

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

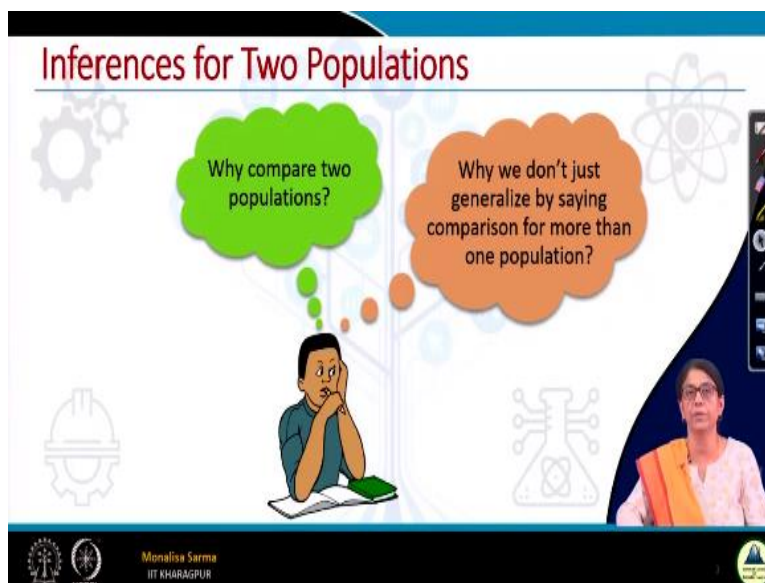
Lecture – 29
Statistical Inference (Part 6)

(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 is titled "Inferences for Two Populations" in red text. Below the title, a green box contains the text "Many interesting applications involve only two populations:". To the right of this box is a faint atomic symbol icon. Below the green box, there are three examples of two-population comparisons:

- On the left, a pink female icon and a blue male icon are shown next to a blue box containing the text "Any comparisons involving differences between the two genders".
- In the center, an orange box contains the text "Comparing a drug with a placebo", with an image of two pill bottles labeled "sugar drug" and "placebo" to its right.
- At the bottom left, two side-by-side images of a street are shown, labeled "BEFORE" and "AFTER", next to a blue box containing the text "Comparing before and after some event etc.". To the right of this box is a faint icon of a person's head.

At the bottom of the slide, there is a black footer bar containing the NPTEL logo on the left, the name "Monalisa Sarma" and "IIT KHARAGPUR" in the center, and another logo on the right. A small inset image of a woman speaking is visible in the bottom right corner of the slide area.

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 is titled "Inferences for Two Populations" in red text at the top. It features a central blue box with two bullet points: "Some of the concepts underlying comparing several populations are more easily introduced for the two-population case." and "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." To the left, a cartoon student is thinking, with two thought bubbles: "Why compare two populations?" and "Why we don't just generalize by saying comparison for more than one population?". On the right, a woman is presenting. The bottom of the slide shows logos for NPTEL and Monalisa Sarma at IIT Kharagpur.

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 somebody 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)

Comparison of Populations

1. The populations are actually different.
2. The populations are a result of an experiment.

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 is titled "Comparison of Populations" in red text at the top. Below the title, there is a blue box containing the text "1. The populations are actually different." To the left of this box is an orange box with a white border containing the text: "Example: Male and female students -- a study involving separate populations is in general known as observational study." To the right of the orange box is a diagram showing two rows of green human icons. The top row consists of six female icons, and the bottom row consists of six male icons. In the bottom right corner of the slide, there is a small video feed of a woman with glasses, wearing an orange and white sari, who appears to be presenting. The slide also features a logo of an atom in the top right and logos of IIT Kharagpur and NETS at the bottom.


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.



 Menalisa Sarma
 IIT KHARAGPUR

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 $2n$ 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: Inference on Difference Between Means


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.

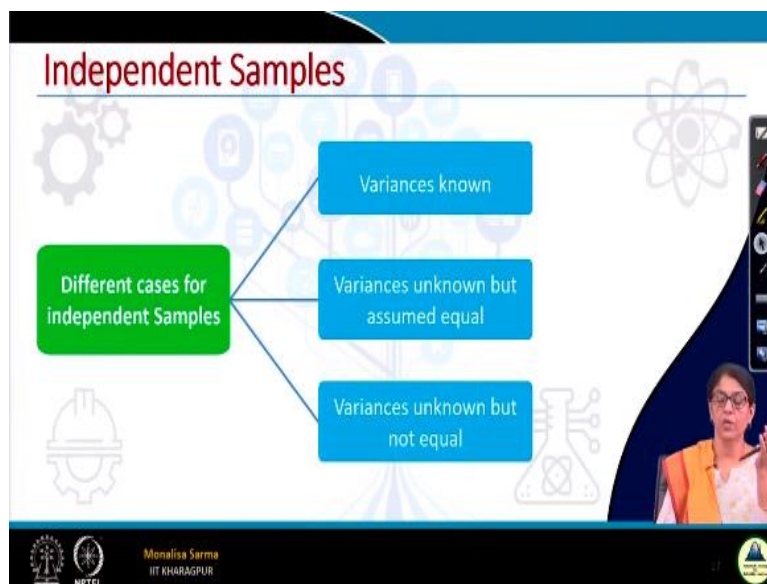
All my condition is that I want to check whether my both the population means are equal then my null hypothesis I can also write it as $\mu_1 - \mu_2 = d_0$ or I can also write as $\mu_1 = \mu_2$, that is my null hypothesis. Because then my alternate hypothesis whatever I want to check whether I want 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 $\mu_1 \neq \mu_2$. 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)

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 $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

\therefore The statistic $Z = \frac{\bar{X}_1 - \bar{X}_2 - d}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$ has the standard normal distribution

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

Monalisa Sarma
IIT KHARAGPUR

So, to make inference on the difference of 2 means, as I told you, this is my test statistics. So, now, this test statistics $\bar{X}_1 - \bar{X}_2$ what it will have $\bar{X}_1 - \bar{X}_2$, 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, $\bar{X}_1 - \bar{X}_2$ will have a normal distribution and what will be the mean of \bar{X}_1 mean of that distribution that is the sampling distribution.

Mean of the distribution is $\mu_1 - \mu_2$ is not it? And variance is $\sigma_1^2/n_1 + \sigma_2^2/n_2$ is not it? σ_1^2/n_1 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 $\sqrt{\quad}$ 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 $\bar{X} - \mu$ by σ by \sqrt{n} this is my z value is not it? Similarly here it is where \bar{X} is the mean of the sample. So, now here what is the mean of my sample? Mean of my sample is $\bar{X}_1 - \bar{X}_2$ I am trying to find out the difference of means, that is why my mean of the sample is $\bar{X}_1 - \bar{X}_2$.

And what is my μ ? μ is the population mean is not it? So, here what I have hypothesis? I have hypothesis that $\bar{X}_1 - \bar{X}_2 = d$ 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 by \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 $Z = \frac{\bar{X}_1 - \bar{X}_2 - d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ 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 $\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \sigma \sqrt{n}$. So, this is $\sigma \sqrt{n}$, 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 $\bar{x}_1 = 81$. A second random sample of size $n_2 = 36$, taken from a different normal population with a standard deviation $\sigma_2 = 3.4$, 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

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 $\mu_1 \neq \mu_2$ 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:

$$H_0: \mu_1 = \mu_2$$

Handwritten notes: $\mu_1 - \mu_2 = d$
 $\mu_1 - \mu_2 = 0$
 $\mu_1 \neq \mu_2$

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 d 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

$$Z = \frac{81 - 76}{\sqrt{\frac{(5.2)^2}{25} + \frac{(3.5)^2}{36}}} = 4.22$$

Handwritten notes: $z = \frac{x_1 - x_2 - d}{\sigma / \sqrt{n}}$
 4.22

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 $\mu_1 \neq \mu_2$. 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 5 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

$$Z = \frac{81 - 76}{\sqrt{\frac{(5.2)^2}{25} + \frac{(3.5)^2}{36}}} = 4.22$$

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 $\sigma_2 = 3.4$, 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.

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.

In fact, 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 Sarma
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

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 deviation 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 instead of σ_1 S_2 instead of σ_2 .

We can very well use S_1 and S_2 . So, can we really use t distribution using the 2 variance estimate S_1 S_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 may be 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.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\frac{\bar{x} - \mu}{S/\sqrt{n}}$$

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 $n_1 - 1 s_1^2 + n_2 - 1 s_2^2$ 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 $\bar{X} - \mu$, $\bar{X} - \mu$ whatever it is μ is the population mean then divided by $s_1 / S / \sqrt{n}$, is not it?

So, said same as $\bar{x}_1 - \bar{x}_2 - d_0$, 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 $\bar{x} - \mu / S / \sqrt{n}$ this is the formula for t distribution for single population now, instead of $\bar{x} - \mu$ 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 $\sigma_1^2/n_1 + \sigma_2^2/n_2$ remember. So, similarly here we are using S_p^2 S_p^2 S_p 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 statistics 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.

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.

$n_1 = 12$
 $n_2 = 10$

Monalisa Sarma
IIT KHARAGPUR

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 buy 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 population 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.

Handwritten notes in red ink:
 $\mu_1 = 85$
 $\mu_2 = 81$
 $12 + 10 = 22$
 20

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

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad v = 20, \quad S_p = 4.470$$

$$\Rightarrow t = 1.04$$
- 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.

Monalisa Sarma
IIT KHARAGPUR

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.

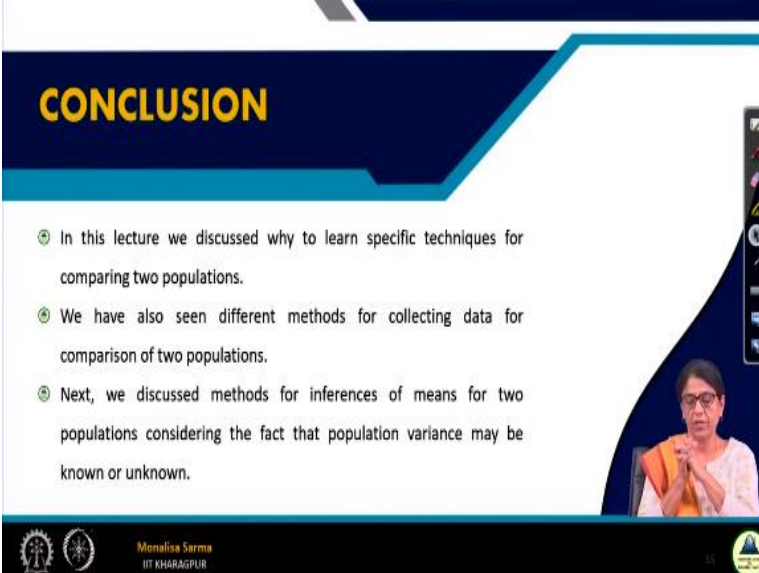
But what happens in the forest where the size of trees are more, that means it mean we will be more mean is more usually since the variance is also more here, the mean is small variance is also small. But in that case, if you so you can tell that there are two population variances different that may 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)



CONCLUSION

- In this lecture we discussed why to learn specific techniques for comparing two populations.
- We have also seen different methods for collecting data for comparison of two populations.
- Next, we discussed methods for inferences of means for two populations considering the fact that population variance may be known or unknown.

Monalisa Sarmah
IIT KHARAGPUR

So, in this lecture what we have seen we have learn 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

Four book covers are displayed in a row:

- Probability & Statistics for Engineers & Scientists** (5th Edition) by Walpole, Myers, Myers, & Ye. Publisher: Prentice Hall.
- STATISTICS** (2nd Edition) by Winer, Brown, & Michener. Publisher: Wiley.
- Probability and Statistics** (3rd Edition) by Ross. Publisher: John Wiley & Sons.
- Statistical Methods** by Freund & Wilson. Publisher: Academic Press.

At the bottom of the slide, the text reads: **Monalisa Sarma**, IIT KHARAGPUR.

So, with that I end this lecture. Thank you. Thank you guys.