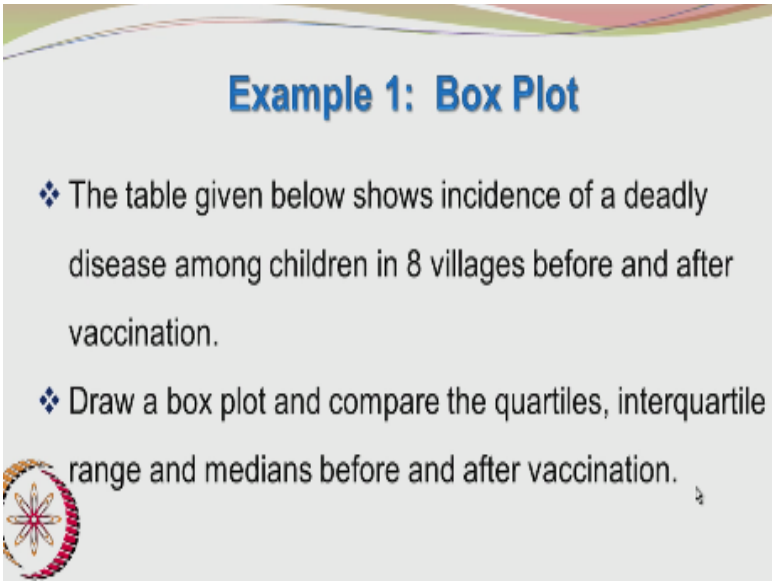
Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology – Madras

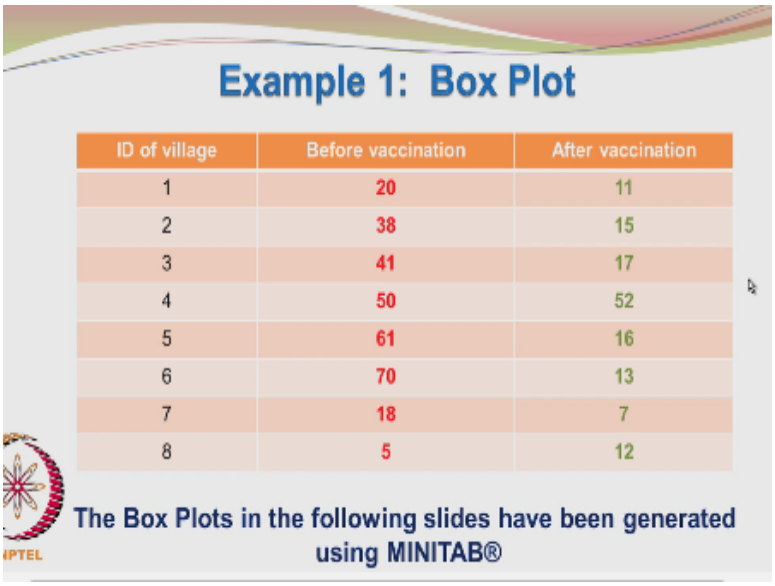
Lecture – 09
Example Set - III

(Refer Slide Time: 00:39)



Hello, again, welcome back to the course lectures, today we will be looking at the third example set, I hope by now you are having a pen calculator and paper with you as you are going through the example problems. In this lecture, we will be looking at the data representation, so the first example involves the box plot construction, you do not have to really construct the box plot manually, you can use standard software, find out how to create such kind of plots.

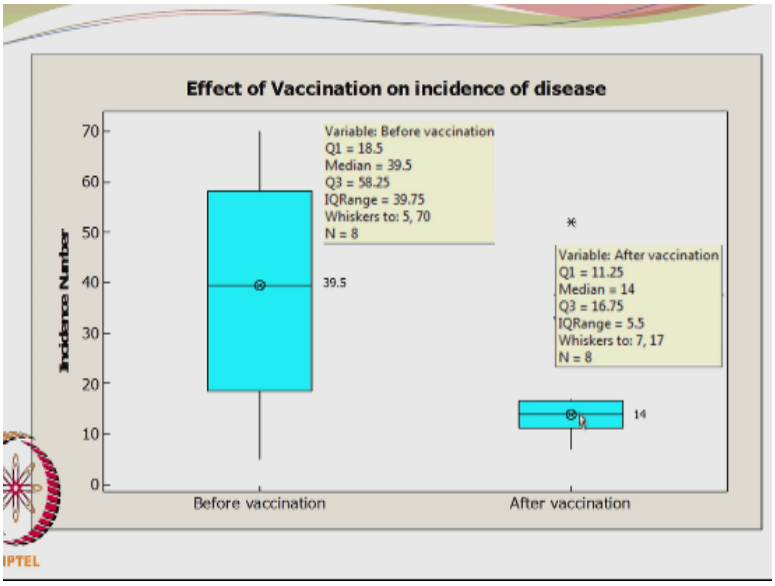
(Refer Slide Time: 01:38)



But even otherwise, if you have no access to these software's, you can easily construct the plots yourself. The example problem goes like this, the table given below shows the incidents of a deadly disease among children in 8 villages before and after vaccination and draw a box plot and compare the quartiles, interquartile, range and medians before and after the vaccination. So, you have the ID number of the village 1 to 8.

And then, you have the incidence of the disease before vaccination that the disease could be polio, fortunately it is more or less eradicated. So, before vaccination, the incidence was 20, 38, 41, 50, 61, 70, 18 and 5 in different villages, it may be different villages have different levels of hospital facilities or hygiene and so on. So, the number is varying from one village to another village.

(Refer Slide Time: 02:55)



In fact, unfortunately here the number of incidences of the disease, the number of incidence of the disease after vaccination has in fact gone up from 50 to 52. In other cases, it has reduced and here, there is another situation, where it has almost doubled in fact, it has increased by 2.4 times okay, so these are data with us and we have to compare them with the help of a boxplot. So, the box plot was generated using Minitab and it shows before vaccination and after vaccination.

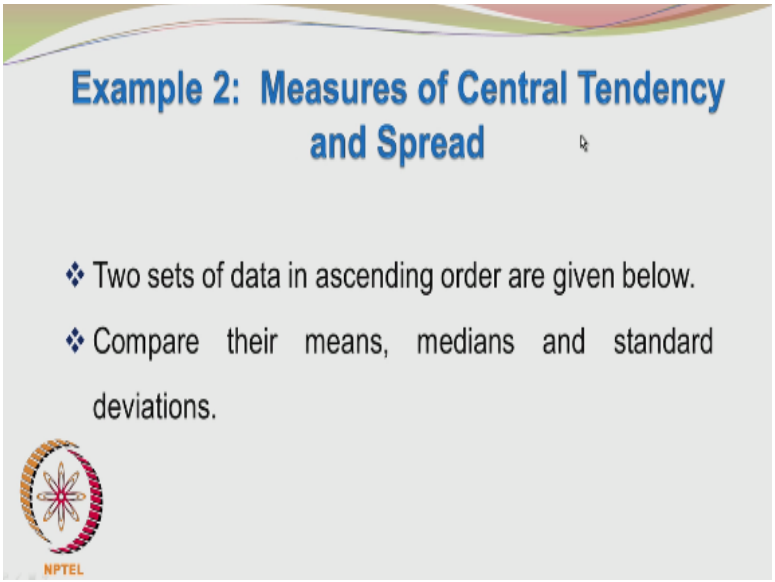
This shows the median value; the median value is same as the mean value in this particular case. This is the third quartile, second quartile, first quartile and this is the whisker. Here, again you have a whisker and you have another whisker here, which you may not be able to see and the median has considerably reduced from 39.5 to 14 and this point here interestingly is an outlier.

We know that whiskers are drawn up to a data point, which is 1.5 times the interquartile range, so you do not have a whisker extending beyond this and so we have an outlier at this particular point. This obviously corresponded to the 52, okay which we saw in the previous table, this 52 here is an unusually high number, so rather than trying to figure out the numbers with respect to the scale, you can also have the data plotted in the software.

So, the variable before vaccination, the first quartile is 18.5, median is 39.5 and the third quartile is 58.25, the interquartile range is the difference between the third quartile and the first quartile and that comes to 39.75 and the number of data points is equal to 8. Here, the first quartile is 11.25, earlier it was 18.5 and the first quartile is 11.25, the median is 14 and the third quartile is at 16.75, the interquartile range is reduced from 39.75 to 5.5.

Whiskers are drawn from 7, that is the lower whisker up to 17, you cannot see 17 here because it is almost matching with the third quartile and you have an outlier 52, so what we can definitely say is the incidence number has considerably reduced and the average or the median value has gone down from 39.5 to 14, so hence vaccination was really effective and also earlier, you are having the large spread, different villages were having different numbers of incidences of the disease.

(Refer Slide Time: 06:19)



And after the vaccination has been carried out, the spread has considerably reduced. Now, what we have to do is; look at the measures of central tendency and spread, you have 2 sets of data in ascending order, you have to compare their mean values or average values, medians and standard deviations.

(Refer Slide Time: 06:39)

Sl. No.	Data Set 1	Data Set 2
1	22	5
2	24	10
3	26	15
4	28	25
5	32	35
6	34	45
7	36	50
8	38	55

So, you have the data set 1 and you have the data set 2, let us look at the data. The data points here are ranging from 22 to 38, okay, the range is 38 – 22, which is 16, here the data values are ranging from 5 to 55, so the range is 50. So, immediately you can say that this is a broader or a wider spread of data points than this okay, so these may be marks in class 1 and marks in class 2, you can see that the marks in class 1 are more bunched together when compared to class 2.

If you look at the average, first let us count the total, it is not too difficult to count this 46, 72, 100, 132, 166, 202, 202 + 38 is 240, so total is 240/ 8 is average would be 30 and then, if you count this 15, 30, 55, 91, 35, 185, 185 + 55 is 240, so again you have 240, so when you divide 240/8, you will get 30, so what we are seeing is the 2 data sets have the same average. So, even though you are taking the data from 2 different classes okay, the average value; the average values are the same for both the classes.

(Refer Slide Time: 08:31)

Sl. No.	Data Set 1	Data Set 2
1	22	5
2	24	10
3	26	15
4	28	25
5	32	35
6	34	45
7	36	50
8	38	55
Mean	30	30
Median	30	30
Standard Deviation	5.86	19.09

So, the mean value is 30 for both the classes, the median values also 30 for both the classes. How do you find the median? You arrange the data in the ascending order, which has already been done for you and you have an even number of data points, so you have 2m data points, where 2m is = 8, so m is equal to 4, you have to find the fourth and the fifth data point in the set.; 1, 2, 3, 4; fourth and the fifth data point in the set.

So, you take the average of these 2 numbers; 28 + 32 is 60; 60 divided by 2 is 30, so you have 30, so that is also matching with the mean value. If you look here, when you take the average between average number of 25 and 35, the average is again 30, so the median is again 30, so both the data sets have the same mean and they have the same median. The standard deviation on the other hand for data set one is 5.86, whereas it is 19.09 for the second data set.

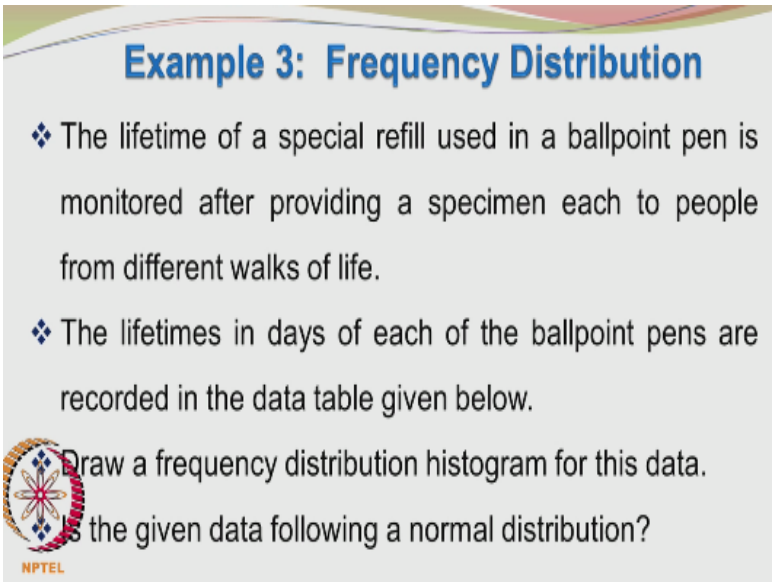
So, the standard deviation here is almost 3 times higher than the first data set and so you can immediately conclude that there is a larger spread in this second data set. By now, I hope, you know how to calculate the standard deviation, to do that you have to take the individual data set

and individual data set value, let us say 5 and then you have to subtract the mean value from that. So, 5 - 30 is -25 and then you have to square that value.

So, 25 squared is 625, similarly you take 10 – 30, which is -20, when you square it, you will get 400; 50 – 30 is -15, you square it, you get 225. So, you can add up all these numbers and you divide it by n – 1, okay n - 1 in that case would be 8 – 1, which is = 7 and that will give you the variance. So, you divide the sum of squares of the deviations with respect to the mean by n – 1, where n is the number of data points.

So, you will get the variance, you take the square root of the variance, you will get the standard deviation. The standard deviation for the second data set is considerably higher than the standard deviation for the first data set, so even if 2 data sets how the same mean and the same median, they may have different standard deviations. Data of different spreads can have the same mean and median.

(Refer Slide Time: 11:55)



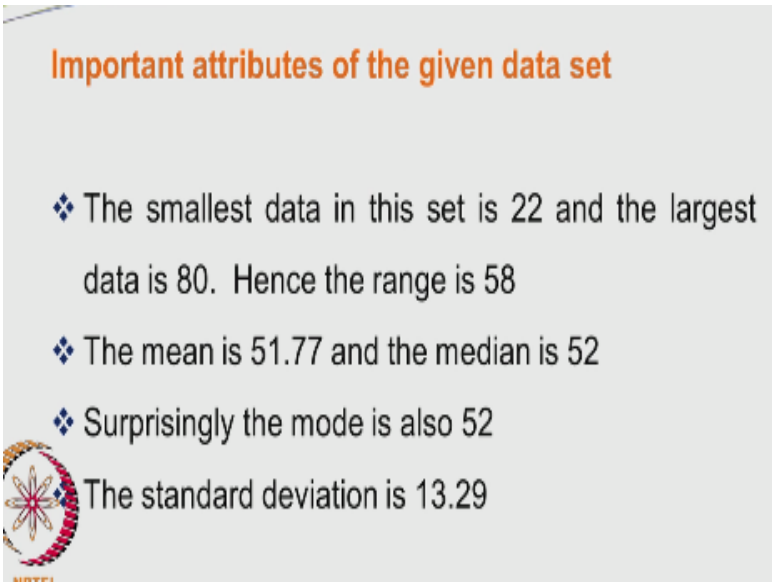
So, let us go to the third example, where we discuss the frequency distribution, I have set up all these problems on my own, I hope there are no mistakes in any of these problems. If there are any mistakes, kindly send me the feedback. The lifetime of a special refill used in a ballpoint pen is monitored after providing a specimen each to people from different walks of life. The life times in days of each of the ballpoint pens are recorded in the data table given below.

(Refer Slide Time: 12:25)

47	42	52	53	33	48	63	78	63	36
50	39	26	53	63	34	47	52	55	40
52	27	40	33	60	69	73	37	57	66
54	44	50	69	79	41	62	57	50	55
56	73	51	27	80	45	24	49	40	65
58	68	37	64	57	34	53	52	51	48
61	66	52	34	29	56	53	54	56	35
62	58	22	66	46	75	35	69	50	69
6	48	46	49	62	44	64	62	43	49
	62	55	38	39	56	63	29	66	41

Draw a frequency distribution histogram for this data, is the given data following a normal distribution. So, you are given a huge data set, there are 100 data points here, so you have data; 47 days, 50 days, 52 days so on and here we have so many data points, so we have to draw a frequency histogram of this data, we will do some analysis of this data set to get a feel for it.

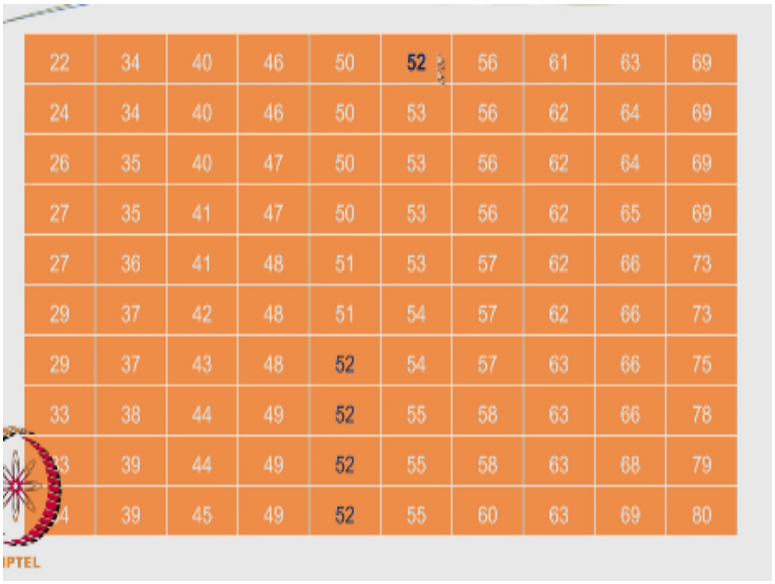
(Refer Slide Time: 13:01)



So, the important attributes of the given data set are the smallest data in this set is 22 and the largest data is 80, okay. So, the range is 80 – 22, which is 58, the mean value is 51.77 and the median is 52. How did you find the mean? We add up all the numbers in this collection and divided by 100, so the mean value is 51.77 and the median value is 52, the median value is pretty close to the mean value.

Even more surprisingly, the mode is also 52, the mode by definition is the number, which appears most frequently okay and that comes to be 52, so we can see as far as this data set is concerned, the mean is equal to median, well almost equal to; you can approximate 51.77 to 52 without too much of a complaint and the mode is also equal to the median, so all the 3 parameters; mean, median and mode are matching.

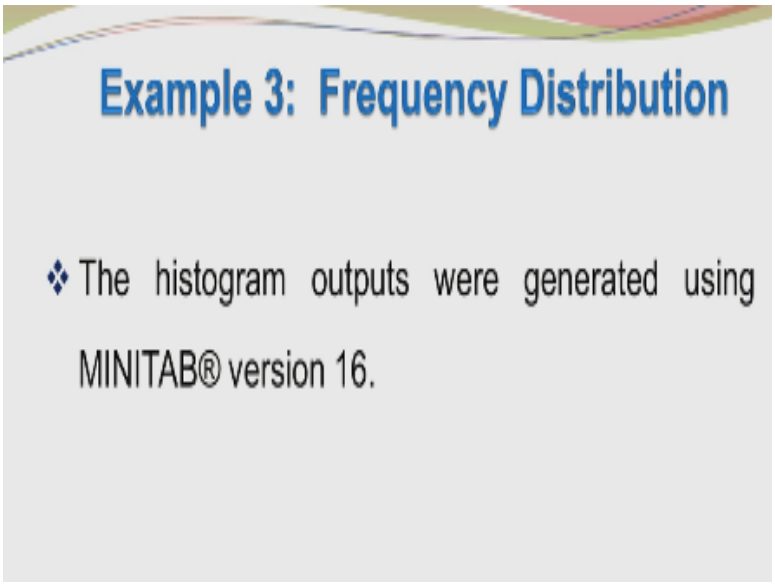
(Refer Slide Time: 14:27)



And the standard deviation is 13.29, what I have done here is put the data in the ascending order and if you look at the data, you can see 52 appearing most frequently, so no other data is appearing 5 times here, 52 is appearing the most number of times in this given data set and so you have the mode as 52. Well, you can also show that the median is 52, you have an even number of data points, so 2m is = 100, so m is = 50.

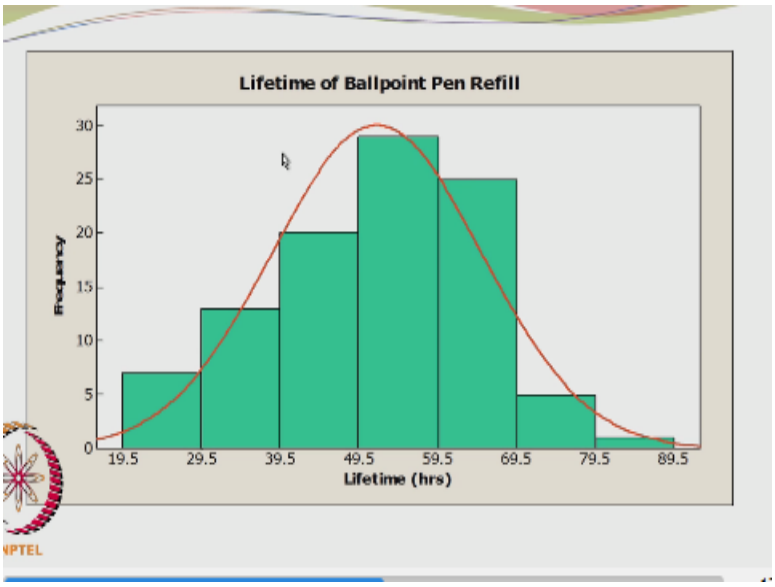
And so you have to find the average of the 50th and the 51st data point. How many data points you have here? 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, okay, so you have 10 data points, you have to look at the 50th in the 51st data point, 10, 20, 30, 40, 50; 50 and 51st, okay. So, the 50 data point is 52 and the 51st data point is also 52 and so the average is very easy to calculate, the average is also 52, so the median is also 52, the mode is 52 and the mean was 51.77.

(Refer Slide Time: 15:56)



Interesting, if you start looking at numbers from a statistical view point, you can pick up lot of interesting relations between the numbers and their different parameters associated with these numbers, so we can generate a histogram and I have used the Minitab version 16 and the histogram looks like this.

(Refer Slide Time: 16:24)

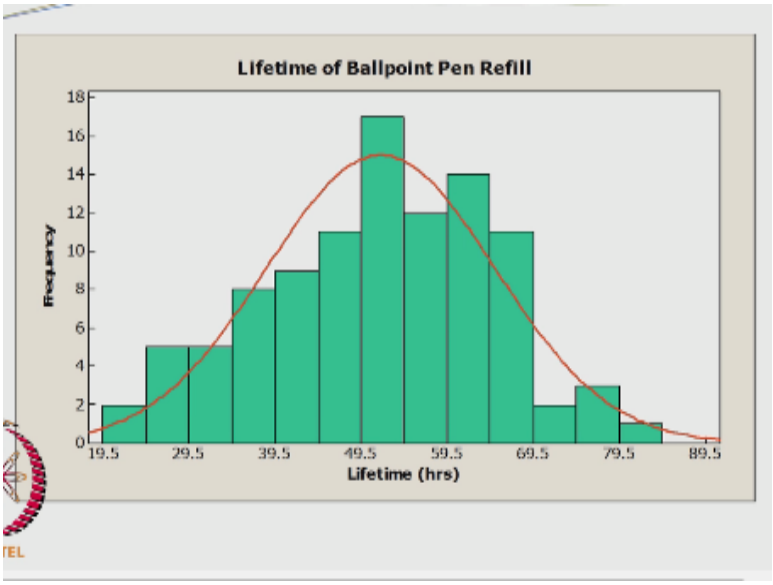


You can see that I have taken how many cells; 1, 2, 3, 4, 5, 6, 7 bins or 7 cells I have taken, the smallest number was 22 and the largest number was 80, so I have actually started from 19.5 and gone up to 89.5 okay and I am having 7 divisions, how did I get the number seven? There were a few recommendations for the suggested number of bins, if you recollect the earlier lecture; we had the Sturge's formula and also the number recommended by Montgomery and Runger okay.

Square root of the number of observations; square root of 100 would be 10, so about 10 bins, okay, if you look at Sturge's recommendation, it may give a slightly different value. So, I have chosen 7 bins here and so between 19.5 to 29.5, you have how many data points; maybe 5, above 5 definitely, maybe 6 or 7 data points between 19.5 to 29.5; 19.5 to 29.5, so numbers below 33, so it will be 1, 2, 3, 4, 5, 6, 7, okay.

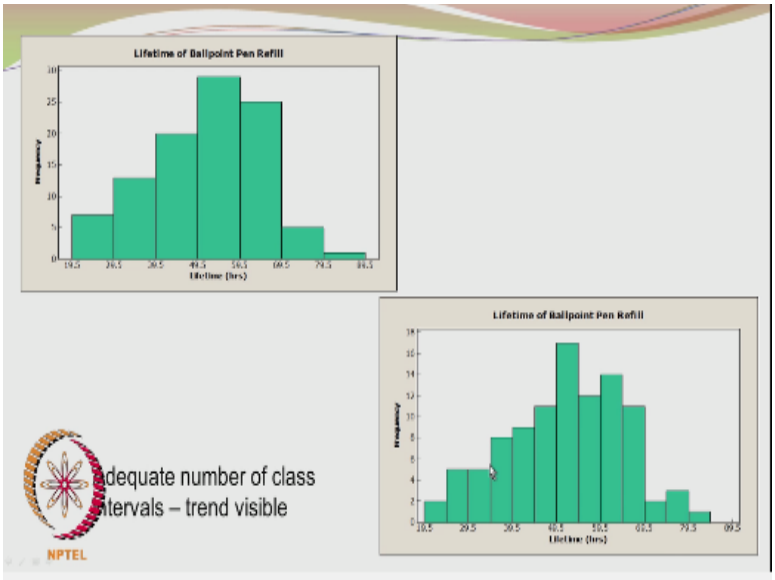
So, you have 7 data points between 19.5 to 29.5, this is 7, between 29.5 to 39.5, this maybe 13, between 29.5 to 39.5, so we do not include 29, we start with 33, so this is 3 and we are saying 39.5, so 39 is below 39.5, the entire collection is 10; 10 + 3 is 13, so you have 13 numbers between 29.5 to 39.5 and that is what is shown in the histogram here, you have 13 as the frequency corresponding to the interval 29.5 to 39.5.

(Refer Slide Time: 19:27)



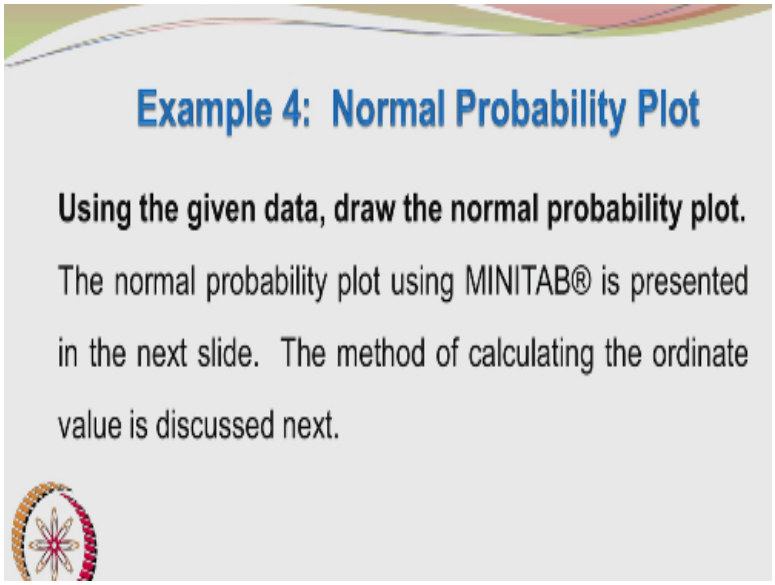
So, you can count the numbers between each interval and then, these kind of bar diagrams are created and histogram is formed, this is the normal distribution fitted to this distribution of data. Suppose, I had increased the cell sizes from 7 to 1, 2, 3, 4, 5, 6, so I have increased it to 13 and there is obviously more detail but somehow, at least subjectively to me, it does not look as good as the previous histogram, it looks very compact and shows the trend.

(Refer Slide Time: 20:20)



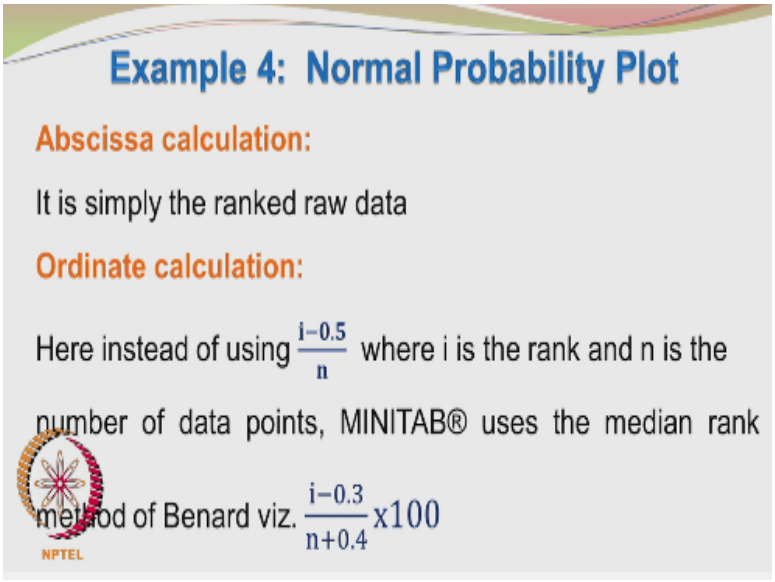
The trend is approximately normal okay, so here the data points are bit more cluttered okay. Well it is subjective and it is up to you but I do feel that 13 cells is far too many even the number suggested by Runger and Montgomery was about 10, okay, so you can see that adequate number of class intervals 7 and the trend is visible, here it is slightly more cluttered okay.

(Refer Slide Time: 20:59)



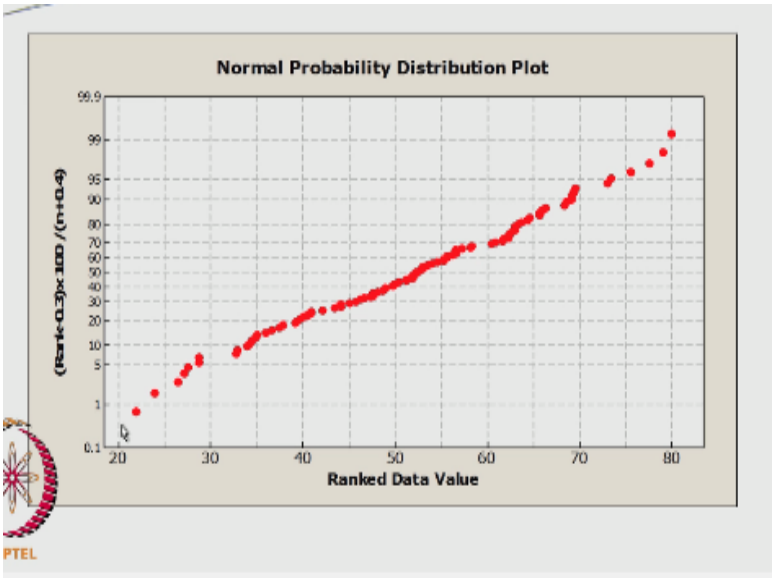
So, interesting; the histogram is meant only for a qualitative interpretation of the distribution associated with the data, we have to really quantify it and prove conclusively that the data have indeed come from a normal distribution, so we can check it by plotting the data in the normal probability plot and see whether our normality assumption is satisfied. Again, I have used the normal probability plot using Minitab.

(Refer Slide Time: 21:21)



And we will see how to calculate the x axis values and the y axis values. Well, there is nothing great in the abscissa calculation, as I told you earlier, you have to rank the data from the smallest one to the largest one and the abscissa marking the abscesses also called as the x axis okay, so the x axis marking is simply the rank raw data. The numbers which are ranked from the smallest to the largest and so you identify the number on the x axis.

(Refer Slide Time: 22:29)



The ordinate calculation is quite interesting and there are several versions for the ordinate or y axis calculation. The most common one at least from what I have seen is i – 0.5/n that i, is the rank and n is the number of data points okay. Well, Minitab uses the median rank method of Bernard and it uses a different formula, it uses i – 0.3/ n + 0.4 * 100, okay. So, if you follow Minitab recommendation, so the rank – 0.3, okay; rank – 0.3 * 100/ n + 0.4 is plotted on the y axis.

And then you have the rank data value, the smallest value if you recollect was 22 and the largest value was 80, so you are plotting 22 here and then you are plotting 80 here but the y axis is based on the rank. For this data point the rank is 1 and 1 – 0.3 is 0.7, so 70/100.4; 70/100.4, we can take it as approximately 0.7, remember, see the scale is starting from 0.1 to 1, so the 0.7 would be somewhere here and that is what is your data point.

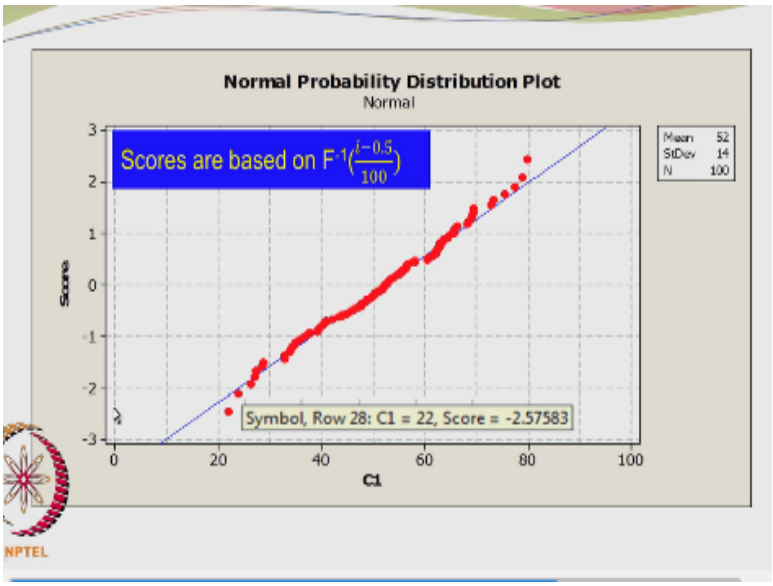
And here that is the 100th data point, the rank is 100, so this would be pretty much close to 1 or pretty much close to 100 because 100 – 0.3 is 99.7 and 100 + 0.4 is 100.4, 99.7/100.4 is pretty much close to 1 and so you have close to 99.2 or something, so that is how you calculate the y axis corresponding to the x axis values. The important thing is you can see that the data points are pretty much falling on a nice straight line.

And so, you can convince your sceptics by saying that the data indeed came from a normal distribution. Well, this is hardly surprising because the original data it chose was generated using the random generator option of the normal distribution, so it randomly picked up values from the normal distribution and gave it to me and then I did the other calculations like ascending order and histogram drawing and so on.

So, it is hardly surprising that the data is obeying the normal distribution but in your experimental work, you may get some data from your equipment or instruments and you may have a model okay, the discrepancy between the experimental observation and your model, will define the residual and the common assumption is you have explained all the controllable factors using the model.

And the difference between the experiment and the model prediction may be attributed to noise or random error and the common assumption that is made is the random error is normally distributed okay, so you can calculate the difference between your experimental value or experimental response on the model prediction, you will get the residual, you rank the residuals in the ascending order and then plot it in the normal distribution plot.

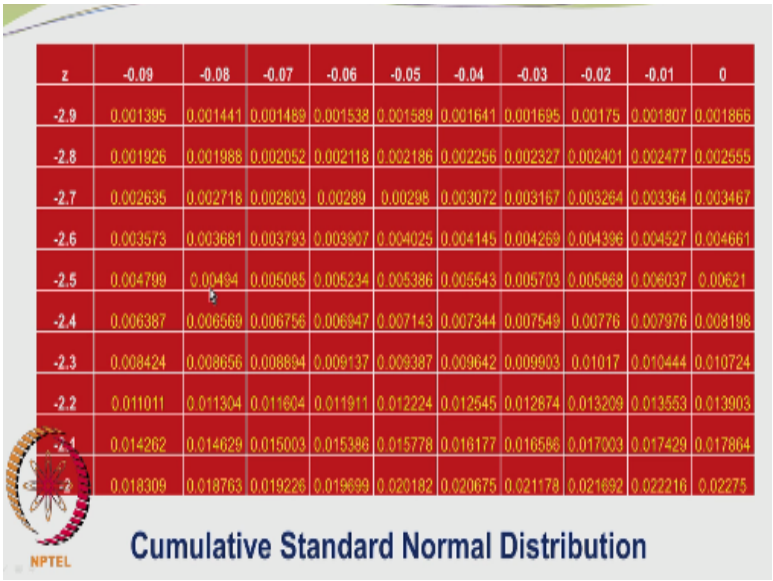
(Refer Slide Time: 26:04)



And see whether the data are lying on a straight line or nearly a straight line, you cannot get all the data points lying on a perfect straight line more or less a straight line and then you can say that my assumption that the errors are normally distributed is justified, right. When you look at this particular graph, again the data points are plotted from 22 to 80 and so these are the values 22 is here and 80 is here and other data points are in between.

How did you find the y axis here okay and for that the value is -2.57583, okay, a rather frightening number, how did you get this number? The rank is of course, 1, so you have 1 – 0.5, which is 0.5 and then 0.5/100 is 0.005, so what is the z value, which gives the probability of 0.005 that is what you have to check and that small probability corresponds to the area on the left hand side of the normal distribution curve.

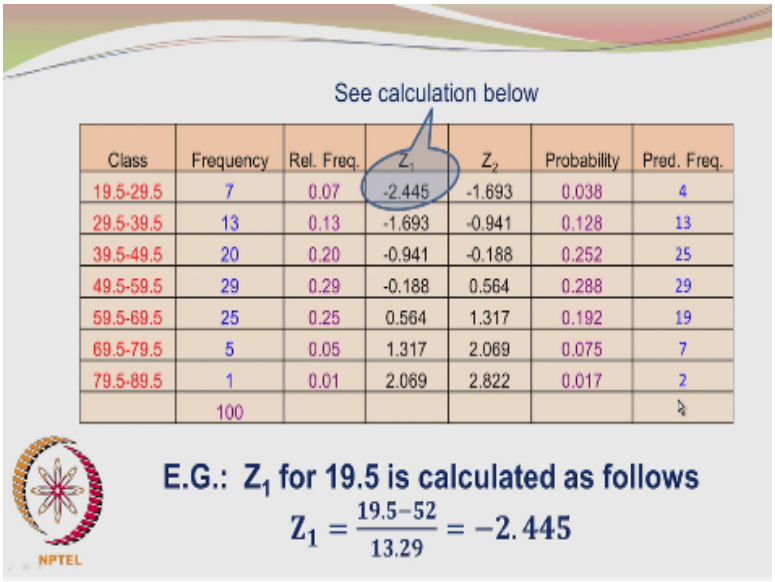
(Refer Slide Time: 27:45)



It is corresponding to the left tail of the normal distribution curve, so let us see what is the probability corresponding to the value of z or to put it in another way, what is the z value, which will give a probability of 0.005 and that value is marked here. Let us go to the normal distribution table, so – 2.57 is giving probability of 0.001 and -2.58 is giving you a probability of 0.00494, so the required z value is somewhere in between okay, it is lying in between -2.57 and – 2.58.

And that is the number, we have since it is done by the software; you are getting a more accurate value. Similarly, corresponding to the next data point, you can find the rank, the rank is obviously 2; 2 – 0.5 is 1.5, 1.5/100 would be 0.015, okay and the z value corresponding to the probability of 0.015 may be found from the table, it is close to -1.; -2.1 or something okay that you can find out from the table.

(Refer Slide Time: 29:33)



So, you can plot all these scores based on the inverse of the cumulative distribution function and you will find that the data points are lying pretty much on a straight line. Now, we are going to do something interesting. The class intervals were divided from 19.5 to 29.5, we had 1, 2, 3, 4, 5, 6, 7 such classes and we find the number of occurrences of the data points in each class interval.

So, we have 7, 13, 20, 29, 25, 5, 1 okay, so between 49.5 to 59.5, you had the maximum number of occurrences that is 29, so the relative frequency or the probability would be obtained by dividing the frequency value by 100, so you have 0.07, 0.13, 0.2, 0.29, 0.25, 0.05, 0.01. So,

how do you calculate the Z value okay? So, I am going to 2Z values; Z1 and Z2, Z1 corresponds to lower interval and Z2 corresponds to the higher interval and sorry.

Z1 corresponds to the lower limit of this interval and Z2 corresponds to the higher limit or the upper limit of this interval, so you can see that the upper limit of this interval becomes the lower limit of the next interval. The upper limit of this interval is -1.6934 for Z and that became the lower limit for the next interval but the main question is; how did you find Z1 and Z2, so these are raw scores okay.

Even though, they are apparently coming from a normal distribution well, we have confirmed that they are indeed coming from a normal distribution because the normal probability plots were showing the linear trend okay, so they are definitely coming from the normal distribution and we have to convert them to the standard normal form and we use the mean and the standard deviation from the data, we easily calculated the mean and standard deviation.

And actually x – Mu/ sigma well, I have used 52 here actually, you should use the value of 51.7, I think that is a more accurate value and then the standard deviation and then you also have; for this you put 29.5 - 52 and then you divided by 13.29, you will get the value as -1.69, okay and then you have these 2 Z values, you have to find the probability that the Z will lie between -2.445 and -1.693.

What is the probability that the Z will lie between these 2 limits? -1.693 and -2.445, so you can find out the probability of the Z value lying below -1.693 and then you can also find the probability of the Z lying below -2.445 and then subtract the 2 probabilities and you will get 0.038, so you multiply 0.038 by the total number, which is 100 and you will get 3.8, can approximately put it as 4.

These 4 values compares well okay with the 7 and then when you do the same thing for the next class interval, the Z1 and Z2 are -1.693 and -0.941, so the probability of the Z value lying between -0.941 and -1.693 on the standard normal curve is 0.128, when you multiply 0.128 / 100, you will get 12.8 and that number is approximately 13 and the actual number of observations between 29.5 and 39.5 was 13 and so the 2 numbers are matching reasonably well.

So, you can see that the other numbers are also matching reasonably well and so we can say that the present set of numbers are distributed normally, so these are the standard normal distribution cumulative charts, you can refer to it or take the values from the book or use spread sheet or statistical software like Minitab to get the probabilities. This concludes our presentation and we have been doing a few problems in the normal and lognormal distributions.

And also we showed some typical problems involving the exploratory data analysis, we saw the histogram, we saw the normal probability plot, we saw the box plots, they are very useful for concise presentation of data, where you can express a lot of conclusions in a single location. If you start showing 3 or 4 diagrams and explaining the trends with those 4 diagrams would be quite difficult.

Rather, if you can summarize all the data in a compact form for example, the box plot form your conclusions and presentations would be more effective. Many of these calculations do not really require statistical software, so even if you do not have access to it, you can with the pen pencil paper the standard probability normal probability charts on the calculator, you can do pretty much all the calculations and present the graph okay.

So, this concludes the continuous probability distributions and the data representation will be now slowly moving on to the next aspect of our statistical analysis. Thank you.