

Practitioners Course in Descriptive, Predictive and Prescriptive Analytics
Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture – 14
Normal Distribution

Good evening students, welcome to yet another lecture of practitioners approach in descriptive predictive and prescriptive analytics. And we have been looking at different aspects of analytics and we are building the foundation and the growing up upon which we can actually learn bigger tools and important tools. And today we are going to discuss one of the most important concepts in the analytics which is called as a normal distribution.

You might remember that in the previous class in an earlier class, we have gone through the probability then probability density function, probably continuous cumulative density functions and extra like that. But today what we are going to talk about it is more from how do you; how do you look at the normal distribution which is the cornerstone of many of the analytics tools from a practitioners viewpoint.

(Refer Slide Time: 01:09)

HISTOGRAMS

→ Pictorial representation of the Frequency distribution. Switch groups data into classes.

- Histogram is really nothing more than a bar chart used to chart quantitative data
- Consider the data collected on lunch time taken by coworkers

1	2	3	4	5	6	7	8	9	10	11	12
30	35	35	40	40	40	45	45	45	45	50	50
50	50	50	55	55	55	55	55	55	55	60	60
60	60	65	65	65	65	70	70	70	75	75	80

12 x 3 = 36
 3 rows
 90, 90, 90, 100, 100, 105
 30-34, 35-39

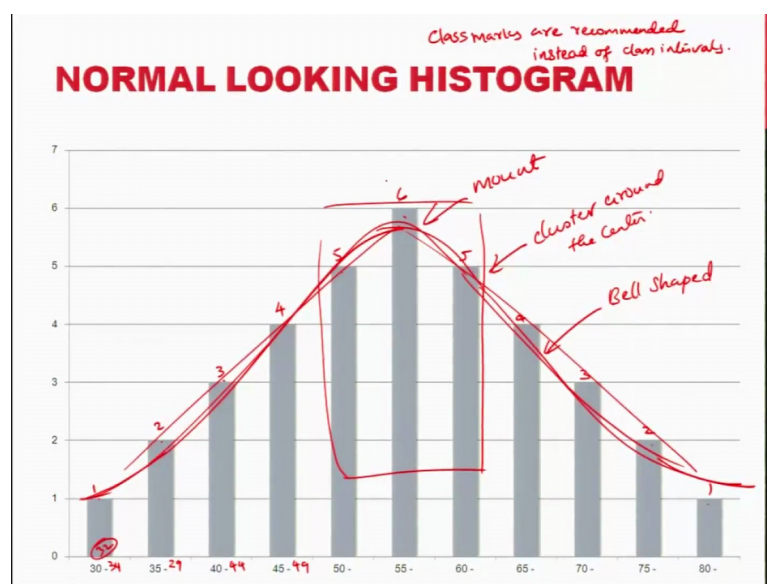
- Create a frequency distribution with a class width of 5
- Using the counts in the frequency distribution, create a histogram
- What is the shape of the histogram?

So, let us first talk about histograms you already talked about what is a histogram and we have also defined that histogram is a pictorial representation of the frequency distribution. So, histogram is really nothing more than a bar chart used to chart quantitative data. So, your quantitative data you are trying to chart; so, in another way it is a pictorial representation of the frequency distribution.

And the frequency distribution is which groups data into classes; we have seen this in the earlier lecture. So, histogram is a pictorial representation of the frequency distribution which groups data into classes; the quantitative data is footage into classes So, we had already seen this data this lunch data we have been lunchtime taken lunch time of coworker data, we have already seen there are 1 data values 1, 2, 3, 4; this is 1, 2, 3 all the way up to 11, 12 and there are 3 rows like this and So, we have 12 times 3; 36 data values are part of this ah.

Remember we try to talk about how many classes and other things, but let us create a frequency distribution with a class width of 5 class width of 5 means you can think about starting it does 30 to 35 or actually 30 to 34 if you think about it 30. So, it will be 30, 31, 32, 33, 34; then 35 to 39 like this that type of a classes ok. And let us then count using the counts on the frequency distribution; we will create a histogram and let us see what is the shape of the histogram let us see what how does a histogram look like ok.

(Refer Slide Time: 03:09)



So, if we take a look into the data you can see that this is the 30 to 34 and then this is 35 to 39; 40 to 44, then 45 to 49; like this. Remember this is one of the reasons why I earlier told that class marks can be used; class marks are recommended instead of class intervals ok. So, you can see here as 32, 34 instead of that you can basically represent it as 32 which will actually is the class mark for this particular class rather than trying to do this interval ok.

So, you can see that the 30 to 34 interval you have one data point, 35 to you have two then you have 3, then you have 4, 5, 6 then you have again 5, 4, 3, 2 and 1 ok. So, you have one value here there are two values, the 3 here 4 observations at 5 observations, there are 6 observations then there are 5, then there are 4, then 3 and the 2 and 1 ok. So, when you think about a histogram like this when you when you plot a histogram like this because this is a unique set of data.

(Refer Slide Time: 04:28)

NORMAL DISTRIBUTION

Why we like normal distribution?

\bar{x} = Mean = Sample average
 $\frac{\sum x_i}{n}$
 \bar{x} = mid point

- If we plot the data values from a relatively large sample or population, the data form a "mound" or "bell" shaped distribution, clustered around the center of the dataset – called as normal distribution
- In normally distributed datasets, mean, median, and mode are equal; and other values are distributed symmetrically around the center
- Having normal distribution of the dataset usually allows the usage of parametric statistics to analyze the data
- Refine what we said before: Use of parametric statistics is based on quantitative data that are normally distributed

*If data is normally distributed \Rightarrow parametric stats is possible.
 mean, std-dev \Rightarrow form conclusion about data \Rightarrow process from data.*

Practical Insight

not so perfect normality in data.

So, it actually looks very nicely, but when you have a system like this; if you plot the data values from a relatively large sample or population, the data forms a mound or a bell shaped curve or bell shaped distribution, clustered around the center of the data set which is called as a normal distribution.

So, what we are saying is that it forms a mound this is the mound that we are talking about the mound ok. And it forms say if you think about it; it is kind of this is what we call as the bell shape ok; this is the bell shaped and this is the cluster around the center

ok. So, that type of a distribution where you have lower law going down tails and a particular curve where there is a hump or a mount at the middle of the data which most of the data values are clustered around it. So, that type of a data set is supposed to be said as something that their data is coming from the normal distribution.

So, why is normal distribution important? The fundamental question is why we like normal distribution or why statisticians fall in love with the normal distribution? The big reason is in normally distributed data sets the mean is the \bar{X} median \tilde{X} and mode is the most common occurring value occurring value ok; they are all equal ok.

So, the central tendency of the data is lying around the mean, median and the mode mean is the average mean implies sample average countered by $\frac{\sum x_i}{n}$ ok; \tilde{X} median is the midpoint of the data. So, the \bar{X} which is the mean \bar{X} \tilde{X} and the mode they are all equal. And other values the values that are other than this mean, median and mode they are symmetrically distributed around the center.

So, it is also a well behaved system. So, that is why the other values that are distributed symmetrically around the center where the mean median and the mode are equal. Having normal distribution of the data set usually allows the use of parametric stat to analyze the data. So, if your data is normally distributed; if data is normally distributed, parametric statistics is possible is possible.

I mean what parametric stats means is you can use the parameters that are taken from the data like the mean, standard deviation etcetera to form conclusions about the data; form conclusion about data. You can form conclusion about the data; that means, you can form conclusion about the process from which the data came in process from data ok. You can do that; if it is not normally distributed the normal distribution is not possible then we use what we call us on parametric statistics, but parametric statistics there have quite a lot of tools and it is quite well developed system. So, if you can get normal distribution; it makes our analysis or analytics life quite easy ok

So, refining what we have said before use of parametric statistics is based on the quantitative data that are normally distributed ok. So, if the data is normally distributed quantitative data numerical data that are normally distributed data then you can use parametric statistics. So, another practical rule this is a practical insight is; even if the data is not perfectly normal. If the curve is not like this if it is not like the perfect bell

shaped curve it could be. So, if you draw the histogram it could look something like this; if you see that histogram like this ok, it is not really perfectly shaped normal, but you can still see there is a hump bun stuff like that and the most of the values are around the center stuff like that. The inferential statistics or our parametric statistics is powerful enough to allow some flexibility or it allows for some of those deviations in the data. So, even if you have a not so, perfect normal normality in data, you can still use parametric statistics.

(Refer Slide Time: 09:58)

THINGS AFFECTING DISTRIBUTION LOOKS

ideal normal distribution is symmetric around the middle value.

- **Skewness** – means that the data distribution is “stretched out” to one side or other more than what is expected from a normal distribution
 - When we have more values greater than the mean, the distribution is positively skewed or skewed to the right
 - When we have more values than expected less than the mean, we say that it is negatively skewed or skewed to the left
- To illustrate this, add the values of 90, 90, 90, 100, 100, and 105 to the lunch time data and redo the histogram

positively skewed distribution (more values to the right side of mean).

negatively skewed distribution.

Now, what are the things that affects the look of the distribution or the way the normal distribution looks? The first factor that affects it is called as the skewness of the distribution and the skewness means that the data distribution is stretched out to one side or the other more than what is expected from a normal distribution. So, what we said about the normal distribution was that; normal distribution is symmetric around the middle value right.

The mean, median, mode are externally the same and then there is symmetrically distributed with most of the values clustering around the mean that is what we talked about ok. So, an ideal normal distribution is symmetric; symmetric around the middle value which is typically the mean. But if the distribution is moved to you know one side or the other; if it is more to one side and other than then we can say that the distribution is stretched out. For example, if the distribution is something like this then it can be told

as a skewed distribution ok. So, the skewness is a measure of the symmetry of the distribution.

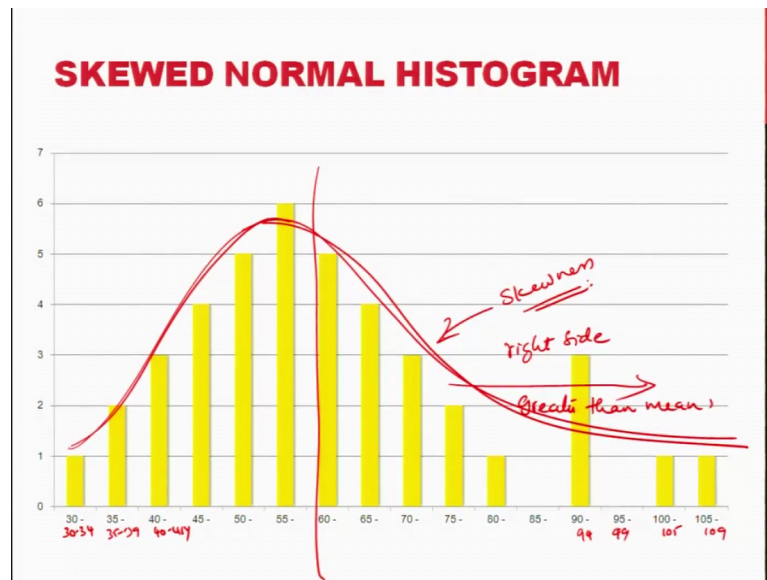
So, when we have more values that are greater than the mean ok. So, if you take an example ah; if this is the mean the center line is the mean if you have more values that are greater than the mean; that means, the values are to this side this side more than the greater to the mean, then the distribution is said to be positively skewed or skewed to the right ok. So, a distribution that has a shape like this more values to the mean, this can be called as a positively skewed distribution or more values to the right side right side of mean ok.

Now when we have more values than expected less than the mean or which is said that is negatively skewed or skewed to the left. So, same way if you have a distribution where it is something like this, where this is the mean in both cases this is the mean, but more values are to the left side of the mean this is the left side of the mean then this is called a say negatively; negatively skewed or skewed to the left side.

So, the values that are less than the mean will give you a left skew; values that are greater than the mean will give you a right skew or the positive skew. So, to illustrate this example or to illustrate this concept; let us add the values 90, 90, 90; 3; 90's and 100 and 100; 2 100 and a 105 to the lunchtime data which was in the previous example that we talked about this is the lunchtime data.

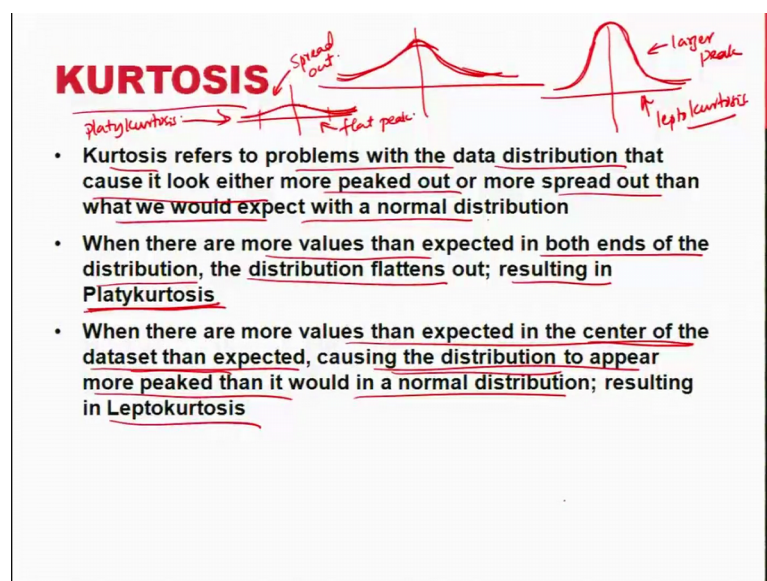
So, what we are doing here is we are adding values of 90, 90, 90, 100 100 and 105. So, you are done 1, 2, 3, 4, 5, 6 more data points are added; so, now, you have 42 data sets ok. When you add that much of values to this and then you go back and let us plot the histogram again ok.

(Refer Slide Time: 13:53)



We do the histogram with the same 5; so, this is the 30 to 34, 35 to 39; 40 to 44 like this ok. And we do this histogram and you can see that; obviously, yes in 90 to 94, 95 to 99; 100 to 105, 105 to 109 kind of a thing, you can see that there is more observations, it is not perfectly symmetrical you have more to the right side right side or greater than mean; the average value would probably be somewhere here. So, there is more observations to this side; so, this kind of a histogram which kind of creates something you can call it as like a longer tail, it is not symmetric but it actually creates this one.

(Refer Slide Time: 14:57)



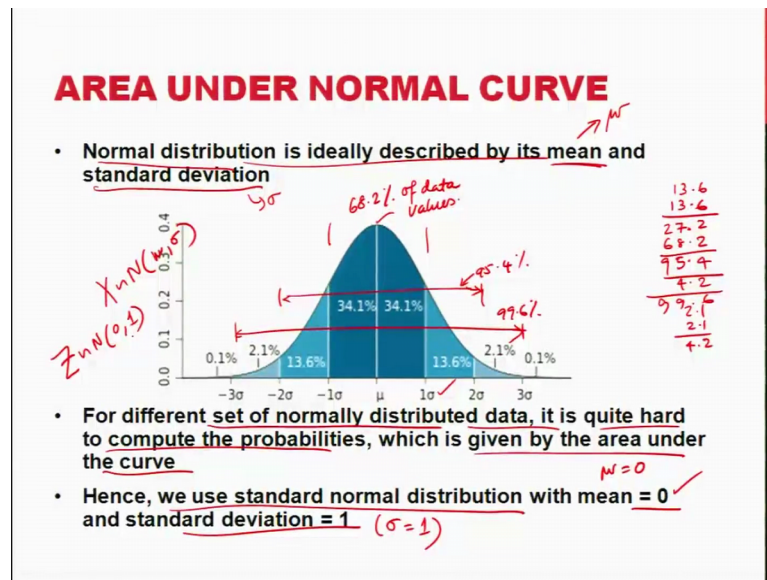
So, this is a skew skewness ok; so, this is a skewed normal distribution in this regard. The other one is a concept that we talked about it does the kurtosis one thing what we talk about is the skewness where it is talks about the symmetricity of this. The kurtosis on the other hand kurtosis, it refers to the problems where the data distribution that cause it looks like either more peaked out or more spread out than what would we expect in a normal distribution.

So, what happens is you might assume that the normal distribution is something symmetric like this ok, but instead of this distribution you could think about a scenario where it could be it could come more peak to like this you can think about a scenario like this more peaked. Or you could think about another scenario where it is more flat like this anyway this is slightly it should be more peaking than this, but anyway my normal distribution drawing skills are quite bad.

So, this can be like a larger peak this is a flat peak ah. So, this is where it is a it is more spread out ok, but this is more picked up ok. When there are more values than expected, when there are more values than expected at both ends of the distribution the distribution will flat out. So, when there are more values at the ends then it will flatten out; this results in a scenario called as platy kurtosis ok. So, this is plain platy kurtosis becomes plainer ok.

So, this is the platykurtosis when there are more values than expected at the center of the observations; the if the reverse is true if more observations are found in the center of the observations of the dataset than expected, it causes the distribution to appear more peaked than it would be resulting in an leptokurtosis. So, this is what is called as a leptokurtosis; where you have you find more than expected observations at the center of the data or near the center of the data alright.

(Refer Slide Time: 17:15)



Now, the other part is that the area under the normal curve which is because as I said the normal distribution is a symmetric distribution and the people like normal distribution because its if the mean, median and the mode are exactly the same values and this produce also quite symmetric. So, things that are symmetric are easy to estimate; so, normal distribution by if you are to describe the normal distribution this ideally described by its mean the average and the standard deviation ok.

So, the mean is typically denoted by the letter mu and the standard deviation is denoted by the Greek letter sigma ok. So, the ideally speaking the normal distribution says that about 34.1 plus 34.1 which basically gives you 68.2 percent of data values ok, they will lie within this one standard deviation of the mean that is one observation of the normal distribution.

So, now you have 13.6 and 13.6, 13.6, 13.6 that is 2; 6, 7; 27. So, 27.2 percentage more gets added to the 68.2; so, that is 95.4. So, 95.4 percent somewhere here 95.4 percent will lie between the two standard deviations and then 99.8 percentage will lie between the 3 standard deviations this is what the 2.1 and 2.1; 2.1; 2.1 will get added together is to 4.2. So, if you add 4.2 with this then you will get it as 99.6 percent of the data will fall within the 3 standard deviation 99.6 percentage of the data ok.

So,, but the issue here is that for each data set you will have its own mean and standard deviation. So, for different set of normally distributed data; it is quite hard to compute

the probabilities which is given by the area under the curve. So, if you have one set of data, you will calculate the mu, you will calculate the standard deviation and from there you will create the curve.

Once the mean and the standard deviation changes you have to draw a new curve. So, to avoid that what we do is we use a standard normal distribution with the mean of 0 and the standard deviation of 1 ok. So, what we do is we have a reference distribution with mean of 0; this mu is equal to 0 this is one case and standard deviation equal to 1, which is sigma is equal to 1 this is the say this is called as a standard normal distribution.

So, if we called this as a random variable X sorry not X bar; X X is a random variable, X is a normally distributed random variable mu and sigma as parameters mu and sigma, then z is a standard normal distribution is distributed normal with 0 and 1; the mean and sigma being 0 and 1. So, the random variable z is denoted for z; e is denoted for standard normal distribution.


(Refer Slide Time: 20:45)

WHY IT IS HARD TO FIND NORMAL PROBABILITIES

- Normal distribution is a continuous probability distribution whose probability density function is given by:

$$P(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

*P(X ≤ x)
= ∫_{-∞}^x f(x) dx
— CDF.*



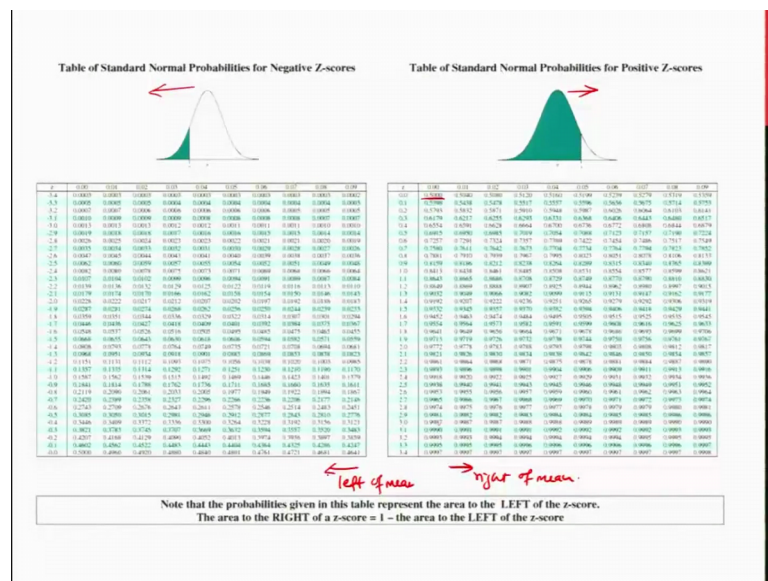
- For any given value or range of values, integrating this function and find the values are quite hard
- Hence, a standard probability table is used – or use the NORMDIST() function in excel

And the question; obviously, is that I mean why is it. So, hard to find the area under the normal probabilities or why is it hard to find the area under the normal distribution; because normal distribution number one is a continuous probability distribution and the probability density function is given by this ugly looking equation; sigma tiles 1 by root of 2 pi e to the power of whatever it is ok.

So, for any given range as we suggested earlier the CDF if you want to find probability of X less than or equal to little x; it was given by minus infinity to x; f of x dx, this is what our CDF was Cumulative Density Function was. So, in this case what happens is we have to keep on integrating this function; for all different set of values to find this ok. So, you keep on integrating and so, when you draw this curve all you are doing is you are integrating it at many places and these integrals get added to get the final curve.

So, hence what we do is instead of doing this; we use a standard probability table and then from there you calculate the value; otherwise you use the normal distribution function in excel to calculate the normal probability. Either you use a table to check it if you are doing it manually or use the norm distribution function from the excel to find the probability values ok.

(Refer Slide Time: 22:07)



So, the standard normal distribution tables are shown here. So, what happens here is there are two parts; one is the left side of the mean the other is the right side of the mean the positive values. So, if you look at here you can see that you know minus 0.009 is 4.61 and you can see the exactly the 0; 0.00 value is 0.5; the middle value.

So, this is the left of mean and this is the right of mean and you can say that these values actually tally between. So, this is the ah; so, when you say what is 3; 3.0; 3.0 is 0.9; see you see the this is border 3.0; this is the 3 standard deviation right. And if you look at the

minus 3; it is it will be right here minus 0.13. So, will be 99.7 percentage would be exactly what we will talk about ok.

(Refer Slide Time: 23:06)

EXAMPLE PROBLEMS

- A study suggests that the average height of male in United States is 69 inches and the standard deviation is 3 inches. The data follows normal distribution. Find the probability that any given person will fall within the height of 65 inches to 75 inches?
 Use Excel: $NORM.DIST(75, 69, 3) = P1$
 $NORM.DIST(65, 69, 3) = P2$
 $P1 - P2 = 0.95$
- What would be the height below which 95% of the USA males will fall under, using the information above.
 $P(Z \leq 0.95) = 1.65$
 $Z = \frac{X - \mu}{\sigma} \rightarrow 1.65 = \frac{X - 69}{3} \rightarrow X = 69 + 1.65 \times 3 = 73.95 \text{ inches}$
- What percentage of males will have a height more than 74 inches?
 $Z = \frac{X - \mu}{\sigma} = \frac{65 - 69}{3} = \frac{-4}{3} = -1.3$
 $P(-1.3) = 0.0968$
 $Z = \frac{X - \mu}{\sigma} = \frac{75 - 69}{3} = \frac{6}{3} = 2.0$
 $P(2) = 0.9772$
 $0.9772 - 0.0968 = 0.8804$

So, one simple example or we will try to say is that a study suggests that the average height of a male in United States is 69 inches and the standard deviation is 3 inches. The data follows normal distribution, find the probability that any given person will fall within the height of 65 inches to 75 inches.

So, what we are talking about is there is a normal distribution something like this with the average being 69 inches; this is the 69 inches; the average height ok. And the standard deviation is the sigma equal to 3 inches and what we are asked to find is what is the probability that the height of the person will be between 65 to 75 inches, this is 75 inches, this is 65 inches. So, to find that what do we do is; we find first this area and then we find this total area and then we subtract the both these areas. So, to do this on a one way to do it is you can do use excel and find NORMDIST of 75 with mean of 69 and standard deviation of 3.

And same way norm distribution of 65 with mean of 69 and 3 what is the probability that you will get the P 1 and you get the P 2 this is the cumulative probabilities and so, P 1 minus P 2 will give you the resultant probability. Other option for us to do it is; we convert this into the standard normal curve what we call as z is equal to X minus mu divided by sigma ok.

So, what we are giving here is the X is the value that we want which is 65; the μ is the average which is 69 and σ is 3 ok. So, $65 - 69$ that will be minus 4 divided by 3; $4 / 3$ will be minus 1 point 33 that is 1 point 33 before t right that is 1.33 ok. So, we go back to the table you can see that minus 1; this is minus 1.0 ok; 1.33 will be this; this value 0.1515 ah; 1 point 33 sorry; not 1.03 my bad sorry I gave you the wrong idea 1.33; that is what we want. I mean there is another one, but we are only looking at this 1.33; so, 0.0968 ok.

So, that value will be the probability of this will be 0.0968; so, let us call it as 0.1 ok. Similarly the other case the z is equal to ok; $X - \mu / \sigma$, where it is $75 - 69 / 3$; that is again $6 / 3$ right $69 + 1$; $70 / 3$ plus 1; $6 / 3$ which is 2 ok. So, the probability of 2 in this case is we go back to the table and you find where is 2.0; this is the 2.0 2.0 right here 0.9772 ok. So, we get the value as 0.9772; so, let us call it as 0.98 for this case. So, the probability resulting probability in this case is $0.98 - 0.1$; which is we can call this 0.88. So, there is an 88 percent probability that a person any given person will fall between the height of 65 to 75 inches.

Similarly, using the same approach you can calculate the same for the what is the height below which 95 percent of the US males will fall under ah. One way to solve this problem is you we go back to the table and we find out where is 0.95 is ok. So, you can see that somewhere here you will get 0.95; so, let us call this as 0.95 which is 1.65 let us approximated ok. So, probability of z you know less than or equal to 0.95; gives us 1.65 ok. So, in this case we can find that z is equal to $X - \mu / \sigma$. So, we just need to find what is X ? Because we know it is 1.65 is equal to $X - 69 / 3$; 69 is the mean it is this.

So, your X will be given by X is equal to $69 + 1.65 \times 3$ ok. So, then there you will get a value that value is the value below underneath which is 1.65×3 will be 4.95; 69 + 4.95 will be 73.95 inches; so, 73.95 inches is the height at which below which 95 percent of the American males should fall under. So, using the same problem we can same approach we can keep on solving these problems for us ok.

So, I would request you all of you guys to go through this normal distribution presentation and get yourself comfortable with the a normal distribution tables. And more than that learn how to use normal distribution under the excel so, that you can use

excel to find the values of this I mean this is a cute lecture because the next thing you are going to learn soon is the hypothesis testing and the perception of normal distribution and t distribution stuff like that.

So, this should help you to understand the applied side of normal distribution why it is important that we take normal distribution. So, that we can use parametric statistics do inferential statistics or testing of the hypothesis because testing of the hypothesis is the fundamental aspect of analytics. The difference between data mining and data analytics is that data analytics has a hypothesis, where data mining does not have a hypothesis it just looks for a inherent patterns in the data.

With this, we will conclude today thank you very much and we will see you in the next class.

Thank you.