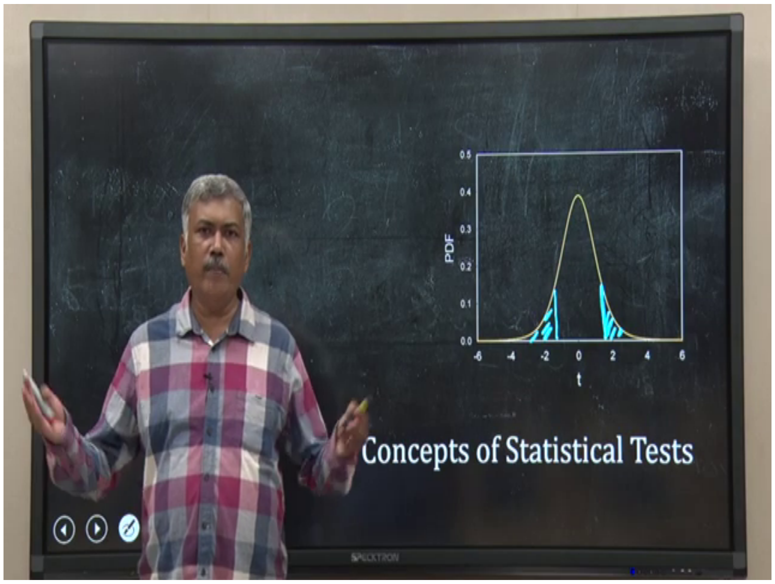
Data Analysis for Biologists
Professor Biplab Bose
Department of Biosciences & Bioengineering
Mehta Family School of Data Science & Artificial Intelligence
Indian Institute of Technology, Guwahati
Lecture: 7
Concept of Statistical Tests

(Refer Slide Time: 00:32)



Concepts of Statistical Tests

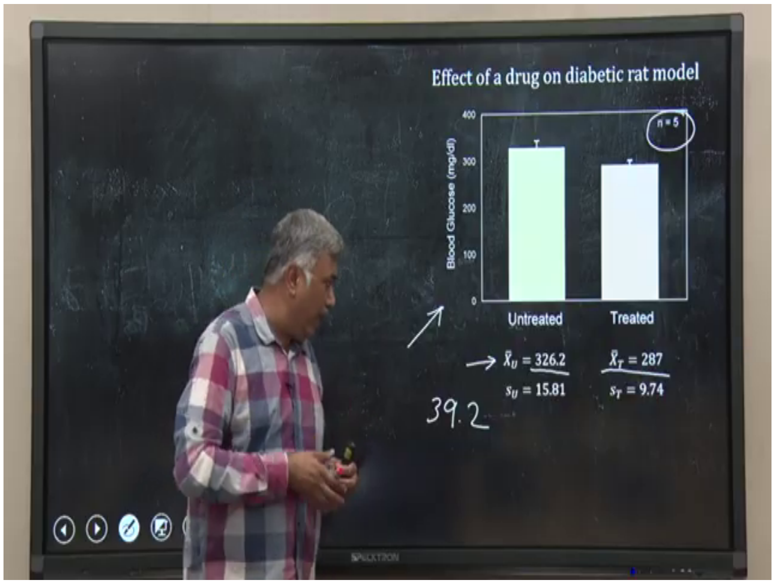
Welcome back. Statistics is an integral part of biology. Every time you do a experiment, you have to use some sort of statistics. For example, from data you calculate a mean, variance, standard deviation, those are actual statistics of your data. Then suppose you are doing a population-based study, for that you may have to decide the sample size.

So, you have statistical tool to decide the sample size or suppose you have done a, created a large data set from a population and then you want to segregate the outlier and remove them from your analysis. Again, you have Statistical tools for that, and one of the commonest statistics that most of the experimental biologists use somehow is test of hypothesis, test of significance.

So, statistics is something which is integral in biology and I will always advise every student of biology to learn statistics, this course is not a biostatistics or statistics course, there are courses available on statistics and biostatistics, I will advise you to join those course and go through that. But, even though this course is not a statistics course, there are, certain statistical tool that we have to use for data analysis in the subsequent days.

So, in this lecture, what I will do, I will discuss about one of the very crucial thing in statistics in biology that we use is the concept of statistical test, hypothesis testing. So, I will focus more on, not on all possible statistical tests of hypotheses that as they are and people use them, but rather I will take one example T test and through that example, I will try to convey the basic key concepts of test of hypothesis. So, just like any other lecture, let us start with a data set.

(Refer Slide Time: 02:15)

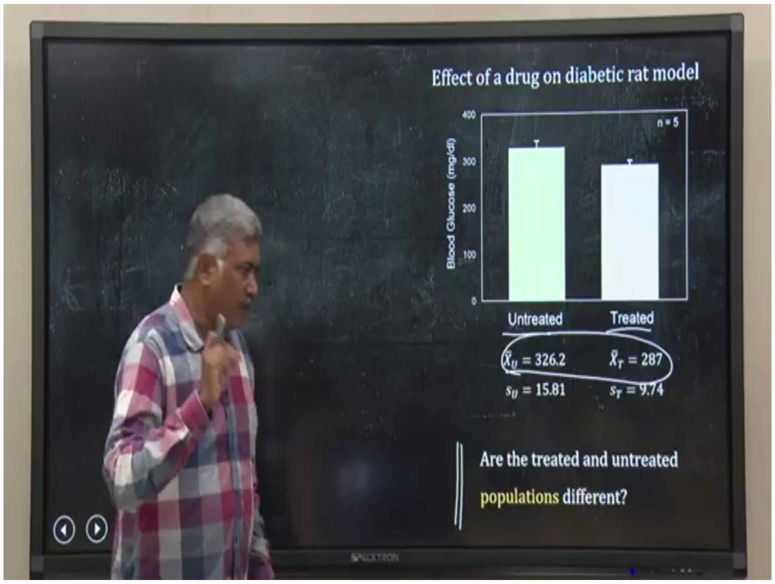


Effect of a drug on diabetic rat model

$\bar{X}_U = 326.2$ $\bar{X}_T = 287$

$s_U = 15.81$ $s_T = 9.74$

39.2



Effect of a drug on diabetic rat model

$\bar{X}_U = 326.2$ $\bar{X}_T = 287$

$s_U = 15.81$ $s_T = 9.74$

Are the treated and untreated populations different?

$$\overline{X}_U = 326.2 \quad \overline{X}_T = 287$$

$$s_U = 15.81 \quad s_T = 9.74$$

So, suppose we are assaying a set of molecules, which are supposed to be having anti diabetic effect. So, we have a diabetic rat model and we have taken one molecule and we are treated those, a group of diabetic rat with that and we have a control one which is untreated.

So, the data is shown here right, I have measured the blood glucose level and I have made a bar plot conventionally that type of data you will see in papers and presentation we have untreated and treated group. Now, I have also reported the mean, remember this is called sample mean that is why X u bar we have, sample mean for blood glucose level for untreated is 326.2 whereas, for treated group it is 287.

So, how much will be the difference, difference will be around 39.2 possible you can do the calculation. So, I have seen the difference, there is a difference in the treated and untreated group, the treated group has slightly lesser blood glucose level. If I show you this data, obviously, you can ask me a question, is it a fluke experimental error or is real?

I can go back to my lab and can repeat the experiment again, 2-3 times and again come with a similar data set to you with bar plot and I can try to convince you but remember, that does not solve the problem because this question that you are asking whether this result is real or is just a fluke is connected to something more basic.

Remember the idea of population and sample, this experiment has not been done on a population of rat, this experiment has been done on a sample of n, n equal to 5, so, I have taken 5 rat, those rat are diabetic, and we have not done anything to them, we call them untreated group. Whereas I have taken another set of diabetic rat 5 in numbers.

And we have treated them with drug we call them treated group. In both the cases, these groups and these 5 rats in each group are nothing but samples. But what we are trying to know what you are asking me for after looking at data, you are asking that whether at the population level, these difference between treated and untreated group will hold or not.

So, you are asking me at the population level, will we see any difference. Now, remember I do not know what is the population, population is very large, we cannot do the experiment with the whole population. We do not know the mean variance, all these things of the population either. As we do not know those statistics of the population, those parameters of the population

We are doing experimental samples and from the sample I want to project to population. So, the key question that people will ask when they will see this data is that, are the treated and untreated populations are different, not the samples. So, how should we answer it?

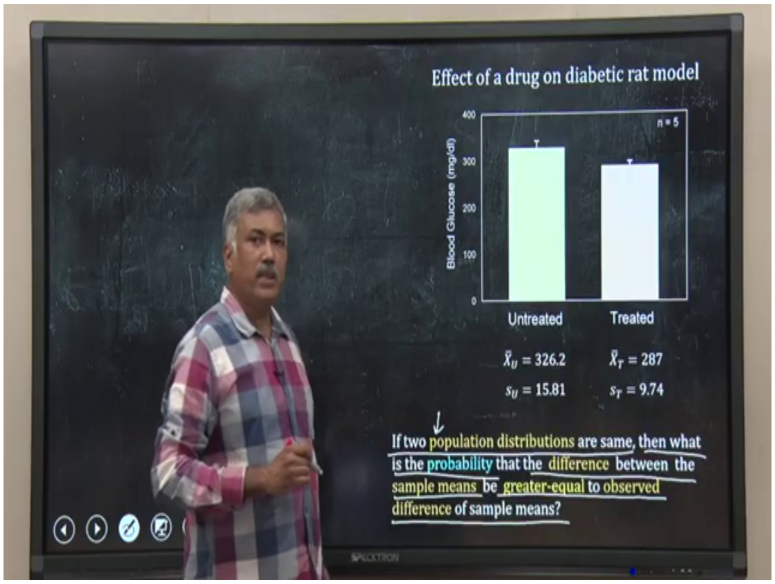
As I said, I cannot do the experiment on the whole population I do not know even the size of the whole population. So, statisticians try to answer this question using a probabilistic method. And the tricky part is they do not answer this question directly rather they pose a bit twisted question, and that is where the tricky thing is.

So, follow me carefully, what how I pose the question how I will change this question. Assume that at the population level the blood glucose level for these untreated and treated groups at population level, their population distributions are same identical, they are not different. Now, you make fresh sampling, one fresh sampling from the untreated group and another sampling from the treated group.

Now, you calculate the mean blood glucose level for these two samples untreated and treated and find that difference in the mean between these untreated sample and treated sample. So, you will have again a new difference value will come. Now, the question is what is the probability that this new difference coming from the new samples that you have taken will be greater equal to the difference that I have seen already in this experiment?

The difference I have seen in this experiment is something 39.2. What is the probability that the new difference that is coming from new samples that have sampled again will be greater equal to that 39.2. We want to calculate that probability. If that probability is very high, then my assumptions that at population level the distribution of blood glucose in both treated and untreated group are same will hold. If that probability becomes very small, comes very small, then I will reject that hypothesis that assumption. It's bit tricky. We have inverted the problem almost. So, just try to follow it once again.

(Refer Slide Time: 07:22)



Effect of a drug on diabetic rat model

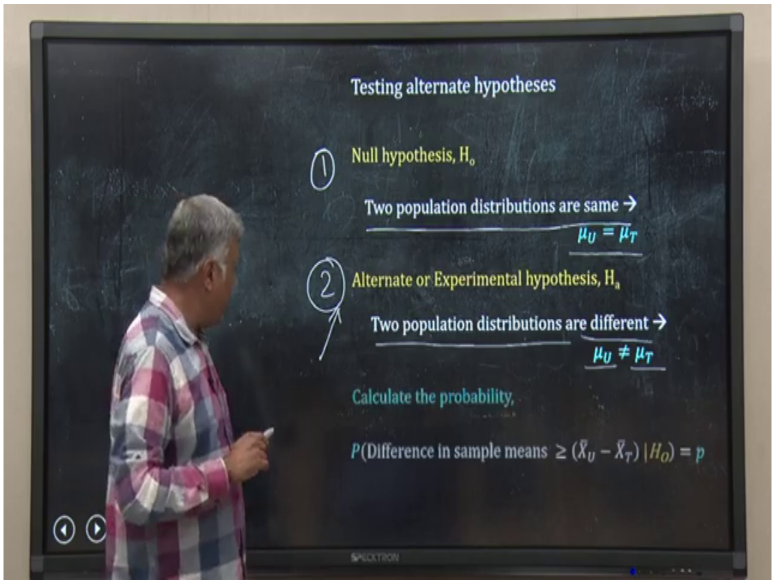
$\bar{X}_U = 326.2$ $\bar{X}_T = 287$

$s_U = 15.81$ $s_T = 9.74$

If two population distributions are same, then what is the probability that the difference between the sample means be greater-equal to observed difference of sample means?

What I am saying if two population distributions are same, this is my assumption, then what is the probability you have to calculate the probability that the difference between the sample means you have to take a new sample and you have calculated difference in their means, the difference between the sample means be greater equal to observed differences of sample means? That you want to calculate and based on that probability, either you will consider this assumption that two population distributions are same or you will reject that assumption. So, let me put it symbolically that will be easier to understand.

(Refer Slide Time: 08:08)



Testing alternate hypotheses

① Null hypothesis, H_o

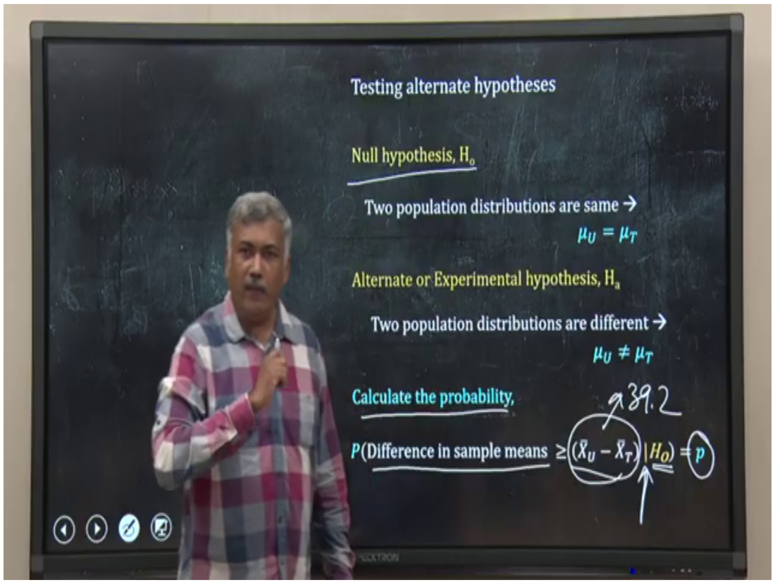
Two population distributions are same → $\mu_U = \mu_T$

② Alternate or Experimental hypothesis, H_a

Two population distributions are different → $\mu_U \neq \mu_T$

Calculate the probability,

$P(\text{Difference in sample means} \geq (\bar{X}_U - \bar{X}_T) \mid H_O) = p$



Testing alternate hypotheses

Null hypothesis, H_o

Two population distributions are same → $\mu_U = \mu_T$

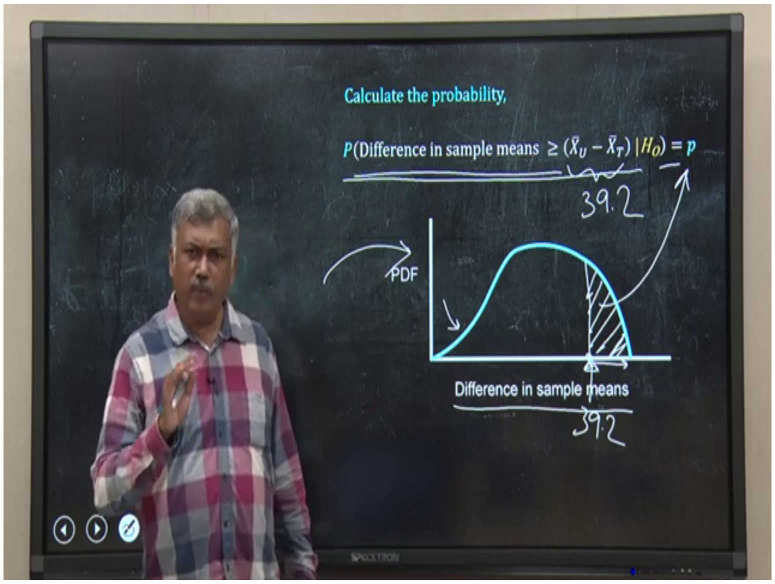
Alternate or Experimental hypothesis, H_a

Two population distributions are different → $\mu_U \neq \mu_T$

Calculate the probability,

39.2

$P(\text{Difference in sample means} \geq (\bar{X}_U - \bar{X}_T) \mid H_O) = p$



Calculate the probability,

$P(\text{Difference in sample means } \geq (\bar{X}_U - \bar{X}_T) \mid H_0) = p$

$$Null\ hypothesis,\ H_a \Rightarrow \mu_U = \mu_T$$

$$Alternate\ or\ Experimental\ hypothesis,\ H_a \Rightarrow \mu_U \neq \mu_T$$

$$P\left(\frac{Difference\ in\ sample\ means \geq \left(\bar{X}_U - \bar{X}_T\right)}{H_0}\right) = p$$

So, what I am doing here, I have two hypotheses, one is called Null hypothesis. The other one is called Alternative or Experimental hypothesis. Let us first take the alternate or experimental hypothesis because that is in line of the experiment that we have done in the data we have seen. In this case, we consider that the two-population distribution remember distribution of blood glucose level at the population level not sample, the two-population distribution are different.

That is what we believe that is what I want to prove possibly. So, in that case, Mu of u that is the population, mean Mu is population mean. Mu of u is not equal to Mu of T. Now, what is the null hypothesis of this, the null hypothesis is the two population distributions are same that means Mu of u is equal to Mu of T. Just notice this one, the null hypothesis and the alternate hypothesis are actually binary.

And they are mutually exclusive. And null hypothesis does not mean that it has some value 0 or something like that. It is essentially negation of the other one, this alternate hypothesis or the experimental hypothesis or the belief that you possibly started the experiment with. So, once I have these two hypotheses.

Then I want you to calculate a probability, what probability? I want to calculate the probability that the difference in sample means, means you are again doing sampling from the population will be greater equal to the observed difference. What is the observed difference in this case? That is 39.2 possible If I am not doing a wrong calculation, 39.2 given, this is important, given H0 is true, that means null hypothesis is true. I have to calculate this probability.

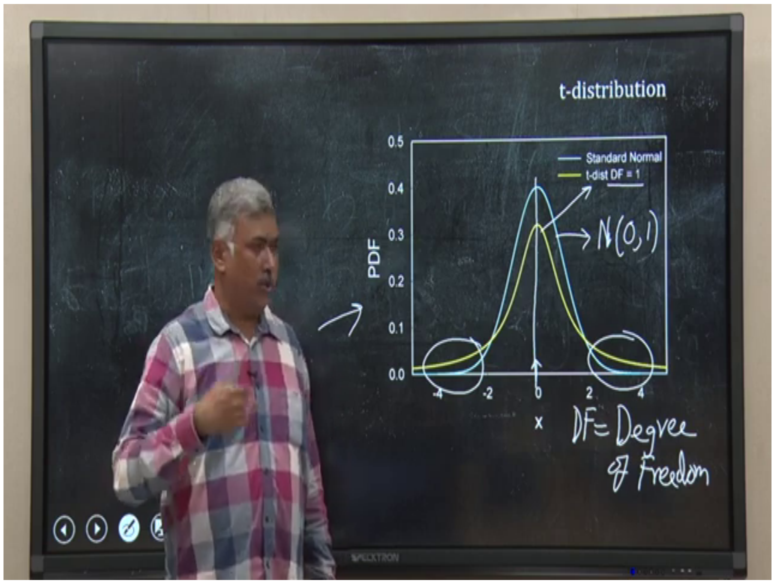
And we will call this p value, this probability p value. Now, when we are talking about calculating a probability, we have to require some sort of probability distribution. And here we are calculating the probability of difference in sample means. So, the variable is a continuous variable. So, we need some sort of probability density function.

Suppose I know that for the time being, we assume we know that. So, this is the PDF given to me, I know that and this blue shaped this is the PDF function and I have in the horizontal the difference in sample means. And now I have to calculate this probability, how should I calculate, suppose this value is 39.2 then somewhere here suppose, that value comes 39 and just 0.2.

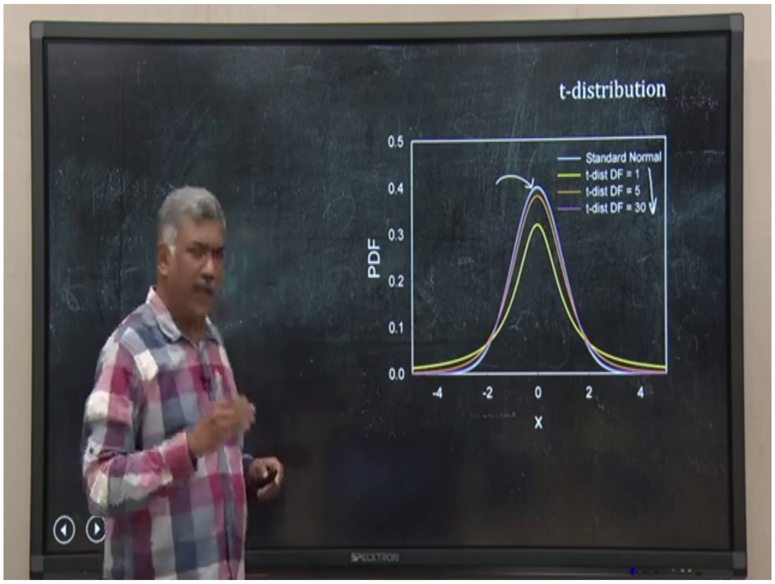
So, then the probability that this probability P will be actually I can draw a line this area under the curve, this is greater equal to the observed difference, this is the observed difference, and greater equal to mean this side. So, this is the area under the curve that is the probability I am asking you to calculate and that will be the p value and based on that probability, we have to make a decision about the null hypothesis that we have made.

Now, it turns out that, the probability density function of earlier distribution that we have to use for this particular problem of blood glucose level that I started with, and the test that I will do will be called a T test, we have to use t distribution, we have studied about normal distribution in continuous random variable case. So, just let me introduce is briefly what is t distribution.

(Refer Slide Time: 11:51)



t-distribution



t-distribution

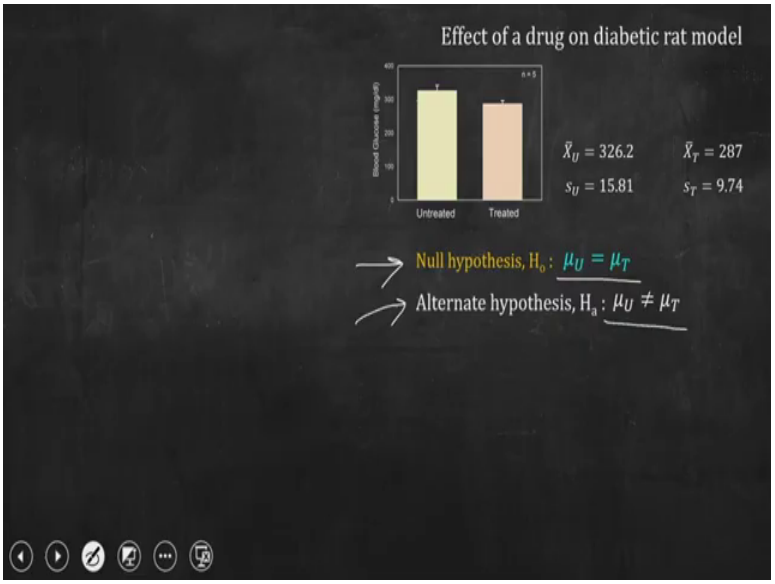
T distribution is almost similar to normal distribution, what I have plotted here in this graph I have plotted standard normal versus t distribution. So, standard normal is this blue one. So, standard normal if you remember the mean will be 0, and the variance will be 1. So, this is where the mean is.

So, the peak is at, for the blue curve, is at 0, and the yellow curve is for t distribution with a Degree of Freedom, DF is Degree of Freedom, I am not going to details of how Degree of Freedom we calculate and all these things for the time being we have to remember the Degree of Freedom is a parameter in the PDF of t distribution and it is related to the sample size if sample size is large degree of freedom will be also large. So, what if you see this diagram you can see actually, the t distribution is also symmetric around this mean 0 and it is falling just like exponential though it is not a complete exponential fall, the only difference is that this is bit fatter on the tail sides. That is the only difference.

Now, the most interesting feature of t distribution is that if I keep on increasing the degree of freedom, the parameter for t distribution, it will move towards standard normal distribution. So, let me show you this in the graph. So, I am increasing degree of freedom from 1 to 30 and the pink line is for degree of freedom 30 and you can see it is almost converging with the blue line that is the standard normal.

So, in a way, I can say as the degree of freedom for t distribution becomes very, very large t distribution becomes close to a standard normal distribution. Fine, we have learned t distribution that is good enough for this lecture and subsequent data analysis also. So, let me go back to the original blood glucose problem.

(Refer Slide Time: 14:01)



Effect of a drug on diabetic rat model

$\bar{X}_U = 326.2$ $\bar{X}_T = 287$

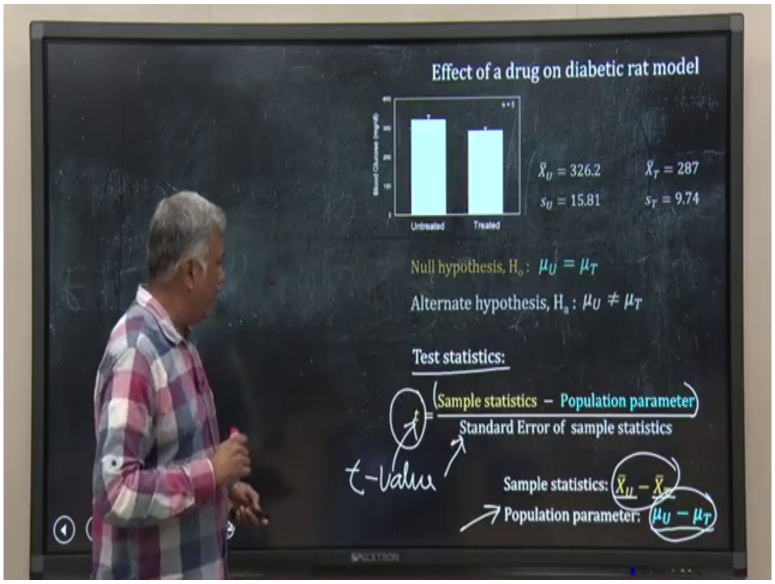
$s_U = 15.81$ $s_T = 9.74$

Null hypothesis, H_0: $\mu_U = \mu_T$

Alternate hypothesis, H_a: $\mu_U \neq \mu_T$

$$Null\ hypothesis,\ H_a \Rightarrow \mu_U = \mu_T$$

$$Alternate\ or\ Experimental\ hypothesis,\ H_a \Rightarrow \mu_U \neq \mu_T$$



Effect of a drug on diabetic rat model

$\bar{X}_U = 326.2$ $\bar{X}_T = 287$

$s_U = 15.81$ $s_T = 9.74$

Null hypothesis, H_0: $\mu_U = \mu_T$

Alternate hypothesis, H_a: $\mu_U \neq \mu_T$

Test statistics:

$$t = \frac{(\text{Sample statistics} - \text{Population parameter})}{\text{Standard Error of sample statistics}}$$

Sample statistics: $\bar{X}_U - \bar{X}_T$

Population parameter: $\mu_U - \mu_T$

$$t = \frac{(Sample\ statistics - Population\ parameter)}{Sandard\ Error\ of\ sample\ statistics}$$

$$Sample\ statistics:\ \bar{X}_U - \bar{X}_T$$

$$Population\ parameter:\ \mu_U - \mu_T$$

So, what I have done till now, I have made a null hypothesis, where I have considered both the population distribution as same. So, their population means are identical. And my alternate hypothesis is they are not identical. And remember, we are not going to do anything on the alternate one we want to play with the null hypothesis.

Now, to move further, we have to make something called test statistic or a new variable we will introduce and calculate the value of that and in this case, this is called the t value. That is what I have written as small t. So, what is the t value, t value is the difference between sample statistics and population parameter.

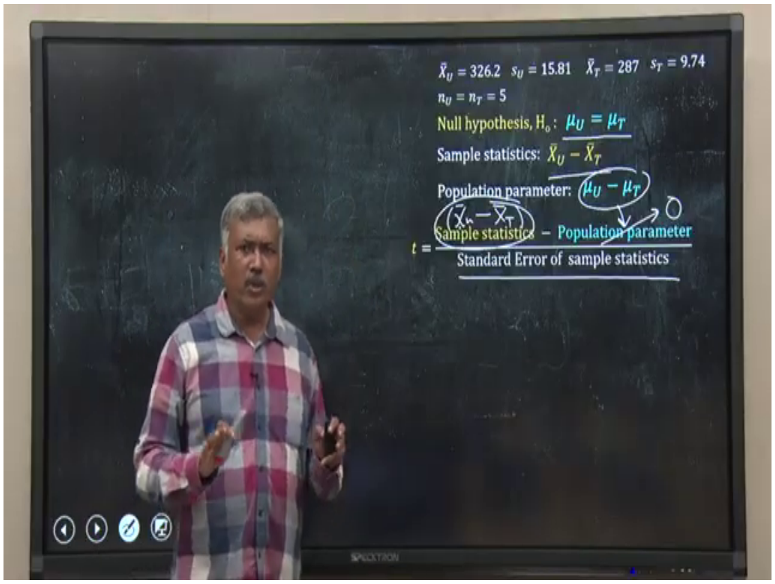
What is sample statistics in this case? Sample statistics is the mean of my sample, untreated sample minus mean of my treated sample. So, that is I am calling sample statistics, what is my population parameter, my population parameter is Mu u minus Mu T. So, please notice this one that they are actually similar thing, one is for sample

The other one is for the population, both of them are difference between means of two groups, one is at the population level another one is the sample level the sample that you have observed you have to experimented with. So, for samples we call it statistics, whereas, for the population we call it parameter, because we do not know those values and we are trying to project the value of parameter, we are trying to make a guess about the parameter unknown parameter form the sample statistics.

So, this is what I know, but usually you will not know these unless until you make some hypothesis. So, I have sample statistics minus population parameter in numerator and in denominator we have standard error of sample statistics. I will come to that later. Now, one interesting thing is that, I will not go into the details of the mathematics if we assume, if we assume that at population level this bad glucose thing is that measuring variable the variable that we are measuring at population level it follows normal distribution.

If that assumption is correct, then it can be shown that this t value this test statistic t the way we have defined will follow t distribution. So, that makes our life easier, because we know the t distribution we can do calculations on that. So, with this let me move further and calculate the t for this particular problem.

(Refer Slide Time: 16:38)



$\bar{X}_U = 326.2$ $s_U = 15.81$ $\bar{X}_T = 287$ $s_T = 9.74$

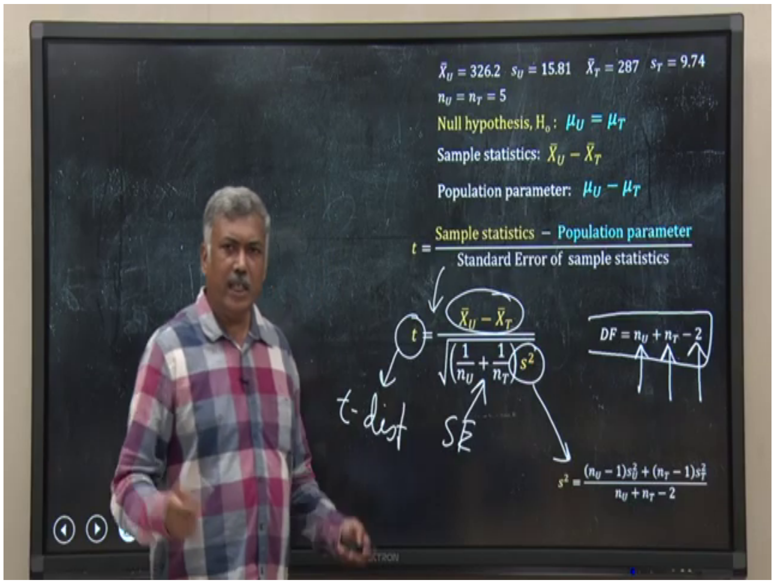
$n_U = n_T = 5$

Null hypothesis, H_0: $\mu_U = \mu_T$

Sample statistics: $\bar{X}_U - \bar{X}_T$

Population parameter: $\mu_U - \mu_T$

$$t = \frac{\text{Sample statistics} - \text{Population parameter}}{\text{Standard Error of sample statistics}}$$



$\bar{X}_U = 326.2$ $s_U = 15.81$ $\bar{X}_T = 287$ $s_T = 9.74$

$n_U = n_T = 5$

Null hypothesis, H_0: $\mu_U = \mu_T$

Sample statistics: $\bar{X}_U - \bar{X}_T$

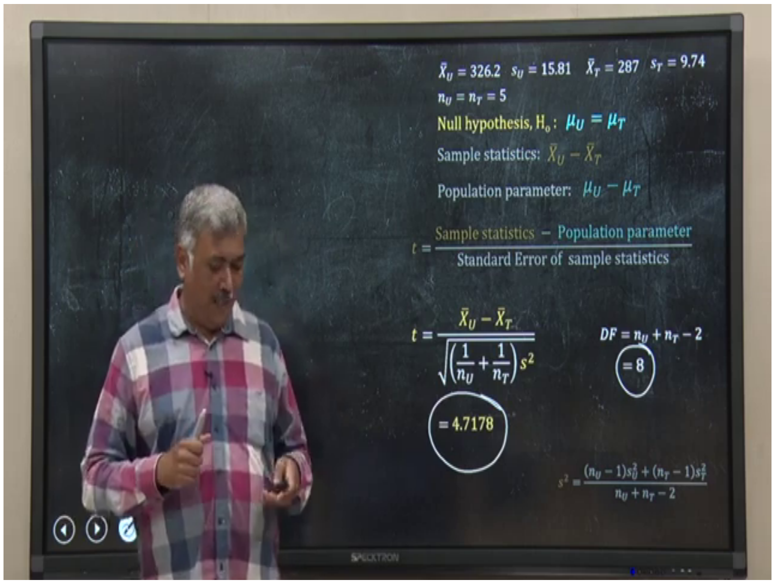
Population parameter: $\mu_U - \mu_T$

$$t = \frac{\text{Sample statistics} - \text{Population parameter}}{\text{Standard Error of sample statistics}}$$

$$t = \frac{\bar{X}_U - \bar{X}_T}{\sqrt{\left(\frac{1}{n_U} + \frac{1}{n_T}\right) s^2}}$$

$$DF = n_U + n_T - 2$$

$$s^2 = \frac{(n_U - 1)s_U^2 + (n_T - 1)s_T^2}{n_U + n_T - 2}$$



$\bar{X}_U = 326.2$ $s_U = 15.81$ $\bar{X}_T = 287$ $s_T = 9.74$

$n_U = n_T = 5$

Null hypothesis, H_o: $\mu_U = \mu_T$

Sample statistics: $\bar{X}_U - \bar{X}_T$

Population parameter: $\mu_U - \mu_T$

$$t = \frac{\text{Sample statistics} - \text{Population parameter}}{\text{Standard Error of sample statistics}}$$

$$t = \frac{\bar{X}_U - \bar{X}_T}{\sqrt{\left(\frac{1}{n_U} + \frac{1}{n_T}\right) s^2}}$$

$$\boxed{= 4.7178}$$

$$DF = n_U + n_T - 2$$

$$\boxed{= 8}$$

$$s^2 = \frac{(n_U - 1)s_U^2 + (n_T - 1)s_T^2}{n_U + n_T - 2}$$

$$t = \frac{\overline{X}_U - \overline{X}_T}{\sqrt{\left(\frac{1}{n_U} + \frac{1}{n_T}\right)s^2}}$$

= 4.7178

$$DF = n_U + n_T - 2$$

=8

$$s^2 = \frac{(n_U - 1)s_U^2 + (n_T - 1)s_T^2}{n_U + n_T + 2}$$

So, sample statistics what it will be it will be X u bar minus X t bar that is given here, population parameter will come here. Now, what is my null hypothesis, my null hypothesis is Mu u is equal to Mu T population distribution is same for untreated and treated group. So, that means that identical that means this population parameter difference in mean at population level must be equal to 0.

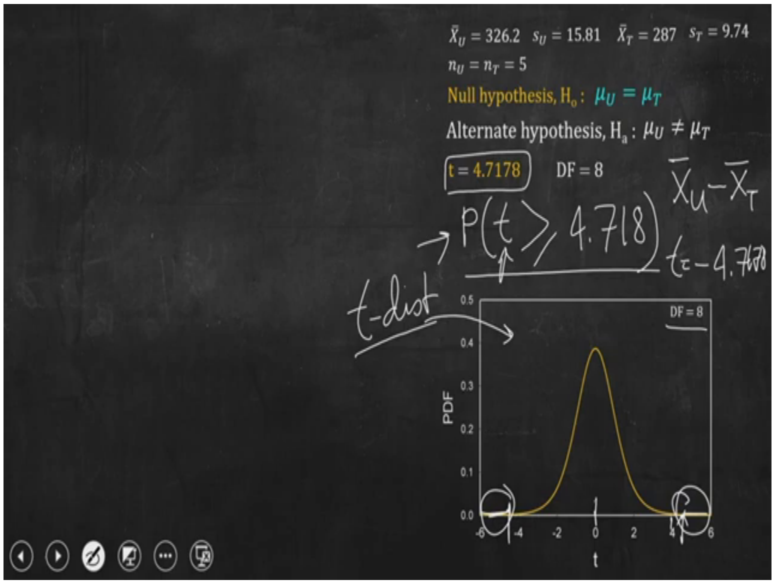
So, I am left with only two terms X u bar minus X t bar divided by standard error of sample statistics means standard error of this one, this difference. The derivation of the standard error is bit lengthy. I will skip those. And you do not also need to go in detail of that. But you should know that we can do that.

Because anyway in a statistical tool or function in R or somewhere, we will do the same thing we do not need to go and manually do that. So, eventually I get this formula, t equal to, I have the difference in mean from the sample. And this one is the standard error of that difference. And I have a square term which is given here and remember, this will follow t distribution.

So, t distribution has a parameter, it depends on a parameter that is called degree of freedom. So, it can be shown for this case the degree of freedom will be sample size of the untreated plus sample size of a treated minus 2. So, let us plug the numerical values because we have the numerical value from the experiment and we get the degree of freedom will be 8, 5 plus 5 minus 10 minus 2, 8 and the t value comes 4.7178.

So far so good we had null hypothesis from, based on that I have hypotheses using the data of the experiment I have calculated a t value. I know the degree of freedom now, I have to calculate the probability. So, we have to go back to the question.

(Refer Slide Time: 18:50)



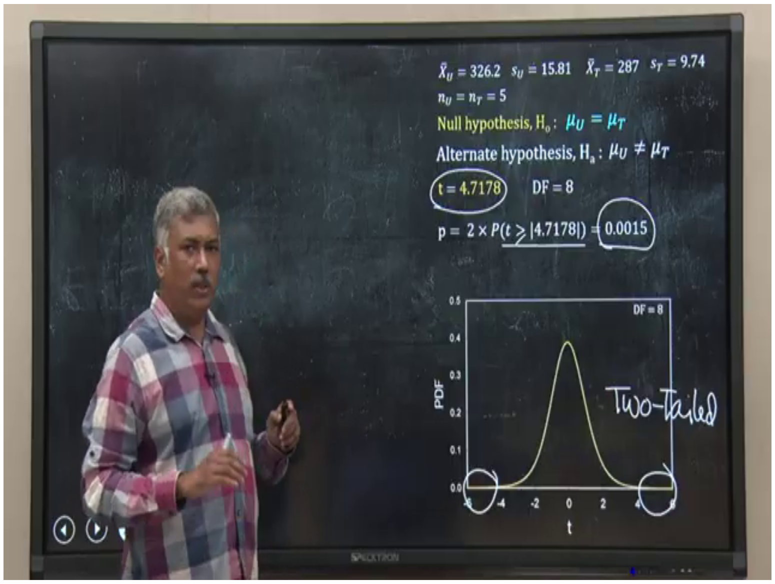
$\bar{X}_U = 326.2$ $s_U = 15.81$ $\bar{X}_T = 287$ $s_T = 9.74$

$n_U = n_T = 5$

Null hypothesis, H$_o$: $\mu_U = \mu_T$

Alternate hypothesis, H$_a$: $\mu_U \neq \mu_T$

$\boxed{t = 4.7178}$ DF = 8



$\bar{X}_U = 326.2$ $s_U = 15.81$ $\bar{X}_T = 287$ $s_T = 9.74$

$n_U = n_T = 5$

Null hypothesis, H_0: $\mu_U = \mu_T$

Alternate hypothesis, H_a: $\mu_U \neq \mu_T$

$t = 4.7178$ DF = 8

$p = 2 \times P(t \geq |4.7178|) = 0.0015$

$$t = 4.7178 \quad DF = 8$$

$$p = 2 \times P(t \geq |4.7178|) = 0.0015$$

So, this is the t distribution, I have just drawn it for analysis, I have not to draw this one just to explain I have draw this is a t distribution for degree of freedom 8, around 0 you have symmetrical on both sides. Now, if I have to calculate probability that a value of t, using the definition of t that we have seen in the last few slides before, will be greater equal to 4.718 that is what I have to calculate.

If I freshly sample from the same population, then again, I calculate do experiment and again get the values and calculate t. So, that is my new t, I have to calculate the probability of getting that t greater equal to what I have done already I have got this one already. So, I have to calculate this probability. So, how should I do that, let us look into the diagram.

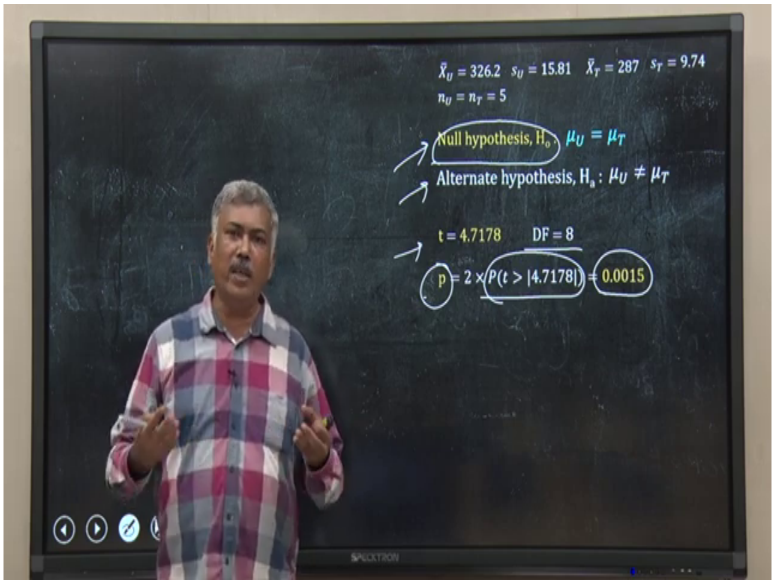
So, this is 4, our t value is 4.718 so, it will be somewhere here I cannot draw it, but somewhere this area, this area will be area under the curve and you can easily see that area under curve will be very small that means, the probability this probability, probability that t will be greater equal to 4.718 will be very small and I will show you the value.

So, it is a very small probability, but there is a issue here. See, how I am calculating the difference in mean, I am taking X u bar minus X T bar, why cannot I do the reverse thing take the XT bar first and subtract Xu bar then, then my all values of difference in means will come in negative for this data set and my t value will come minus 4.7178.

So, let us check what will be that probability, so here this one will be again very small, but as the difference can be both negative and positive, I have to consider both the probability and they will be identical and simply add them I will multiply by 2. So, that is what I have done here, the probability that t is greater than the absolute value, so that I am not bothered about the sign, absolute value of 4.7178 which I have got for my t value in my experimental data, it will be greater equal to that in a way actually. So, that will be and then I multiply with 2 because I have p both side.

So, that will be 0.0015, one thing just to mention here, as I have taken the, calculated the probability from both side this is called Two-Tailed. If I have reasons to believe that one side is good enough that make a physical meaning, there is a meaning biological or physical meaning for that, then that will be One-Tailed. So, this is a Two-Tailed probability I have calculated and that has come to 0.0015, good enough.

(Refer Slide Time: 22:17)



$\bar{X}_U = 326.2$ $s_U = 15.81$ $\bar{X}_T = 287$ $s_T = 9.74$

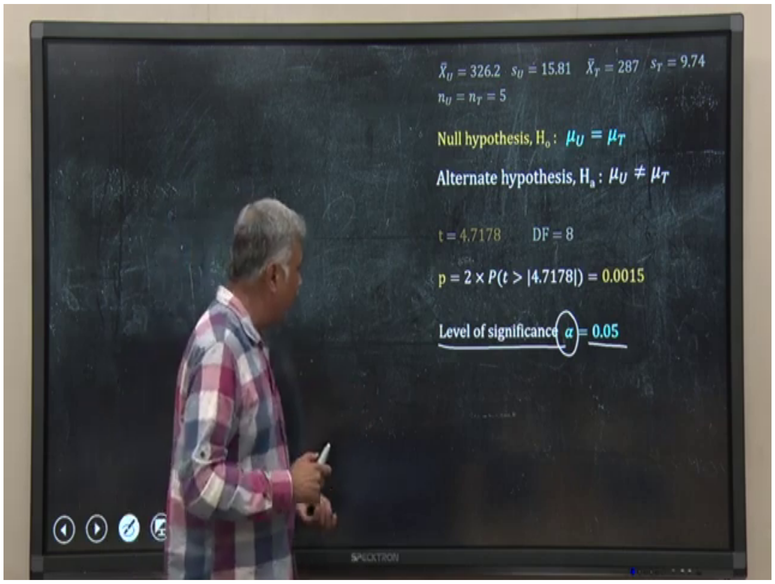
$n_U = n_T = 5$

Null hypothesis, H_0: $\mu_U = \mu_T$

Alternate hypothesis, H_a: $\mu_U \neq \mu_T$

$t = 4.7178$ DF = 8

$p = 2 \times P(t > |4.7178|) = 0.0015$



$\bar{X}_U = 326.2$ $s_U = 15.81$ $\bar{X}_T = 287$ $s_T = 9.74$

$n_U = n_T = 5$

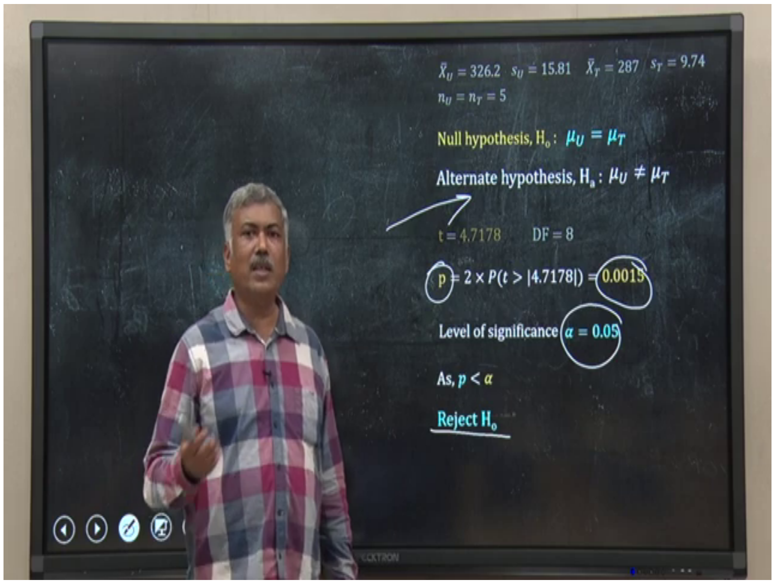
Null hypothesis, H_0: $\mu_U = \mu_T$

Alternate hypothesis, H_a: $\mu_U \neq \mu_T$

$t = 4.7178$ DF = 8

$p = 2 \times P(t > |4.7178|) = 0.0015$

Level of significance $\alpha = 0.05$



$\bar{X}_U = 326.2$ $s_U = 15.81$ $\bar{X}_T = 287$ $s_T = 9.74$
$n_U = n_T = 5$

Null hypothesis, H_o : $\mu_U = \mu_T$

Alternate hypothesis, H_a : $\mu_U \neq \mu_T$

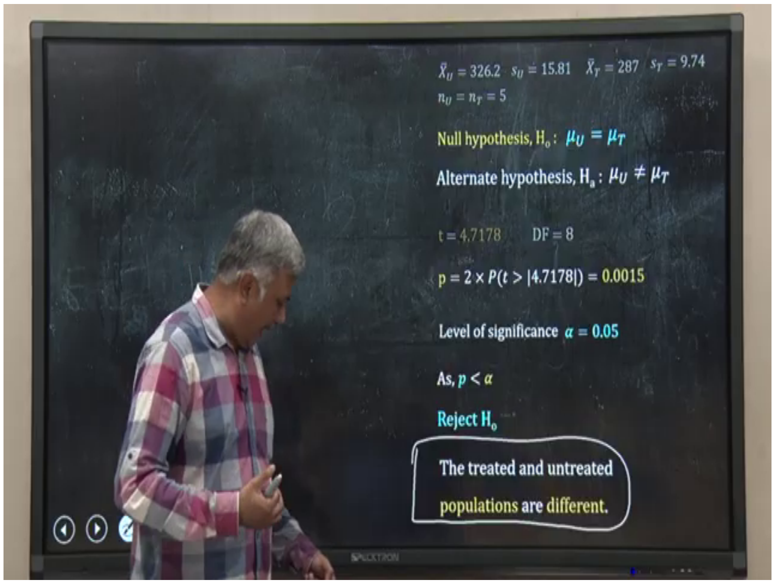
t = 4.7178 DF = 8

$p = 2 \times P(t > |4.7178|) = 0.0015$

Level of significance $\alpha = 0.05$

As, $p < \alpha$

Reject H_o



$\bar{X}_U = 326.2$ $s_U = 15.81$ $\bar{X}_T = 287$ $s_T = 9.74$
$n_U = n_T = 5$

Null hypothesis, H_o : $\mu_U = \mu_T$

Alternate hypothesis, H_a : $\mu_U \neq \mu_T$

t = 4.7178 DF = 8

$p = 2 \times P(t > |4.7178|) = 0.0015$

Level of significance $\alpha = 0.05$

As, $p < \alpha$

Reject H_o

The treated and untreated populations are different.

I started with my data, I made null hypotheses, I made alternate hypotheses and I have got t value, degree of freedom and I have got the p value also. Now, remember my initial question, my assumption my null hypothesis is that a population distribution of blood glucose for these two groups, treated, untreated are identical their mu is the same.

And then I calculated the probability of getting this situation and that probability has come to 0.0015 and as I said at the very beginning, if this probability is very low, I will consider that null hypothesis is wrong, my assumption is wrong. If this probability is very high, then I will be sure that this hypothesis is correct.

Now, larger bigger smaller these are all relative term, there is no absolute small or something like that. So, when I say probability is small what do I mean? How small is small. So, that is why I have to use a cutoff, I have no other option and, in this case, this cutoff is called level of significance and usually represented by a symbol alpha.

Suppose, we have agreed, you and me has agreed that level of significance, that cutoff will be 0.05. Now, how do we agree, it depends upon the convention among the people, scientists who are working in a particular field, it also depends upon how sensitive the assays or experiments that you are doing, if you have a very sensitive assay.

Where you can measure make measurements suppose, up to 10th decimal point obviously, you will go for a lower cutoff value. So, how reproducible the results are all these issues are there based on that we have some thumb rule in every field there is a thumb rule and many a time in many biological thing it is 0.05 some experiment you may consider is 0.001 also.

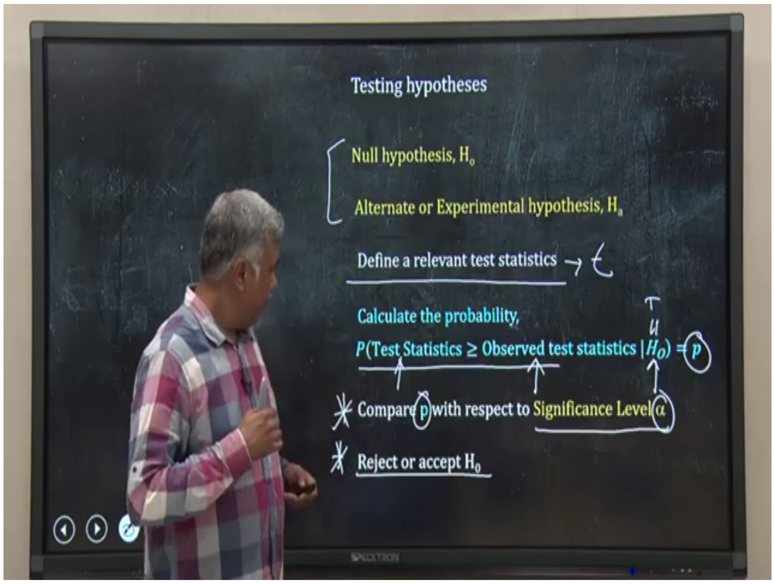
So, this is the cutoff. So, this is called level of significance, and in my case in this particular problem, what has happened, this p value is lesser than, because this is lesser than my cutoff. So, that means, the p value the probability that the t that is statistic that I have calculated for a new sample from the same population will be greater equal to the existing t statistics that I have seen in my experimental data.

That probability is smaller than the cutoff. That means, my null hypothesis has to be rejected. So, I reject null hypothesis. If I reject null hypothesis what we are left with because my hypothesis binary have two binary hypotheses, if one is true, another one is false, if one is false, the other one is true.

So, if H0 is false, then my alternate hypothesis is true. That means, the Mu the population mean of untreated and treated group are different. That means, in other words, the treated and untreated population, not sample only, sample you have seen, that data at population level the treated and untreated groups are different. So, that is my conclusion.

So, till now I have not deal with alternative experimental hypothesis I was still dealing with the null hypothesis, once I rejected the null hypothesis, I reached to the conclusion I want to draw at the very beginning. So, this is how you do t test and hopefully I could communicate the basic essence of hypothesis testing using that.

(Refer Slide Time: 26:02)



Testing hypotheses

Null hypothesis, H_0

Alternate or Experimental hypothesis, H_a

Define a relevant test statistics → t

Calculate the probability,

$P(\text{Test Statistics} \geq \text{Observed test statistics} \mid H_0) = p$

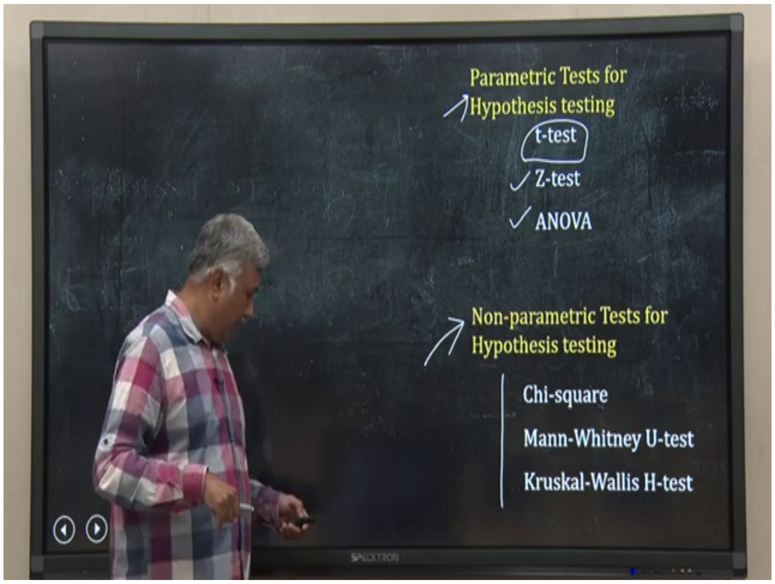
Compare p with respect to Significance Level α

Reject or accept H_0

Let me generalize what you do in a test of hypothesis any sort of test of hypothesis, you first make two alternate binary hypothesis null hypothesis and alternate hypothesis, null hypothesis does not mean that something has to be 0, null is negation of the other one, then you define some sort of test statistics in this our case we have done the t value, it may not be t for something some other tests.

Then what you do you calculate the probability that the test statistics is greater equal to the observed test statistics given the null hypothesis is true, and that you call p value. Now, you compare that p value, you compare that P value with some significance level alpha. And based on that comparison, either you reject or accept the null hypothesis. This is what is test of hypothesis or some people will call a test of significance.

(Refer Slide Time: 27:08)

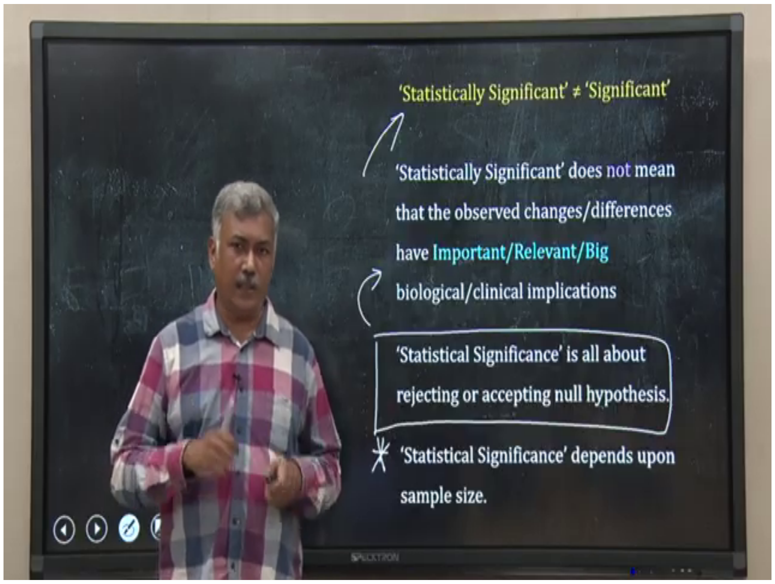


Parametric Tests for Hypothesis testing

- t-test
- ✓ Z-test
- ✓ ANOVA

Non-parametric Tests for Hypothesis testing

- Chi-square
- Mann-Whitney U-test
- Kruskal-Wallis H-test



'Statistically Significant' ≠ 'Significant'

'Statistically Significant' does not mean that the observed changes/differences have Important/Relevant/Big biological/clinical implications

'Statistical Significance' is all about rejecting or accepting null hypothesis.

'Statistical Significance' depends upon sample size.

These are called parametric tests. Because in all these three cases I have given the example here, you have to make some assumption about the population distribution. And that assumption is actually the population distribution is normal. That is what I mentioned also earlier for t test initially.

I assumed that the population distribution of blood glucose in both the treated and untreated group are normally distributed. Whereas, there are certain other hypothesis tests like the Mann-Whitney Test, where you do not make any explicit assumption about the population distribution behind your sample. So, those are called Non-parametric.

And they are many times very useful when actually your population that you are dealing the case you are dealing with where the population may not be normally distributed. Now, we have almost reached the end of this lecture, but before I move to the key points, I want to highlight certain important issue, which many a times we miss while using statistical tests of hypothesis tests of significant is the most crucial thing is that statistical significant is not equal to significant.

What do I mean by that? See something which has come out to be statistically significant, that does not mean that, that observed behavior or observed change that you have seen is important relevant are very big or very small with respect to the biology or clinical implications. For example, in our experimental data that we discussed, we have seen the difference between treated untreated groups blood glucose level is statistically significant.

But that does not tell me whether that difference is biologically significant or biologically relevant or not, whether that will affect the longevity of the rats or not, whether that will affect the metabolism of the rat or not, we do not know all information, all this information are absent to us. So, we cannot conclude about the biological or clinical implications from a statistical set of significance that we have to remember.

That is why it is said whenever you are writing in a report or you are giving a presentation, you are writing a paper, do not say the observation is significant. If you have done a statistical test, for example, t test or ANOVA. And you are talking about statistical significance, then you write the observation or the difference is statistically significant.

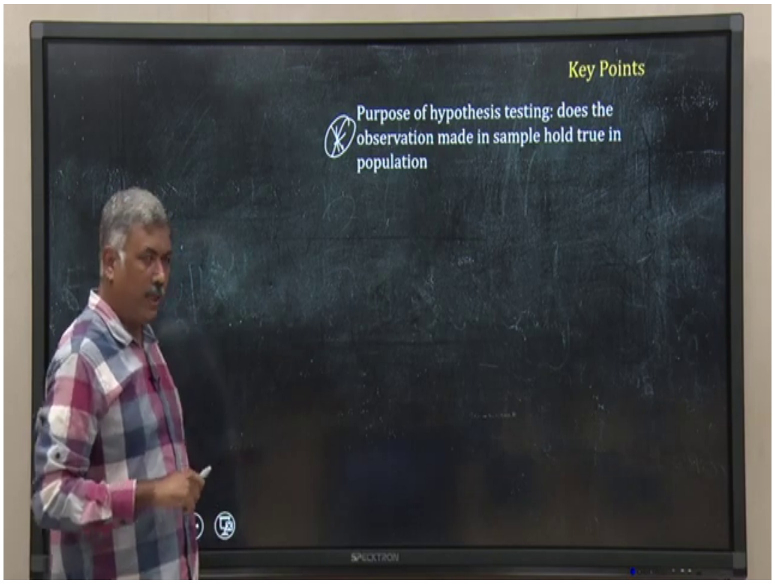
That is all that phrase you use. And if you want to mean that some observation is huge, or some observation has a huge implication in a case of biology or clinicians. So, in that case, you use those type of terms like these are important, this is relevant this is a big, something like that, but do not use the word significant because then people get confused.

The other thing that we have to remember is a statistical significance is all about rejecting or accepting the null hypothesis, this is connected to this other first one, but we have to keep it remind ourselves what we are dealing here is nothing but whether we can reject or accept the null hypothesis. And the last one is that many times we forget in the statistical test of significance, the results depend upon the sample size.

So, if your sample size you are large, and in another case, your sample size is small, the same experiment, one will be statistically insignificant, the result of other will be statistically significant. So, there are lots of ifs and buts in use of test of statistical significance or test of hypothesis. Many a times we forget those ifs and buts those issues.

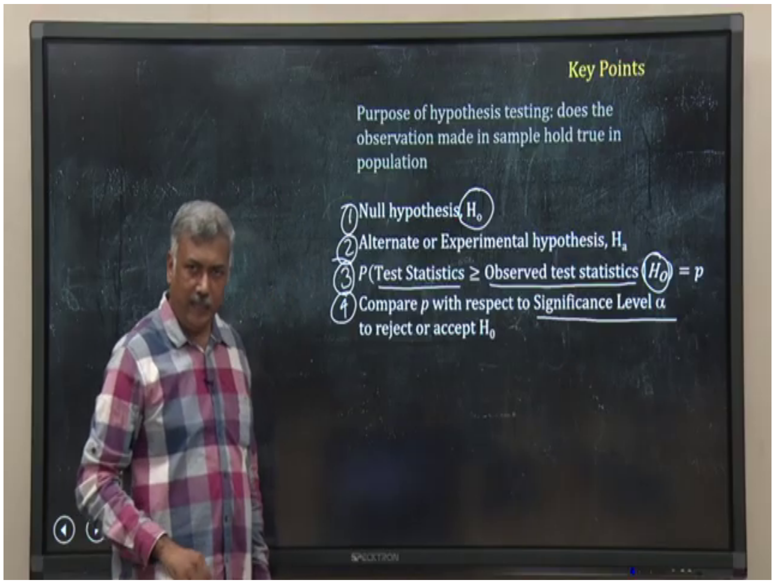
And that is why many a time we use wrong test of hypothesis, we misinterpret the result of test of hypothesis and we should avoid them. And in fact, there are alternatives to statistical test of hypothesis also, for example, there are certain Bayesian method based on Bayes theorem to give a similar course of answer similar questions. So, some people insist that we should not use these tests of hypothesis rather we should use those Bayesian methods, but for the time being test of hypothesis as we have discussed here is going to stay in biology and we should know how to use it we should know how to use it judiciously.

(Refer Slide Time: 31:27)



Key Points

Purpose of hypothesis testing: does the observation made in sample hold true in population



Key Points

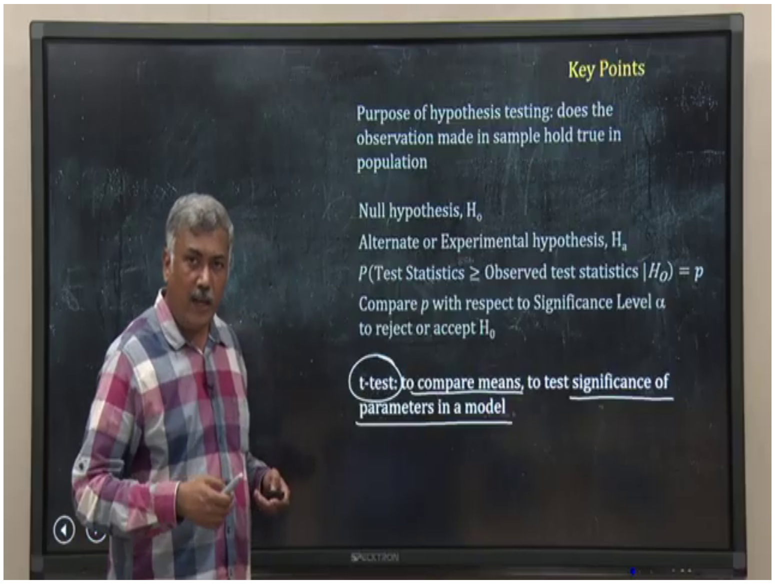
Purpose of hypothesis testing: does the observation made in sample hold true in population

① Null hypothesis, H_0

② Alternate or Experimental hypothesis, H_a

③ $P(\underline{\text{Test Statistics}} \geq \underline{\text{Observed test statistics}} \mid H_0) = p$

④ Compare p with respect to $\underline{\text{Significance Level } \alpha}$ to reject or accept H_0



Key Points

Purpose of hypothesis testing: does the observation made in sample hold true in population

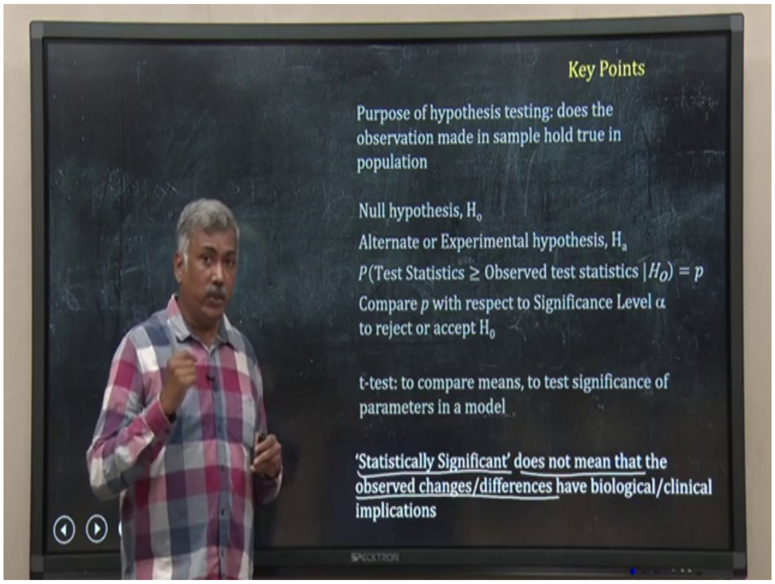
Null hypothesis, H_0

Alternate or Experimental hypothesis, H_a

$P(\text{Test Statistics} \geq \text{Observed test statistics} \mid H_0) = p$

Compare p with respect to Significance Level α to reject or accept H_0

t-test: to compare means, to test significance of parameters in a model



Key Points

Purpose of hypothesis testing: does the observation made in sample hold true in population

Null hypothesis, H_0

Alternate or Experimental hypothesis, H_a

$P(\text{Test Statistics} \geq \text{Observed test statistics} \mid H_0) = p$

Compare p with respect to Significance Level α to reject or accept H_0

t-test: to compare means, to test significance of parameters in a model

'Statistically Significant' does not mean that the observed changes/differences have biological/clinical implications

So, let me jot down what we have learned. The first thing we have learned the purpose of hypothesis testing, the purpose is to know does the observation made in a sample hold true for the population or not that is the main question that we are trying to answer. To answer this question, what we do we create a null hypothesis, then we make a alternate hypothesis these are binary mutually exclusive.

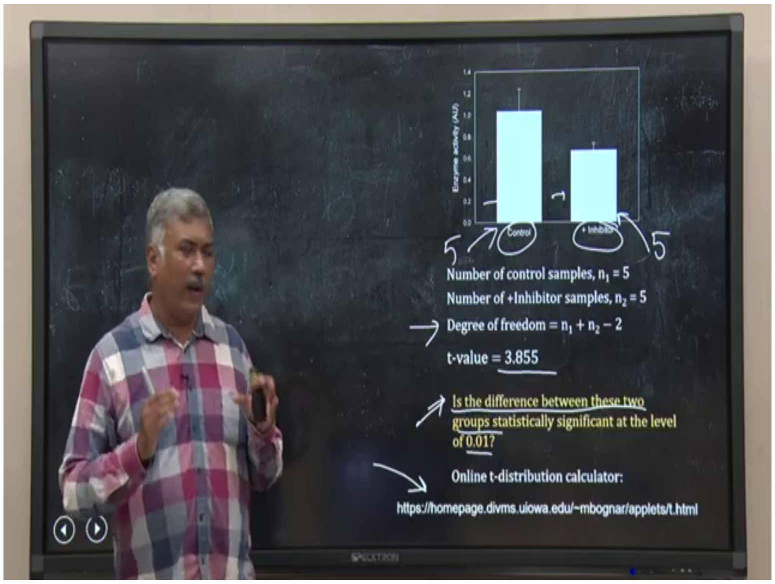
And then we do some create some test statistics depending upon the what type of experimental data we have. And then we calculate the probability that the test statistics is bigger than observed statistics considering the null hypothesis is true, and we get the p value and we compare that p value with a cutoff called significance level alpha and based on that.

Either we accept H0 null hypothesis or reject that, that is how you do all sort of test of hypothesis and we have discussed only t test in this lecture, we have shown t test to compare means of two groups t-test we will use in our subsequent lecture to find the significance of or the other the statistical significance if I say correctly, or parameters in a model.

So, you have data from that you create a model. So, you want to you have some constant term called parameter in that model, and you want to know whether those estimated parameter values are significant or not statistically significant. And we can use t test in that case, in some cases we will use ANOVA.

So, anyway, we will be using test of hypothesis. And the last one that we have to remind ourselves continuously is that statistically significant does not mean that the observed changes or differences have biological or clinical implication. So, I will end this lecture with this important point to keep in your mind it is easy to use t test ANOVA because almost all the statistical software has those just a click away, but we have to keep in mind the implications the we have to understand of this statistical tests and have to understand the result very judiciously.

(Refer Slide Time: 33:37)



Number of control samples, $n_1 = 5$

Number of +Inhibitor samples, $n_2 = 5$

Degree of freedom = $n_1 + n_2 - 2$

t-value = 3.855

Is the difference between these two groups statistically significant at the level of 0.01?

Online t-distribution calculator:

https://homepage.divms.uiowa.edu/~mbognar/applets/t.html

So, I will just like other lecture, I leave you with a problem to solve I have experimental data here. So, what we are doing we are trying to suppose generate some inhibitor for enzyme and we have done an assay we have a control that is untreated enzyme the other one is with inhibitor and the data is shown here.

The number of samples here for control is 5 and this one is also number of samples 5. So, degree of freedom will be sample plus sample number minus 2. And I have already given you the t value I have calculated that is 3.855. So, what you have to do at home, you have to check out is the difference between these two groups, difference between these two groups this one and this one is statistically significant at the level of 0.01 that you have to test.

And to do this test, you have to use some sort of t distribution earlier days people used to use t tables, but now a days in a computer you can get the whole t distribution you can calculate the exact value p value. So, t tables are becoming redundant. That is why I am not insisting you to do use t table but rather you require some tool to calculate the t distribution to calculate the probability.

So, there are online tools also one example I have given here, but if you do a search on the web, you can find something else. So, use that tool and find out that the answer to this question is that difference between these two groups statistically significant at the level of 0.01. That is all for this lecture. See you in the next session. Till then happy learning.