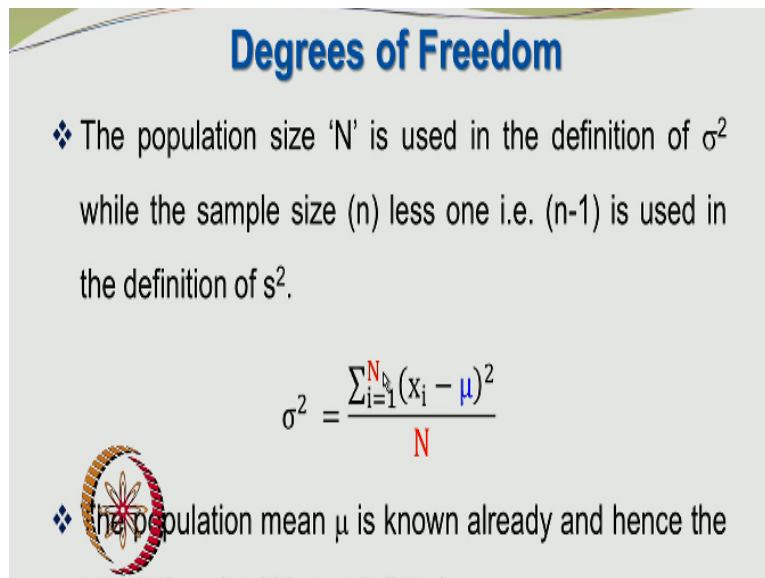


Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology – Madras

Lecture – 07B
Exploratory Data Analysis - Part B

Resuming the discussion on the variance let us say that we want to find the variance of the population, a population of a huge size. Let us assume that the mean of the population is already known to us and we find the variance of the population using a similar kind of formula.

(Refer Slide Time: 00:45)



Degrees of Freedom

❖ The population size 'N' is used in the definition of σ^2 while the sample size (n) less one i.e. (n-1) is used in the definition of s^2 .

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

❖ The population mean μ is known already and hence the

Here we write $\sum_{i=1}^N (x_i - \mu)^2 / N$. μ was already known and we did not find μ from the x_i . So we are able to use N in the denominator. In the case of the sample variance we use the sample itself to find the sample mean then we use the sample mean to find the sample variance and so the N deviations were not independent and we were forced to use $n-1$. There is another reason why we use $n-1$.

(Refer Slide Time: 01:23)

Degrees of Freedom

- ❖ Two reasons are discussed in Montgomery and Runger (2011).
- ❖ The same sample is used to find sample mean.
- ❖ This implies that only $n-1$ deviations about the sample mean are independent and hence there are only $(n-1)$ degrees of freedom.

The two reasons are discussed by Montgomery and Runger. If you look at the sample mean or the arithmetic mean it is known that the sample mean is quite close to the sample values than to the population mean. Since you are calculating the arithmetic mean from the sample these arithmetic average or the arithmetic mean is going to lie closer to the sample values than to the population mean.

This would lead to smaller deviations and smaller sum of deviations than the actual case. In order to compensate for this closeness, the $n-1$ is preferred. If you are using sample mean, then you are going to get apparently least scatter because the sample mean is more closer to the data values. Hence instead of using n in the denominator if you use $n-1$ then a partial compensation occurs.

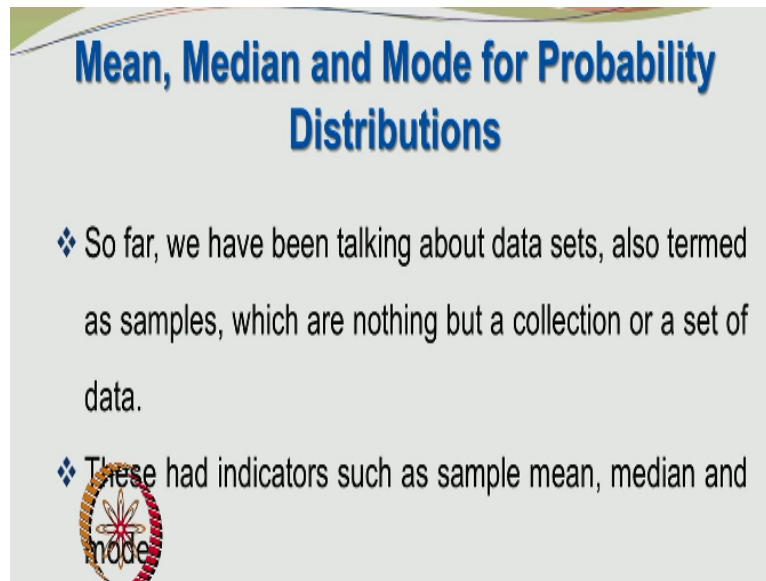
(Refer Slide Time: 02:55)

Degrees of Freedom and Sample Variance

- ❖ If n had also been used in the sample variance calculation, then the estimated value would be lower, leading to a false impression of less scatter.
- ❖ Using $(n-1)$ rather than n is termed as the Bessel's correction (DeCoursey, 2003)

Using $n-1$ rather than n is termed as Bessel's corrections.

(Refer Slide Time: 03:02)

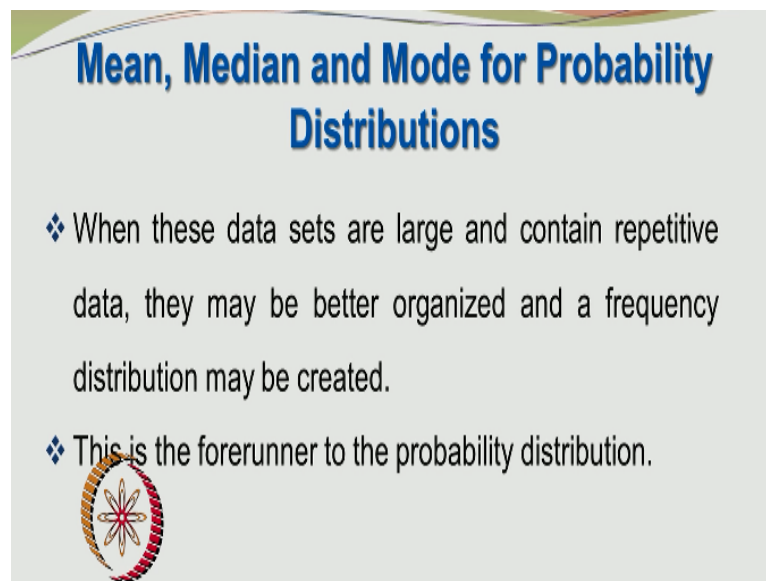


Mean, Median and Mode for Probability Distributions

- ❖ So far, we have been talking about data sets, also termed as samples, which are nothing but a collection or a set of data.
- ❖ These had indicators such as sample mean, median and mode.

So far we have been talking about data sets and these were also termed as samples and there features such as sample mean, median and mode.

(Refer Slide Time: 03:16)



Mean, Median and Mode for Probability Distributions

- ❖ When these data sets are large and contain repetitive data, they may be better organized and a frequency distribution may be created.
- ❖ This is the forerunner to the probability distribution.

When these data set are large and contain repetitive data they may be better organized under frequency distribution may be created. You might have done this in class 9 and class 10 where you created the frequency table. This is the forerunner to the probability distribution.

(Refer Slide Time: 03:44)

Mean, Median and Mode for Probability Distributions

- ❖ Even discrete and continuous probability distributions have mean, median and mode associated with them
- ❖ We have seen how to calculate the mean for these distributions.
- ❖ Let us see how to find median and mode for these

Both the discrete and continuous probability distributions have mean, median and mode. So we will see how to calculate the mean and median and mode for these distributions.

(Refer Slide Time: 03:59)

Mean, Median and Mode for Probability Distributions

- ❖ The probability distribution functions in the discrete case assigns probability values to the individual random variables present in the sample space.
- ❖ When random variables are continuous we have a continuous probability density function
- ❖ Both of these are associated with mean, median and

The probable distribution functions in the discrete case assigns probability values to the individual random variables present in the sample space. When random variable are continuous we have a continuous probability density function.

(Refer Slide Time: 04:20)

Median for Continuous Probability Distributions

The median x_m for a continuous probability distribution is the point within the range of the allowed values that the random variable X may take such that the cumulative distribution value at x_m is exactly 0.5 i.e.



$$\int_{-\infty}^{x_m} f(x) dx = F(x_m) = 0.5$$

The median X_m for a discrete probability distribution is the value within the range of the allowed values that are random variable X may take where the cumulative distribution value is exactly 0.5. In other words, x of F of $X_m = 0.5$. On the same lines we can define the median for a continuous probability distribution function where the integral $-\infty$ to X_m f of x $dx = F$ of X_m is 0.5

(Refer Slide Time: 04:59)

Median for Continuous Probability Distributions

It may be also seen that



$$\int_{x_m}^{\infty} f(x) dx = F(x_m) = 0.5$$

It can be seen that just as you had $-\infty$ to X_m f of x $dx = 0.5$ you also have X_m to ∞ f of x $dx = F$ of $X_m = 0.5$. You locate X_m in such a way that both the integrals have the value of 0.5.

(Refer Slide Time: 05:23)

Quantiles for Continuous Probability Distributions

- ❖ The median x_m for a continuous probability distribution is the second quartile.
- ❖ We may define any **quantile** for a probability density function quite easily.



The median is the second quartile. We may also define a quantile for a probability density function quite easily. The quartile is with the r now we are talking about quantiles with the n.
(Refer Slide Time: 05:41)

Quantiles for Continuous Probability Distributions

- ❖ In general, the p^{th} quantile of the continuous distribution is given by

$$\int_{-\infty}^{x_p} f(x) dx = F(x_p) = p$$

- ❖ Sometimes the term percentile is used.



By definition the p^{th} quantile of the continuous distribution is given by $-\infty$ to x_p of $f(x) dx = p$. The p^{th} quantile the p value here. We also use percentile. Percentile is quantile times 100.

(Refer Slide Time: 06:10)

Quantiles for Continuous Probability Distributions

Sometimes the term percentile is used.

$$\text{Percentile} = \text{Quantile} \times 100$$

Hence,

$$\text{one-quarter quantile} = 25^{\text{th}} \text{ percentile} = \text{first quartile}$$



So you have the one quarter quantile which is the 25th percentile and which is= to the first quartile.

(Refer Slide Time: 06:24)

Mode of a Probability Distribution

In a discrete distribution, the mode is the value taken by the random variable X for which the **probability is a**

maximum.



The mode of a probability distribution we have seen and the case of a discrete distribution the mode is the value taken by the random variable X for which the probability is a maximum.

(Refer Slide Time: 06:40)

Mode of a Probability Distribution

- ❖ For a continuous distribution, the mode is the value of X at which the probability density function attains a maximum value.
- ❖ At this point, we have



$$\frac{df(x)}{dx} = 0 \text{ and } \frac{d^2f(x)}{dx^2} < 0$$

In the case of a continuous probability density function we have the mode as the value taken by the random variable X at which the probability density function attains a maximum. So by the definition of the maximum point with respect to the probability density function f of x . We have $\frac{df(x)}{dx} = 0$ and $\frac{d^2f(x)}{dx^2} < 0$. So this is the criteria for the mode.

(Refer Slide Time: 07:31)

Mode of a Probability Distribution

- ❖ It is possible that the maximum we have detected is only a local one.
- ❖ In a probability density function distribution, if there is an unique maximum, the density function is said to be



unimodal.

It is possible that depending upon where we started or where we are doing the analysis the maximum we have detected is only a local maximum. The distribution may have several peaks and we may have identified only one of them. In a probability density distribution if there is a unique maximum the density function is said to be unimodal.

(Refer Slide Time: 08:04)

Mode of a Probability Distribution

- ❖ If the distribution has more than two peaks, it is said to be multimodal.
- ❖ For e.g., particle size distributions may have more than one peak and are referred to as multi-modal.



If the distribution has more than two peaks it is said to be multimodal. Sometimes the distribution may have 2 peaks under it is referred to it as the bimodal distribution. One example where you may encounter multiple modes this in the particle size distribution diagrams you may get 2 peaks or 3 peaks. The first peak may correspond to the fines and the second peak may correspond to the coarse or particles.

(Refer Slide Time: 08:38)

Histograms

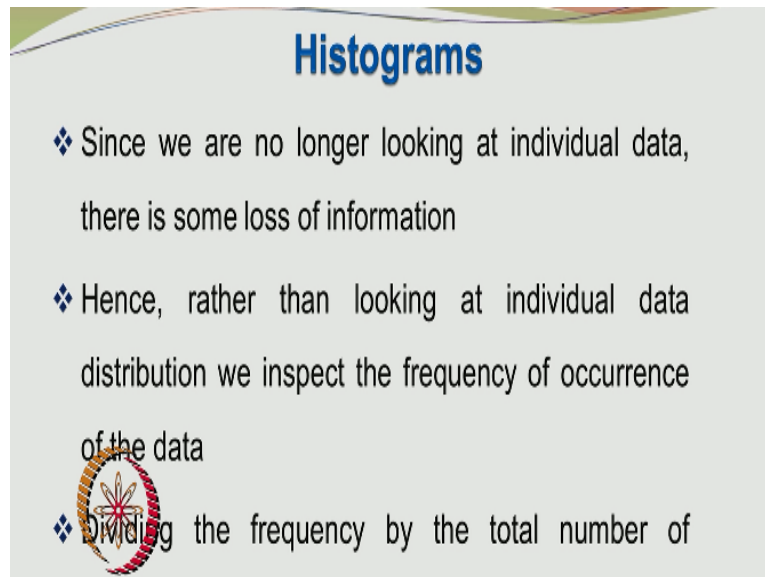
- ❖ It is a visual representation of the frequency distribution (Montgomery and Runger, 2011).
- ❖ They are suited for data that are continuous in nature and hence voluminous as well
(e.g. particle size distributions output from a particle size measuring device).



Now let us come to the next mode of representation of data. We have seen the data being represented in the form of box plots scatter plots now we want to represent the data in the form of histogram. The histogram is a visual representation of the frequency distribution. They are suited for data that are continuous in nature and also voluminous. The particle size distribution output from a particle size measuring device is quite voluminous and we often represent such data in the form of histograms.

Well histogram sort of compiles and combines sections of data before presentation. So there is some loss of information. So you have to step back a bit if you want to look at the overall picture and in doing so you may miss out on some of the finer points. Similarly, when you generate histograms you tend to miss out on some individual data values.

(Refer Slide Time: 10:05)

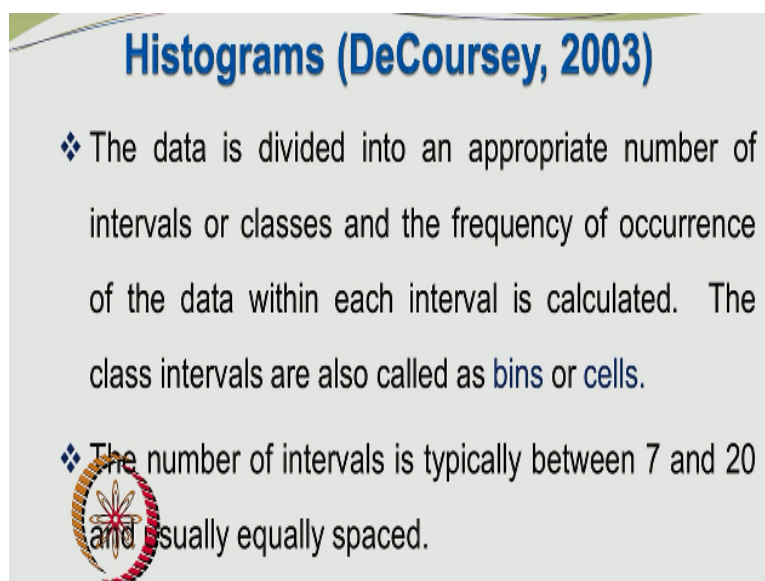


Histograms

- ❖ Since we are no longer looking at individual data, there is some loss of information
- ❖ Hence, rather than looking at individual data distribution we inspect the frequency of occurrence of the data
- ❖ Dividing the frequency by the total number of

And rather than looking at individual data distribution and this may become very cumbersome and tedious if they are too many in number. We inspect the frequency of occurrence of the data. What we have here is the frequency distribution and dividing the frequency by the total number of observations leads to the probability distribution.

(Refer Slide Time: 10:40)



Histograms (DeCoursey, 2003)

- ❖ The data is divided into an appropriate number of intervals or classes and the frequency of occurrence of the data within each interval is calculated. The class intervals are also called as bins or cells.
- ❖ The number of intervals is typically between 7 and 20 and usually equally spaced.

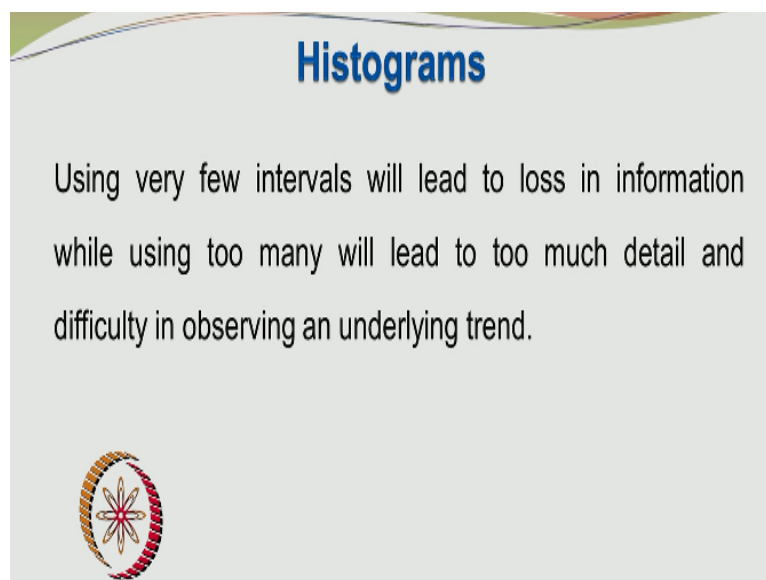
The procedure for doing the histograms is given in several books I am following the

DeCoursey 2003 here. The data is divided into an appropriate number of intervals or classes and the frequency of occurrence of the data within each interval is calculated. The class intervals are also called bins or cells. Suppose you have the distribution of marks and they may be varying from 0 to 100 you may want to generate a histogram of the class performance rather than looking at individual marks you may want to create divisions or bins or cells of size 10.

And for every 10 marks you want to find the number of students following in that particular category. So now you are not looking at individual student's marks, but you are looking at the number of students who have got marks within a certain interval. Let say if the interval is 40 to 50 about 30 students in the class may have got between 40 to 50. How many bins or intervals should you use?

General recommendation is between 7 and 20 and these intervals of bins are usually equally spaced. There may be some situations where you may want to go in for unequally sized intervals. Montgomery and Runger discussed what should be done in such a case. We will continue with the discussion corresponding to intervals of equal sizes.

(Refer Slide Time: 12:35)



Histograms

Using very few intervals will lead to loss in information while using too many will lead to too much detail and difficulty in observing an underlying trend.

If you use very few intervals, then there will be a significant or considerable loss of information if you use too many cells or bins then you are providing too much detail and it will be difficult to sort of pick up an overall trend.

(Refer Slide Time: 13:00)

Histograms

- ❖ Histograms are stable for larger data sets i.e. their appearance does not change drastically with change in bin width.
- ❖ The large data set typically implies more than 75 observations (Montgomery and Runger, 2011).



Histograms are stable for larger data set which means that their appearance does not change drastically with the change in the bin width. So you want to apply the histogram for data sets which are typically having 75 or more observations.

(Refer Slide Time: 13:25)

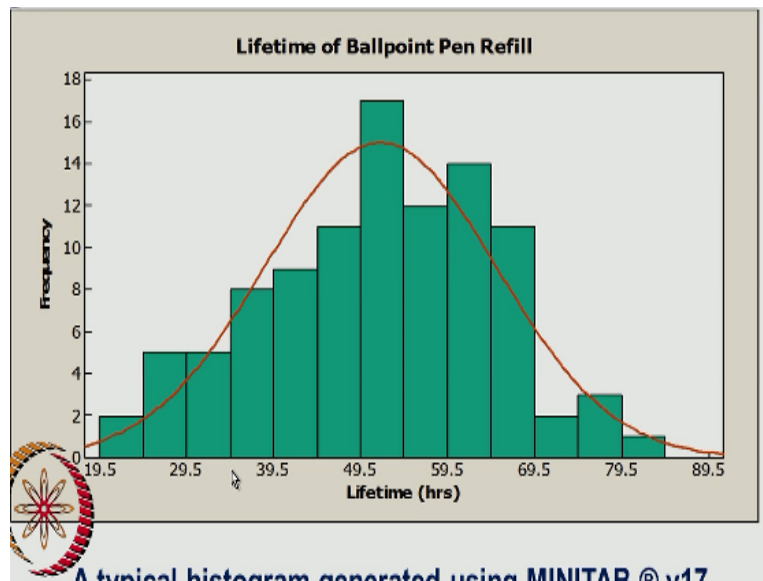
Histograms

- ❖ For large data sets, the histogram shape is a good indicator of the probability distribution best describing them.
- ❖ It also indicates whether the distribution is unimodal or multimodal.
- ❖ It also indicates whether the distribution is symmetrical or skewed.



So to repeat or emphasize the points for large data sets the histogram is a good indicator of the probability distribution best describing them. From the histogram you can see whether the distribution is the unimodal or multimodal. It also indicates whether the distribution is symmetric or skewed.

(Refer Slide Time: 13:45)



So this is the histogram you can see that bins or cells have been created. So each bin is about 5 hours in width you are having lifetime in hours and frequency. So between 19.5 to 24.5 you have 2 occurrences between 24.5 to 29.5 you have 6 occurrences and so on. You have used something like 1, 2, 3, 4, 5, 6, 13 bins here. This is the shape of the distribution in this case the normal distribution best fitting with the given data.

How do you find the number of intervals or classes? Some guidelines or suggestions have been provided. It is not mandatory that you strictly follow or adhere to them. They are just guidelines you may want to tweak the number suggested a little bit to make the appearance of the histogram better from your point of view.

(Refer Slide Time: 15:06)

Histograms (DeCoursey, 2003)

- ❖ The number of intervals/classes (n_i) is given by an empirical formula termed as the **Sturges' rule**:

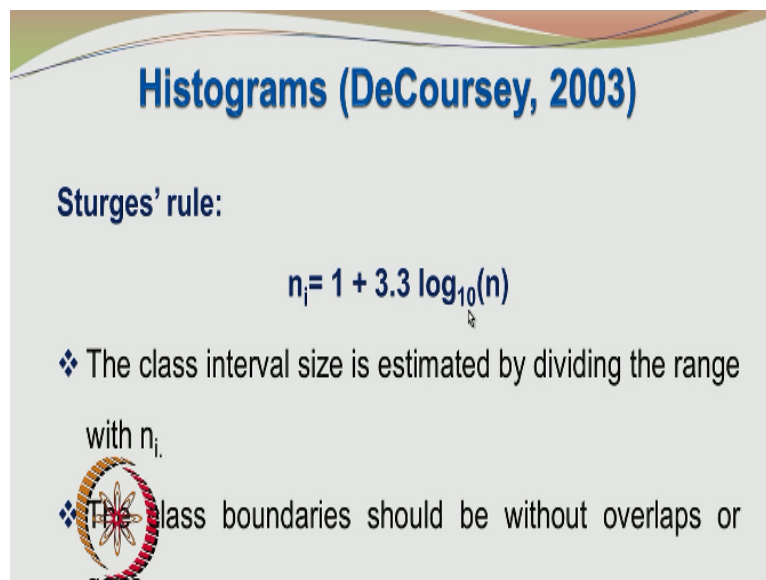
$$n_i = 1 + 3.3 \log_{10}(n)$$
- ❖ Another thumb rule is (Montgomery and Runger, 2011)

$$n_i = (n)^{1/2}$$
- ❖ Here, the square root of the number of observations is

So the number of intervals or classes is denoted by n_i and that is given by $1+3.3 \log n \log to$

the base 10. Montgomery and Runger have another thumb rule they say that the number of intervals should be square root of n where n is the total number of data points. There is a small typo here I will just correct it. The subscript capital I should in fact be small i. So in the recommendation by Montgomery and Runger it is square root of the number of observations is taken for the estimated or suggested number of class interval.

(Refer Slide Time: 16:00)



Histograms (DeCoursey, 2003)

Sturges' rule:

$$n_i = 1 + 3.3 \log_{10}(n)$$

- ❖ The class interval size is estimated by dividing the range with n_i .
- ❖ The class boundaries should be without overlaps or

The Sturges rule is given by $1 + 3.3 \log_{10} n$ as we saw in the previous slide. And the class interval size is estimated by dividing the range with the n_i this is for the number of bins or number of cells and once you have the range you divide that range with the n_i to get the class interval size that is the width of each column in the histogram. Of course the class boundaries should be without overlaps or gaps.

We can see that this histogram is pretty close to normal. We do not see a great deal of asymmetry in this spread of data and we also do not see very distinct multiple peaks we see more or less a single peak in this distribution of course the interpretation of this histogram in terms of number of peaks and symmetry is a bit subjective and it is expected to be so because it is a visual inspection.

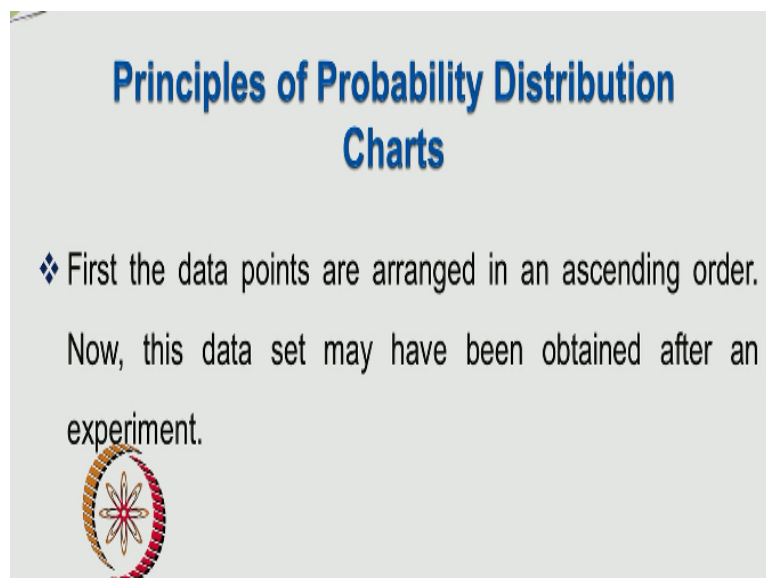
When you are generating the histograms take the lower limit of the smallest class or interval to be slightly lower than the smallest data. And the upper limit of the largest class or interval to be slightly higher than the largest number. What it means is suppose you are having data points like 1 to 100 when you are defining the range of the histogram the lower limit would be slightly lower than the smallest data point.

The smallest data point is 1 you may want to take 0.5 and similarly the upper limit of the largest interval or class should be slightly higher than the largest number. So you may want to put 100.5 even though the largest number is 100 you may want to put 100.5. So the class interval will start from 0.5 the first interval will be drawn using that as the lower limit and the last interval or the class will end at 100.5.

Now we come to another aspect if you look at the histogram we plotted the data in the form of a histogram and then you can see that a normal or the bell shaped curve was drawn to see whether the distribution was normal. We can do even better by using probability charts. In many cases your experiments modeling may be based on certain assumptions. You may assume that the errors in the experiments are random and they are normally distributed with 0 mean and constant variance.

This is a usual assumption. So you want to check whether the errors are distributed normally and for this reason we use probability distribution charts. When we want to find the shape or form of the underlying distribution we have a speculation either based on experience or other worker results. So you want to test whether your data also follows that particular distribution. So before you plot the data you have to do a bit of pretreatment and the full details about the motivation for the pre treatments and all that is given by Ogunnaike.


(Refer Slide Time: 20:39)



Principles of Probability Distribution Charts

- ❖ First the data points are arranged in an ascending order.

Now, this data set may have been obtained after an experiment.



What you do first is to arrange the data points in an ascending order. So the data may have been obtained after an experiment and the way you perform the experiments could have been

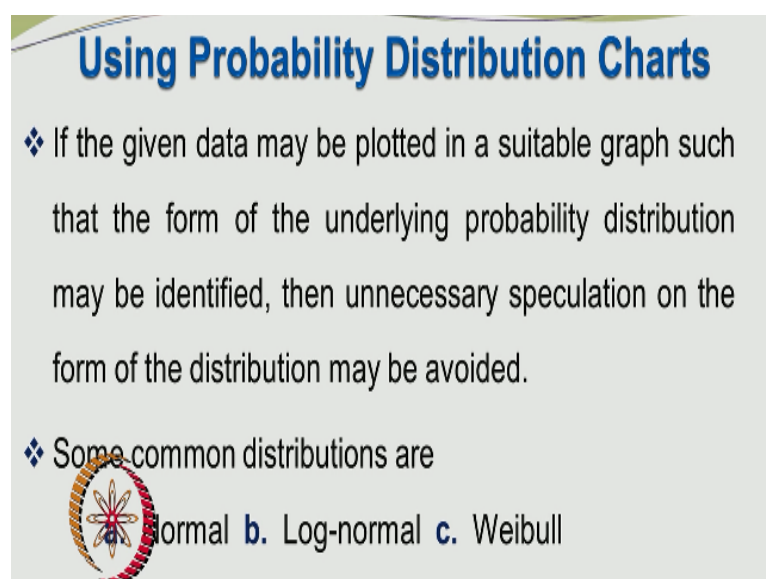
in a random fashion. So rather than doing the experiment with the lowest sitting first and then moving on to the medium sitting and then going to the highest sitting or the largest sitting you might have mixed the sequence of runs.

So that the external influences are sort of evenly spread for all the sittings. So this is called as randomization, but once you have obtained the responses from your experiments you arrange them in an ascending order. Now if you conduct an experiment or series of experiments and note down the data and arrange them in ascending order. You have to see that you have performed these experiments between June and July.

Now you are curious and again perform the same set of experiments between July to August there is no guarantee or assurance that the data you have created in June to July will be identical to the data you have created between July to August. There are so many random factors that may have influenced the outcome of your experiments so the data may not match. However, if the distribution of the data is according to a particular trend in the month of June to July.

You more or less would get a similar kind of trend for the experiment conducted between July to August. So the distribution of the data would be more similar than the data points themselves. So how to go ahead and identify the trend of the distribution. You do not have to always and only test for normal distribution there are also other important statistical distributions and probability distribution charts are available for event those.

(Refer Slide Time: 23:15)



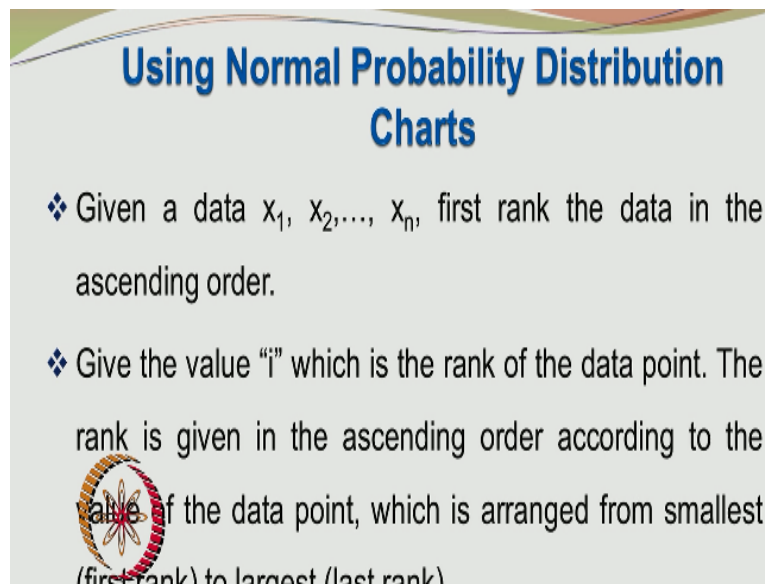
Using Probability Distribution Charts

- ❖ If the given data may be plotted in a suitable graph such that the form of the underlying probability distribution may be identified, then unnecessary speculation on the form of the distribution may be avoided.
- ❖ Some common distributions are
a. Normal b. Log-normal c. Weibull

Some of the more common ones are of course the normal distribution and then the Log-normal distribution, Weibull, Chi squared and gamma. So once you plot the data in a suitable graphical form and then show that indeed the data is following the distribution there would not be any ambiguity. It is a much better way than simply stating the assumption that the errors are normally distributed okay then it becomes a speculation if it is not backed with proper proof.

Now things have become more advanced and you do not even need a pencil and the probability chart and do the so called plotting procedure. There are several software which have now available and the data can be plotted and represented very easily.

(Refer Slide Time: 24:07)



Using Normal Probability Distribution Charts

- ❖ Given a data x_1, x_2, \dots, x_n , first rank the data in the ascending order.
- ❖ Give the value "i" which is the rank of the data point. The rank is given in the ascending order according to the value of the data point, which is arranged from smallest (first rank) to largest (last rank).

So once you have the data you arrange the data points in an ascending order. Ascending means the smallest data comes first and the largest data goes to be end. And after having arranged the data in the ascending order rank the data. The lowest data which was first in the list is given the first rank. Well this is sort of different from the classroom case where the highest marks are given the first rank.


But here we give the data which is the lowest in the lot as the rank number 1 So the data are arranged such that the smallest data has the first rank and the largest data has the last rank.

(Refer Slide Time: 24:58)

Using Normal Probability Distribution Charts

- ❖ The ordered observations are then plotted against their observed cumulative frequency

$$(i-k)/(n-2k+1) = (i-0.5)/n$$

- ❖  n is the number of data points and $k=0.5$ on the appropriate probability paper.

Now there are several ways to plot the data and the data points the order observations are plotted against the so called observed cumulative frequency. The formula is $(i-k)/(n-2k+1)$. K is the parameter in the observed cumulative frequency function. So if you put $k=0.5$ this formula here reduces to $(i-0.5)/n$ where I the rank K is the parameter you can take it as 0.5 as an example.

N is the number of data points in the set. So when you put $k=0.5$ this formula reduces to $(i-0.5)/n$. So you are going to plot the observed cumulative frequency against the ordered observations. You can plot $(i-0.5)/n$ versus the ordered observation directly on the appropriate probability paper. If you do not have the probability paper, then you have to do a bit more calculations.

So we are essentially creating percentiles and these percentiles are also equidistant. There is another formula for finding the percentiles and this is given by $i/n+1$. The formula we are using first is $(i-0.5)/n$ then you have $i/n+1$. So there are several versions for generating these percentiles.

(Refer Slide Time: 26:50)

Using Normal Probability Distribution Charts

Note that $100 \times m^{\text{th}}$ percentile in a data set indicates the value p such that approximately $100 \times m\%$ of the data are below or equal to p and $(1-m) \times 100\%$ data are



So what is meant by this percentile? Please note that the $100 \times m^{\text{th}}$ percentile in a data set indicates the value p such that $100 \times m\%$ of the data are below or equal to p and $1-m \times 100\%$ data are above it. So we calculate the percentiles.

(Refer Slide Time: 27:28)

Using Normal Probability Distribution Charts

See if the values corresponding to the percentiles according to the normal distribution correlate with the actual data values, i.e. whether they are similarly distributed with respect to each other



Now the idea is to see if the values corresponding to the percentiles according to the normal distribution correlates with the actual data values whether they are similarly distributed with respect to each other. So essentially we are going to compare it with the normal distribution percentiles how to do that we will see.

(Refer Slide Time: 27:58)

Using Normal Probability Distribution Charts

- ❖ Obtain the z value for each $(i-0.5)/n$
- ❖ This represents F^{-1} of $(i-0.5)/n$ as well where $F(x)$ represents the cumulative distribution function of the normal distribution

Let us take $i-0.5/n$ as a probability value then you have to find what is the corresponding z value from the normal probability distribution. Usually you are given the value of z and then you find the probability, but here we are doing in a slightly different manner. Here we are giving the value of the probability and finding out the value of z so this is the inverse case. So you have to find the cumulative distribution function inverse of $i-0.5/n$. F represents of course the cumulative distribution function of the normal distribution.

So to sort of repeat we arrange the data actual raw data in the ascending form and rank the data the smallest data has the first rank $i=1$ and the largest data will have the highest rank. Then what you do is you find out for each of the I values $i-0.5/n$ that will be a number. And once you have done that you find the z value from the standard normal probability chart.

(Refer Slide Time: 29:44)

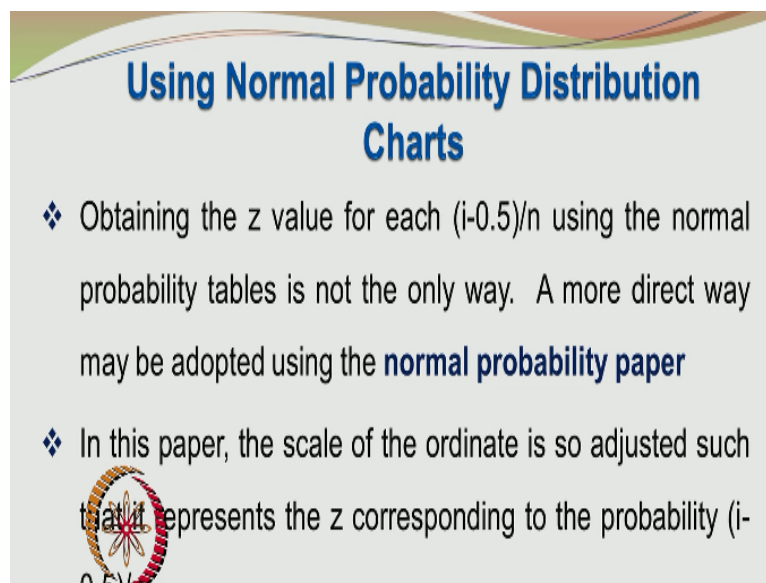
Using Normal Probability Distribution Charts

- ❖ Plot z - values as **ordinate** against the ordered values as **abscissa**. Hence we can do this on a regular graph sheet
- ❖ If the assumed distribution adequately satisfies the data, then the points will fall on approximately a straight line
- ❖ If the points deviate significantly from linearity, the

Once we have the z values plot the z values as or ordinate against the ordered values as abscissa. So the z values which you have obtained by the inverse of the cumulative distribution function corresponding to $(i-0.5)/n$ or plotted against the ordered values in the abscissa. Here we can use the regular graph sheet itself. If we assume the distribution adequately satisfies the data, then the points will approximately fall on a straight line. If the point deviate significantly from linearity the hypothesized model is not adequate.

So you want to get the data points aligned more or less on a straight line. You would not get a perfect straight line if you are showing a general overall trend it is acceptable.

(Refer Slide Time: 30:41)



Using Normal Probability Distribution Charts

- ❖ Obtaining the z value for each $(i-0.5)/n$ using the normal probability tables is not the only way. A more direct way may be adopted using the **normal probability paper**
- ❖ In this paper, the scale of the ordinate is so adjusted such that it represents the z corresponding to the probability $(i-0.5)/n$

0.5V

So obtaining the z values for each $(i-0.5)/n$ using the normal probability table is not the only way. We can also use the normal probability paper. When you find the z value using the probability tables you could then go ahead and use the usual or regular graph paper, but if you are already having the normal probability paper with you then you do not have to identify the z value for each $(i-0.5)/n$.

You go ahead and directly plot the $(i-0.5)/n$ on the graph paper on the normal probability graph paper. What is so special about this graph paper? In this graph paper the scale of the ordinate is so adjusted that it represents the z corresponding to the probability of $(i-0.5)/n$ directly.

(Refer Slide Time: 31:44)

Using Normal Probability Distribution Charts

- ❖ When you plot y vs. x on a log-log sheet, you directly plot y and x on ordinate and abscissa of the logarithmic sheet respectively.
- ❖ You don't need to take logarithm of x and logarithm of y and then plot them.

You can take a simple analogous situation when you plot y versus x on a log-log sheet. You directly only plot a value of x and y on the abscissa and ordinate respectively.

(Refer Slide Time: 32:06)

Using Normal Probability Distribution Charts

- ❖ If the data obeys $y=mx^n$, you will get a straight line on a log-log graph automatically since

$$\log(y) = \log(m) + n \log(x)$$

Here, slope is n and intercept is m (not $\log(m)$!)

- ❖ The scales are automatically adjusted for the

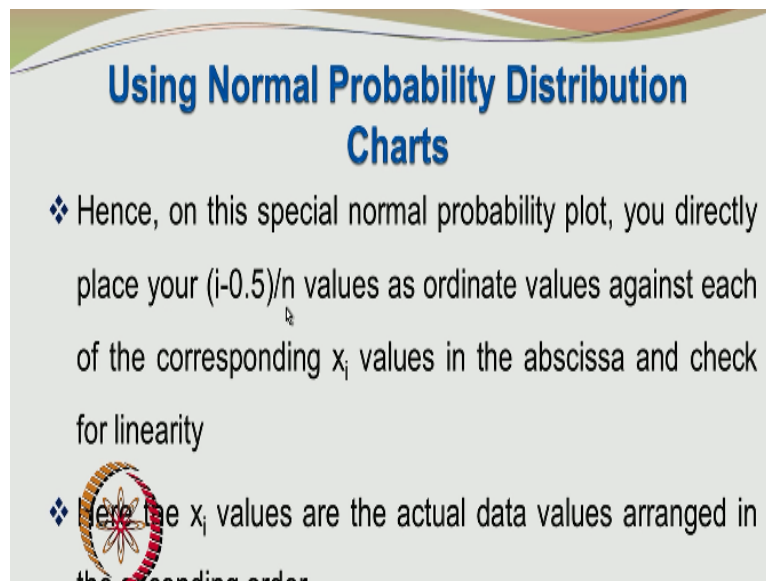
For example, if you want to test the model $y=mx$ to the power of n . You want to check whether this model is correct. So you can take $\log y$ and $\log x$ do these mathematical calculations yourself and then go to the regular or usual graph paper and plot the data value, but if you have the log-log sheet then you can plot \log of y versus \log of x directly. What I am trying to say here is plot y versus x in the log-log sheet.

In the y axis identify the value of y . In the x axis identify the value of x you do not have to take \log of y and then identify that particular value in the logarithmic scale. You plot y and x on the logarithmic scales. When we do that you will find that if your assumption of model is

correct the data points are falling on a straight line with slope n and intercept m remember the slope is n and intercept is m directly not $\log m$ okay that is an important thing to note.

So in the log-log sheet the scales are automatically adjusted in terms of the length for the logarithmic basis. Similarly, in the normal probability chart when you are plotting the ordered x values against $(i-0.5)/n$ you directly use $(i-0.5)/n$ and x values. That is important you do not have to find the z values when you are having the normal probability chart review.

(Refer Slide Time: 34:20)

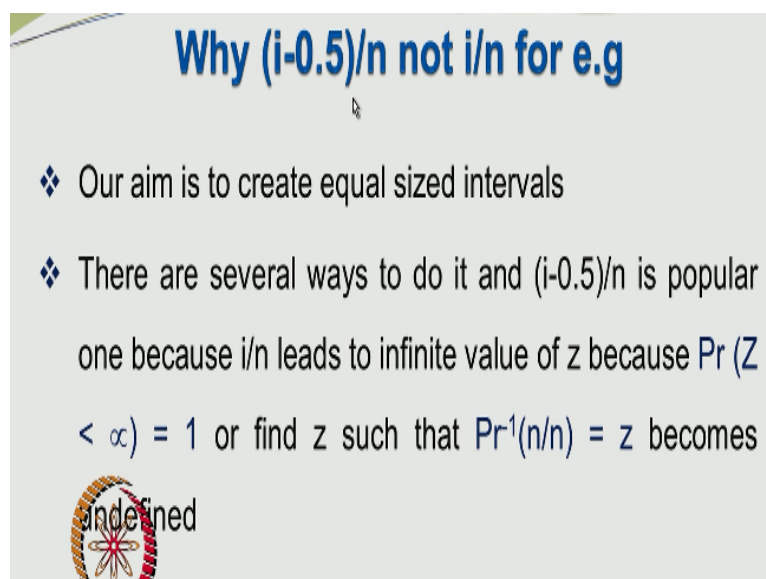


Using Normal Probability Distribution Charts

- ❖ Hence, on this special normal probability plot, you directly place your $(i-0.5)/n$ values as ordinate values against each of the corresponding x_i values in the abscissa and check for linearity
- ❖ Here the x_i values are the actual data values arranged in the ascending order

So on the special normal probability plot you directly place your $(i-0.5)/n$ values as ordinate values against each of the corresponding x_i value in the abscissa and check for linearity. Here the x_i values are the actual data values arranged in the ascending order.

(Refer Slide Time: 34:39)



Why $(i-0.5)/n$ not i/n for e.g

- ❖ Our aim is to create equal sized intervals
- ❖ There are several ways to do it and $(i-0.5)/n$ is popular one because i/n leads to infinite value of z because $\Pr(Z < \infty) = 1$ or find z such that $\Pr^{-1}(n/n) = z$ becomes undefined

Why did we choose $i-0.5/n$ and not i/n . Our aim was to create equal sized interval and $i-0.5/n$ is popular one because if you had chosen i/n it may lead to an infinite value of z because probability of $z < \text{infinity}=1$ and if you want to find the value of z such that the inverse of the probability= 1 . Then the x becomes undefined. You are doing the inverse of the cumulative distribution function.

What is the value of x that will give the required probability if the required probability for the last data point is 1 then the identified value of x will be undefined. So you run into this problem with the usage of i/n in the creation of equal sized intervals. So what you are doing here is instead using $i-0.5/n$.

(Refer Slide Time: 36:00)

Why $(i-0.5)/n$ not i/n for e.g

i	Marks	$i/(n+1)$	z	$i-0.5/n$	z	i/n	z
1	20	0.091	-1.335	0.050	-1.645	0.100	-1.282
2	30	0.182	-0.908	0.150	-1.036	0.200	-0.842
3	37	0.273	-0.605	0.250	-0.674	0.300	-0.524
4	44	0.364	-0.349	0.350	-0.385	0.400	-0.253
5	50	0.455	-0.114	0.450	-0.126	0.500	0.000
6	60	0.545	0.114	0.550	0.126	0.600	0.253
7	70	0.636	0.349	0.650	0.385	0.700	0.524

Let us take a simple example here let us say that marks 20, 30, 37, 44, 50, 50, 60, 61, 70 the marks are already arranged in the ascending order and so the rank is also given here 1, 2, 3 so on to 8 and you can calculate $i/n+1$. You can also calculate i/n , you can calculate $i-0.5/n$. So equal sized intervals have been created 0.05 to 0.15 it is an interval size of 0.1 again 0.1 and so on.

So you have the different interval depending upon what formula you have chosen. Now you find the Z value what is the value of Z that will give the probability value of 0.091. So it is inverse problem. Given 0.091 this is corresponding to the probability or the area under the curve. So corresponding to 0.091 what is Z value that is -1.335. So an z value of -1.335 will correspond to a probability value of 0.091 in the normal distribution diagram.

Similarly, you can find the z values for all other elements in the column given here. Similarly, you have for $i=0.5/n$ 0.05 0.15 so on to 0.75. So corresponding to these probabilities given here what are the corresponding z values you have to take the inverse of the cumulative distribution function. Similarly, when you do i/n you have values ranging from 0.1, 0.2, 0.3 so on to 0.8 you can also find out the z values.

So you can now plot on the rectangular or I am using the word normal, but rectangle or the usual graph paper you can plot z versus the marks and you can check whether they are having a linear relationship.

(Refer Slide Time: 38:37)

raw	Ordered	Rank	Frequency	z value
176	176	1	0.05	-1.645
191	183	2	0.15	-1.036
214	185	3	0.25	-0.674
220	190	4	0.35	-0.385
205	191	5	0.45	-0.126
192	192	6	0.55	0.126
201	201	7	0.65	0.385
190	205	8	0.75	0.674
185	214	9	0.85	1.036
185	220	10	0.95	1.645

Let us take in this example a raw data values which are given to be 176, 191, 214 and so on to 185 these are not ranked. So you arrange them in the order of ascending order. And then you give the rank 1, 2, 3 so on to 10 and use the formula $i-0.5/n$. You have 10 data points so $n=10$ and $i-0.5/n$ would mean I is the rank $1-0.5$ is 0.5. $0.5/10$ is 0.05 $1-0.5/n$ rank is 2 $i=2$, $2-0.5$ is 1.5, $1.5/10$ is 0.15.

Similarly, you can calculate for all these other ranks then you can find out the z value corresponding to these probabilities that -1.645 for 0.05 area under the curve or 0.05 probability is quite famous one. It is familiar to all those people who extensively use the normal chart. Similarly, +1.0645 corresponds to the area under the curve of 0.95 this is because of the symmetry of the normal distribution. So you can plot the z values against the ordered mark and see whether you get a straight line.

One more thing I would like to point out why we should use preferably $i-0.5/n$ and not i/n . For example, if you had used i/n in the last mark would have had a rank of 10/10 would lead to an i/n value of 1 and that would be undefined in the value for z what is the value of z that would lead to a probability of 1 the answer is infinity just as you have the probability of 0 having a z value of $-\infty$ you have the z value of $+\infty$ corresponding to the probability of 1 or the one which is containing 100 percentile of the data.

So since you have this difficulty and if you had used $1-0.5/n$ the last entry becomes only 0.95 and for which you can easily find out the z value. So to summarize what you do is plot the value of z against the marks and then if they are falling on a straight line you then conclude that it is following the normal distribution. So the given data when plotted in terms of z on the y axis and x_i on the x axis shows as linear trend.

So we can safely assume that the data are distributed normally. To summarize we have only looked at few important characteristics of distributions and the presentations of data. We looked at mean and median and we also compared them which is a more robust estimator of the central tendency. In addition to knowing the center point of the distribution it is also important for us to get an idea about the spread the range is quick estimate of the spread of the distribution.

But more reliable one which uses all the data points in the distribution or the distributed data is the sample variance. We also have to correctly use the appropriate degrees of freedom in the different formula we are considering for our analysis. We looked at the box plots, the scatter plot and the histogram method of representing the data. The overall summary and comparison between 2 sets of data could be given nicely by the box plot.

When you have 2 sets of data and we want to compare if at all there is any relation between them then you can go ahead and use the scatter plot. When you have a large data set then you may want to present the overall trend rather than the individual details and then in that case you go for the histogram. By using the histogram, you can see whether the data is symmetric single peak or multiple peak and whether it is following the normal distribution.

Finally, we looked at more concrete way rather than the visual inspection to identify which distribution the given data better relates to and we found that even without using special

probability papers we can check for the distribution. What we have to do is organize the data into ranked data. We put the data in the ascending order then assign ranks to them then use an appropriate percentile creating formula the more popular one and the more convenient one is $i-0.5/n$ where I is the rank and n is the total number of data points.

And what we then did was to identify the z value which gave the probability = $i-0.5/n$ and we plotted the z value against the ranked x values. And if we get a straight line, then we say that data is normally distributed. So we will continue our discussion in the next lecture. We will also work out a few example problems which will drive home the concepts we have covered so far. Thank you.