

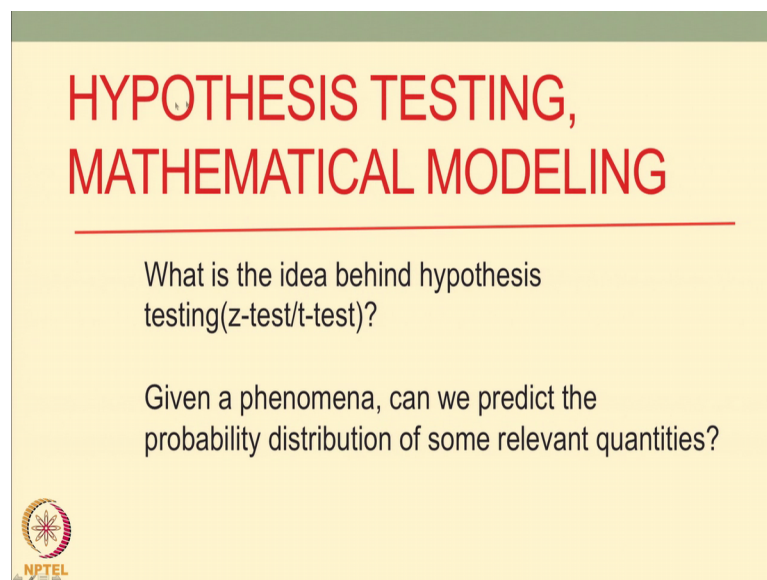
Introductory Mathematical Methods for Biologists
Prof. Ranjith Padinhateeri
Department of Biosciences & Bioengineering
Indian Institute of Technology, Bombay

Lecture - 40
Hypothesis Testing and Mathematical Modeling

Welcome to this lecture on Mathematical Methods. This is towards the end of the course and we learned many new things, we learned how to do calculus which is essentially how to think about a function is a graph and then how to find derivatives and what is the meaning of derivatives which is slopes and so on and so forth. And then we learned applications of this like diffusion and many of the phenomena that we discussed and we learned some statistics and probability distribution.

Now, it was the end of this course we would ask 2 questions. So, that will give us something beyond this course this is an entry this is an introductory course. So, beyond this introductory course if you really want to do something for your research how do we go about what is the direction there that is what this lecture will set.

(Refer Slide Time: 01:23)

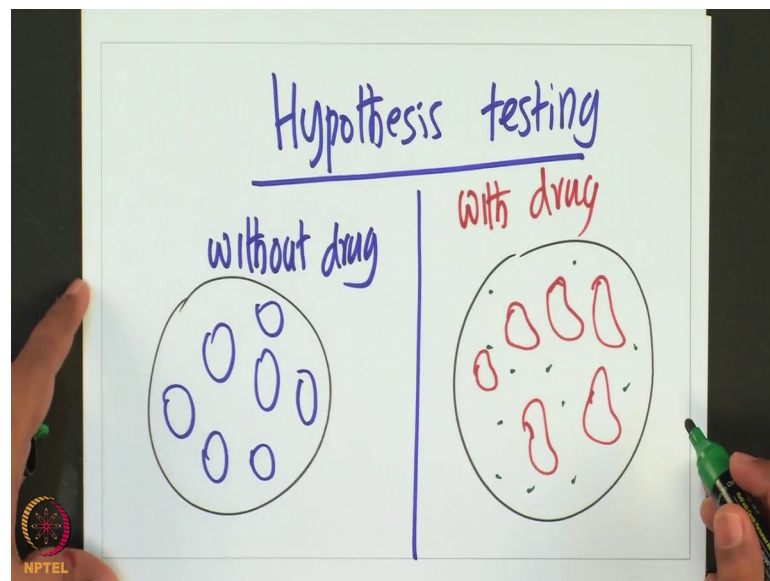


So, the title of this lecture is it has 2 sub 2 topics 1 is hypothesis testing we will discuss and mathematical modeling we will discuss. So, first we will briefly discuss what is the idea behind hypothesis testing? like z-test and t-test. So, will only discuss the idea behind it and of course, the details etcetera would be discussed in a complete biostatistics

course. And in the second part we would discuss given a phenomena can we predict the probability distribution of some relevant quantities.

Because we are always computing things from probability distribution and now can we get the probability how do we know, what is the probability distribution given a phenomena. So, these are the things that we would try to answer.

(Refer Slide Time: 02:14)



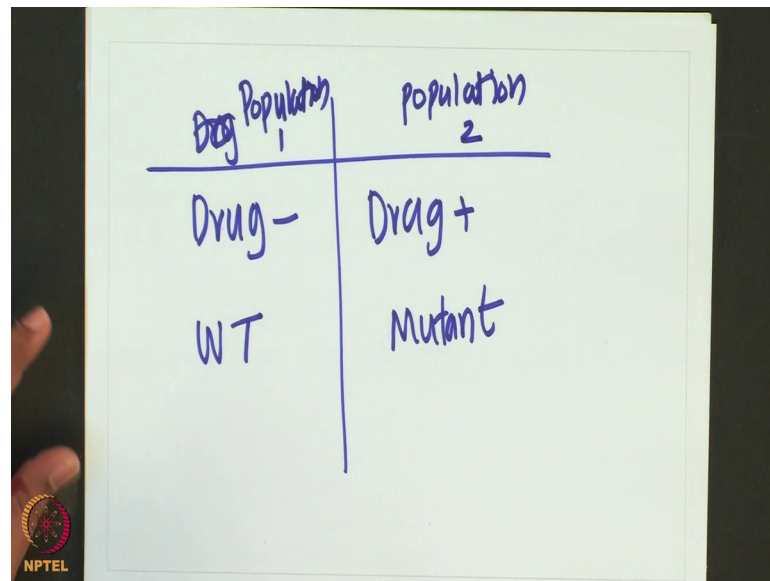
So, first let us discuss this thing called hypothesis testing. So, hypo we will try to understand what is the idea behind hypothesis testing. So, let us think of a typical experimental phenomena that you would do that you will have a set of population of cells or population of individuals and you would want to test something, let us assume that you want to test either a drug effect of a particular drug in a population of cells or some mutations or effect of a drug in a population of individuals. So, this is that something that you would be typically testing.

So, let us think of a population. So, this is a population of cells we are going to draw in this. So, this is population of cells without any drug equivalently you could call it without mutation. So, this could be either without drug and then you could have cells in another petri dish, which will have similar number of cells now after giving some drug. So, you would add some drug here in this some amount of drug and here you will not have any drug just the medium.

Now, if you measure some quantities of this like I say if it is area or any quantity of gene expression of a particular gene in this population versus this population, you would want to test whether they are behaving similarly or differently. So, with and without drug are they behaving differently or similarly this is some obvious question that you want to ask.

Instead of this you could think of a mutation and no mutation right.

(Refer Slide Time: 04:28)



Drug Population 1	population 2
Drug -	Drug +
WT	Mutant

The image shows a hand-drawn table on a whiteboard. The table has two columns and two rows. The first column is labeled 'Drug Population 1' and the second column is labeled 'population 2'. The first row is labeled 'Drug -' and the second row is labeled 'Drug +'. The first row, first column cell contains 'WT' and the first row, second column cell contains 'Mutant'. There is an NPTEL logo in the bottom left corner of the whiteboard.

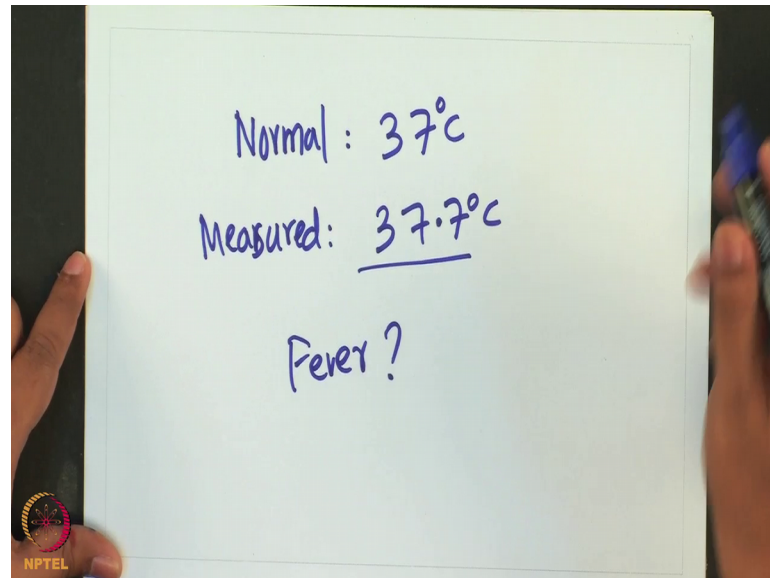
So, you would have a population. So, you could have many many examples as we said you could have drug. So, you could have without Drug, Drug minus and with drug which I let me call this is drug plus and you could have wild type or a mutant. So, there will be always 2 populations. So, let us call this population 1 and population 2. So, we have population they are growing under they are surviving or they are growing under slightly 2 different conditions.

If you measure any quantity let it be gene expression let it be the size let would be anything that we quantitatively measure is there a difference between this and this. So, of course, if you could calculate the probability distribution you can compare them that is 1 way we may not know other probability distribution is so typically what we can do is compare the means.

So, if you just look at the mean area of population 1 versus mean area of population 2. So, this comparison would give us some test and would give us some more information

and whether this 2 population are statistically similar or statistically different, there if there is some difference is the difference is statistically significant or not. We could think of many other examples like also let us say you have let us say we have temperature.

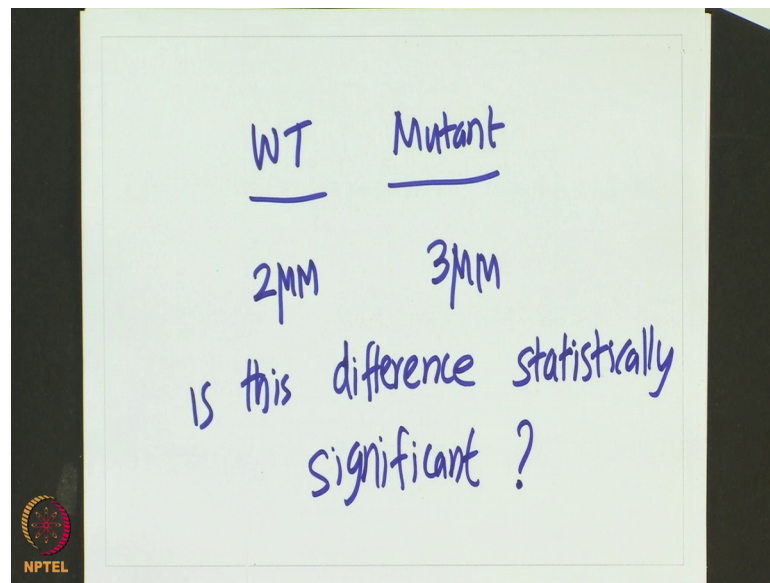
(Refer Slide Time: 06:12)



So, let us say you have we have we know that body temperature is normal is normal body temperature is 37 degree Celsius, let us say and now you have 37.7 degree Celsius when you measured. Will you call this fever or not right is this this difference would you call it fever is this difference statistically significant or not.

Similarly, if you have such population and if you have such 2 population which is drug minus and drug plus you would want to measure some quantities or wild type a mutant.

(Refer Slide Time: 06:58)



A handwritten table on a light green background. The first row has two columns: 'WT' and 'Mutant', both underlined. The second row has two columns: '2μM' and '3μM'. Below the table, the text 'Is this difference statistically significant?' is written in a cursive style. In the bottom left corner, there is a small circular logo with the text 'NPTEL' below it.

<u>WT</u>	<u>Mutant</u>
2μM	3μM

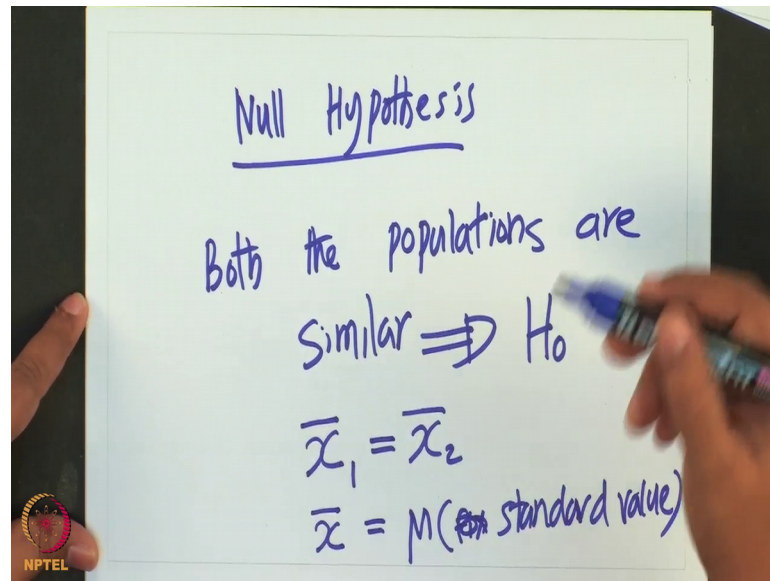
Is this difference statistically significant?

So, is this always you will have a wild type and a mutant and you measured some gene expression and you got here gene expression has 2 micro molar of protein being produced here you got 3 micro molar.

So, is this different is this very different or they 2 and 3 within the error is it is this difference statistically significance is this difference statistically significant or not. This is the question that you would want to answer and for this you could just hypothesis test this review this for this people would use this idea called hypothesis testing.

So, now what is hypothesis testing. So, first the hypothesis in this case we would assume that there are 2 population are similar we have start with the assumption.

(Refer Slide Time: 08:15)



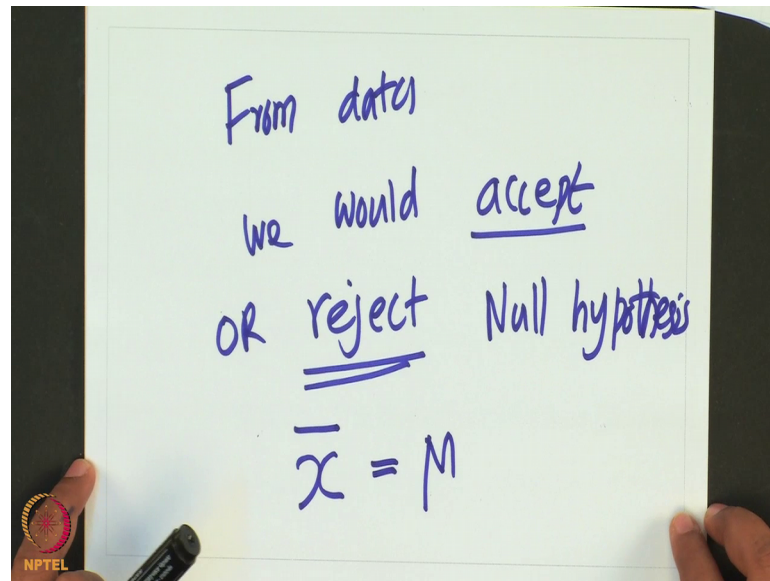
So, that hypothesis called a typically called a null hypothesis. So, you typically would start with something called a null hypothesis and typically what this is it says that there is no difference. So, that is both the populations are a similar there is no difference, that is the null hypothesis typically 1 would start with which is called H_0 which is null hypothesis.

So, both population; so which would also mean which would either would mean the mean from the first population is equal to the mean from the second population or if this is a fix value some time you would compare against a fix value. So, 37 degree Celsius fever is a fix value. So, the \bar{x} you measure the temperature many times and this is equal to μ which is typically a fixed value which is a standard value which is let us say 37 degree which is a typical standard value that you want.

So, either you want to assume that the measure measurement we got is same as the cell similar to the standard value. So, there is no difference or these 2 populations are same. So, if you start with this assumption that is mutant and wild type are the same if you start with this assumption either you can prove or disprove this assumption either you can validate the null hypothesis or you would invalidate you would reject then I had null hypothesis.

So, we would always start with null hypothesis and then we would proceed and either reject the null hypothesis or accept the null hypothesis.

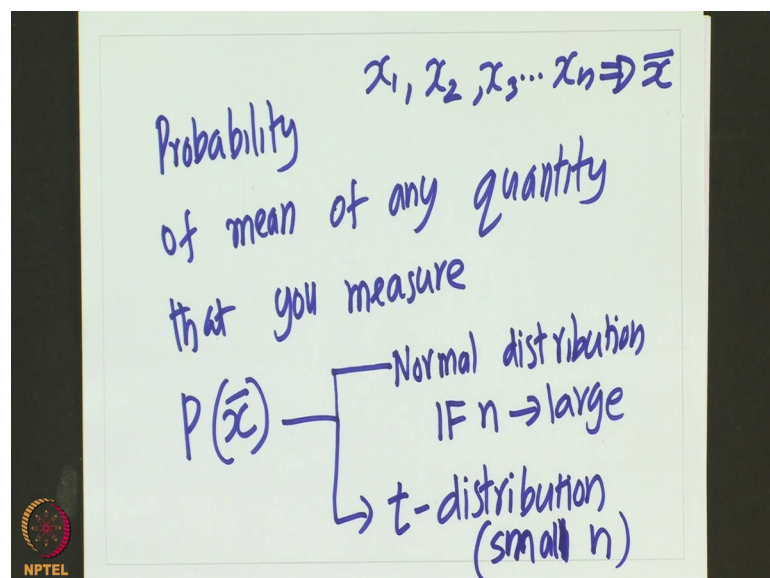
(Refer Slide Time: 10:05)



So, from the data from data we would accept or reject null hypothesis that is what we do.

Now, important point remember is for this what we are calculating we are calculating the mean. So, and then comparing with either a standard value or with some other value of another population. So, we are always starting with mean. So, now, what is the property of mean, we know that it is known that does not matter if you have a quantity whatever you measure the mean of that quantity will have a normal distribution. So, this is something that we mentioned.

(Refer Slide Time: 11:08)

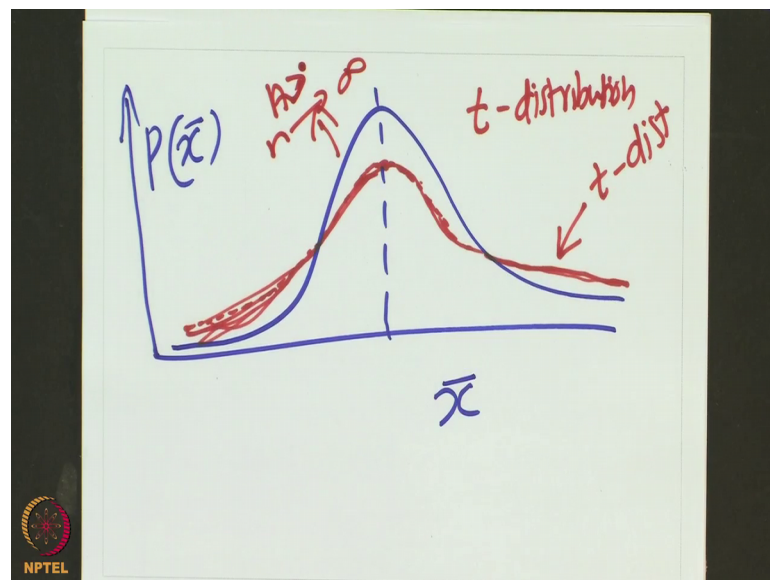


So, probability of mean probability of mean of any quantity that you measure probability of mean of any quantity that you measure so, there is P of \bar{x} . So, you have x is the quantity that you are measuring and so you what we have we have $x_1 x_2 x_3 \dots x_n$ and from which from this we can calculate \bar{x} . So, you do measurement many times and you can calculate the mean or the average which is \bar{x} and if you repeat this experiment many many times.

So, let us say hundred different many different people do this measurement n times and we repeat this measurement and then calculate \bar{x} many people can calculate \bar{x} and if we plot all those \bar{x} you would get a normal distribution, you will get a normal distribution and the assumption is if n is very large if n is very large you would get normal distribution. If n is not very large you will get something a lot t distribution which looks like a normal distribution, but of course, this is slightly different you would get small for small n you will get a t distribution if large and you would get normal distribution.

So, this is the first thing that we should remember that if we have a quantity.

(Refer Slide Time: 13:03)

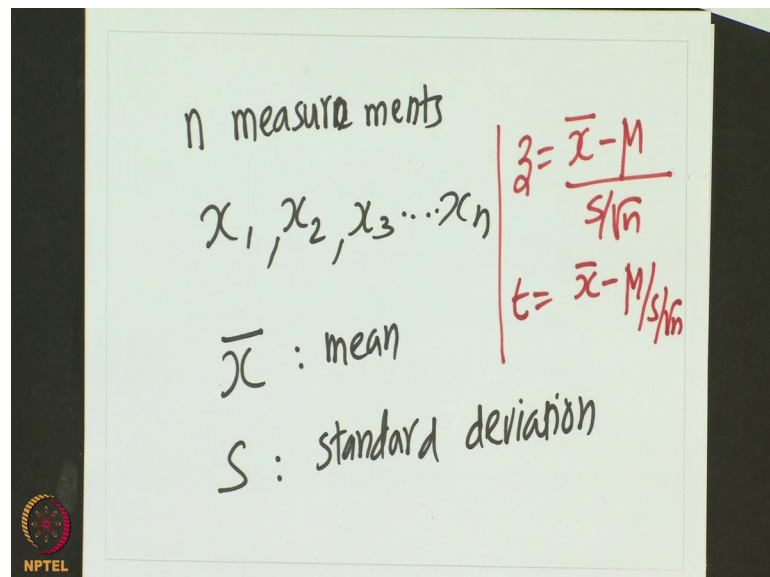


So, this is what does we are plotting here this is P of \bar{x} versus \bar{x} will have a normal distribution if n is very large if n is not that big you will get a t distribution which is looks like this, but it has a slightly bigger tail.

So, which is somewhat like that it is also symmetric if a just pardon me if I will draw it symmetric, but this also is somewhat this is symmetric so, but the property is that this is t distribution t distribution; t the mean will have a normal distribution or a t distribution if this is n is very large. So, n going to infinity if an n is very large you will have a blue curve and n is very small you will have a t distribution.

So, this is the thing that we know. So, let us first think of normal distribution then we will think about t distribution. So, you have. So, let us think of situation first we would do an experiment and we would do n measurements.

(Refer Slide Time: 14:37)



Handwritten notes on a green background:

- n measurements
- $x_1, x_2, x_3 \dots x_n$
- \bar{x} : mean
- S : standard deviation
- Formulas for z and t statistics:

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

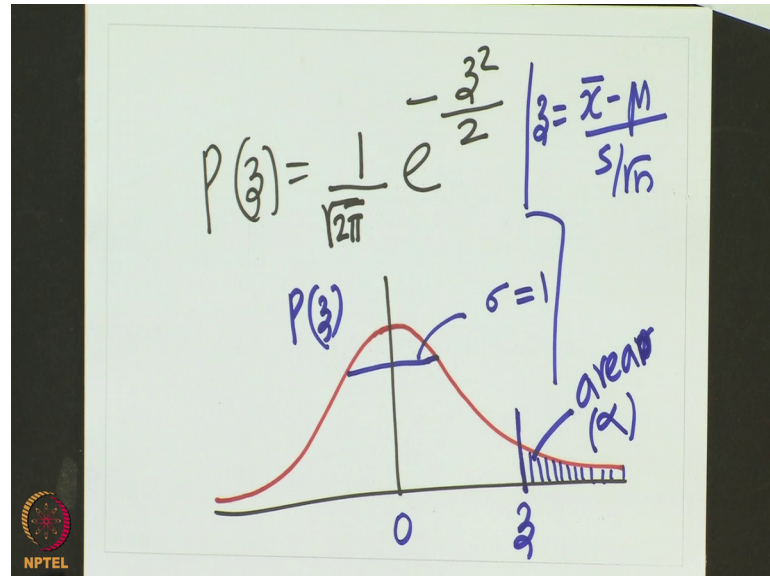
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

So, what are we doing we are doing n measurements. So, you have $x_1, x_2, x_3, \dots, x_n$, which is the mean and you could calculate s which is the standard deviation and given this \bar{x} and this n we can calculate a quantity which is called the z or t, which is z which is \bar{x} minus a fixed number μ which we already will have s by root n.

So, this is so let us say μ is like the 37 degree the control number that we want. So, you want to know that our number is comparable to the number that we have is a our standard number or not. So, we would calculate the z same thing same definition for t and now the distribution of z or t would be a standard normal z the distribution of z would be a standard normal distribution.

So, what is a standard normal distribution? So, it is known that P of z is a standard normal distribution.

(Refer Slide Time: 16:04)



P of z is e power minus z square by 2. So, 1 by root 2 pi. So, would what does it mean integral of this has to be 1 and this is a standard normal distribution; that means, if I just have a distribution it would look like a normal distribution with standard deviation is equal to 1 and. So, this is a distribution which has width which has a width which is a sigma which is 1 and the mean is 0. So, the mean is 0. So, this is z bar. So, the z bar is the mean is 0 and the width is 1.

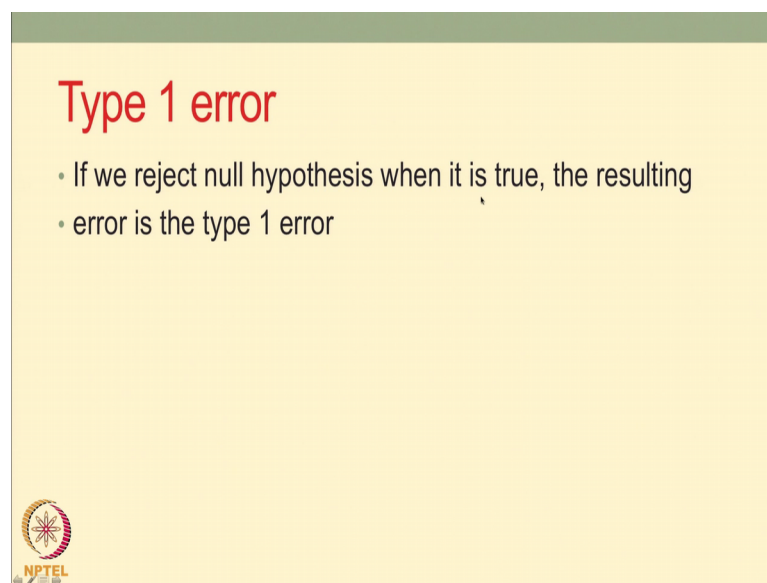
So, this distribution is called a standard normal distribution and P of z will have this kind of a distribution and now we can get the z quantity which is x bar minus mu by s by root n and we can calculate the z and if x bar is equal to mu if or the measured quantity is very close to mu you would get a you would expect a very small set value or it could also be equivalently if you do large number of experiments or and so on and so forth. So, depending on various things here if x bar is very close to mu z will be very small. So, you think about this when z would be very small so, but if z is very large typically it would mean that x bar is much larger than very different much larger than mu and therefore, z is very large.

So, depending on if the set is large or small you would accept or reject the null hypothesis. So, if z is very small you can accept the null hypothesis if z is very large you

would reject the null hypothesis. So, look at this again what is plotted here what is plotted here is P of z. So, now, what are we saying we are saying that z could have any of this value there is a probability of getting any of this value, but if you get a very large value of z if you get a value of z which is very large then which is very likely to be far away from the mean very likely that your hypothesis null hypothesis can be rejected, but even if you rejected there is some there is some probability of getting bigger z values these values.


So, your rejection of null hypothesis could be wrong and there is likely to be in error in rejecting the null hypothesis even if it. So, that error is this area. So, this area beyond z value the that you would get area this area beyond z is an important quantity call alpha which is called a type 1 error. So, this area is called a type 1 error.

(Refer Slide Time: 19:30)



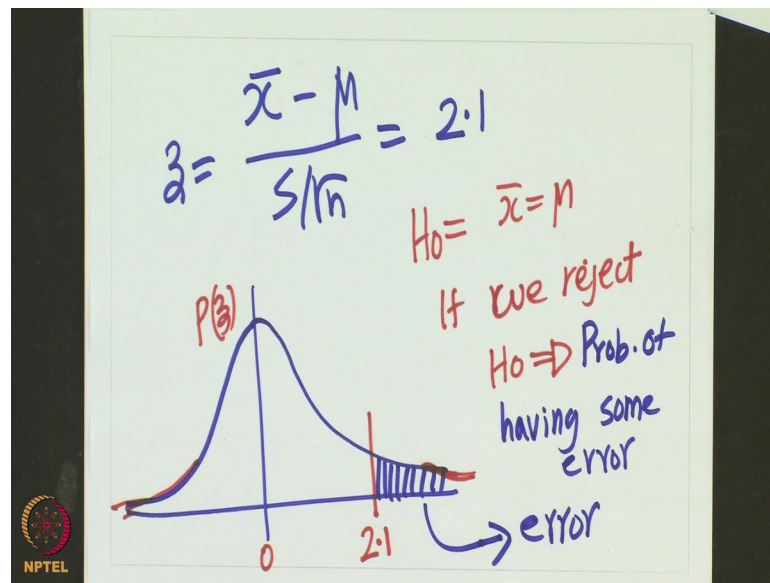
Type 1 error

- If we reject null hypothesis when it is true, the resulting
- error is the type 1 error

 NPTEL

So, what is type 1 error? So, let us look at here if we reject null hypothesis when it is true the resulting error is the type 1 error. So, let me let us take some numbers and explain this quickly.

(Refer Slide Time: 19:44)



So, we do some experiments and get an \bar{x} and we want to compare with some μ value. So, we calculate this quantity z . So, let us say you got 2.1 as the z value. So, this is my z and I got 2 point 1 and what we know is that the distribution of the z is a normal distribution and this is 2.1 which is the z value 2.1 and since this is 0 since 2.1 is much larger than 0, since 2.1 is much larger than 0 very likely that \bar{x} is very likely that \bar{x} is far away from μ .

But this whole distribution the claim of the distribution itself that z could have any of this value the probability of z having this value is small, but there is some finite probability. So, the probability for this z to have some value is 0 very small, but finite. So, there is a finite probability of having large z values as well therefore, if we reject this hypo null hypothesis that H_0 , which says that \bar{x} is equal to μ if you reject it if we if we reject there is likely to be some error, if we reject the null hypothesis there is some probability of getting some error that probability.

So, the probability of having some error there is some probability of having some error. So, that error and that error is nothing, but this area. So, the area beyond z is this error the area beyond. So, the smaller this area beyond z the smaller will be the error in rejecting the null hypothesis. So, if this area is very small if that is if z is very large we can reject the null hypothesis.

So, typically this area you want to be like less than 5 percent or this is nothing secret about 5 percent that is typically what I would want. So, that is what typically I measure in this kind of hypothesis testing there are a lot of details which of course, cannot be discussed in 1 lecture, but I am just indicating you to the use of normal distribution how do we use what is the area beyond some value what does it mean of course, there is something an alternate hypothesis and depending on the alternate hypothesis I would decide you should take area on both sides 2 sided or 2 tails we should consider and so on and so forth. Those are the details which you would learn in a statistical statistics course which is dedicated statistics course which is a subject in itself.

But the idea here is just to point you that if we understand normal distribution properly and the area under the normal distribution these are the ideas that is used in statistical testing hypothesis testing and we would want a z value which is large enough. So, that the area beyond that z value is very small the area here is very small therefore, error in rejecting this null hypothesis so the type 1 error which is the area.

So, if we reject the null hypothesis when it is true the resulting error is the type 1 error and this type 1 error we want to be as small as possible and therefore, we would want z to be as this possible, but it is up to us to decide how much error we can tolerate if we can tolerate 5 percent error we would be happy with it if we can only if you want error to be super small like 1 percent, if you want if you would reject it only if the z is very big such that the area beyond that z value is less than 1 percent or so on and so forth.


So, this is the basic idea just remember the aim of this short introduction about this hypothesis testing it is not to teach you the everything about hypothesis testing, but just you to point out point you towards what is going on just to make you guide you towards it where it should read more about this hypothesis testing in detail.

Then the last part I would want to describe is equations for probability distribution given a phenomena How do we get the probability distribution?

(Refer Slide Time: 24:27)

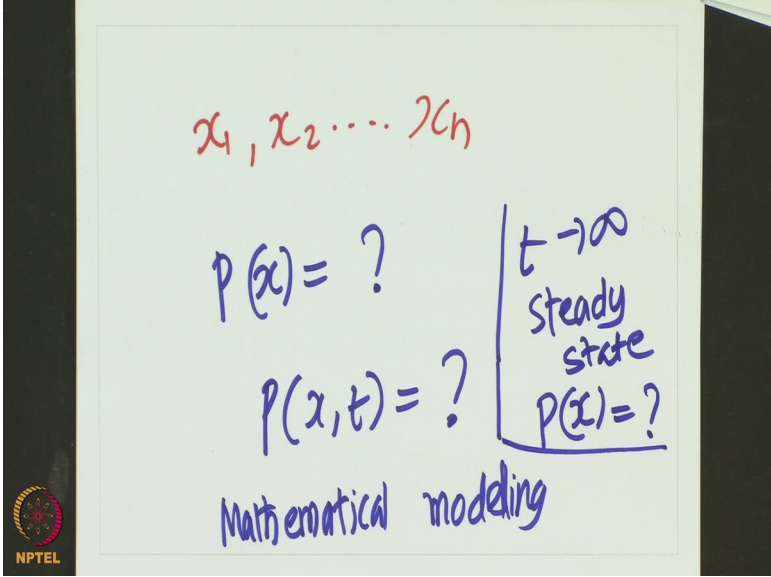
Equations for probability distribution

- Given a phenomena, how do we get probability distribution?



So, we do many may various experiments and in all these experiments we do not know what is the we measure some quantity then we do not know we do not know what is the probability distribution of that quantity, we are measuring some quantity x 1 many times we measure x n .

(Refer Slide Time: 24:43)




x_1, x_2, \dots, x_n

$P(x) = ?$

$P(x, t) = ?$

Mathematical modeling

$t \rightarrow \infty$
steady state
 $P(x) = ?$



So, some quantity x is what we measure and we repeated many times and we measure n of this, we do not know what is P of x a priori we do not know we do not know we are not sure, whether it will be a normal distribution whether it will be a whatever what

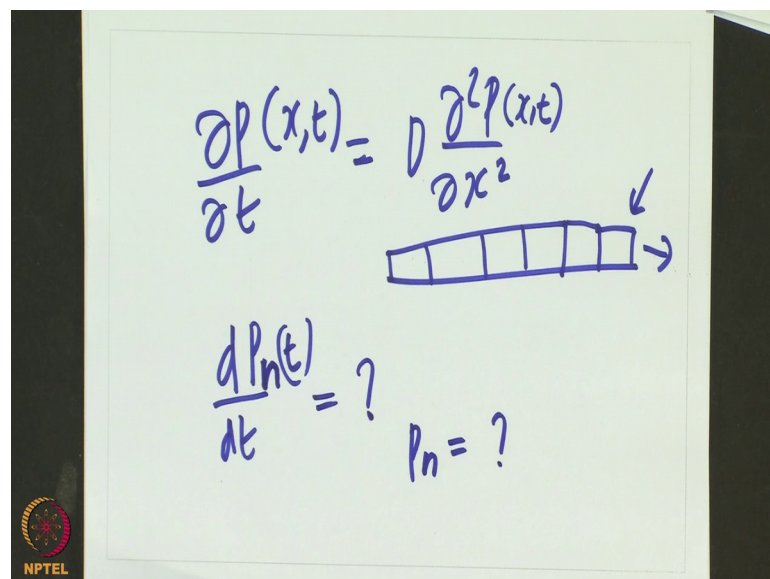
would be a distribution it would be exponential what would be the distribution we do not know a priori.

But can we know this how do we know this can we use mathematics to know this, it turns out that it is possible and people write equations for probability distribution the way to do that is so a lot of mathematical modeling is done to understand to predict what would be the probability distribution given a phenomena.

So, the aim of this mathematical modeling is to predict P of x comma t to predict can we predict this given some phenomena, this is the aim of this is what lot of mathematical ultimately from lot of mathematical modeling lot of mathematical modeling is to can we get this or can we get at least 4 very large t . So, when t is very large that if you wait long enough when in a steady state where things will be independent of t many things are in a steady state can we get P of x . So, these are the either can we get P of x for any given time or can we get P of x in the steady state steady state means t is very large and it is independent of time it does not it is not independent of the initial condition that you start with it is just independent of time itself.

So, can we get this and how do we do that people write differential equations for t . So, x p so typically what you would write is we would write equations like $\frac{\partial P}{\partial t}$ by $\frac{\partial^2 P}{\partial x^2}$.

(Refer Slide Time: 26:45)



The image shows handwritten mathematical expressions and a diagram on a green background. At the top, the equation $\frac{\partial P(x,t)}{\partial t} = D \frac{\partial^2 P(x,t)}{\partial x^2}$ is written. Below it is a diagram of a horizontal chain of six squares, representing a lattice. An arrow points to the right from the rightmost square, and another arrow points downwards from the top of the rightmost square. Below the diagram, the equation $\frac{dP_n(t)}{dt} = ?$ is written, followed by $P_n = ?$. In the bottom left corner, there is a small circular logo with the text 'NPTEL' below it.

So, we could write equations like this for example, diffusion equation like equations are typically used for random moments random walks are something is randomly moving and there were some quantities randomly changing, then I would write equations like this which is like the diffusion equation there are equations like the master equations. So, I would write $\frac{dP}{dt}$ for some quantities. So, if you have a polymer which is polymerizing and depolymerizing and let n be the size of the polymer p_n of t is the probability of having size n what is the probability of finding polymers with n monomers and we can write equations for such probabilities and so all those equations and get P_n .

So, this would give us by. So, solving such equations I would get P_n . So, such equations are typically called master equations. So, mathematical modeling is basically writing equations for such p_n and probability distributions and then solving them to predict to understand what would be the probability distribution given certain phenomena and these equations will reflect the phenomena. So, what is this equation that you would write down it depends on what is the phenomena that we want to describe.

So, this is what this I would do in an advanced mathematical modeling course where I would learn how to write this equations and how to describe a phenomena given, how to write down equations given a phenomena and how to understand how to describe the probability distribution etcetera in terms of P of n or the P of x

With this summary I would just stop this lecture and this is towards this is the end of this course where we want to conclude by saying that mathematics is again like a language where, we would learn how to describe the various how to describe various phenomena that we see around us and the nature in general like let it be whatever be the physically I which is the it should to be in living science life sciences or any other branch of science and engineering.

We could use mathematics is a tool to describe any phenomena that we see around us in a quantitative manner and the things that we learned is the primary fundamental lessons on start thinking about how to use this math mate this mathematics is a tool to describe things around us with this I will stop here, Goodbye good luck.