When we are telling we are unable to reject the null hypothesis means there is still scope for further experimentation to find out whether whatever result we got is correct or not. Because why we are trying we are testing it because some doubt has come to our mind is not it? That why only we are testing it. So, in this result, we found that our doubt is illogical, but then again if a probability p value is very less than we make do the experiment again.

(Refer Slide Time: 39:33)

Variances Unknown but Not Equal

How to handle the Variance Inequality ?

Variance inequality maybe handled by:

1. making "transformations" on the data
2. If both n_1 and n_2 are large (both over 30) we can assume a normal distribution
3. If either sample size is not large, and if the data come from approximately normally distributed populations, a reasonable (and conservative) approximation is to use the degrees of freedom for the smaller sample.

So, now how to handle the variance inequality? The trick is that all variance known, another is variance unknown, but can be assumed as equal, another is when the variance is unknown, but not equal. So, one way is by making transformation on the data transformation on the data means, like there are some cases when what happens when we are trying to compare 2 different

types of populations, the population in fact seen that when a mean is more accordingly the variance is also more.

So, maybe, when we try to compare the 2 different variance, maybe they are same only, but then this one sample is mean is only more the size of the things we are comparing suppose some 2 different forest we are trying to compare in one different forest there are some trees are a very small size and other different forest the trees are bigger size, then what happens maybe the variance in both the population may be same.

But what happens in the forest where the size of trees are more, that means it mean we will be more mean is more usually since the variance is also more here, the mean is small variance is also small. But in that case, if you so you can tell that there are two population variances different that may be wrong. So in this case, is what we can do, we can transform the data, maybe we will transform the population data of the where the; what to say trees are a bigger size.

And but we can do transformation means we can say we can do some log transformation, and then we can do the test then if we find a variance equal than equal, if not equal, they are not equal. That is one way. And another way is when both n_1 and n_2 are large over 30 we can assume a normal distribution as well. As I told you remember in the first one, we are talking t distribution for when the sample size is bigger t distribution can be estimated by normal distribution as well.

So, when the sample size is bigger, instead of t distribution, we can use a normal distribution normal distribution is pool, is not it? We do not have to worry about that. So, if the either sample size is not large, and if the data comes from approximately normally distributed population, a reasonable and conservative approximation is to use the degrees of freedom for the smaller sample. If the population is the data comes from approximately normal distribution only we know the population is almost normal only.

Then what happens even if the sample size is not large, then still we will use t distribution but in t distribution, we have to use only one degree of freedom. So, here what we can use we can use the degrees of freedom of the smaller sample this is also one way.

(Refer Slide Time: 42:07)

The slide has a dark blue header with the word 'CONCLUSION' in yellow capital letters. Below the header is a white section containing a bulleted list:

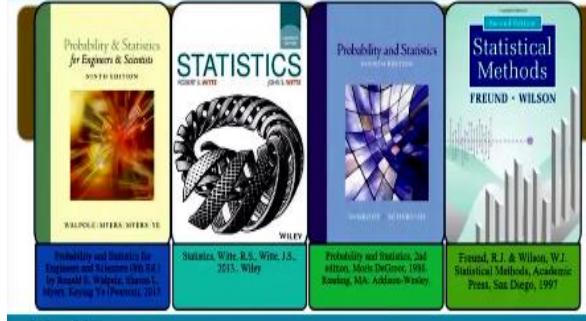
- ④ In this lecture we discussed why to learn specific techniques for comparing two populations.
- ④ We have also seen different methods for collecting data for comparison of two populations.
- ④ Next, we discussed methods for inferences of means for two populations considering the fact that population variance may be known or unknown.

On the right side of the slide, there is a video player showing a woman with glasses and a white shirt, gesturing with her hands while speaking. The video player has a play button and other control icons. At the bottom left of the slide, there is a logo and the text 'Monalisa Sarma IIT KHARAGPUR'. At the bottom right, there is a small circular logo with the number '15'.

So, in this lecture what we have seen we have learned specific techniques for comparing 2 different populations. We have also seen different methods for collecting data for comparison of two populations, we have seen 2 different methods one is independent sample one is dependent sample next we have discussed method for inference of means for two population considering the different fact the population variance may be known may be unknown what happens if it is unknown? Can we assume it equal if it is equal then what a case if it is not equal then what is the case?

(Refer Slide Time: 42:35)

REFERENCES



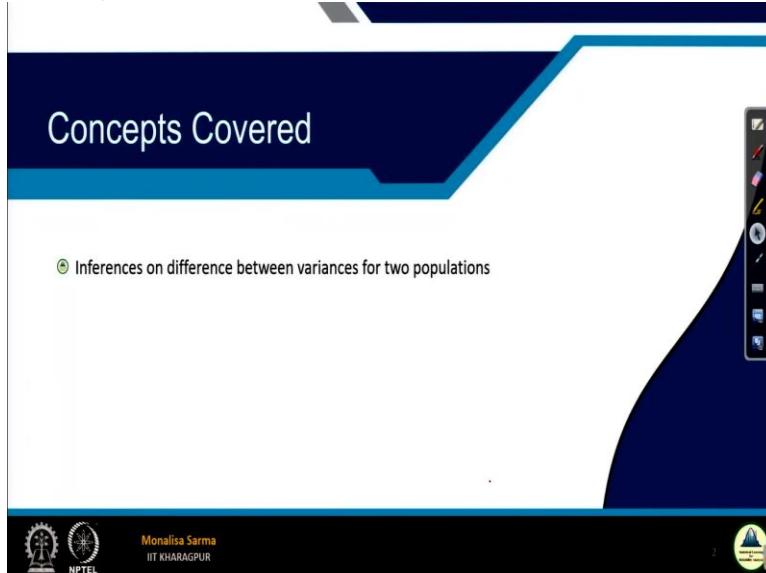
Monalisa Sarma
IIT Kharagpur

So, with that I end this lecture. Thank you. Thank you guys.

Statistical Learning for Reliability Analysis $\chi\beta\infty\alpha\sqrt{\mu}\sigma$
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology - Kharagpur

Lecture - 30
Statistical Inference (Part - 7)

(Refer Slide Time: 00:31)



Hello guys. So, again today we will also learn some other topics on statistical inference. So, in the last class, we have seen inference on difference between variance or difference between means of 2 populations remember, so, for that we have considered different cases. So, in today's lecture, we will see inference on difference between variances for 2 populations. So, that means, we want to compare the variance of 2 populations.

Now, can you from the last lecture, can you just find out where there is some application of this in the last lecture also, can you just think for a moment. And if you guys got the answer very good, if you did not get the answer, like when we discussed statistical inference on the difference between 2 population mean so, we have considered 3 different scenarios remember, for independent samples, we have only discussed in our independent samples, for independent samples, we have discussed 3 differences, 1 is variance known that was not an issue at all.

Very equally we could use this z distribution, another is variance is unknown, when variance is unknown, then what we have done when the variance unknown, again there are 2 cases for

variance unknown 1 is that we have assumed that the variances are equal and accordingly we have assumed that the variances are equal and from that we tried to find out the pool variance estimate remember.

And so, that test that we use for that we call it pool t test. And the statistics was pool t statistics. So, there what is our assumption was that both the variances are equal, but directly can we assume that both the variances are equal it will be because if the variances are not equal and if you assume the variances are equal by just when we are assuming the variance are equal, we are pooling the 2 variance by finding out a weight and giving out a pooled variance estimate.

If the pool estimated variances are not equal and you were assuming it is equals the result which you will get the result may be very incorrect. So, our total statistical analysis will go wrong in that case. So, we will have to find out whether the 2 population variances are equal or not, is not it? So, that is here we see inference on the difference between variances for 2 populations, this is how we will try to find out in this class we will see that.

(Refer Slide Time: 02:52)

Independent Samples: Inferences on Variances

Inferences on variances of two populations are important in different applications.

$$S_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

To determine whether a pooled variance may be used for inferences on two population means.

In many quality control experiments, it is important to maintain consistency, and for such experiments inferences on variances are of prime importance

Monalisa Sarma
IIT KHARAGPUR

So, as I mentioned just now, what I mentioned when I am trying to find out a pooled variance estimate. So, the way to determine whether the pool variance may be used for the inference of 2 population means. So, this is 1 reason why we may have to infer the variance and 2 populations to determine whether we can use the pool variance. Another is that you know lesser variance

means that any system or equipment is of good quality any material if you see if there are lots of variances to here.

Then definitely it is not meeting the specification I suppose, I bought a material some cloth material I expected certain density of it and I got an next slot again suppose I liked the material and I bought a same cloth but then that cloth is totally different thing different density. So, that is not expected, is not it? So, this there in different again I have given the cool drinks example also, if I am buying a cool drink can of 1 litre I expect it to be 1 litre only I do not expect that I am buying it 1 litre.

And I am getting 900ml or 950ml I do not expect that if I get more that is merrier, but for me it is merrier, but for the manufacturer, it is a loss forget about the manufacture, look the consumer point of view also, Smaller variance will okay do 1 litre means slight variation I can agree but more variance definitely it is not acceptable. So, for quality experiment also it is important to maintain consistency and for such experiments inference and variance are of prime importance.

So, inference on variances, it has many applications we have seen these 2 applications these are the main applications.

(Refer Slide Time: 04:46)

Inferences on Variances

- Null Hypothesis:** $H_0: \sigma_1^2 = \sigma_2^2$ or $H_0: \sigma_1^2 / \sigma_2^2 = 1$
- Alternate Hypothesis:** $H_1: \sigma_1^2 \neq \sigma_2^2$ or $H_1: \sigma_1^2 / \sigma_2^2 \neq 1$
- Independent samples of size n_1 and n_2 are taken from the two populations to provide the sample variances s_1^2 and s_2^2
- Compute the ratio: $F = s_1^2 / s_2^2$
- This value is compared with the appropriate value from the table of the F distribution, or a p value is computed from it.

Handwritten notes on the slide:

- $F = s_1^2 / s_2^2$
- $(n-1)s^2 = SS$
- $F = \frac{s_1^2}{s_2^2}$
- $F = \frac{S_1^2}{S_2^2}$

Video feed of Monalisa Sarma:

So, how we go about same case how we have inferred on variance of a single population same way we will go, but when we have discussed inference on a single population, remember which

distribution we have used when we have tried to infer on the variance of a single population we have used chi square distribution and we try to find out when the test statistics that we calculated it remember $n - 1 S^2$ by σ^2 , $n - 1 S^2$ this is also we call it also sum of square by σ^2 , this is my test statistics.

And if this test statistics falls within the acceptable region, acceptable region means within the 95% of the whole chi square distribution, then I say that is whatever is the null hypothesis is accepted remember, so, now here when we are trying to compare 2 different populations. Then 2 different population in the way and we are comparing the 2 different population variance, we cannot use chi square distribution chi square distribution is found for only 1 population.

Then we can we will be using f distribution we have already discussed this when we have discussed sampling distribution as I told you statistical inferences backbone is the sampling distribution only. So, what are my hypothesis my null hypothesis is $\mu_1 = \mu_2$ or I can read write $\mu_1^2 / \mu_2^2 = 1$, it is 1 the same thing then my alternate hypothesis $\mu_1^2 \neq \mu_2^2$ or $\mu_1^2 / \mu_2^2 \neq 1$.

But usually and most of the cases we will find when we are trying to infer on variance we always try to check whether my particular variance is greater than a particular value. Always in variances, most of the cases it is a single tailed and we try to compare the greatness of the variance because if the variance is less that is good, why and this is some in my null hypothesis is equal that is something which I want and less is also something which I want. So, why will again go and test it for less. So, usually when we are testing we are testing it for greater than.

So, when you are trying to find out any inference of the variance at a single population or double population, usually we go for 1 tailed alternate hypothesis that is greater than but that does not mean we go for 2 tailed hypothesis. So, here is an example of 2 tailed hypothesis. So, independent sample of size n_1 and n_2 are taken from the 2 populations to provide the sample variance s_1^2 and s_2^2 .

So, from that we will compute the ratio what is my F remember, $F = s_1^2 / s_2^2$ is not it? This is my F ratio when I have taken s_1^2 / s_2^2 equals to 1 that means, this book I am writing it properly my handwriting is not very good actually s_1^2 / s_2^2 , s_2^2 / s_1^2 . So, this portion is equal to 1 that means, what is remaining is my test statistics is s_1^2 / s_2^2 . So, this is my test statistics.

So, this value is compared with the appropriate value from the table of the F distribution or a p value is calculated anything we can do either from whatever we can either we can find out the p value corresponding to whatever F we get or we can find out the rejection region for a particular specific a significance level we find that clear rejection region, here F distribution is not symmetric remember F distribution and then chi square distribution is not symmetric.

So, we will have to find out a rejection region for both the side both lower tail and the upper tail. So, we find a critical region and if the value of F falls within the critical regions, that means not means what to say if it is less than the critical regions, then we say that the variances are equal that means, the null hypothesis is accepted if it is false in the critical region, then we say that null hypothesis rejected.

(Refer Slide Time: 09:06)

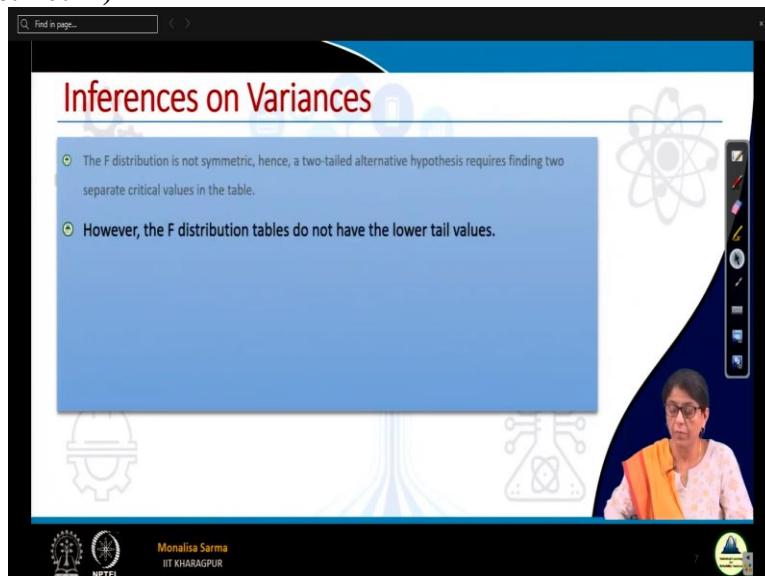
The F distribution is not symmetric, hence, a two-tailed alternative hypothesis requires finding two separate critical values in the table.

Now, the F Distribution is not symmetric hence a 2 tailed alternative hypothesis requires finding 2 separate categorical values in the table. So, if we see the F distribution it is something some this is not an exact figure of F distribution you can see a table my drawing is very bad actually.

So, it will be something this sort of this is 1 critical region, this is 1 critical region this critical region is probability of α to this.

So, it is F of α the value corresponding to this F of α . So; any value greater than F of α will fall in the critical region if this is F of α . So, this is what will be F of $\alpha / 2$ $1 - \alpha / 2$. So, if it falls in the left of this then also this falling in the critical region. So, if my value is less than this value F of $1 - \alpha / 2$, or if my value is greater than F of $\alpha / 2$, then it falls into the critical region, then we reject the null hypothesis.

(Refer Slide Time: 10:12)



However, F distribution do not have the lower tail values, in the F distribution table, you see you do not have the lower tail value, lower tail means my α is a 5%. So, I will go to this figure this is α is 5% so, this is $\alpha / 2$ is what will be the $\alpha / 2$ 0.025 is not it? F of 0.025 if this is 0.025, this is my lower tail value will be $1 - 0.025$ this value, this F of $1 - 0.025$ this we do not have in that table.

In there for F table we have values for 0.01 0.05 0.025 0.005 something this for very smaller α value we have values, but then for bigger α value, we do not have the values in the F table. But then we have one theorem we have seen that theorem when discussing the sampling distribution; we can use that theorem for our rescue.

(Refer Slide Time: 11:08)

The F distribution is not symmetric, hence, a two-tailed alternative hypothesis requires finding two separate critical values in the table.

However, the F distribution tables do not have the lower tail values.

These values may be found by using the following relationship:

$$F_{(1-\frac{\alpha}{2})}(v_2, v_1) = \frac{1}{F_{(\frac{\alpha}{2})}(v_1, v_2)}$$

So, what was the theorem remember, so F of $1 - \alpha / 2$, but here the degrees of freedom changes, if F of $1 - \alpha / 2$ degrees of freedom F of $v_2, v_1 = 1 - F$ of $\alpha / 2$ v_1, v_2 . What does v_1, v_2 means that the numerator I have the sample size first what to say variance of the first population that is s_1^2 and the denominator I have the second variance of the second populations. So, the degrees of freedom just when I try to find out the upper tailed value, we will see with an example.

(Refer Slide Time: 11:47)

Point estimate for $\sigma_1^2 / \sigma_2^2 = s_1^2 / s_2^2$

Interval estimate of σ_1^2 / σ_2^2 can be established using the statistics

$$F = \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2}$$

So, this is so, interval when I tried to find out the interval estimate, so, what is my corresponding point estimate, my corresponding point estimate is that this is my corresponding sorry my corresponding point estimate is s_1^2 / s_2^2 so, we will use this F statistics.

So, we can write if α is the significance level what is my confidence coefficient? Confidence coefficient is $1 - \alpha$. So, to fall within my confidence interval F value is this value, this should fall within this range, it should fall within this means it should be less than this value and it should be greater than this value it should fall in this, this is my confidence interval, is not it? So, this is my confidence interval. So, now, it is a replace F by this value.

(Refer Slide Time: 12:53)

Inferences on Variances

Point estimate for $\sigma_1^2 / \sigma_2^2 = s_1^2 / s_2^2$

Interval estimate of σ_1^2 / σ_2^2 can be established using the statistics

$$F = \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2}$$

We can write

$$P \left[f_{1-\frac{\alpha}{2}}(v_1, v_2) < F < f_{\frac{\alpha}{2}}(v_1, v_2) \right] = 1 - \alpha$$

From above equation, we get the lower and upper critical regions

Lower critical region: $f_{1-\frac{\alpha}{2}}(v_1, v_2)$
 ⇒ Reject null-hypothesis if $f_{1-\frac{\alpha}{2}}(v_1, v_2) > F$

Upper critical region: $f_{\frac{\alpha}{2}}(v_1, v_2)$
 ⇒ Reject null-hypothesis if $f_{\frac{\alpha}{2}}(v_1, v_2) < F$

Monalisa Sarma
IIT KHARAGPUR

And then do some simplification, after simplifying we can find out what is the boundary for σ_1^2 / σ_2^2 simple simplification if we do simple simplification, we will see the lower critical region this is of course, we have seen the lower critical region will reject the null hypothesis if it is less than this we have already seen, we will reject the null hypothesis it might add value is greater than this value we have already seen.

(Refer Slide Time: 13:18)

The confidence interval for the point estimate s_1^2 / s_2^2

$$P \left[f_{1-\frac{\alpha}{2}}(v_1, v_2) < F < f_{\frac{\alpha}{2}}(v_1, v_2) \right] = 1 - \alpha$$

$$\Rightarrow P \left[f_{1-\frac{\alpha}{2}}(v_1, v_2) < \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2} < f_{\frac{\alpha}{2}}(v_1, v_2) \right] = 1 - \alpha$$

$$\Rightarrow P \left[\frac{s_1^2}{s_2^2 f_{1-\frac{\alpha}{2}}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2 f_{\frac{\alpha}{2}}(v_1, v_2)} \right] = 1 - \alpha$$

We know that, $f_{1-\frac{\alpha}{2}}(v_1, v_2) = \frac{1}{f_{\frac{\alpha}{2}}(v_2, v_1)}$

NPTEL
Monalisa Sarma
IIT KHARAGPUR

Now, to find out the significance level, so, we have simplified F with this value and we will do simple simplification and this is the confidence interval, this is the upper confidence interval, this is the lower confidence interval. Now, here in the upper confidence interval what I have in the denominator I have F of $1 - \alpha / 2$ v 1 v 2 instead of F of $1 - \alpha / 2$ v 1 v 2 cannot I replaced it because this value I will not get it from the F table I will need some value which will get it from the F table.

So, this value I can replace it F of $1 - \alpha$ v 1 v 2 I can replace it with this value is not it? Simple simplification what I have done as a simple simplifications dividing first step what I have done first step I have divided all the expression by $s_1^2 / 2^2$. That is the first step I have done then second step my denominator was σ_2^2 and numerator σ_1^2 , I have made σ_1^2 numerator σ_2^2 denominator then accordingly my inequality also getting changed.

So, accordingly I got this value. Now, what is my upper confidence by sending this like this is the value I got. So, this is my upper confidence limit, this is my lower confidence limit.

(Refer Slide Time: 14:49)

Example-1

An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 81 with a sample standard deviation of 5. Assume the populations to be approximately normal with equal variances.

- a) Can we conclude at the 0.05 level of significance that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units?
- b) Are we justified in making the assumption that the two populations variances are equal?

So, now a simple example so, in the last class, remember we were trying to find out the average² of 2 different laminated material where we are trying to compare the means that the 2 different materials for which we are trying to find out the weight of we have taken from 2 different population we have taken a sample 2 different types of samples of size if I remember correctly 1 was of size then another was of size 12 something like that.

And then the variance of the population was not given. So, what we have we assumed? We have assumed the there is we have assumed it to be equal and that way we found out the pool variance estimator we found out S_p^2 . Now in this question, but we will try to everything else remains same. Here what we try to do is that are we firstly let us read the question, an experiment was performed to compare the abrasive wear for the 2 different laminated materials, 12 piece of material 1 were tested by exposing each piece to a machine measuring wear fine.

10 piece of material 2 were similarly tested. In each case, the depth of the wear was observed the sample of material 1 gave an average wear of 85 units with a sample standard of 4, standard deviation of 4 while the sample of material 2 it gives an average of 81 with a sample standard deviation of 5. So, n_1 is given n_2 is given for sample 1 size, standard deviation is given, mean is given, \bar{x}_1 is given, s_1^2 is given, \bar{x}_2 is given, s_2^2 is given, n_1 is given, n_2 is given.

So, from that we found out the S_p^2 , what is S_p^2 ? Pooled variance estimate that means, we try to find out the weighted variance of both the sample so, can we conclude that 0.05 level of

significance that average wear of material 1 exceeds that of material 2 by more than 2 units. This problem we have already solved is not it? For this what was our hypothesis testing remember how hypothesis testing null hypothesis $\bar{x}_1 - \bar{x}_2 = 2$ is not it?

Sorry $\mu_1 - \mu_2 = 2$ and an alternate hypothesis is $\mu_1 - \mu_2$ is greater than 2 that was my alternate hypothesis and we have tested it and when we have found it that my null hypothesis is not rejected. So, in there we have assumed here it was then we have assumed that the population has equal variances. Now, are we justified in making the assumption that the 2 population variances are equal.

So, that is the same question first we have assumed it is equal actually we will do other way around before assuming first we will do this test whether the 2 population variances can be considered equal, first we will do this and then if it is can be considered equal then we will assume then we know it is equal then we will find out a pool variance estimate. First this is done then that is done now, in this first since we have introduced that so, first we have done the mean there we have done this assumption now, we are trying to prove that whether the 2 population variances are equal.

(Refer Slide Time: 17:55)

The slide is titled "Example 1: Solution". It features a green header bar with the text "Solution for a)" and a blue sidebar containing a question about material wear. The main content area is orange and contains the text "Question 1.a) was solved in previous lecture." The sidebar text reads:

Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 81 with a sample standard deviation of 5. Assume the populations to be approximately normal with equal variances.

a) Can we conclude at the 0.05 level of significance that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units?

The slide also includes a video player showing a woman speaking, and various decorative icons and logos.

So, that is why I have taken given the same question here. So, question 1 was solved in my previous lecture.

(Refer Slide Time: 18:01)

Example 1: Solution

Solution for b) : Hypothesis Formation

Let σ_1^2 and σ_2^2 be the population for the abrasive wear of material 1 and material 2, respectively.

The hypotheses are:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

$$\alpha = 0.10$$

Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 81 with a sample standard deviation of 5. Assume the populations to be approximately normal with equal variances.

- b) Are we justified in making the assumption that the two populations variances are equal?

So, here let σ_1^2 and σ_2^2 be the population variance for the abrasive wear of material 1 and material 2 respectively. The hypothesis are, this is my hypothesis that is the significance level is given in the question, can we conclude that 0.05 level of significance that abrasive wear material, significance level is here in this question significance level is not there and we have taken his $\alpha = 0.10$.

(Refer Slide Time: 18:36)

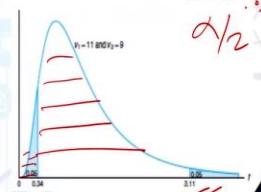
Example 1: Solution

Solution for b): Critical Region

From F table, we get that $f_{0.05}(11,9) = 3.11$, and

$$\text{we get, } f_{0.95}(11,9) = \frac{1}{f_{0.05}(9,11)} = 0.34$$

$$\alpha = 0.1$$



So, there is a mistake anyway, I will correct it in the slides, but you will get it basically. So, here we are taking the significant levels 0.05. So, it is 0.05, α is 0.05. So, now, what we will do is that we will from the table we will try to find out from the F table we will try to find out what is the F of 0.05 corresponding to 11, 9. So, this is what we will do is that for 11, 9 this is the value 3.11.

So, if this is 3.11 then how we can find out the corresponding value for this? What to say this value this critical region is nothing but $1 / 0.05$ 9 11 will give me 0.95 11 9.

So, it is 0.34, see here, here we have taken a significance level of because this is 2 tailed, 2 tailed means it is $\alpha / 2$ is not it? So, if my α is 0.05, then my α should be 0.025 so, here I have considered 0.05 I am sorry for the mistake I will correct it in my slide this that means I have considered a significant level of $\alpha = 0.1$ only this slide was right 0.1 but the question is wrong a equation I have put it is 0.1 I will correct it no, that means my significance level is 0.1.

So, it is to tail it is $\alpha / 2$ these are minor things you can understand that is no, you already know this concept. So, it will be different, it will be easier for you to understand. So, when it is 2 tailed and we consider it $\alpha / 2$, so, it is 0.1 my $\alpha / 2$ is 0.05. So, I found out F of 0.05 11, 9 remember we took the same on the numerator, but we took numerator we always take the sample with the larger variance.

So, this we got it 3.11. So, once we get to 3.11 corresponding the value for what to say lower tailed, we can find out is in this formula and we got 0.34. Now, if my F value lies within this range, then my null hypothesis is accepted what is accepted that the both the population variances are equal.

(Refer Slide Time: 21:07)

Example 1: Solution

Solution for b): Critical Region

From Figure, we see that $f_{0.05}(11,9) = 3.11$, and
we get, $f_{0.95}(11,9) = \frac{1}{f_{0.05}(9,11)} = 0.34$

Therefore, the null hypothesis is rejected when $f < 0.34$ or
 $f > 3.11$, where $f = \frac{s_1^2}{s_2^2}$ with $v_1 = 11$ and $v_2 = 9$ degrees of freedom.

So, I calculated my F value. So, what is my F value here? F value is 0.64, null hypothesis will be rejected if it is less than 0.34 and if it is greater than 3.11 if it is greater than 3.11 or if it is less than 0.34 it will be rejected with v 1 = 11 and v 2 is also 9 degrees of freedom.

(Refer Slide Time: 21:36)

Example 1: Solution

Solution for b): Decision on null hypothesis

Given

$$s_1^2 = 16, s_2^2 = 25,$$

and hence

$$f = \frac{16}{25} = 0.64$$

Decision: We should not reject the null hypothesis (H_0).

Conclude that there is insufficient evidence that the variances differ.

Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 81 with a sample standard deviation of 5. Assume the populations to be approximately normal with equal variances.

b) Are we justified in making the assumption that the two populations variances are equal?

Monalisa Sarma
IIT KHARAGPUR

Now, for to calculate the F value I need to find out s_1^2 and s_2^2 from it is given here. So, I found my F value is 0.64 very well lies within this range. So, my null hypothesis is accepted what is my null hypothesis? That is both the variances are equal. So, we should not reject the null hypothesis H_0 conclude that there is insufficient evidence that the variances differ because it is asking are we justified in making the assumption that the 2 population variances are equal yes. We are justified because we are not being able to reject the null hypothesis.

(Refer Slide Time: 22:13)

④ In this lecture we learnt inferences on difference between variance for two populations, which are important in different applications

Monalisa Sarma
IIT KHARAGPUR

NPTEL

25

So, in this lecture, we learn inference on difference between variance of 2 population we have seen what are the usefulness of it, mainly, I have specified 2 different cases where it is useful once for finding other inference and population means wherever variances are not known. In that case, we will have to see the variances are equal then our life become cool, then we can directly use the t distribution.

And however, after finding out the variance in 2 population, we found that the variance are not equal, then we cannot use t distribution then, what we will have to use them first 1 way is that we will see if we can transform the data if transformation data is at all feasible, then we can transform the data and we can find out the variances are estimated if the variances are equal. Another way is that if the sample size is bigger that means, if it is more than 30 say then we can very well use the normal distribution.

If we can use the normal distribution, there is no issue at all of the degrees of freedom then we can very quickly use that. Then other is that if the value that if the parent population, if it is known to us, that the parent population is very closer to the normal population. Then we can also use the t distribution where we will use the degrees of freedom of the smaller variances as the smallest sample as the degrees of freedom. So, that is what we have seen.

(Refer Slide Time: 23:35)



So, these are the references and thank you guys. Thank you.

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)

Dependent Samples

Example: Consider two different methods for conducting a experiment to find the effect of a special diet on weight gains

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.

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.

This method of data collection is called independent samples

This method of data collection is called dependent samples

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 x_1 bar 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 x_2 bar. Now if I try to find out x_1 bar - x_2 bar 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)

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 individuals' differences in weight before and after the special diet are then a more precise indicator of the effect of the diet.

Monalisa 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

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.

Monalisa 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 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 $t = \frac{d - d_0}{\sqrt{s^2/d}}$
- The t statistic is usually called the "paired t statistic."

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 σ_1 σ_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 $d_{\bar{}} - \lambda_0$ 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_d^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	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

Monalisa Sarma
IIT KHARAGPUR

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 give 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 were 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 Sarma
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.

Monalisa Sarma
IIT KHARAGPUR

24

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,

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 > \mu_2$$

$$\bar{d} = 0.1417, s_d = 0.190$$

$$t = \frac{0.1417}{0.190/\sqrt{12}} = 2.48$$

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.

Monalisa Sarma
IIT KHARAGPUR

25

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?

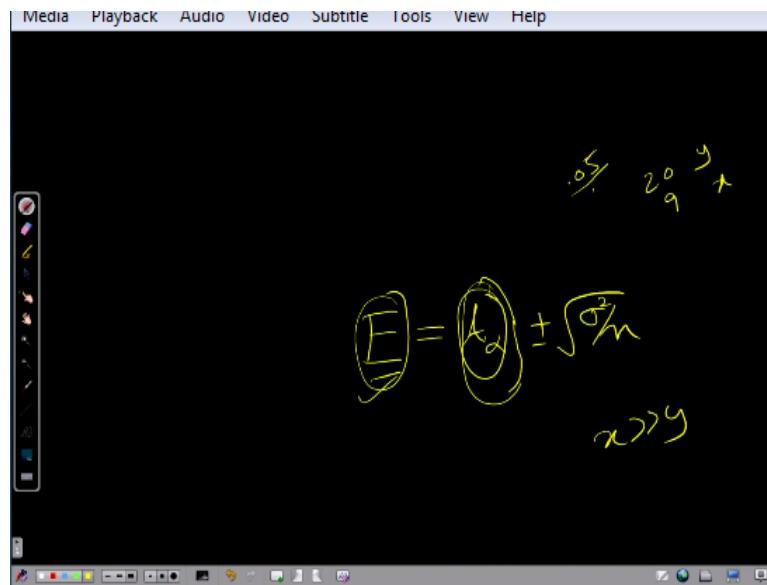
We will use independent or dependent both has its own pros and cons. When we use paired samples, though, we within sample variance is less, but what we have we sacrifice on the degrees of freedom, sacrifice on the degrees of freedom means see since my sample size is n my degrees of freedom is $n - 1$, say my sample sizes $n - 1$ my sample sizes $n - 1$ 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)



When my t value is more remember when we talk this error of estimation what is my error of estimation? Error of estimation is $t(\alpha/2) + \sigma^2 / \sqrt{n}$ 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, $t \alpha / 2$ so, here what is my $t \alpha$ 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 let us quickly go for the summary. So, compress 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.

NPTEL
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.

NPTEL
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 we 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 KHARAGPUR

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)

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

- ④ In this lecture,
- ④ We covered the topic of inferences on mean for dependent samples of two populations
- ④ We also introduced the idea of using non-parametric methods for those scenarios where the assumption of normality does not hold
- ④ In next lecture, we will discuss about inferences on more than two populations.

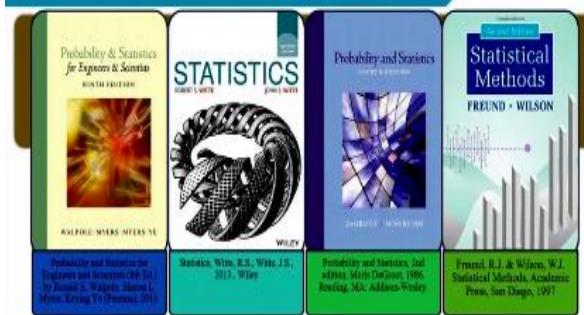


Monalisa Sarma
IIT KHARAGPUR

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 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



Monalisa Sarma
IIT KHARAGPUR

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

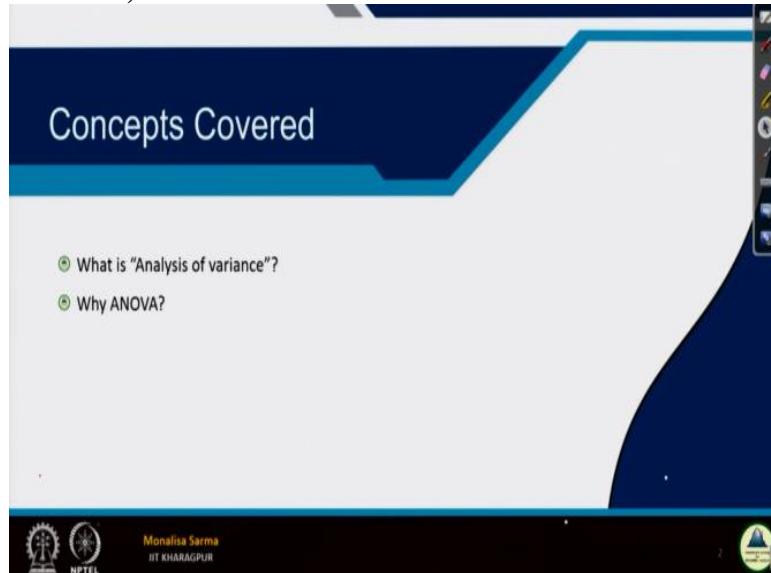
Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology - Kharagpur

Lecture - 32
ANOVA - I

Hello everyone so, today we are starting a new topic that is analysis of variance in short we call it ANOVA. So, it is not exactly a new this is also another statistical inference technique, but here this is used for when we try to compare more than two population as sincerely then we use analysis of variance and the thing is that this topic before I start it is a bit complicated than the topics which we have covered previously.

So, you will have to listen to it very minutely and you may have some doubts no issue that doubts can be cleared in the doubt clearing sessions there is no issue and even you can read through the textbooks also I have mentioned the textbook specific textbook from where I have taken this.

(Refer Slide Time: 01:15)



So, now, coming to the things what I will be covering in this class first is what is analysis of variance in short what is ANOVA and then second, why ANOVA we will be covering basically these 2 topics in this lecture.

(Refer Slide Time: 01:31)

What is analysis of variance?

Single Population

Multiple Population

Monalisa Sarma
IIT KHARAGPUR

So, till now, what we have seen is that we have tried to infer what to say about a single population, we try to infer mean of a single population variance of a single population, then again we have tried to infer the mean of two populations and variance of two population whether I we did not try to infer what mean of two population we have when we used two population.

Basically two population we used to compare with the mean of two population are same greater less and then similarly, variance also we used, we tried to compare the variance of two populations for that we use app distribution if you can remember now, what we will do is that we want to compare multiple proportion more than two populations.

(Refer Slide Time: 02:16)

What is the issue?

Are the statistical inferences valid?

μ σ

Monalisa Sarma
IIT KHARAGPUR

So, whatever concepts we have used while comparing two population or single populations like we have we tried to find out the mean of the sample from there we try to infer about the

population mean similarly, we try to find out the variance of the population and from there we try to infer about the population variance all those the way we have done is it now valid we can we do it the same way.

But we have done for single population and two population can we do it the same way? Let us see that now.

(Refer Slide Time: 02:50)

The slide is titled "Example 1". On the right, there is a video frame showing a woman with glasses and a pink shirt speaking. To her left is a green speech bubble containing text. Above the speech bubble are three orange boxes with icons: a clipboard labeled "CLAIM", a person thinking labeled "What if it affected the results of the students in a negative way?", and a brain with gears labeled "What kind of music would be a good choice for this?". Below these is another green speech bubble containing the text "Need to design a specific experiment to have some proofs that it actually works or not". At the bottom left of the slide, there is a logo for NPTEL and the name "Monalisa Sarma IIT KHARAGPUR".

So, for that let us start with an example first. So, what is the example? A recent study maybe in the newspaper it has come or somewhere it has come like, a recent study claims that using music in a class enhances the concentration and consequently helps students to absorb more information. So, study has been done study means on a sample of students basically, some organizations is trying has done a study.

And from there what it has found is that the conclusion is that so, if we use music in the class, it enhances concentration and when it increases the concentration then it consequently helps students to absorb more information students performs better. So, study has concluded there concluded that now, we are not convinced with this study, we want to test it, is it really true? So, we are also interested in finding out what if it affected the results of the students in a negative way?

The study has told that this music has improved the concentration increase the concentration and students have started performing better, but then is it really true? Is it we if we put music in the class, will it really affect the performance? It is affecting the performance but is it

effective in a positive way or negative way? That is the first thing. Second thing, what kind of music would be a good choice?

Music means there can be any music. So, what kind of music will help the students will have an impact positive impact or what kind of music will lead to a negative impact. So, for that, we need to design a specific experiment to have some proof that it actually works. So, study claim that we are not satisfied with the study, we are not sure of the study so we wanted to find out.

So, for that we really need to do an experiment to have proof that actually music actually increases the concentration.

(Refer Slide Time: 04:49)

The teacher decided to implement it on a smaller group of randomly selected students from three different classes.

Step 1

1. Three different groups of ten randomly selected students from three different classrooms were taken.

So, what is the design of experiment the teacher has planned? The teacher decided to implement it on smaller groups of randomly selected students from 3 different classes of the same level. So, what is teacher how he or she decided to carry out the experiment he has designed experiment in such a way what he has decided what he or she has decided to implement it on a smaller group of randomly picked students from 3 different classes are the same level.

So that is the first step what is the first step? First step is create different groups of 10 randomly selected students from 3 different classrooms were taken, that is the first step.

(Refer Slide Time: 05:39)

Example 1: Design of Experiment

The teacher decided to implement it on a smaller group of randomly selected students from three different classes.

Step 2

2. Each classroom was provided with three different environments for students to study:

A. Classroom A had constant music being played in the background.

Monalisa Sarma
IIT KHARAGPUR

Then and so, we have taken 10 students from 3 different classes in 1 class, what the teacher do? The teacher has constant music being played in the background, while the classes are going on, at a certain scene that a constant music is constantly playing in the background. So, that is the first group he has total 3 groups. So, in the first group that is done.

(Refer Slide Time: 06:07)

Example 1: Design of Experiment

The teacher decided to implement it on a smaller group of randomly selected students from three different classes.

Step 2

2. Each classroom was provided with three different environments for students to study:

B. Classroom B had variable music being played in the background

Monalisa Sarma
IIT KHARAGPUR

And in the second group, the class B had variable music in the background, sometimes very slow, sometimes loud music sometimes very soft music variable music is played in the background that is the second group of students.

(Refer Slide Time: 06:22)

Example 1: Design of Experiment

The teacher decided to implement it on a smaller group of randomly selected students from three different classes.

Step 2

2. Each classroom was provided with three different environments for students to study:
- C. Classroom C had no music being played at all



Monalisa Sarma
IIT KHARAGPUR

And third group of students had no music played at all. So that is how the teacher has designed the experiment, the teacher has picked randomly picked 10 students from 3 different classes of the same level. And the first group he has played the music constantly in the background, and the second group, the teacher has played random music and for the third group, the teacher has played no music at all.

(Refer Slide Time: 06:49)

Example 1: Design of Experiment

The teacher decided to implement it on a smaller group of randomly selected students from three different classes.

Step 3

3. A test was conducted for one month for all the three groups and their test scores were collected after that.



Monalisa Sarma
IIT KHARAGPUR

So, test was conducted for 1 month, for 1 month testing is continued. And after that, the teacher has taken some sort of tests and for all the 3 groups and the test scores were collected after that.

(Refer Slide Time: 07:05)

Example 1: Test results

	Test scores of students (out of 10)										Mean
Class A (constant music)	7	9	5	8	6	8	6	10	7	4	7 ✓
Class B (variable music)	4	3	6	2	7	5	5	4	1	3	4 ✓
Class C (no music)	6	1	3	5	3	4	6	5	7	3	4.3 ✓
Grand Mean											5.1

So, this is the test scores for 10 different students, this is for classroom A, these are marks out of 10 teacher has taken a test conducted a test out of 10 and this is the marks obtained by classes with the constant music is played these are the marks which is obtained by the class B where variable music, and this is the these are the marks which the teacher got where no music has been played.

And we see if we find out the mean, for the class A it is 7 class B it is 4 class C it is 4.3. So, you know how to find out the mean, we have already seen it, it is basically the average. So, from this mean, what we have seen is that directly if we try to infer from the mean, what we see is that class C has performed much better than class B and class C, class B and class C, we can see there is not much of difference, no significant differences there.

But class A has performed significantly different from class B and class C, if we see the mean from the mean, we can give that conclusion.

(Refer Slide Time: 08:13)

Example 1: Observations

Observations from the Results

It is noticed that the mean score of students from Group A is definitely greater than the other two groups, so the treatment must be helpful.

But what if we happened to select the best students from class A, which resulted in better test scores??

Remember, the selection was done at random.

Monalisa Sarma
IIT KHARAGPUR

So, it is noticed that a mean score of students from group A is definitely greater than the other 2 groups so the treatment must be helpful. We have seen the results from the results, we have seen class A where we have played a constant music constant score, what they got the students what they got the mean score is significantly better than the other 2 groups and the class B and class C.

That means is the music really helpful? Means it makes us think, maybe the treatment is actually helpful. If we play music background in the thing, then maybe it is really helpful, maybe it is true. But what if we happen to select the best students from class A, because we have randomly picks the students is not it? Random means it is the probabilities is not it? We have just picked.

So, when we picked might be that we have selected all the best students for class A, that may happen, is not it? Or the other way around might be we have picked all the weaker students for class B and C might be the music is not having any effect at all just that we have picked very good students, some of the very good students for class A or maybe for class B and C, we have selected some very poor students, there may be some outliers in class B and C some students who are very, very poor.

That might have had the effect of bringing the mean to such a low value or class A we have there may be some outliers or some 2 or 3 students have scored really good that is again an outlier maybe which has bought as mean quite high, is not it? While we can trying to find out

the mean we do not analyze the marks what they are getting, we do not analyze all the data points. We have seen that when we find out a mean or variance.

We do not analyze each and every data points we directly look at the mean value or the variance value wherever we are interested, whichever we are interested. So, that may happen that may be the case our selection may not be proper in a random selection, so, since the selection was done very much at random.

(Refer Slide Time: 10:18)

Example 1: Observations

Questions from the Observations:

- ① How do we decide that these three groups performed differently because of the different situations and not merely by chance?
- ② In a statistical sense, how different are these three samples from each other?

The solution to this is ANOVA!!

Monalisa Sarma
IIT KHARAGPUR

So, how do we decide these 3 groups perform differently because of different situation and not merely by chance. So, how do we decide that these 3 groups that have performed differently it is therefore, perform differently it is because of the different treatments and not because of mere chance, when I will call it what to say this groups that perform differently merely by chance.

Maybe when the students which are picked are not a uniform kind of students, it is actually the treatment that did not have any effect, we are getting different score because the students that we have picked may not be uniform students maybe that is why we got these different results, is not it? So, we just by seeing the mean we cannot be sure that the difference is because of the different treatment that we have done or because of this simple random chance.

In a statistical sense, how different are these 3 samples from each other there is a very important thing we need to find out. So, the solution to this is ANOVA, so, how can we find

this this sort of things, how can we find out? The solution to this ANOVA like let me give you a 1 more example suppose we are interested in finding out the there are suppose there are 3 different treatment type available for a certain disease.

Like if let us take cancer, for cancer, there are different treatments for 3 different types of treatments are available. Now, what we and we do not know which treatment is better, which among this, which treatment is better, we do not know, we need to find out which treatment is better. So, maybe how will you do from the population of all the people who are suffering this and suffering the particular disease for which the treatments are there.

3 treatments are there for we will randomly pick 3 different samples, 3 different groups in each group, we will apply these 3 different treatments, and maybe the treatment which is efficient maybe we can find that by curing time, but now the time it takes to cure the disease, that can be our criteria to find out the efficiency of the disease. Time it takes to cure make the person free from the disease.

So, now, we have given this treatment, we have suppose this treatment, we have used the treatment for say around 2 months and after that, we are seeing the curing time how much percentage has been cured or it is totally cured or not whatever and based on that, we can now the condition is based on that directly suppose treatment a treatment B and treatment C similar to whatever example we have seen just now, this music example.

So, now suppose we have seen for treatment A maybe the person the curing time was quite less for treatment B curing time may be bit higher for treatment, C maybe a bit more higher or maybe same as B or whatever it is. Now, just from this, can we directly say that treatment A is better than treatment B and C similar to the example that we have discussed? Maybe yes, maybe no? Why maybe no?

Like when we have picked the people, there may be chances that among that, there are some patients who are very old, who have some other ailments. So, then for them, disease it really takes time to cure. So, maybe for B or C whichever has taken a long time to cure or so they are maybe they are there while picking the patients maybe we have picked such patients who has already some other ailments.

Because of that, maybe the curing time it is taking a longer time or maybe in the sample A for the sample for the group A for the first group where we have used treatment A maybe in that sample, there are some patient maybe who are already taking some other medication because of that maybe their what is the disease is almost partially cured before starting this treatment. If it is partially cured before starting this treatment, then what happens when the treatment starts they will it will cure soon.

I am not talking about all maybe some because some will have an effect on the overall mean. So, this sort of things may be there. So, directly if we see the mean time it takes to cure we cannot directly come to the conclusion that this mean this time is different only because of the treatment. There may be other factors to that. So, this we can find out with the help of ANOVA.

(Refer Slide Time: 15:06)

Analysis of Variance (ANOVA)

ANOVA: Definition

ANOVA is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.

ANOVA checks the impact of one or more factors by comparing the means of different samples.

ANOVA was invented by Sir Ronald Aylmer Fisher (1921), and is often referred to as Fisher's ANOVA.

Monalisa Sarma
IIT KHARAGPUR

Now, the question is what is ANOVA? ANOVA is a statistical technique that is used to check if the means of 2 or more groups are significantly different from each other significantly different I have used this term significant before when I have discussed statistical inference I will not be repeating that. So, ANOVA is a statistic that is used to check if the means of 2 or more groups are significantly different from each other.

Now, what are the difference what we have seen in the recent music example? Are they are significantly different, now, what does significantly different here means we will have to see. So, ANOVA basically tries to find out if the means of 2 or more groups are significantly different from each other. So, what do we mean by significantly different we will come to

that. So, ANOVA checks the impact of 1 or more factors by comparing the means of different sample impact of 1 or more factors now, what is this factors?

It checks the impact of 1 or more factors by comparing the means of different sample here in the music example what is the factor? Factor is the music we are playing different type of music, the constant music, very random music and no music the factor is music different type of music no music can also be we can call it as a null music, so that is a factor. So, again with their maybe the second factor.

Second factor maybe suppose in this example, instead of taking the class of the same level, if we take some say class, some students have class 10 and some student have say class 2, so means how music affects elder students how music affects younger students. So, there are 2 factors maybe music and age. So, that is what ANOVA checks the impact of 1 or more factors by comparing the means of different samples.

It was invented ANOVA was invented by Sir Ronald Fisher and is often referred to as Fisher's ANOVA by his name it is also referred to as Fisher's ANOVA, so, now, the question is why we ANOVA means analysis of variance very well. Now, here we have already spoke that, that used to check the means of 2 or more groups ANOVA is trying to check the means of 2 or more groups are significantly different from each other.

So, we are talking of means we are talking about comparing means of 2 or more groups. Then, at the same time, we are talking analysis of variance, how by analyzing variance, we can talk about the difference of 2 or more groups of mean that is something we need to see and we will gradually come to that now, before going to that first thing is that, now, this question when our objective is to compare the means of different more than two population that is our objective.

We are not from an objective we have nothing to do with a variance we just have to find out the we had just to compare the means of more than two populations and music example there were 3 population even the disease example what I have taken there also there are 3 population we had to compare 3 population mean of 3 population then why not use t test?

(Refer Slide Time: 18:35)

Some Important Questions

- Why not use t-test ?
- Why analysis of variance for comparing means?

Monalisa Sarma
IIT KHARAGPUR

We have used details say we have seen t test details we could use for comparing 1 population details we have used for comparing two population as well is not it? So, why we are using t test? Why we are using analysis of variance? For comparing means this is something very confusing.

(Refer Slide Time: 18:55)

Using t-test

For which purposes t-test is used?

t-test is used:

- to infer mean of a single population
- t-test can be used to compare two populations

Our task here is to compare mean of more than two populations

Monalisa Sarma
IIT KHARAGPUR

So, t test where we use t test is used to infer meaning of a single population t test can be used to compare two populations. Now, t test we can use to compare more than two population as well. What is the big task here?

(Refer Slide Time: 19:09)

Extending the two population procedure

Extending the two population procedure

- Construct pairwise comparison on all means.
- For 5 populations, $\Rightarrow 10$ possible pairs.
- Considering $\alpha = 0.05$,
probability of correctly failing to reject the null hypothesis for all 10 tests is
 $0.95^{10} = 0.60$, assuming that the tests are independent
- Thus the true value of α for this set of comparison is at-least 0.4, instead of 0.05
- It inflates the Type 1 error.

Monalisa Sarma
IIT KHARAGPUR

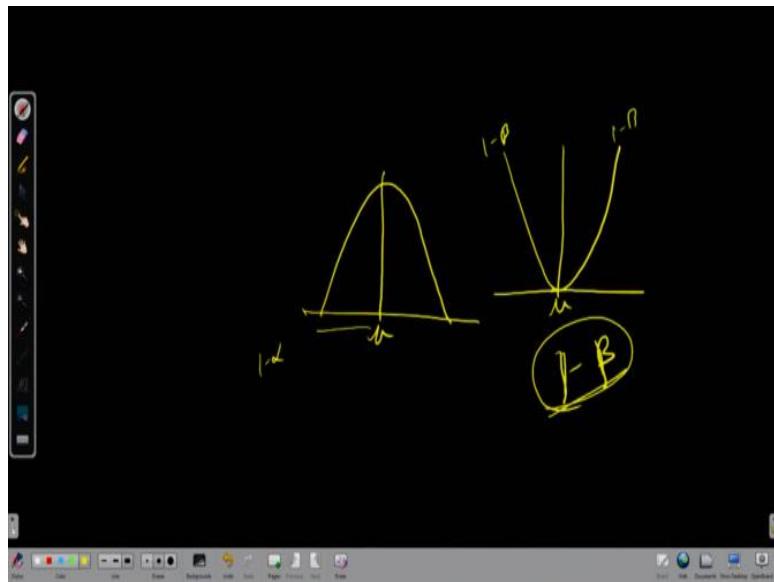
So, we will see what is the big task here? Suppose if you see if you are interested in comparing 5 population then how because t test we can use for comparing 1 population or at the most two population not more than that t test is designed in such a way now if you are interested in comparing mean of 5 population, so 5 population if you see all possible combination, it will be total 10 different possible pairs.

So, if you want to compare 5 population means of 5 population then we will have to do total 10 t test that is okay. If we have enough time it has, we will do 10 t test, that is not a problem, if we have enough time, but there are some other problems. So, what is that problem? Suppose we consider α is 0.05 that means our significance level is 0.05. So, if the significance level is 0.05.

Then what will be my β what will be my what to say maximum β maximum because β we have seen we have seen the β curve remember β curve is something this is not it? Here at this is basis of the mean where we will just see what is my β curve? B curve was something like this is not it? B is maximum at this point, when it is very near to the null hypothesis value.

That at that point this is β is maximum β will be minimum that is goes further away from the null hypothesis value, we have seen that and what is this? This is power curves.

(Refer Slide Time: 20:44)



So, this is my β curve so, if I talk about my power curve, this is my power curve. So, this is that this equals to μ 1 my value is very near to μ and what happens my β is highest, but when it goes further away from μ then my β goes gradually reduce and it goes to around $1 - \alpha$ is not it? So, that is why higher β will also not happen it will not have a very negative impact we have seen that.

And this is my power curve my power curve my power is very less when it is very near to when my hypothesis value is almost equal to the actual value then my power is less and gradually my power increases is not it? So, at this point what is my power curve what is my value of the power it is $1 - \beta$ this is my maximum power I am gradually it will sorry, this is my maximum power my maximum power is $1 - \beta$ this is $1 - \beta$ and gradually it is coming down.

So, now, if we see here, so, if we consider for 5 pairs if there are total 10 possible pairs for each pair if the α is 0.05. Then what happens what is the power of each test power of each test will be 0.95 is not it? Power of each test will be point 0.95 so, for all t tests is so then power of all t tests total there will be 10 tests or what will be the power of total t tests all these tests will be independent is not it?

All these tests are independent we are independently trying to test different tests. So, since all the tests are independent, so, when we are interested in power of the total test, because the total test requires us 10 possible 10 tests the power of each test in total tests will be 0.95 to

the power 10 and that is 0.60 and that will be the 0.60 is the power of my whole test. Now, if this 0.60 is the power of the whole test, then the true value of α is at least 0.4.

Because my power will be more than this when my power be more than this my value will come this will be this is at least 0.5 when α is at least 0.4. Then that means my α can be more than this also it can be α can be 0.5 also 0.6 also, when it will be this when it will become more when this will become more or less is not it? When the power curve is as I saw new power curve is like this.

This way it comes, comes, comes and power is very less at this point when the power is very less what happens my α will become more, more than 0.4. So, whereas I have started with an α of 0.05 see, that is the how my type 1 error gets inflated type 1 error a 0.4, it is too much to bear. And that is at least if my value can be more than that, because this 0.6 power is the maximum power and as I have seen power curve it this is the maximum at this point and gradually it comes down.

So, my power can come down to when my power will be lesser than that my α will go more and more up it will inflate the type 1 error and for 5 population we need 10 possible pairs for 6 population again more number much more for 10 population you can see imagine the number of pairs. So, you can imagine the type 1 error that we will have. So, t population will not be useful at all.

(Refer Slide Time: 24:34)

Set 1			Set 2		
Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
5.7	9.4	14.2	3.0	5.0	11.0
5.9	9.8	14.4	4.0	7.0	13.0
6.0	10.0	15.0	6.0	10.0	16.0
6.1	10.2	15.6	8.0	13.0	17.0
6.3	10.6	15.8	9.0	15.0	18.0
$y = 6.0$	$y = 10.0$	$y = 15.0$	$y = 6.0$	$y = 10.0$	$y = 15.0$

So, now that is why we have seen the t test is not at all suitable. We have seen that t test is really not effective when we try to compare more than two populations because it really inflates our type 1 error. So, what is the next so for that, we will try to see an example for that we have developed some data. So, the 2 sets of contrived data contrived data means we have made some data just for this example purpose only just consider there are 2 sets of data this is set 1 this is set 2.

So, set 1 has some from this set 1 we have 3 samples of 3 different populations. Similarly, for set 2 again we have 3 samples of 3 different populations. So, now what we have seen if we have tried to find out a mean of this sample, you see the mean of sample 1 of set 1 is 6 and mean of sample 1 of set 3 is also 6. Similarly, mean of sample 2 of set 1 is 10 here also it is 10 it is 10 here so, it is 10 here if you see this 15 here also it is 15.

So, if we just see the mean of all the both the samples, that means, we can do maybe we can take that both the sets, we have picked the sample from the same population means sample 1 of set 1 and sample 1 of set 2 we have picked on the same population again sample 2 of set 1 and sample 2 of set 2 also we have picked from the some other population. So, population B again sample 3 from set 1 sample 3.

From set 2 we have picked say from other population C this is this 2 are from the same population these 2 are from the same population, these 2 are from the same population if we just see the mean we can come to the conclusion, but is it actually true.

(Refer Slide Time: 26:13)

Example : Why ANOVA

Set 1			Set 2		
Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
5.7	9.4	14.2	3.0	5.0	11.0
5.9	9.8	14.4	4.0	7.0	13.0
6.0	10.0	15.0	6.0	10.0	16.0
6.1	10.2	15.6	8.0	13.0	17.0
6.3	10.6	15.8	9.0	15.0	18.0
$y = 6.0$	$y = 10.0$	$y = 15.0$	$y = 6.0$	$y = 10.0$	$y = 15.0$

Example: Observation from Means

Observations

- Looking only at the means, we can see that they are identical for the three populations in both the sets.
- Using the means alone, we would state that there is no difference between the two sets.

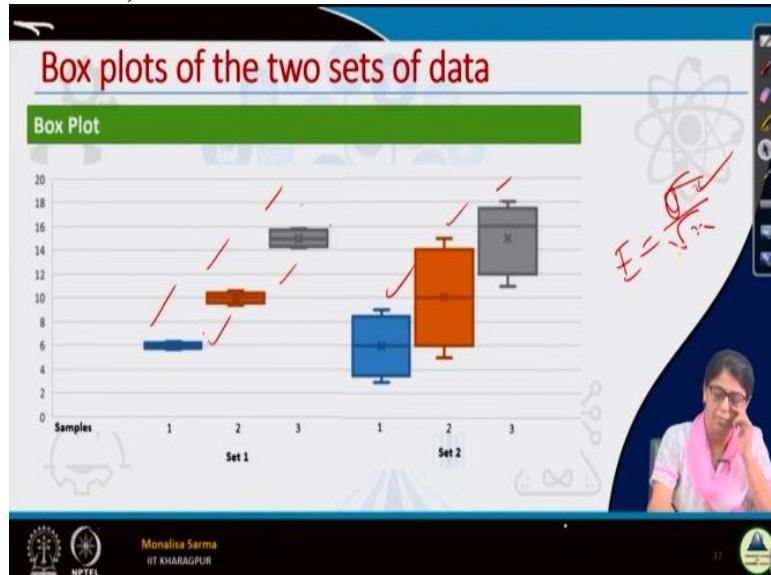
Monalisa Sarma
IIT KHARAGPUR

See, looking only at the means, we can see that they are identical for the 3 population in both data sets. Using the mean alone, we would state that there is no difference between the 2 sets. If we just see the main we can say that there is no difference between the 2 sets our intention while we are taking these 2 sets of data, our intention here is to find out is this forget about this data set.

If we just in 1 set our intention here is to find out whether this 3 populations are different the mean of 3 population are different, if we just see the mean here we can see that mean of this 3 population are really different 1 is 6 10 15 it is it looks to be significantly different, some say 2 also we are interested in finding out the mean of the 3 populations are different here also we are getting similar to set 1 we can say this 3 populations are actually the mainstream significantly different.

So, maybe that is 3 populations are different. If we can conclude this direct listing from the mean and from the mean we can say set 1 and set 2 are identical if we just see the mean.

(Refer Slide Time: 27:17)



Now, what we have done, I have drawn a box plot of this 2 set of graphs I have drawn the blocks plot, remember how we draw the boxplot I have discussed in the first lecture box plot, you please go see the lecture again if you have forgotten that thing so I will not explain it here. So, we have drawn the box plot of this 3, 2 sets of data. So, this is the box plot of set 1 for 3 different samples.

This is the boxplot of set 2 for 3 different samples. Now here we proceed from the box plot if we see the means are same here also the mean is 6 here mean is 10, here mean is 15, 2 sets mean are same, but you see here in set 1 the data's are bunched together it is very near or very close the variance of this data is very different here. But here it you will see that data's are very much variant remember, when we have discussed central limit theorem.

So, what I told you what was the if we take a sample from a population with mean μ and variance standard deviation σ , if you have taken a sample from a population with mean μ and standard deviation σ than the if we have the sampling distribution of the sample when we see the mean of the sampling distribution is the mean of the population mean and what is the variance of the sampling distribution variance of the sampling distribution was σ / \sqrt{n} is not it?

And here also we have pointed out if the variance of the parent population was is quite more if the variance of the parent population is very high, then what happens that mean what we can that the mean is the on an average sample mean is equal to the population mean that particular statement is not a precise statement for that population statement, we cannot say this is at all a precise statement, there is real reliability of that statement is very, very low, when our standard deviation is very high it means when a variance is very high, is not it?

When a variance is very high, this is the standard error. So, the standard error of σ is very high means my error will be very high. What does the variance of the sampling distribution of mean indicates variance indicates the variance between the different sample means, if we pick different samples from the same population, and if we try to find out the mean the difference between this means is quite high.

If the difference between this mean is quite high, that we cannot just tell mean of all this mean is equals to the population mean we cannot tell claim that so when the variance is high, that is reliable of that statement comes down is not it we have seen that. So, here if we can use this in this 3 data you will see it is punch to get a variance is very less but variable here variance is very high.

(Refer Slide Time: 30:03)

Box plots of the two sets of data

Observation from the Box Plot

- It appears that there is stronger evidence of differences among means in Set 1 than among means in Set 2.
- The observations within the samples are more closely bunched in Set 1 than they are in Set 2

Monalisa Sarma
IIT KHARAGPUR

So, if we see this sets of data, it appears that there is a stronger evidence of difference amongst state 1 then among the means state of 2 in set 1 we can directly see that means of this data, there is a stronger evidence that the means of state 1 are really different with stronger evidence that this district population means are really different, but here though the mean is same to the first set.

But we cannot say that there is no strong if not at all a strong evidence that a mean of this 3 population are different, maybe this 3 populations, this 3 data belongs come from the same population, because you see the variance the observation within the sample are more closely bunched in state 1 than they are in state 2.

(Refer Slide Time: 30:54)

Box plots of the two sets of data

Observation from the Box Plot

- Thus, although the variances among the means for the two sets are identical, the variance among the observations within the individual samples is smaller for Set 1 and is the reason for the apparently stronger evidence of different means.
- This observation is the basis for using the analysis of variance for making inferences about differences among means.

Monalisa Sarma
IIT KHARAGPUR

Thus, although the variance among the means for the 2 sets are identical, the variance among the observation within the individual sample is smaller for set 1 the variance among the

observations within the samples, the variance within the observation it is very small, is not it? The variance among the observation within the individual sample is smaller for set 1 and this is the reason for apparently stronger evidence of different means.

Now, we can understand why we need to use analysis of variance when we try to find out the compare the means of more than two population this observation is the basis for using the analysis of variance for making inferences about differences among means, we have seen this in the first set from the because of the variance, but what does it mean such thing, but in the first set, because the variants are very less.

The data's are really placed apart the from if we see the boxplot from the box plot, we could see that as a really plays a part it has stronger evidence that this the means that this 3 population, this 3 samples come from really 3 different populations, because the variants are very less from that only we could find out that this 3 cannot be from a single population, but in the other case set 2 it maybe we have it is the same type of population.

We have picked the data from the same type of population, but the data we have collected that is giving a different mean maybe because if we see the mean, if you see the variance it is so much the data there is so much varied as if it is it belongs to all the data belongs to 1 single population. So, we use analyzes we have used these variance within simple sample to find a way to tell something about the mean of the population that is the concept behind.

So, this observation is the basis for using the analysis of variance for making inference about the difference among means.

(Refer Slide Time: 33:01)

Idea of ANOVA

The analysis of variance is based on the comparison of the variance among the means of the populations to the variance among sample observations within the individual populations.

So, basically, what is the idea of ANOVA analysis of variance is based on the comparison of variance among the means of the populations, so, this is 1 population this is 1 populations. So, comparison a variance means, what is the variance among this mean suppose, this is this mean is here, this means here, this means here, so, difference between this main difference between this mean.

So, is based on a comprehension of the variance among the population to the variance among sample observation, variance among the sample observation among the samples. So, well rank for in ANOVA what we do we try to find out 2 different means, or 2 different variants, one is variance within the different means and one is variance within 1 sample.

See, this is one sort of variants within the different means, and one is variance within the samples there is variance within this sample.

(Refer Slide Time: 34:20)

CONCLUSION

- ④ In this lecture we had a basic introduction about the applications where ANOVA is required and why t-test is not applicable in those cases
- ④ In next lecture we will learn more about ANOVA

Monalisa Sarma
IIT KHARAGPUR

So, long way to go, so, now, it is almost, we have already exceeded the time, so I will stop this lecture here. So, in this lecture we have a basic introduction about the application way ANOVA is required and why t-test is not applicable in those cases and in the next lecture we will learn more about ANOVA.

(Refer Slide Time: 34:40)

REFERENCES

DESIGN AND ANALYSIS OF EXPERIMENTS
EIGHTH EDITION

④ Design and Analysis of Experiments (8th Edition),
Douglas C. Montgomery, John Wiley & Sons, 2013.

Monalisa Sarma
IIT KHARAGPUR

And this is the reference which I talked about and thank you guys.

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

Lecture - 33
ANOVA-II

Hello guys, so in continuation of our lecture on analysis of variance, so today what we will see?

(Refer Slide Time: 00:30)

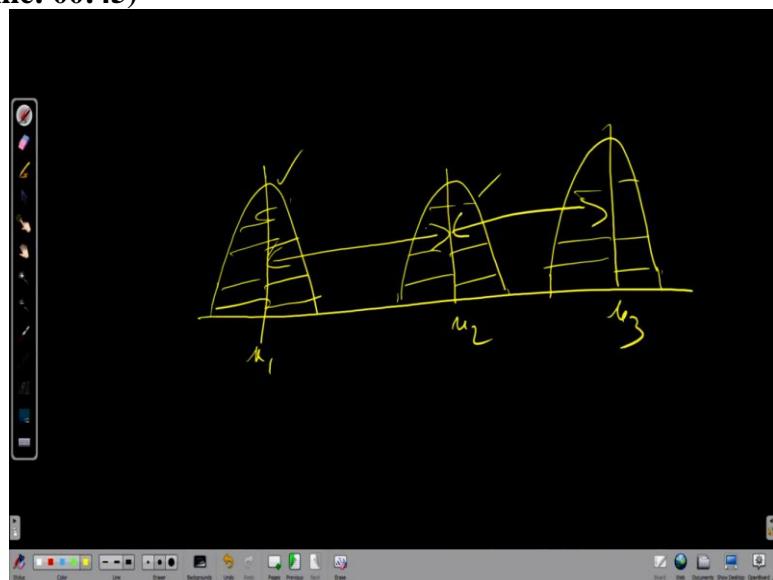
The slide has a dark blue header bar with the title "Concepts Covered". Below the header, there is a list of two items:

- ④ Understanding the difference between group variability and within group variability
- ④ Concept of factors and levels of factors

In the bottom right corner, there is a video feed showing a woman with glasses and a pink shirt speaking. The video interface includes a play button and other controls. At the very bottom of the slide, there is a footer bar with the NPTEL logo, the name "Monalisa Sarma", and "IIT KHARAGPUR".

So today, we will see basically, we will try to understand the difference between group variability and within group variability and the concept of factors and level of factors, why I am talking about group variability and within group variability.

(Refer Slide Time: 00:45)



So, remember when I talked about last class when I talked about this, what is an ANOVA basically means, what ANOVA based on what I discussed that it is based on a comprehension of the variance among the means of the population suppose we have 3 different population from the my 3 different populations, this is my 3 different populations is sampling, sampling of 3 different populations.

So, what is the mean of this my diagram is very bad and very poor in drawing actually. So, this is the mean of this population 1, this is μ 1, this is μ 2 this is μ 3. So, if when we are interested in the we are talking a variance when variants we are interested in variants within this treatment, because, why I am calling it treatment like, let us recollect this example, what we have taken the first example of music being played in the background.

So, what is the treatment? Treatment is the music and how in the music that means, one is with constant music another one is random music another one is no music, suppose this is with constant music. So, this is the mean this is with random music, this is the mean this is with no music, this is the mean. So, mean of this, this is the variance for this variance of 2 different samples, this is the variance of the sample 2 and sample 3.

So, we are interested in this variance as well as we are interested in variance within one samples, this is the variance this variance here also this is the variance. So, basically in ANOVA we say we are interested in finding out a variance within the treatment means as well as the variance within the samples, you know, what is variance you know, right I do not have to discuss that so, now, coming to that, so basically for that we will understand the difference between group variability.

What does group variability indicate that means a variant between 2 different treatments is not it. And within group variability within group variability means variance within a sample.

(Refer Slide Time: 02:59)

So, now, so, consider this distribution of these 2 different samples, this is one sample this is the second sample consider this 2 different samples, but we have seen these samples are quite overlapped. As these individual means would not differ by a great margin, these samples are overlap if we try to find out the individual means, what is this mean? This means, it may not be very much different means.

Because it is overlap there will be difference definitely, but it will not be much difference the difference between the individual means and grand mean can be significant enough now, what is grand mean? It is again a new term we are hearing here, individual mean we know what is the mean of a sample you know that grand mean when we let us go to the same example where I have discussed about music being played in the school.

So, for each class I have taken 10 data. So, 10 data for class A constant music is being played and 10 data for class B well random music, 10 data for class C but no music. So, totally how many data we have? Totally we have 30 data, so, mean of all this 30 we call it as a grand mean. So, if our data is in this form, where it is overlap and the samples overlap, then what happened there individual mean it will not be much different from the grand mean, grand mean understand mean of all the datas.

When the samples overlap there we will find that is obvious the grand mean will be divisible mean will not be much different from the grand mean. So, some what is simple mean we already seen. So, there are 2 kinds of means that we use in ANOVA calculation, one is

separate sample mean that is $\mu_1 \mu_2 \mu_3 \mu_4$ based on the number of population and the grand mean μ grand mean is the mean of all the data.

(Refer Slide Time: 04:51)

Now Consider this Case:

Consider these two sample distributions:

- ⦿ The samples differ from each other by a big margin
 - ⇒ Their individual means would also differ
 - ⇒ The difference between the individual means and grand mean would also be significant

Between Group Variability

Afore-mentioned variability between the distributions called Between-group variability or variance among the means of the populations.

Monalisa Sarma
IIT KHARAGPUR

So, now you considered these 2 distribution sample distribution and again these 2 are the 2 different sample these are the distribution is, so these are distribution that quite apart earlier it was overlap, now it was quite apart no now since it is wide apart the mean what I get from this data, the mean what I get from this data. So, this mean will be different and moreover this mean will be quite different from the grand mean as well, grand mean will be sum of all this.

So, their individual mean would also differ, the difference between the individual mean and grand mean would also be significant in these cases. So, this is there is always a between group variability now, we are discussing between group variability is not it, this is one group this is another group that we are trying to find out the variability between this group, how we are finding out the difference of means is a variable variance upon the groups is not it?

So, afore-mentioned variability between the distribution is called the between group variability or variance among the means of the population. So, this is called between group variability or variance among the means of the populations.

(Refer Slide Time: 06:00)

Between Group Variability

Between Group Variability

Afore-mentioned variability between the distributions called Between-group variability or variance among the means of the populations.

How Between Group Variability is Computed?

Each sample is looked at and the difference between its mean and grand mean is calculated to calculate the variability.

Monalisa Sarma
IIT KHARAGPUR

So, now how between group variability is computed when we are interested in finding the between is sample is looked at and the difference between its mean and the grand mean is calculated to calculate the variability, how we calculate the variability? We find out the grand mean from the grand mean we subtract each mean that is the variability is not it, that is called between group variability.

(Refer Slide Time: 06:23)

Between Group Variability

Point to Remember

1. If the distributions overlap or are close
 - ⇒ The grand mean will be similar to the individual means
2. If the distributions are far apart
 - ⇒ Difference between means and grand mean would be large

Monalisa Sarma
IIT KHARAGPUR

Now, so some points to remember if the distribution overlap are very close the grand mean will be similar to individual mean not exactly similar but similar. So, if the distribution of far apart distribution difference between means and grand mean would also be large.

(Refer Slide Time: 06:47)

Within Group Variability

Variance among Sample Observations

Consider the given distributions of three samples.

Monalisa Sarma
IIT KHARAGPUR

So, now, we are interested in finding out within group variability we have seen what is between group variability between means between 2 groups. Now, we will see within group variability, so, consider given distribution of 3 different samples here we have 3 different samples, this is the distribution of sample 1, this is the distribution of sample 2, this is the distribution of sample 3 see and see the mean of first is x_1 bar, mean of second is x_2 bar, mean of third is x_3 bar.

(Refer Slide Time: 07:22)

Within Group Variability

Variance among Sample Observations

As the spread (variability) of each sample is increased, their distributions overlap and they become part of a big population.

Monalisa Sarma
IIT KHARAGPUR

As the spread of each sample is increased, since the variability is quite high this so, what we see that distribution overlap, their distribution is overlapping and it is as if we have got it from a single population it is as if they are not a different populations. But it is as if it is from a single populations.

(Refer Slide Time: 07:46)

Within Group Variability

Variance among Sample Observations

Now consider another distribution of three other samples but with less variability:

Monalisa Sarma
IIT KHARAGPUR

Similarly, now, consider another distribution of 3 other samples, see the means same x 1 bar, x 2 bar and x 3 bar. Here also what we got mean x 1, bar x 2, bar x 3 bar here variability is less but the mean is almost same.

(Refer Slide Time: 08:03)

Within Group Variability

Variance among Sample Observations

Although the means of samples are similar to the samples in the above image, they seem to belong to different populations.

Monalisa Sarma
IIT KHARAGPUR

So, although the means of the sample are similar to the samples in above the means, they seem to belong to different populations, here it is as if we got it from one population as if all the 3 samples belong to sample because there is overlapping there is no demarcation we cannot really distinguish that this population is different from this, again this is different from this we cannot really demarcate here, it is so much overlap.

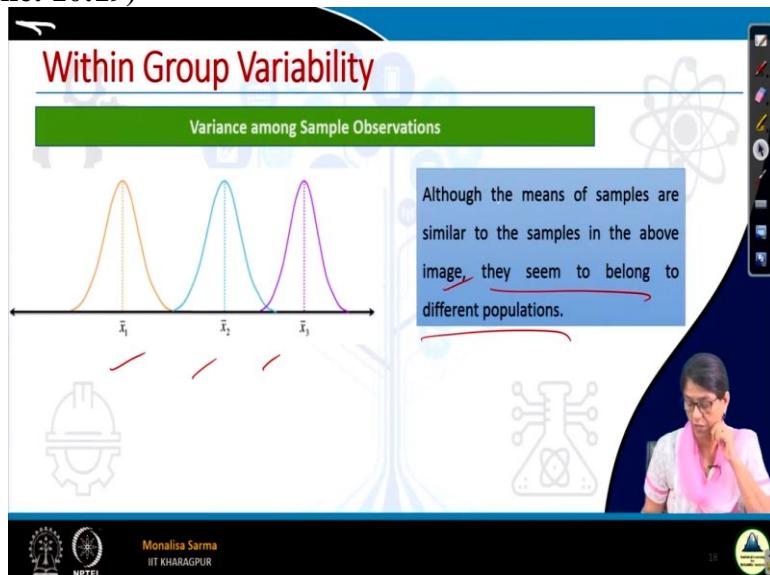
Why it is overlap because of lots of variants within a sample means the samples are very much varied like form very much varied example, let me give you an example like suppose, we are interested in finding out the mileage of a car same type of suppose we are interested

we are trying to see the miles of Verna car for one car we got suppose it is mileage is 20 kilometre and for another we got it is 15 for another we got a 10 another we got it 5 another we got it 8 is variance, it is a very much varied, the data is very varied.

That is what we call it is high variance. So, similarly, we took an example for another car for, for another make not this time to get Verna next time say we took it for another make say I20. So, again here also we try to see them and see the mileage we got a 25 20 15 10 22. So, as if this is belonging to the same population as if it is not 2 different cars, but the same car because the data are so varied, it is not that for Verna.

We are getting that lies within and around 18 say 16 17 18 19 15 very much nearby and for say I20 we are getting say 24 25 23 21 26. It is not that but it was varied data as it derived overlapping as if we are looking for, as if we are considering the mileage of a single make of the car it was we got that feeling for the data. So, this is the case when the data are very much varied. Similarly here when the data says see here we got the same mean, but the datas are not varied.

(Refer Slide Time: 10:19)



So, this particular this belongs this seems to belongs to different populations, but the previous one it does not seem to belong to a different population. So, this is within group variability.

(Refer Slide Time: 10:31)

Within Group Variability

Variance among Sample Observations: Comparison

As the spread (variability) of each sample is increased, their distributions overlap and they become part of a big population.

Although the means of samples are similar to the samples in the above image, they seem to belong to different populations.

Monalisa Sarma
IIT KHARAGPUR

So, although the spread as the spread of each sample is increased their distribution overlap and they become part of a big population, although the means of the samples are similar to the samples in the HAVOC means, they seem to belongs to different population here as if they belong to different population different A, B and C here as it the belongs to simple single populations. So, this is within group variability.

(Refer Slide Time: 10:59)

Example

Problem

Suppose in an industrial experiment an engineer is interested in how the mean absorption of moisture in concrete varies among 5 different concrete aggregates. The samples are exposed to moisture for 48 hours. It is decided that 6 samples are to be tested for each aggregate, requiring a total of 30 samples to be tested. The data are recorded in Table:

Aggregate	1	2	3	4	5	Total	Mean
551	595	639	417	563			
457	580	615	449	631			
450	508	511	517	522			
731	583	573	438	613			
499	633	648	415	656			
632	517	677	555	679			
Total	3320	3416	3663	2791	3664	16,854	561.80
Mean	553.3	569.33	610.50	465.17	610.67	561.80	

Monalisa Sarma
IIT KHARAGPUR

Now, let us see an example, suppose in an industrial experiment, an engineer is interested in how the mean absorption of moisture in concrete barriers among 5 different concrete aggregates, so, what it has is it has 5 different concrete aggregates basically 5 different concrete aggregates means there are 5 different concretes in 5 different aggregates it has used different maybe different chemicals, 5 different chemicals to the aggregates, concrete aggregates.

So, use different chemical store on their total same, we have taken some concrete aggregates, and they we have used some chemicals, we use 5 different chemicals, that is why we have got 5 different concrete aggregates. And what we are interested in finding out the samples are exposed to moisture for 48 hours it is decided that 6 samples are to be tested for each aggregate requiring a total of 30 samples to be tested the data are recorded in the table.

So, we are interested in finding out what is the moisture absorption rate of each aggregate. So, more the moisture observation more washes that we do not want that kind of concrete. So, absorbed more and more moisture if it absorb more moisture than what happened with the dampness will be there in the walls. So, we do not want that type of concrete way which absorbs more and more moisture.

So, that is why we have done treatments in 5 different treatments 5 basically different chemicals, we have used we are interested in finding out is the by using this chemicals which one will give lesser absorption of moisture. So, we have tested for that accordingly we have got this data, these are the data for different 5 different types of concrete is concrete 1 2 3 4 5, these are the moisture absorption of moisture in 48 hours. We found the mean and then we have found the total and then we found out the mean now this is the grand mean as I was talking about.

(Refer Slide Time: 13:06)

Statistical Test

Hypothesis

We may wish to test

$H_0: \mu_1 = \mu_2 = \dots = \mu_5,$

$H_1:$ At least two of the means are not equal.

Monalisa Sarma
IIT KHARAGPUR

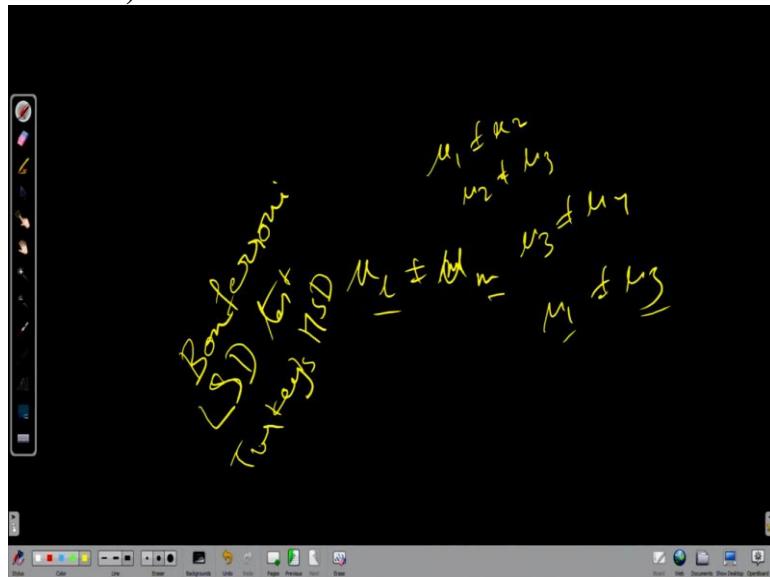
So, now we are interested in finding out in this question we are interested in finding out whether the absorption of moisture in all these 5 concrete types is different or not, whether any because of this different type of mixture we are using because of that will absorption

moisture will be different or it is the same. So, when we are interested in finding out this sort of thing definitely we will do hypothesis test what we have already seen when we are doing for single population or double populations.

Now here what will be the hypothesis test? The hypothesis test the null hypothesis is that all the population has equal mean that is our null hypothesis all the population will have equal mean now, then what will be the alternate hypothesis? Now, alternate hypothesis is at least 2 or more means or at least 2 or more means are not equal that means, here our null hypothesis we are telling that all the means are equal and alternate hypothesis at least 2 of the means are not equal there may be 3 not equals, 4 not equals, 5 not equals.

But at least 2 of the means are not equal this will tell that there is some difference in the treatment that we have done.

(Refer Slide Time: 14:27)



Now, the question is, if we asked me here the question is like if we asked me what we are trying to find out at least 2 of the means are not equal. That means I can write it as that is μ_1 not equals to μ_m , where 1 and m maybe any of the variables any of the samples that we have considered, 1 n maybe may come from any of the samples that belong to any of the 2 sample means, like say $\mu_1 \neq \mu_2$ or $\mu_2 \neq \mu_3$, or maybe $\mu_3 \neq \mu_4$ or maybe μ_1 not equals to μ_ϕ like that any 1 n m can take any value from 1 to 5.

So, now, the thing is that in this test what we have studied we will be able to only find out if at least 2 of the means are not equal, but then we will not be able to find out which mean is

not equal or which 1 or which 2 or which 3 means are not equal, that is basically for doing that, there is a different type of experiment for that, which is beyond the scope of this lecture, I will not be discussing that.

Just so, you can just note it down the different what to say techniques for doing this is one technique is called Bonferroni approach than another it is let me I will just write it Bonferroni approach if I am writing it f e double r r, Bonferroni approach another is a least significant this at least significant difference test least significant difference test least significant LSD least's significance different tests.

Another is tukeys t u k e y tukeys HSD, HSD represent honest significant difference. So, these are the 3 diff there are more there these are different type of tests by weeks, we can find out among the tests like among the different means, which mean is actually different or which 2 are which of the means maybe one main difference of 2 or 3 which are the means are actually different in analysis of variance we will not be able to tell which are the means are different.

We will be just able to tell where all the means are equal or there is any difference among them means, we will analyse the variance over just limited to that whether all the means are equal or any of the means are not equal. If you are some specifically find wants to if we found out suppose in this case, our alternate hypothesis is true that is at least 2 other means are not equal.

Then we will have to further do some further exponent which I have just mentioned, I have written it down also you can just note it down. So, we will further some tests to find out exactly which of the means are not equal anywhere. In most of the experiment that is necessary then we will do have to do that and most of experiment when we really do not need to do we are just it is sufficient enough that all the populations are same or all the populations are different.

Analyse ANOVA restrict to that just finding out whether all the populations are equal or all the population are not equal. So, that is why this is our H_0 is a null hypothesis alternate hypothesis, at least 2 of the means are not equal.

(Refer Slide Time: 17:44)

Statistical Test

Variation among the Aggregate Averages

In ANOVA procedure, it is assumed that whatever variation exists among the aggregate averages is attributed to:

- Variation in absorption among observations within aggregate types, and
- Variation among aggregate types, that is, due to differences in the chemical composition of the aggregates.

NPTEL
Monalisa Sarma
IIT KHARAGPUR

So, in the variation among the aggregate averages, we have seen in this example, among the aggregate averages, there is quite a variance that is 553 569 610 465 there are quite variance so, if the variance among the aggregate right is variance exist and another procedure it is assumed that whatever variance exists among the aggregates where we have seen variance among the aggregates it may be attribute due to first one is variation in absorption among observation within aggregate types.

In one aggregate type only chemical suppose we have used chemical A, B, C, D, E in one set of samples we have use chemical A only but there also in one sample only where we have used chemical A only there supposed to we have 10 products using chemical A we have use a sample size of 30, 30 or 5 whatever we have used here suppose here we have used 30 samples when 30 samples we have used chemical A.

Among these 30 also then all moisture absorption may be different some countries may have absorbed more moisture, some countries may have absorbed less moisture, but why variation among aggregate type that is due to the difference in one. This difference in variation may be due to the variation among observation and all variation among aggregates type that is due to the difference in a chemical composition.

This variance may be due to 2 different factors 2 different reasons one is variance because of the different chemical composition, in one type of population we use A another B, another C, another D and another E this variance may be due to the different chemical composition you have used that means 1 variance that is we call it between group variability and another

variance may be within the same group. Variation in absorption among observation may be different, why?

(Refer Slide Time: 19:47)

The within aggregate variation is, of course, brought about by various causes:

- Perhaps humidity and temperature conditions were not kept entirely constant throughout the experiment.
- There was a certain amount of heterogeneity in the batches of raw materials that were used.
 - At any rate, the within-sample variation is considered to be chance or random variation.

To determine if the differences among the 5 sample means are due to random variation alone or, rather, due to variation beyond merely random effects, i.e., differences in the chemical composition of the aggregates.

There are reason for that I have written I think I have the causes here. The within aggregate variation is of course brought about by various causes perhaps humidity and temperature conditions were not kept entirely constant for the experiment. Maybe because of that some concrete has absorbed the same chemical, chemical A we have used for all the 30 but some concrete have used less moisture, some concrete have use more moisture.

Maybe because the condition was not constant because we have conducted experiment for 48 hours condition maybe not constant throughout or maybe some concrete I have kept in one room. So, another concrete I have kept in another room and maybe the temperature is not, environment is not same in both the room and there may be certain amount of heterogeneity in the batches of raw materials.

The raw materials that we; observe the sample that we have taken up 30 samples for each different 5 categories 30×5 . So, this 30 samples what we have taken so, 30 samples well we have used maybe the cement we have bought similar type of cement for different company. So, the different companies so there will be some slight variance.

So, there will be certain amount of heterogeneity in the batches of raw materials for that maybe chemical absorption may be more or less by using the same chemical treatment at any rate that within rate sample variation is considered to be chance or random variation. So,

within sample variation we call it a chance variation or random variation. And variation within the variation between groups we call it what it is? It is a variation because of the treatment different treatment.

So, to determine now, what we have to determine if the difference among the 5 sample means are due to random variation alone or rather due to variation beyond merely random effects that is difference in the chemical composition of the aggregates. Consider the music example, the difference in the test scores is only because that we have picked some intelligence students in group A.

And in group B and C we have might be we have picked some dull student, is it the differences because of that or the difference is because of the different type of music we have played in the different classes. Similarly here, so is this difference in the mean that we have seen is due to the random variation alone, random variation what we have discussed, random variation maybe the moisture content with the same type of chemical may be different.

Because of the heterogeneity of the raw material, because of the environment where we have put. So, is it a variation what we have seen is the variation is because of the random variation or because of the different treatment that we have given. So, we need to determine that. So that is what we will do in ANOVA. ANOVA does essentially that.

(Refer Slide Time: 22:43)

The slide has a blue header bar with the title 'Some Terminologies'. Below the title are two green boxes with definitions:

- What is Factor?**: A characteristic under consideration, thought to influence the measured observations.
- What is Level (also called treatment)?**: Level is a value of the factor.

To the left of the text, there is a gear icon with the text 'Typical data for a Single-Factor Experiment:' below it. In the center, there is a table:

Level	Observations			Total	Mean
1	y_{11}	y_{12}	...	y_{1n_1}	
2	y_{21}	y_{22}	...	y_{2n_2}	
...	
...	
...	
a	y_{a1}	y_{a2}	...	y_{an_k}	

A red circle highlights the first two columns of the table, specifically the headers 'Observations' and the first two data points y_{11} and y_{21} . To the right of the table, there is a video feed of a woman speaking, and at the bottom, there is a footer with the NPTEL logo and the name 'Monalisa Sarma IIT KHARAGPUR'.

So, when we discuss and what the first thing we need to know, what is a factor, as I have already told you that this music, no music, constant music that is what that is factor we can

call it also treatment or else we are changing one factor that is music, either we are playing constant music, random music or no music. And this just recent example, we are using there is one factor, factor is what can we use it type of chemical.

Chemical A, B, C, D, E. So, in the first case music example, there is one factor how many levels are there, there are 3 levels 1 2 3 levels no music, constant music, random music 3 levels here, there is one factor that is we have used chemical, chemical is one factor, how many levels are there? There are 5 different levels 5 different chemicals we have used level is this basically the value of the factor.

So, if we see a typical data set for a single factor experiment, so this is what we have discussed is always a single factor. So, single factor, we can have different levels for each level will have different set of data, so this is a typical data set we can say.

(Refer Slide Time: 23:55)

Variants of ANOVA

Based on the number of Independent Variables and Dependent Variables considered for the study, there are different variants of ANOVA

- One-Way ANOVA:** Only one independent variable (factor) with greater than 2 levels on one dependent variable.
- Two-Way ANOVA:** Two independent variables (i.e., factors) on one dependent variable.
- Three-Way ANOVA:** Three independent variables (i.e., factors).
- Multivariate ANOVA:** It is used to test the significance of the effect of one or more independent variables on two or more dependent variables.

Now, the ANOVA we will go to details of ANOVA first before that, there are basically different variants of ANOVA based on the different dependent variables and independent variables. Now in our example on music, which is the dependent variable and which is the independent variable, our dependent variable is the test score. The test score is dependent on something the dependent variable is a test score for independent variable what? Independent variable is the music.

Music we are changing, we are giving this, we are trying to we are playing on that, that is independent, that is not depending on something we are not changing depending on variances

independently changing that. So that is the independent variable and dependent variable is our score. So, in that example, what we have? We have the music exam example we have one independent variable that is the music and one dependent variable that is the test score.

So, this is called one way ANOVA. Only one independent variable or factor with greater than 2 levels, we have 3 levels there and one dependent variable that is called one way ANOVA, two way ANOVA 2 independent variables on one dependent variables 2 where 2 independent variables. Suppose the same example music example I have to, as I have already mentioned before.

Suppose I used A is also as a factor, I have tried this music experiment on the older students as well as younger students, here also A is independent, I am just changing the A's, I am bringing in whole class 10 students or class 2 students I am just saying it is not dependent on anything. So, likewise, music age is also has a independent factor. So, based on these 2 music, and A s how it is affecting the test score, test score is dependent on these 2 music as well as age.

So, this is the one dependent variable, 2 independent variables. So, this is called two way ANOVA similarly, we have three way ANOVA, 3 independent variables and one dependent variable. So, similarly, we have multivariate ANOVA it is used to test the significance of the effect of one or more independent variables on 2 or more dependent variables and it is called multivariate variables 2 or more dependent variables.

Let us take the recent example what we have taken that concrete mixer. Now here we are interested in, we are mixing some chemical and we are interested in finding out the absorption of moisture suppose, we are also interested in finding out the setting time, setting time of this concrete we are mixing some chemical, chemical A B C D and E how this chemical is affecting the moisture absorption as well as how this chemical is affecting the setting time.

So, here independent variable is 1 that is the chemical treatment that we are using, but we have 5 different chemical treatment level and dependent variable is setting time as well as the moisture absorption. So, similarly there can be more than one independent variable and more than one dependent variable. So, that is called multivariate ANOVA. In this lecture, basically

in this course, we will be discussing only one way ANOVA, two way ANOVA and three way ANOVA it is very complicated to discuss in this class.

Once we know the integrity of one way ANOVA it will be very easy for you to just Google it and learn by yourself actually two way ANOVA and our 3 way ANOVA it is very difficult to do it manually also you will let it really need to do using computer. So, that is discuss in this class two way ANOVA, three way ANOVA and multivariate ANOVA which I assure you once you know one way you will be able to learn it by yourself.

(Refer Slide Time: 27:30)

CLAIM

A recent study claims that using music in a class enhances the concentration and consequently helps students absorb more information.

What if it affected the results of the students in a negative way?
Or,
What kind of music would be a good choice for this?

Need to design a specific experiment to have some proofs that it actually works or not

Monalisa Sarma
IIT KHARAGPUR

So, now, we will just quickly revisit some example, in the context which have already just mentioned in the context of one way, two way and three way, the recent study about the music what I told. So, I have already explained here like what is if I call it is a one way ANOVA does what we have discussed now that is a one way ANOVA that is music as an independent variable with 3 different levels.

(Refer Slide Time: 27:51)

Example 1: Design of the Experiment

- ④ The teacher decided to implement it on a smaller group of randomly selected students from three different classes.
- ④ Three different groups of ten randomly selected students from three different classrooms were taken.
- ④ Each classroom was provided with three different environments for students to study.
 - ④ Classroom A had constant music being played in the background
 - ④ Classroom B had variable music being played in the background
 - ④ Classroom C was a regular class with no music playing
- ④ A test was conducted for one month for all the three groups and their test scores were collected after that.



Monalisa Sarma
IIT KHARAGPUR



And if I consider A is also then that is again one way ANOVA but sorry that is two way ANOVA where I have 2 independent variable and one dependent variable.

(Refer Slide Time: 28:03)

Example 2

- ❑ A car magazine wishes to compare the average petrol consumption of THREE models for car and has available SIX vehicles of each model.
- ❑ There are THREE populations
- ❑ There are samples each of size six from each population

Model 1	Model 2	Model 3



Monalisa Sarma
IIT KHARAGPUR



Similarly this if a car magazine wishes to compare the average petrol consumption of 3 models of the car and has available 6 vehicle of each model. So, we are interested in comparing the petrol consumption of 3 different models model 1 model 2 model 3 petrol; consumption we are interested in finding out the petrol consumption. So, what is the dependent variable here? Our dependent variable is the petrol consumption.

How much is the petrol consumption that is our dependent variable, it is depending on what? It is depending on the model, model 1 model 2 model 3 3 different models. So, here it is again one way ANOVA our independent variable is the model dependent variable is the petrol consumption suppose, again here again I have used 3 different drivers one is a very

young and another is sitting driver one is competitively old driver, one is a very old person basically and another is a very new driver for drive.

So, there again I have introduced 2 independent variables that drivers have 3 different levels, so models and driver becomes 2 independent variable and the petrol consumption is one dependent variables. So, this is again a two way ANOVA.

(Refer Slide Time: 29:15)

Example 3

Aggregate	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Mean	553.	569.33	610.50	465.17	610.67	561.80

Problem
Suppose in an industrial experiment an engineer is interested in how the mean absorption of moisture in concrete varies among 5 different concrete aggregates. The samples are exposed to moisture for 48 hours. It is decided that 6 samples are to be tested for each aggregate, requiring a total of 30 samples to be tested. The data are recorded in Table:

Monalisa Sarma
IIT KHARAGPUR

Similarly this example which I have already discussed in the context of one way ANOVA as well as the multivariate ANOVA, if I consider setting time as well as the moisture absorption it becomes one independent variable, 2 dependent variable So, we call it a multivariate ANOVA.

(Refer Slide Time: 29:31)

CONCLUSION

- ④ In this lecture we learnt about
- ④ The difference between group variability and within group variability
- ④ Concept of factors
- ④ Levels of factors
- ④ In the next lecture we will learn about One-Way ANOVA

Monalisa Sarma
IIT KHARAGPUR

So, in this lecture we have learned about difference between group variability and within group variability we will also learn so, what is a factor, what are the independent variable, dependent variables and levels of the factors also called treatment mind it. In some book may get it as factor in some Google get it as treatments. So, it is a concept of treatment or factor and the different levels of this treatment are different levels of factor.

(Refer Slide Time: 29:59)

CONCLUSION

- ④ In this lecture we learnt about
- ④ The difference between group variability and within group variability
- ④ Concept of factors
- ④ Levels of factors
- ④ In the next lecture we will learn about One-Way ANOVA

Monalisa Sarma
IIT KHARAGPUR

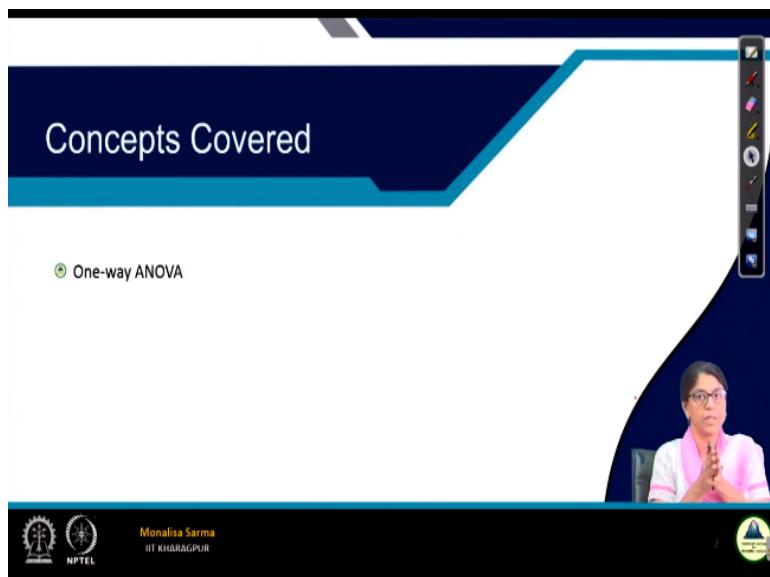
In the next lecture, we will learn about one way ANOVA. So, there is the reference and thank you guys

Statistical Learning for Reliability Analysis
Dr. Monalisa Sarma
Subir Chowdhury of Quality and Reliability
Indian Institute of Technology – Kharagpur

Lecture – 34
ANOVA - III

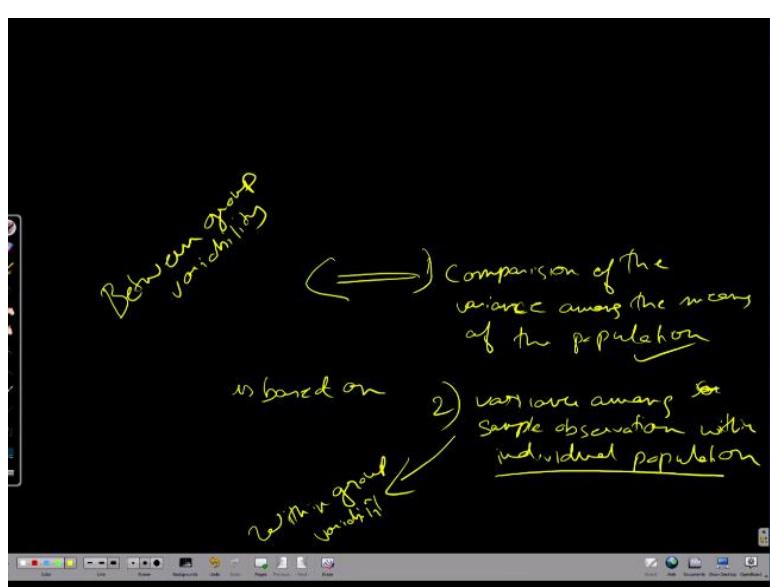
Hello guys. So, we were discussing analysis of variance in short ANOVA so in today's discussion and continuation of our earlier discussion, we have already taken 2 classes on analysis of variance today is the third class on that and continuation of that.

(Refer Slide Time: 00:43)



Today we will be discussing one way ANOVA before that, let us take a quick recap let me use the board here.

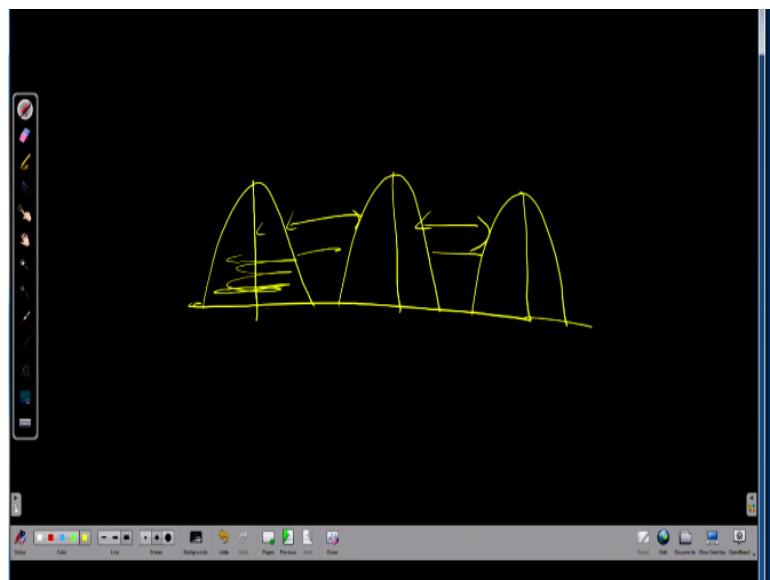
(Refer Slide Time: 00:56)



So, remember what we have seen the analysis of variance is based on let me write it is basically based on 2 factors we have seen what are those 2 factors? Let me write it here one is the comparison of the variance among the means of the population first let me write it and then the second one is like variance among sample observation within individual population we have seen this analysis of variance is based on what one is the comparison of the variance among the means of the populations, this is what we call it this thing?

We call it between group variability or between treatment variability, anything this we call is between group variability and this second one that is variance among the sample observation within individual population, this is what this is called within group variability.

(Refer Slide Time: 03:06)



So, if we basically draw the diagram, if I draw the diagram say this is one sample this is another sample, I am giving the distribution of the sample basically this is one sample. So, this is the mean this is the mean this is the mean. So, first one is within when I tell is between group variability as between group variability means, this variability between this group variability between these 2 groups, variability between these 2 groups this is called between group variability and another is within group variability.

Within group variability this variability within the group that mean then what is the standard deviation basically, standard deviation is what is the how much it deviates from the mean each and every data how much it deviates from the mean. So, this is called within group variability and these things are called between group variability. So, we had analysis of