Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 05
Normal Distribution

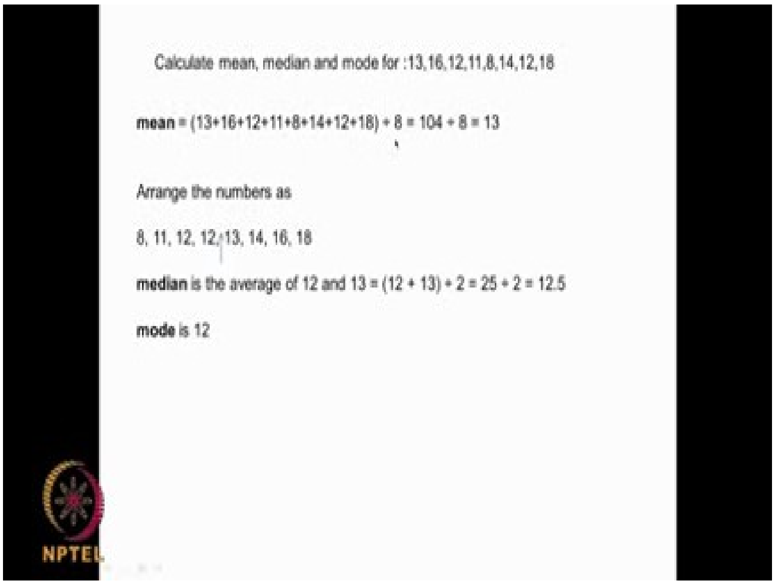
So far we have looked at discrete distributions like Poisson, Binomial. Let us now look at Continuous Distribution and the most important one out of that is Normal Distribution. Normal Distribution is also called Bell shaped curve, Gaussian curve and so on. And generally we assume any data behaves like a normally distributed that means, if I take 10 or 20 students in a class and measure their heights, generally it will fall Normal Distribution. That means there will be an average, there will be some students who will have less height than the average, some students will have greater than the average and the proportion will be almost same.

(Refer Slide Time: 00:57)

Calculate mean, median and mode for :13,16,12,11,8,14,12,18

There are certain terminologies which we need to learn and they are quite simple you must have studied long time back also. They are called mean, median, mode. What is a mean? Suppose I have a set of data and I want to find the mean of this data set. Mean is nothing but taking the average. So it is quite simple.

(Refer Slide Time: 01:18)



Calculate mean, median and mode for :13,16,12,11,8,14,12,18

mean = (13+16+12+11+8+14+12+18) ÷ 8 = 104 ÷ 8 = 13

Arrange the numbers as

8, 11, 12, 12, 13, 14, 16, 18

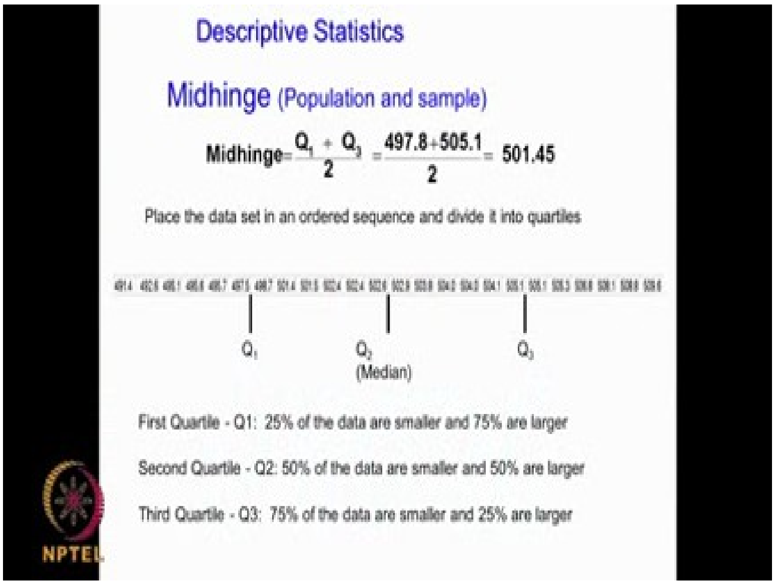
median is the average of 12 and 13 = (12 + 13) ÷ 2 = 25 ÷ 2 = 12.5

mode is 12

We add all these and divide and then we get the mean. Now what is the median of this? Suppose we arrange the data set in this fashion. The middle point is called the median. If you have even number of data so the middle point will be average of these two whereas, if we have odd set of data the middle point will be the center data point. In this particular case we have 8 data sets and the median will be average of 12 and 13. Now what is mode? Mode is nothing but the value which comes more often. For example, it is centered around that value, if you look at these data set 8, 11, 12, 12, 13, 14, 16, 18 we find it 12 as more common or it the distribution is centered around this so the mode is 12 in this case.

So this particular data set we have a mean of 13, median of 12.5 and mode as 12 and generally for a Normal Distribution the mean, median, mode are the same. That is why Normal Distribution is called a uniformly distributed data set and look like a bell shaped well distributed data set.

(Refer Slide Time: 02:43)



Descriptive Statistics

Midhinge (Population and sample)

$$\text{Midhinge} = \frac{Q_1 + Q_3}{2} = \frac{497.8 + 505.1}{2} = 501.45$$

Place the data set in an ordered sequence and divide it into quartiles

[illegible]

Q_1
Q_2
(Median)
Q_3

First Quartile - Q1: 25% of the data are smaller and 75% are larger

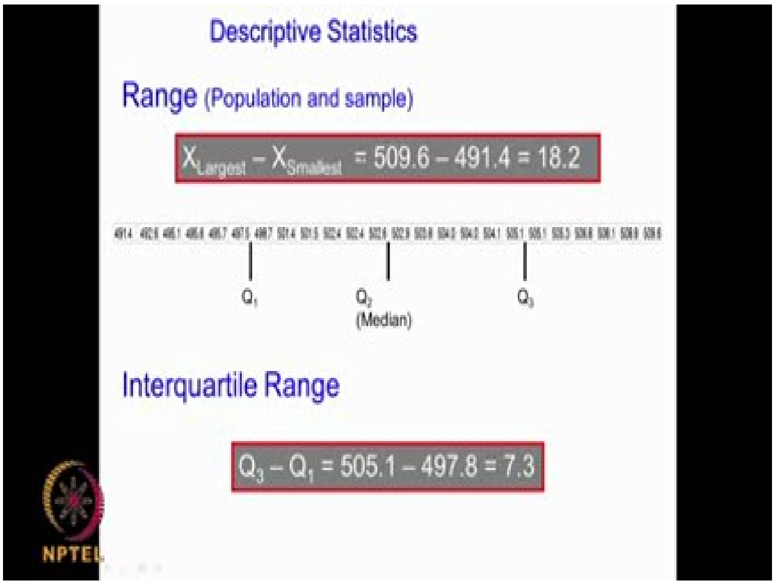
Second Quartile - Q2: 50% of the data are smaller and 50% are larger

Third Quartile - Q3: 75% of the data are smaller and 25% are larger

Then there is something called Midhinge. Suppose you have a quartile 1. What is quartile 1? You have a large data set assume that you can divide this data set into 4 quarters or quartiles. This is called the first quartile and this is called the third quartile. We have divided it into 4 data sets so you have the Q_1 here. So 25% of the data will be smaller than this Q_1, 75 % of the data will be larger than this Q_1. You have a Q_2 that is a quartile 2. So 50 % of the data will be smaller than this and 50 % of the data will be larger than this. Then we have the third quartile Q_3, 75 % of the data will be below and this Q_3 and 25 % will be above this. If it is very uniformly distributed you will have each of these quarters same but if it is not uniformly distributed you will have some variations between Q_1 and Q_3 and so on actually.

So, Midhinge is nothing but the middle point like that is $Q_1 + Q_3 / 2$. In this particular case we have here and we have here take an average it comes to 501.45. So generally the quartiles are very useful to determine whether the data set is uniformly distributed or is it skewed in one particular range, whether Q_1 is not same as Q_3, maybe it is skewed and so on actually. That is the advantage of looking at the quartiles in data set.

(Refer Slide Time: 04:29)



Descriptive Statistics

Range (Population and sample)

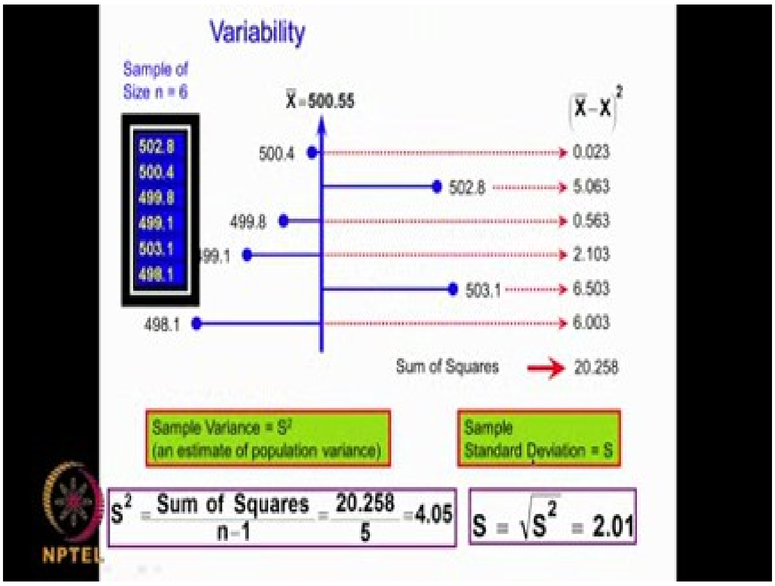
$X_{Largest} - X_{Smallest} = 509.6 - 491.4 = 18.2$

Interquartile Range

$Q_3 - Q_1 = 505.1 - 497.8 = 7.3$

Range, range is nothing but the largest point minus the smallest point. If you have a large data set like this. The range of the range of the data is 509 to 591. If I am measuring the fermentation yield between 50 and 20 °, I would say 50 is one end, 20 is another end the range is 30 °. If I am measuring the growth of an organism between pH 3 and 8 so the range will be 5. This is very obvious and we have been using that then there is also inter quartile range that is nothing but Q_3- Q_1. You know Q_3 you know Q_1. You find the difference that is called the inter quartile range. Now, let us look at the variability of the data.

(Refer Slide Time: 05:17)



Variability

Sample Variance = S^2 (an estimate of population variance)

Sample Standard Deviation = S

$$S^2 = \frac{\text{Sum of Squares}}{n-1} = \frac{20.258}{5} = 4.05$$

$$S = \sqrt{S^2} = 2.01$$

The mean is ok, median is ok, mode is ok. But then we would like to know, how these data set varies with respect to this average? That is a very, very important point because that gives you an idea about the spread of the data. Suppose I am measuring the height of the student in my class I am getting an average of 5.5. Do all the students have their height very close to 5.5? Or is there a large difference from this average of 5.5? Do we have students with 6? Do we have students with 5? So that will give a very large spread and that is going to give you a very large standard deviation. Whereas if the height of the students are very close to 5.5, 5.6 or 5.4 or 5.3 the variations are very going to be very small then the standard deviation of this data set is also going to be very small.

How do you calculate that? You must have studied long time back. If $\overline{X}$ is the average of the data, suppose I have 500.4, 502.8, 499.8, 499.1, 503.1, 498.1 as the data set I taken an average which is this. So,

$$\left(\overline{X} - X\right)^2$$

that is I take the difference with respect to the average square it up then, I add all of them. This is called sum of squares. This is called sum of squares and this is also called variance. So the sample sum of squares / n - 1, n is the number of data points. So

sum of squares / n - 1 is called the sample variance. I take the difference between the $\overline{X}$, which is the mean and the data point square it up, I add up there I get this something called sum of squares. If I divide this sum of squares by n - 1 we get something called sample variance. It is denoted by a square.

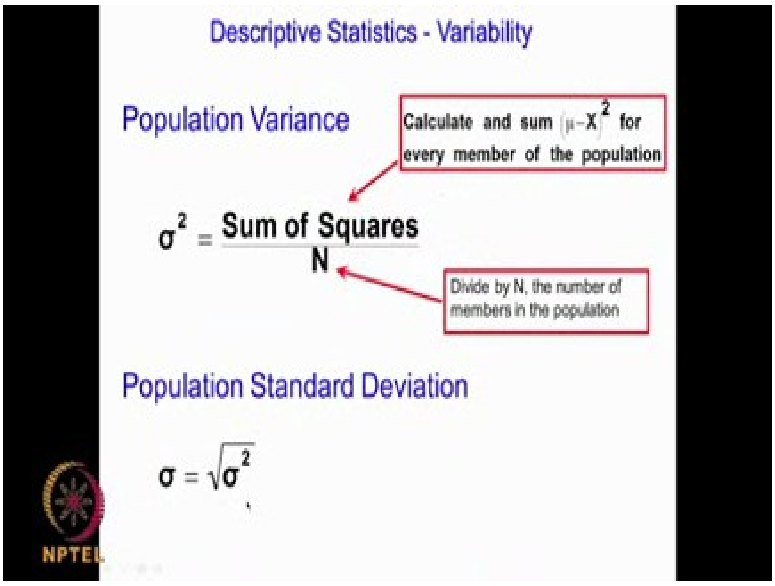
Why is this called sample? Because you have taken a small set of data, so the sample standard deviation, we take a square root of that, that gives you the sample standard deviation now this variance and the standard deviation gives you an indication of the spread of the data. That means, how much the data is spread, If this variance is very large or the standard deviation is very large, you can tell the data spread with respect to the mean or the average is also very, very large. If the standard deviation is very small, then we can say the spread of the data with respect the mean is also very small.

So that is the advantage of this and variance is very, very important as I mentioned in my first class that variation is part of any data and so understanding this variation, the reasons for these variation is very, very important in the area of statistics and identifying what are the causes? What are the reasons for this particular variance? Is very, very important, so this standard deviation or the variance of the sample is the way this is how you calculate,

$$\left(\bar{X} - X\right)^2$$

that is come given as some squares then divided by n - 1.

(Refer Slide Time: 08:52)



Descriptive Statistics - Variability

Population Variance

Calculate and sum $(\mu - X)^2$ for every member of the population

$$\sigma^2 = \frac{\text{Sum of Squares}}{N}$$

Divide by N, the number of members in the population

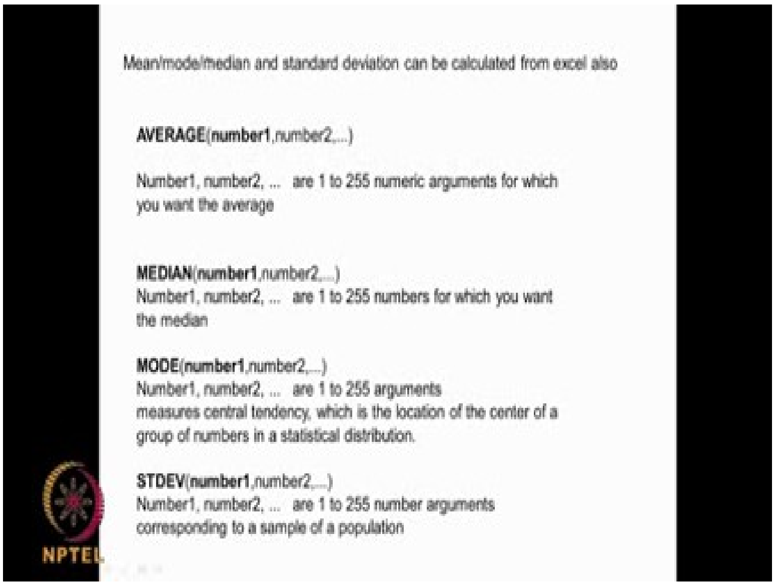
Population Standard Deviation

$$\sigma = \sqrt{\sigma^2}$$

σ Now, in the previous class I mentioned about population and sample. Population is something very, very big. You cannot even comprehend, it is a very large data point set of data. It is like telling the height the average height of an Indian is 5.5 feet. That means, it involves billions of Indians that is a population, whereas if I take about 10 people or walking down the street and take their average height than that is called a sample. That will be represented generally as $\bar{X}$,and whereas when I look at the average height of an Indian I will call that as new. Similarly, generally we say as the population variance and s as the sample variance.

So analogous to a square, we also have σ^2 , where we say sum of squares divided by N as you might have noticed, instead of n -1 which we have it here. Here, we are just having n because the N population the number of data points are huge that does not make much difference whether we takeN or n-1. The population standard deviation of course, is square root of this.

(Refer Slide Time: 10:12)



Mean/mode/median and standard deviation can be calculated from excel also

AVERAGE(number1,number2,...)

Number1, number2, ... are 1 to 255 numeric arguments for which you want the average

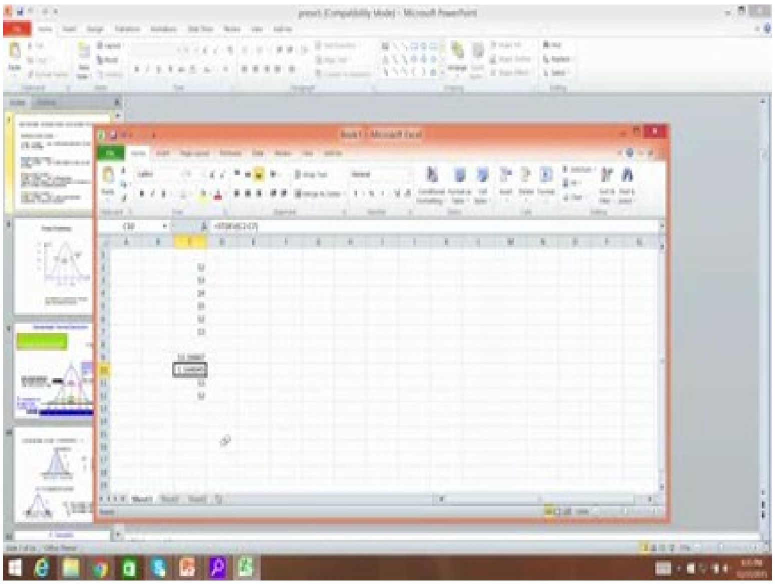
MEDIAN(number1,number2,...)
Number1, number2, ... are 1 to 255 numbers for which you want the median

MODE(number1,number2,...)
Number1, number2, ... are 1 to 255 arguments
measures central tendency, which is the location of the center of a group of numbers in a statistical distribution.

STDEV(number1,number2,...)
Number1, number2, ... are 1 to 255 number arguments corresponding to a sample of a population

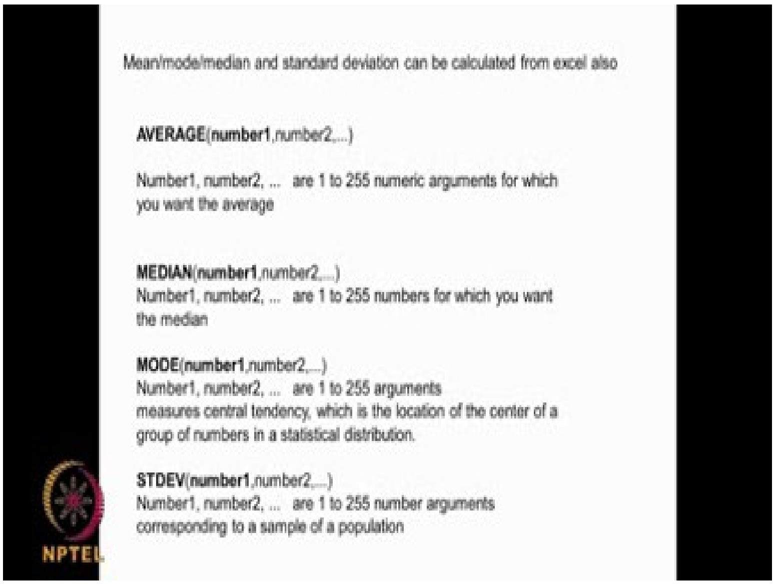
Now, we can calculate all these mean, mode, median standard deviation from Excel also right? There are some commands like average. Suppose if I have a large set of data, I use this function to calculate average. If I have a larger of data, I can use this function median to calculate the median, if I have a large set of data, I can use this function called mode to calculate the mode or the central tendency. If I have a large set of data I can use this command called standard deviation, to calculate the standard deviation of the data set. For example, let us just look at Excel.

(Refer Slide Time: 10:49)



Suppose assume that I have some data points and just giving randomly some data points. I need calculate the average. I put a v e r a g e, average. I put all these points here and I get an average of this, so easy. If you want to calculate standard deviation of this sample set I just write s d e v and then I mark all these I get the standard deviation. Suppose I want to calculate the median, I just say median. Median is nothing, but the midpoint right. So 13 is the median. As you can see 12, 12, 13, 13, 14, 15 the middle point is median. Now mode will be is a central tendency. We have mode as 12 here, we have average or the mean here, we have the median here, we have the mode here, we have the standard deviation. These are quite simple commands, which Excel also has it and we can calculate all these in Excel also. It is very simple.

(Refer Slide Time: 12:13)



Mean/mode/median and standard deviation can be calculated from excel also

AVERAGE(number1,number2,...)

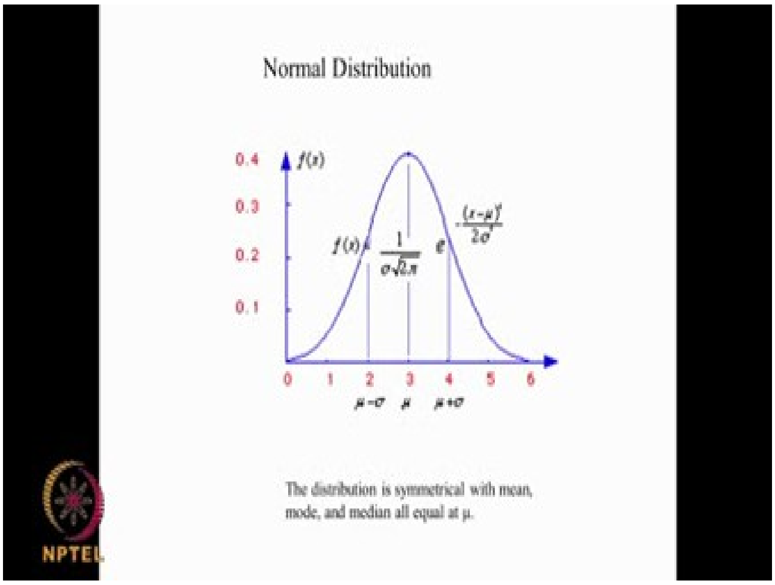
Number1, number2, ... are 1 to 255 numeric arguments for which you want the average

MEDIAN(number1,number2,...)
Number1, number2, ... are 1 to 255 numbers for which you want the median

MODE(number1,number2,...)
Number1, number2, ... are 1 to 255 arguments
measures central tendency, which is the location of the center of a group of numbers in a statistical distribution.

STDEV(number1,number2,...)
Number1, number2, ... are 1 to 255 number arguments corresponding to a sample of a population

(Refer Slide Time: 12:28)



Normal Distribution

The distribution is symmetrical with mean, mode, and median all equal at μ.

Now, let us look at Normal Distribution. It is a most important in statistics, as I said we assume of many systems behaves in a normal fashion but of course, there are some tests which we have to perform to find out whether it is the data set follows a Normal Distribution. So if does not follow then we have to be very careful to use some of these statistical analysis and statistical test we need to remember them. Normal Distribution is very uniform, it is like a bell shaped, what the area under the left hand side is exactly equal to the area under the right hand side the equation is given like this, *f(x)* is equal to that is probability of this function x is equal to

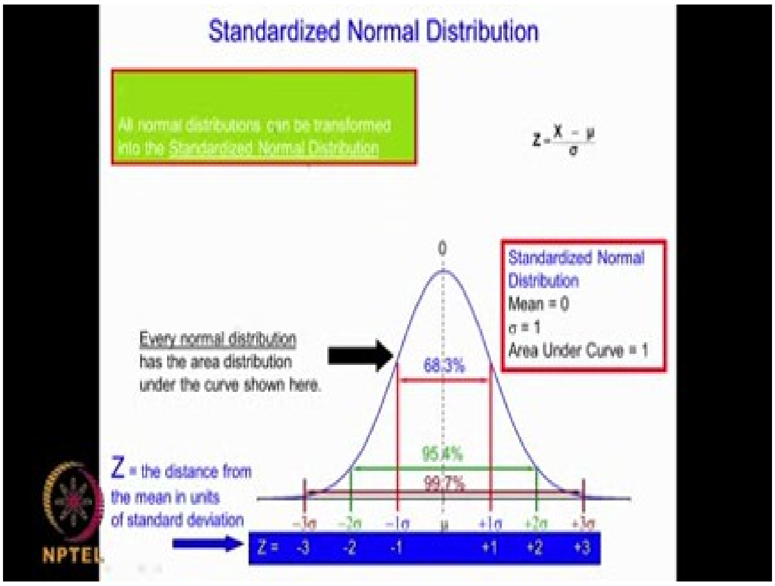
$$\frac{1}{\sigma\sqrt{2.\pi}}$$

$$\frac{(x-\mu)^2}{\;}$$

So μ is the mean or the average and sigma is the standard deviation.

Normal Distribution is symmetric. We know that in a Normal Distribution the mean, mode and median will all be equal to μ. Especially in a Normal Distribution mean is equal to mode is equal to median equal to μ.

(Refer Slide Time: 13:36)



Standardized Normal Distribution

All normal distributions can be transformed into the Standardized Normal Distribution

$z = \frac{X - \mu}{\sigma}$

Standardized Normal Distribution
Mean = 0
σ = 1
Area Under Curve = 1

Every normal distribution has the area distribution under the curve shown here.

Z = the distance from the mean in units of standard deviation

There is something called Standardized Normal Distribution. The Normal Distribution we can convert it into Standardized Normal Distribution. That is generally represented as Z how do you convert that? We take $Z = \frac{X - \mu}{\sigma}$, μ is the mean of the population σ is the standard deviation. When we do that?

What will happen the mean will become 0 that means you are sort of transforming it and you are shifting it. So that the mean become 0 and the area under the curve becomes 1 and sigma becomes 1. So mean becomes 0 that means, you have shifted your curve and then you have adjusted your curve. So that the standard deviation is 1 and the area under the curve is exactly 1. This is very, very useful because instead of handling problems where the averages and standard deviations are differing wide apart. When we use the Standardized Normal Distribution, we will know that the mean is 0 and the area under the curve is always 1. So that is very useful to use. We can convert most of the problems into Standardized Normal Distribution and there are tables which talk about area under the curve for different values of X actually. We will do some problem and then that.

This is a Standardized Normal Distribution where we have the mean as 0, area under the curve is 1 and σ is 1. So we have $\mu + \sigma$ $\mu - \sigma$ $\mu + 2\sigma$ $\mu - 2\sigma$ $\mu + 3\sigma$ $- 3\sigma$. So corresponding to that Z if you see it will become 1 σ will become 1, 2 σ will become + 2, 3 σ will become + 3. So - 1 σ will become - 1, - 2σ will become - 2, - 3 σ will become - 3.

All you have to do is here is substitute there $\mu = 0$, $X=-3\sigma$. So Z will become - 3.

Now as I said this is area under this curve it is equated to 1 in a Standardized Normal Distribution. If you have plus or minus 1 σ this particular area is 68.3 % of the total area that means, it will be 0.683 plus or minus 1 σ is 0.683 plus minus, 2 σ, it is 95.4 %. That means, approximately 0.95. This area spanning the plus or minus 2 of Z will be 0.954. Similarly plus or minus 3 σ will span 99.7 % of this area. So plus or minus 1 sigma will span 68.3 % of the area or it will have value of 0.683 or plus or minus 2 σ will have a value of 95.4 or 0.954 area plus or minus 3 sigma will be 0.997 and so on we can have plus or minus 4 sigma 5, 6 and so on actually because this is an exponentially decaying. As we go along we will add little bit of the area, because the area as we go long becomes smaller and smaller. But still it will try to span as much of the areas possible.

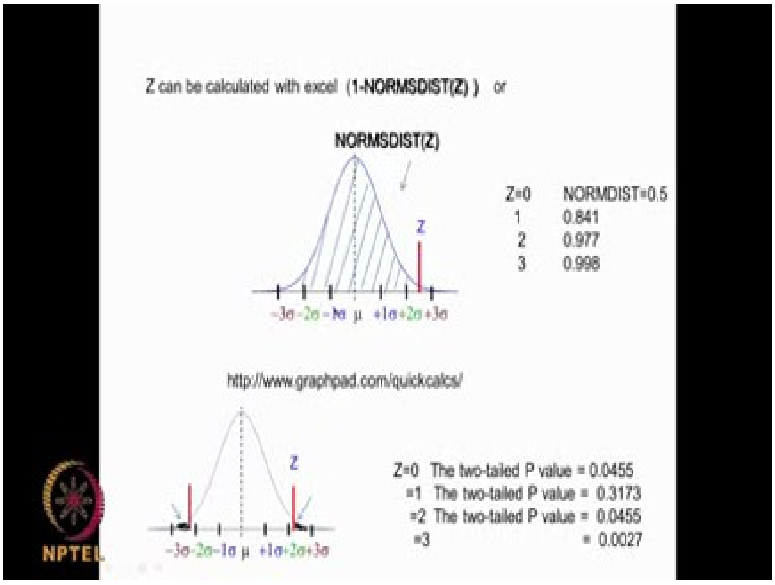
When you say plus or minus 1 σ this area is 68.3. That means, the remaining area this plus this is going to be approximately 32 %. Similarly plus or minus 2 sigma this area is 95 % say when you say it 95 % the remaining area is 5 %. That means, this side will be 2.5 % and his side will be 2.5 % assuming need to be symmetric now similarly plus or minus 3 sigma if we call this as 99 % the remaining outside will be totally 1 %. That means, this side will 0.5 % this side will 0.5 %. So, plus minus, 2 σ will be 0.95 area outside will be 0.5. This side will be 0.25 other side will 0.25 similarly plus or minus 3 sigma will be approximately 99 percent. So outside area will be 1 %. That means, this side is 0.5 %, other side 05 % or 0.005 and 0.005. That is the advantage of converting data the set of a Normal Distribution to Standardized Normal Distribution. So What you do is if I know the μ and if I know σ all I do is

$$Z = \frac{X - \mu}{\sigma}$$

because you are shifting the curve. So that the X become 0, sorry and the area under the curve, becomes 1 and σ becomes 1. When I say plus or minus σ a Z = 1 minus 1. When I say plus or minus 2 σ Z will be plus 2 and minus 2. When I say plus or minus 3 σ Z will be plus 3 and minus 3 now these numbers are also very important when you say plus or minus 1 σ area is 68.3 % plus or minus 2 σ 95 % plus or minus 3 σ 99 %. These numbers will become very important because later on we are going to you will keep on looking at these 95 % 99 %.

So when you say 95 % you are talking in terms plus or minus 2 σ. When you are talking 99 %, we are talking in terms of plus or minus 3 σ. Generally, in statistics most of these significant analysis is done around 95 % That is 2 plus or minus 2 σ or 99 % that means plus or minus 3 σ We are looking at data spreading around a average with plus or minus 2 σ which is 95 σ or plus or minus 3 σ which is 99 %. Many in the future we are going to use these 2 numbers 95 and 99 and now you understand what it mean? 95 %means it is spanning a plus or minus 2 σ area 99 % means plus or minus 3 σ area.

(Refer Slide Time: 20:18)



Z can be calculated with excel (1-NORMSDIST(Z)) or

NORMSDIST(Z)

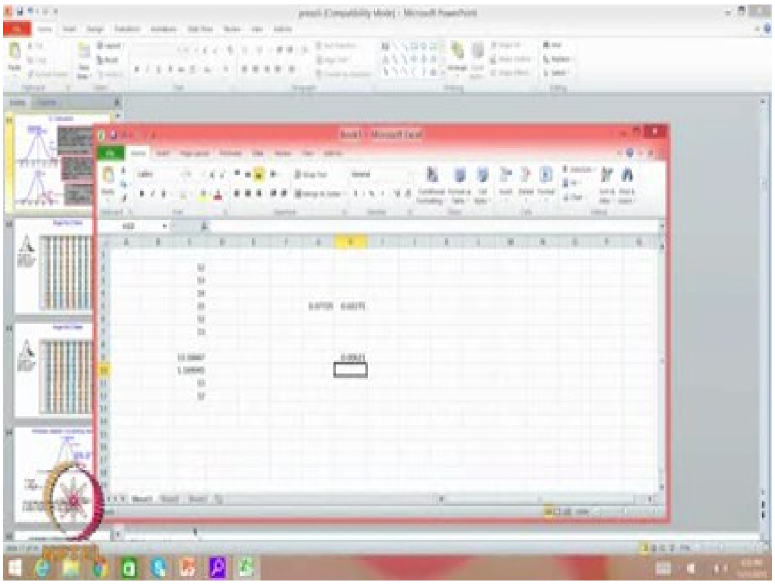
Z=0 NORMDIST=0.5
1 0.841
2 0.977
3 0.998

http://www.graphpad.com/quickcalcs/

Z=0 The two-tailed P value = 0.0455
=1 The two-tailed P value = 0.3173
=2 The two-tailed P value = 0.0455
=3 = 0.0027

Now Z can also be calculated with Excel. There is a command called NORMSDIST Z. NORMSDIST Z, and I said the area under the curve is 1. If I want to calculate, what is this area? And what is this area? At this place suppose I give a value of Z here and I want to calculate this area and I want calculate this area. I can use this particular command, 1 minus NORMSDIST Z NORMSDIST this particular area. If we want to know what this area is I can just say 1 minus NORMSDIST. When I put Z is equal to 0 in NORMSDIST it will give me us 0. 5 that is this area correct because this total area is 1. We can say this area is 0.5. When Z is equal to 1 here, that means here. So this area is equal to 0.841. When Z is equal to 2 this area is 0.77. The remaining area will be 1 minus 0.977, that means 0.023 and if we put here Z is equal to 3 it will give me as 0.998. If you want to calculate remaining area I put 1 minus NORMSDIST. Let me do it for you here, I just say NORMSDIST, oh sorry NORMSDIST it is, yeah that is 0.5.

(Refer Slide Time: 21:52)



So when I put NORMSDIST is equal to 1 that is 841 that means, what I am saying is when I put it here, this side of the area is 0.84 when I put it here Z is equal to 2 this area is 0.977 and 0.84. So when I put it as 2 then 97, the remaining whatever is 1 the right side if you want to calculate I put 1 minus this that is equal to 0.2275. That is whatever on the right hand side is given by 0.02275. Similarly, when I put Z is equal to 3, it gives me 0.998 as there is this area. If you want to know what is this area on the right side I will say 1 minus this. Using Excel also we can do and the command here is NORMSDIST and you can also use the graph pad also to do the same thing actually you know. But the graph pad gives it to you in another form, when you put z it gives you the area on both sides outside area actually.

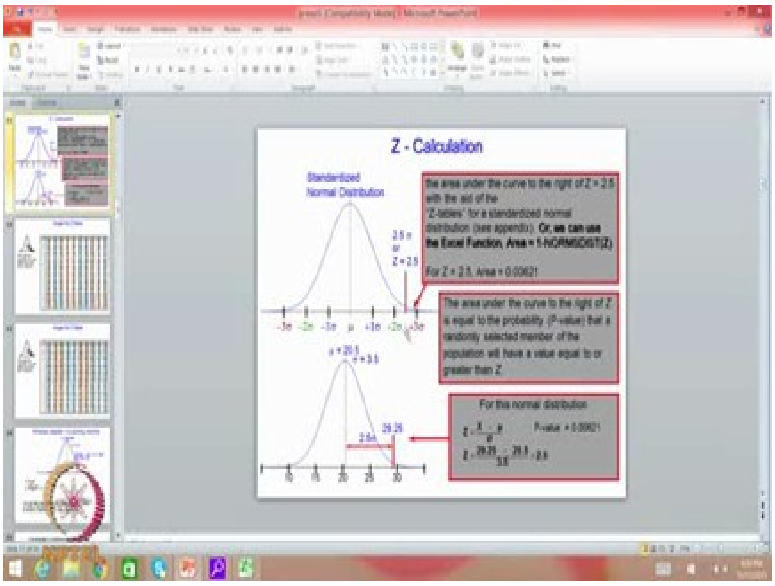
Whereas Excel gives this area graph pad gives you the area outside, both sides that is called two-tail, the two-tail. So when I give Z is equal to 1 it gives me this area. When I give Z equal to 1, it gives me this area and so on actually, here it is giving these 2. Suppose if we want to know only one side of it, I just divide by 2 to get the area on only one side of it, understand?

We can use the GraphPad also to calculate as you can see it is tells you how to calculate different parameters here statistical to calculate. We can say, here we have so we can give a number suppose give a number as 0. It is giving here as p value that is whatever is outside. If I give a number as Z is equal to 1 as giving as 0.3173 that is the outside and so on actually. Even I give Z is equal to 2. So it is giving 0.0455. That means, it is giving this area 0.0455 is almost that is this one it is giving it as 0.0455 that is approximately 0.5 and so on actually. Actually there is a mistake here, this should be 1 here. We can use either the NORMSDIST command in Excel or we can use GraphPad to calculate Z and you can use numerically using a calculator also from this formula is

$$Z = \frac{X - \mu}{\sigma}$$

This is very useful because when we convert any data into a Standardized Normal Distribution we are shifting it so that the mean comes out to be 0 and the area under the curve comes out to be 1. So we will lot of problems as we go along using this type of command. For example, if I want to calculate 2.5 σ. What is Z? It is very simple. All I do is I will put a 2.5 σ μ will become 0. So Z will become 2.5. Now for Z is equal to 2.5 what s the area? I can use one of these, I can use NORMSDIST to calculate this area and then subtract from 1 to get this area. So how do we do that? I will put go to Excel I will do 1 minus NORMSDIST will give 0.00621.

(Refer Slide Time: 27:19)



Z - Calculation

The area under the curve to the right of Z = 2.5 with the aid of the "Z-tables" for a standardized normal distribution (see appendix). **Or, we can use the Excel Function, Area = 1-NORMSDIST(Z)**

For Z = 2.5, Area = 0.00621

The area under the curve to the right of Z is equal to the probability (P-value) that a randomly selected member of the population will have a value equal to or greater than Z

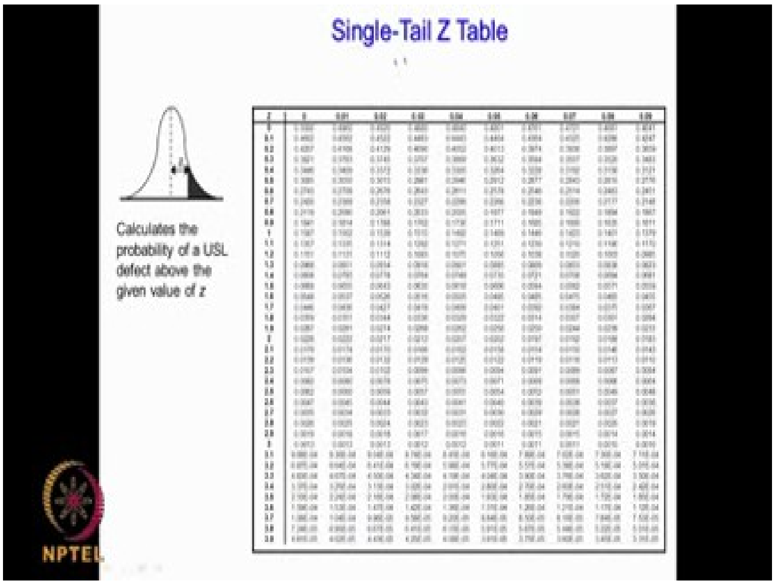
For this normal distribution

$z = \frac{X - \mu}{\sigma}$ P-value = 0.00621

$z = \frac{29.25 - 20.5}{3.5} = 2.5$

That is this area is 0.00621 whereas, if you want the whole area that is what NORMSDIST give actually. That is the advantage of converting a Normal Distribution into Standardized Normal Distribution and we are going to do many problems using this particular command. If we look at this table this is called a single tail Z table.

(Refer Slide Time: 27:46)



Single-Tail Z Table

Calculates the probability of a USL defect above the given value of z

It is single tail because we are looking at only this. When Z is equal to 0 that means if it is here, this area is 0.5 that what this gives. When Z is equal to 0 here, when Z is equal to 1 here this area will be 0.1587. If we are looking at 2 tails but that means, if we are looking at both the sides all you have to do is multiply 0.1587 with 2. So you will get you will get the about 0.316 correct. In fact, that is what this is 0.316 that is both the sides area are double tailed. Whereas this table gives you the single tail here and similarly if you are looking at for Z is equal to 2 what is the area on this side? It will be going down 0.0228. If we want two-tail then, I multiply 0.0228 with 2 that comes around 0.0456 and that is what we have here the graph gives 0.0456 because graph pad gives you on both the sides it is called the two-tail.

I am introducing one more terminology that is called single tail and two-tail one side of it single tail. If you are looking at a situation where you have both sides of it that is called two-tail, we will be using this terminology quite often. So GraphPad gives you area for both the sides. If you want to calculate only one side I will divide by 2 or if I use this table this table gives you area outside on only one side. So, if you want to calculate both sides then I multiply by 2. For 2, I am getting 0.228, for 3 it x 0.0013. If I want a two tail, it will become 0.0026. That is what graph pad gives 0.0026.

We can use different approaches. We can use this table, we can use the NORMDIST which gives in a different way and you need to convert 1 minus and then if you want two tail you multiply by 2 or we can use this graph pad calculator which anyway straight away gives both the sides. So, many different approaches by which one could calculate the area under the curve either internally area or the 1 minus area for a given Z. This is called a Standardized Normal Distribution. In the next class we will look at some problems related to this Standardized Normal Distribution and how useful it is you will see when you start doing problems in this case.

Thank you very much for your time.

Key words- Continuous Distribution, Normal Distribution Sigma, Mu, Mean, median, Mode, Normal Distribution, Excel, Graph Pad, NORMDIST, Normal Distribution, sample variance