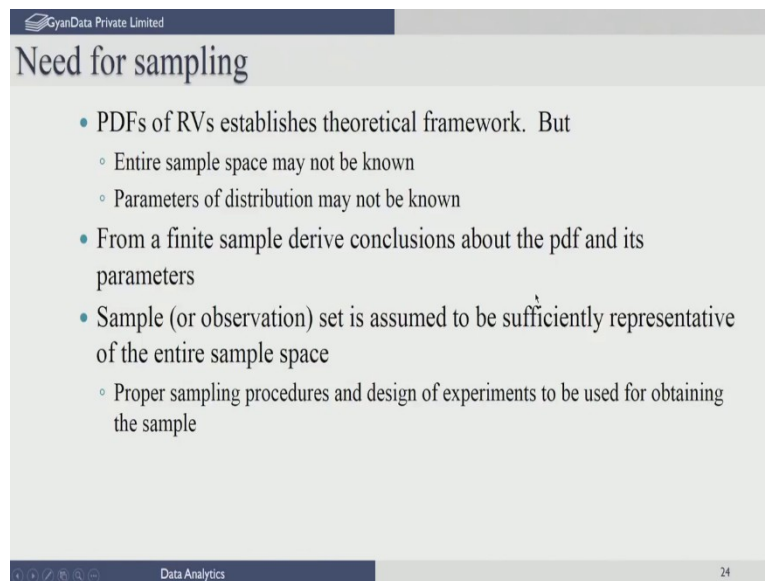


Data Science for Engineers
Prof. Raghunathan Rengaswamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 21
Sample Statistics

In the preceding two lectures, I introduced the concepts of probability. Probability provides a theoretical framework for providing, for performing statistical analysis of data. Statistics actually deals with the analysis of experimental observations that we have obtained. So, in this lecture I will introduce you to a few measures statistical measures and how they are used in analysis.

(Refer Slide Time: 00:47)



The slide is titled "Need for sampling" and is part of a presentation by GyanData Private Limited. It contains a bulleted list of points explaining the need for sampling. The slide is numbered 24 and has "Data Analytics" in the footer.

- PDFs of RVs establishes theoretical framework. But
 - Entire sample space may not be known
 - Parameters of distribution may not be known
- From a finite sample derive conclusions about the pdf and its parameters
- Sample (or observation) set is assumed to be sufficiently representative of the entire sample space
 - Proper sampling procedures and design of experiments to be used for obtaining the sample

So, what is the need for performing statistical analysis when we have already talked about probability density functions and so on?

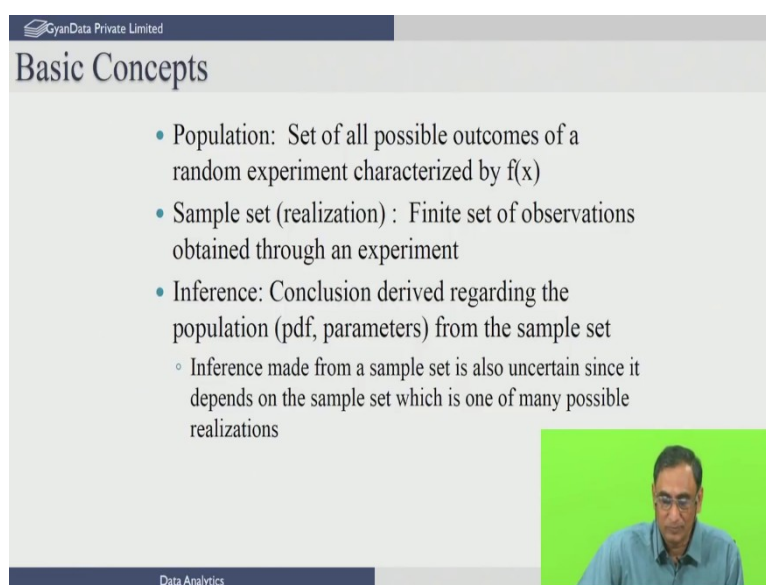
Typically, when we are actually doing analysis we do not know the entire sample space. We may also not know all the parameters of the distribution from which the samples are being withdrawn. Typically we actually obtain only a few samples of the total number of avail population. So, from this finite sample we have to derive conclusions about the probability density function of the entire population and also infer, make inferences about the parameters of these distributions.

So, the sample or observation set is supposed to be sufficiently representative of the entire sample space. Let us take an example.

Suppose you want to actually find out the average height of people in the world, you cannot go and sample just people or take heights of American people alone because they are known to be much taller compared to the Asian people.

So, when you take samples you should take examples from let us say America, from Europe from Asia and so on, so forth. So that you get a representative of the entire population of this world. So, this is called proper sampling procedures and these are dealt with in the design of experiments. We will assume that we have obtained a sample; you have done the due diligence and obtained the representative sample of whatever population you are trying to analyze.

(Refer Slide Time: 02:21)



GyanData Private Limited

Basic Concepts

- Population: Set of all possible outcomes of a random experiment characterized by $f(x)$
- Sample set (realization) : Finite set of observations obtained through an experiment
- Inference: Conclusion derived regarding the population (pdf, parameters) from the sample set
 - Inference made from a sample set is also uncertain since it depends on the sample set which is one of many possible realizations

Data Analytics

A small video inset in the bottom right corner shows a man with glasses and a blue shirt speaking.

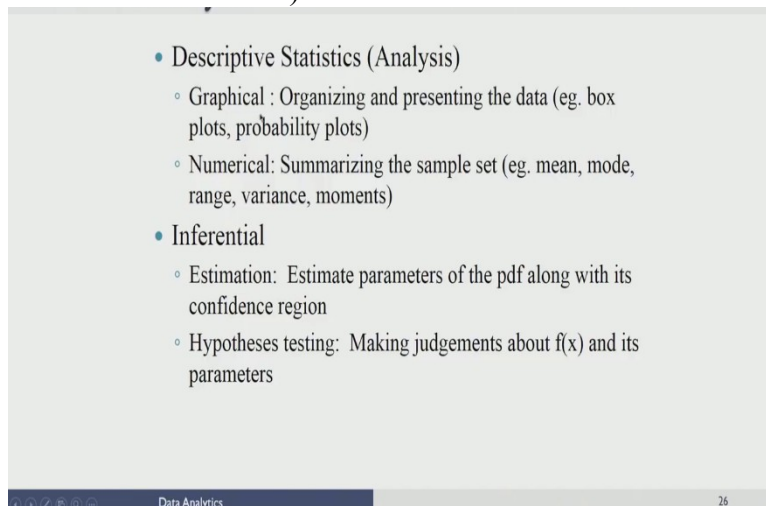
Now, with basic definition of population is the set of all possible outcomes of this random experiment. We have already defined this as the sample space. What you are going to obtain is just a few examples or samples and these are called the sample set. It's an infinite set of observations obtained through whatever experiment that you are going to conduct. Now from this sample set you want to make inferences which are conclusions that you derive regarding the population itself, which you do not know its either the probability density function of the population or the parameters of the population.

You have to note that when you actually lose such inferences, your inference is also stochastic or uncertain because the sample that you have drawn are also uncertain; they are not representative of the entire population. So, you should expect that your inferences are also uncertain and therefore, when you provide the answers, you should

actually provide also the confidence interval associated with these estimates that you have deriving.

So, that is one of the reasons that we actually studied probability density functions because then you can characterize all the estimates that you have obtained from the sample in terms of this confidence interval and so on or the probability that you will obtain this value.

(Refer Slide Time: 03:40)



- Descriptive Statistics (Analysis)
 - Graphical : Organizing and presenting the data (eg. box plots, probability plots)
 - Numerical: Summarizing the sample set (eg. mean, mode, range, variance, moments)
- Inferential
 - Estimation: Estimate parameters of the pdf along with its confidence region
 - Hypotheses testing: Making judgements about $f(x)$ and its parameters

Data Analytics 26

So, let us actually look at some typical analysis statistical analysis. We can divide the statistical analysis into two parts, the graphical part or graphical analysis where we use plots and graphs in order to have a visual feel of the entire data. The other way of doing is to actually do quantitative computations or numerical computations, where you try to summarize the entire sample sent by a few parameters. Example we will talk about mean and variance and so on. Notice that we have taken hundreds of data points or experiments, you cannot go and tell somebody all the hundred values, you cannot reel off all these values that will not be possible for somebody to digest.

Summary statistics that we do numerically allows you to get a feel for the entire data set that you have obtained without knowing the individual observations. And that is why the these are very useful and they are also called summary statistics. Now inferential statistics deals with two kinds of problem estimation problem, where we try to estimate parameters of the probability density function.

We did talk about parameters such as the expected value or the first moment and the second moment and so on, and different distributions are different number of parameters and how do we estimate these parameters from a small sample that we obtain. And how do you give a confidence region for these estimates that is called estimation and the

other kind of decision making that we want to do is; we want to judge whether particular value is 0 or not. The parameters of the distribution and such decision making that we do from a sample is called hypothesis testing.

We want to know whether a customer will continue to remain with you or will leave you for another vendor, based on whatever offers that you are making. So, these are come under the category of hypothesis testing. We will first deal with the descriptive statistics and in the next lecture we will deal with inferential statistics.

(Refer Slide Time: 05:46)

Measures of Central Tendency - Mean

- Represent sample set by a single value
 - Mean (or average): $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$
 - Best estimate in least squares criterion
 - Unbiased estimate of population mean: $E[\bar{x}] = \mu$
 - Affected by outliers
 - Eg: Sample heights of 20 cherry trees
[55 55 59 60 63 65 66 67 67 67 71 71 72 73 75 75 78 81 82 83]
• Mean = 69.25 (population mean used to generate random sample was 70)
 - Mean = 71.75 (after a bias of 50 was added to first sample value)

GyanData Private Limited

Data Analytics

27

So, some of the summary statistics that we can define for a sample or what we call measures of central tendency, it is the kind of the center point of this entire sample you might say. And let us define what is called the mean, these are measures that you are familiar with from your high school courses in mathematics.

So, let us recap some of these. The mean of a sample is defined as the summation of all the data points that you obtain divided by the number of datapoints that you have. So, that is also denoted by the symbol \bar{x} and its also called the mean or the average of the sample. This particular thing as I said can be viewed as a central point of the entire sample that you have got.

And we can show that this estimate that we obtained of the sample the average is the best estimate in some sense. Later on, we will set up what is called the least squares method of estimating parameters. And if you set up this particular criterion for estimating parameters you will find that \bar{x} is the best estimate that you can get from the given sample of data. We can also show some properties of this estimate. For

example, we can prove that this \bar{x} represents an unbiased estimate of the population mean μ which you do not know anything about.

What do we mean by the unbiased estimate? Expectation of \bar{x} is μ . This can be analytically proven for any kind of distribution. And in order to prove understand what this means you say that suppose you take a sample of N points and you get an estimate \bar{x} . And you repeat this experiment and draw another random sample from the population of N points and get another value of \bar{x} .

And you average all these \bar{x} s that you get from different experimental sets, then the average of these averages will tend to this population mean. That is a way of interpreting this statement that its an unbiased estimate. There are other properties of estimates we will see that we want the demand, but this is an useful and important property of estimates that you should always check.

The one unfortunate aspect of this particular statistic or mean is that it is if there is one bad data point in your sample, by mistake you have made a wrong entry, then your estimate of \bar{x} can be significantly affected by this bad value. The bad value is what we call an outlier and even a single outlier in your data can give rise to a bad estimate of \bar{x} .

Let us take one example we have taken 20 cherry trees and we have measured the heights of the cherry trees in terms in feet and we got let us say the set of bunch of numbers; generated these randomly from a normal distribution with some mean which is 70 and the standard deviation of 10. So, the population mean is 70 and the population standard deviation is 10 and I got these values. You can use `rrnorm` for example, in `r` in order to generate such data points.

Now, if you take this sample of 20 points and compute the mean, you get a value of 69.25 which is very close to the population mean. So, you see that it is a good estimate of the population mean, even though you did not know what that value was until I told you. Now on the other hand if I take the first data point 55 and add a bias wrongly enter it as 105 let us say by adding 50 to `ta` and then recompute this mean, I will find that the mean becomes 71.75.

It starts deviating from 70 you see more significantly. A single bias in this sample actually caused your estimate to become poorer. That is what we mean by saying that \bar{x} will get affected by outliers in the data. We can de ne other measures of central tendency which are robust with respect to the outliers, even if the outlier exists, it does not change by much and we will see what that such a measure is.

(Refer Slide Time: 10:13)

GyanData Private Limited

Measures of Central Tendency – Median

- Represent sample set by a single value
 - Median: Value of x_i such that 50% of the values are less than x_i and 50% of observations are greater than x_i
 - Robust with respect to outliers in data
 - Best estimate in least absolute deviation sense
 - Eg: Sample heights of 20 cherry trees
[55 55 59 60 63 65 66 67 67 67 71 71 72 73 75 75 78 81 82 83]
 - Median = 69 (population mean used to generate random sample was 70)
 - Median = 69 (after a bias of 50 was added to first sample value)

Data Analytics 28

Another measure of central tendency is what is called a median. The median is a value such that 50 percent of the data points lie below this value and 50 percent of the experimental observations are greater than this value. So, you like to find out that value below which half the data points lie and above which half the data points lie. For doing this you have to order all the observations that you have got from smallest to highest and then find out the middle value. Let us do this through an example.

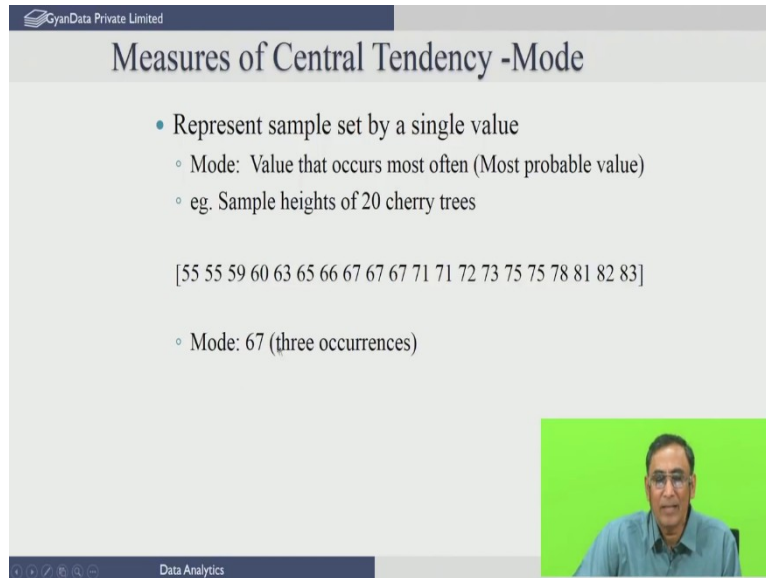
So, same 20 cherry trees data I have looked at, this data point I have ordered from the smallest to the largest. And if you look at it the tenth point 1, 2, 3, 4, 5, 6, 7, 8, 9, 10; tenth point is 67 because there are even number of points, the eleventh point is 71 and you take the average between this and call that the median. If there are odd number of points then you can take the middle point just as it is because there are even number of points, you take the average of the mid midpoints, in this case the tenth and the eleventh point and that gives you a median of 69.

Suppose we add a bias in the first data point as before and make this 105 and then reorder the data and find out the again the median; we find that the median has not changed.

So, the presence of an outlier in this particular case has not affected the median at all and that is why we call this a robust measure even if there is a bad data point in your samples. You can also show that this estimate is the best estimate in some sense. In this case the merit that you are using is what is called the absolute deviation. That is you are asking what is the estimate which deviates from the individual observations in the absolute sense to the least extent? And it turns out that the median is such a point, such an estimate. So, when there are

outliers typically we would like to use this as a central measure rather than the mean.

(Refer Slide Time: 12:25)



GyanData Private Limited

Measures of Central Tendency - Mode

- Represent sample set by a single value
 - Mode: Value that occurs most often (Most probable value)
 - eg. Sample heights of 20 cherry trees

[55 55 59 60 63 65 66 67 67 67 71 71 72 73 75 75 78 81 82 83]

- Mode: 67 (three occurrences)

Data Analytics

A mode is another measure of central tendency and this value is the value that occurs most often or what is called the most probable value. And if you take the example of this 20 cherry trees data, we find that the most probable value, the value that repeats more often, is 67.

Again you said this is 3 consecutive, 3 occurrences of this as compared to any other data point and that is called the mode. And in a distribution if it is a continuous distribution, this represents the highest value of this maximum value of the density function. And you should expect most of the data to be clustered around this most probable value. Sometimes distribution may have two modes. What is called a bimodal distribution in which case if you sample from such a distribution, you will find two clusters one clusters around the one of the modes and another cluster around the second mode. So, you should interpret the mode as that value around which you will find most of the data points.

(Refer Slide Time: 13:35)

GyanData Private Limited

Measures of Spread

- Represents spread of sample set
 - Sample variance : $s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$
 - Unbiased estimate of population variance : $E[s^2] = \sigma^2$
 - Standard deviation is sqrt of variance
 - Mean absolute deviation : $\bar{d} = \frac{1}{N} \sum_{i=1}^N |x_i - \bar{x}|$
 - Range : $R = x_{max} - x_{min}$
 - Eg. Sample heights of 20 cherry trees
 $s^2 = 70.5132$ and **212.25 with outlier**
 $s = 8.392$ (population std used for generating numbers was 10)
 $MAD = 6.85$ and **9.5 with outlier**
 $Range = 83 - 55 = 28$

Data Analytics 30

The another measure which characterizes a sample set is what we call the measures of spread and tells you how widely their data is ranging. So, one of the measures is what to call the sample variance denoted by the symbol s squared and that is defined as the data point x_i - the sample average that you have already computed. This deviation of the observation from the sample mean u square.

And add over all the data points n data points and divide the this particular sum squared value by $N - 1$, such a measure is called the sample variance. And again you can prove that the sample variance is an unbiased estimated estimate of the population variance and the square root of the sample variance is also known as the standard deviation.

Now, just like the mean, the sample variance happens to be also a very susceptible to outliers. So, if you have a single outlier, the sample variance, our sample standard deviation can become very poor estimate of the population parameter. So, we define another measure of spread which is called the mean absolute deviation somewhat similar to the median. In this case instead of taking the sum squared deviation, we take the absolute deviation of the data point from the mean; you can also take it from the median if you wish. So, deviation of the observation from the mean or the median, you take the absolute value of this deviation sum over all the end points and divide by N and that is what is called the mean absolute deviation.

Again whether you should divide by N or $N - 1$ is a point to be noted. Typically we divide by $N - 1$ to indicate that if you have only one data point; it is not possible to estimate s squared. For example, if

you have one data point, s^2 will turn out to be 0 because the mean will be equal to the point. So, really speaking you have only $N - 1$ data points to estimate the spread. So, that is why we divide by $N - 1$ to indicate that one data point has been used up to estimate the sample mean or the median whatever the parameter that you are actually estimated.

So, similarly here also you can divide by $N - 1$ to indicate that only $N - 1$ data points were available for obtaining the mean absolute deviation. A third measure of spread is what is called the range that is basically the difference between the maximum and minimum value. All of these give you indication of how much the data is spread around the central measure which is the mean or the mode or the median as the case may be.

So, let us take an example of the 20 cherry trees we have computed the variance from the given data. And we find out that its actually 70.55132 and if we take the standard deviation; we will find its 8.4. As I told you that I had used a standard deviation of 10 to generate this data from a normal distribution. And we find that the sample standard deviation is a reasonably good measure of the population parameter which we did which was unknown.

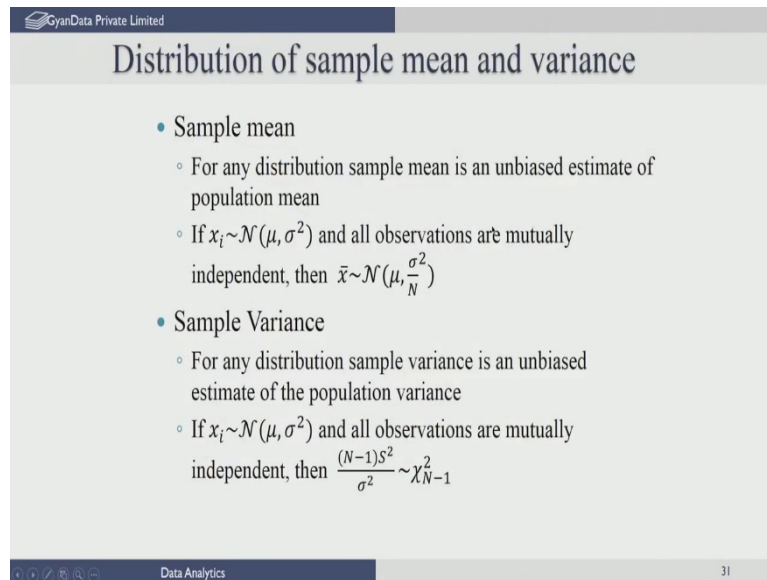
On the other hand, if I add outlier of 50 units to the first data point and recompute s^2 and s , it turns out s^2 turns out to be 212 and you can see if I take the square root it might be around 14, which is significantly deviating from the population parameter 10. So, a single outlier can cause the standard deviation and the variance to become very poor and therefore cannot be trusted as a good estimate of the population standard deviation or variance.

On the other hand, let us look at the mean absolute deviation. In this case I have, if we do not have an outlier, we get a mean absolute deviation of 6.9, which is not too bad compared to 10. The moment you have an outlier, the mean absolute deviation shifts to 9.5, it comes closer that is not what is important, but it does not change much just because of the presence of the outlier. So, this is a much more robust measure. In fact, if you take the mean absolute deviation from the median, it would be even better in terms of robustness with respect to the outlier. The range of the data can be obtained as the maximum and minimum value and I have just simply reported it.

So, these are measures of spread. So, even if I do not give you the entire 20 data points and I tell you the mean is, let us say 69 and the standard deviation is 8.5, then you can say that the data will spread typically between 69 ± 2 times the standard deviation which is 16. So, the lowest value will be about 53 and the highest value will be about 85 and it turns out if you look at the highest and maximum value and that is what it turns out to be.

So, + or - 2 times the standard deviation from the mean would represent about 95 percent of the data points if the distribution is normal. For other distributions you can derive these kind of intervals if you wish, but just giving two numbers allows me to tell you some properties of the sample and that is the power of these sample statistics.

(Refer Slide Time: 19:16)



Distribution of sample mean and variance

- Sample mean
 - For any distribution sample mean is an unbiased estimate of population mean
 - If $x_i \sim \mathcal{N}(\mu, \sigma^2)$ and all observations are mutually independent, then $\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$
- Sample Variance
 - For any distribution sample variance is an unbiased estimate of the population variance
 - If $x_i \sim \mathcal{N}(\mu, \sigma^2)$ and all observations are mutually independent, then $\frac{(N-1)S^2}{\sigma^2} \sim \chi_{N-1}^2$

Now, there are some important properties of the sample mean and variance which we will use in hypothesis testing. So, I want to recap some of these. If you have observations drawn from the normal distribution with some population parameter μ and population variance σ squared. And let us say you draw N capital N observations for all from this distribution; let us assume these draws or samples that you are drawn are independent, it does not have a bias in any manner.

And if you compute the sample average from this set of samples independent samples, then you can prove that \bar{x} is also normally distributed with the same mean population mean μ . Which means the expected value of \bar{x} is μ as I told you before and the expected variance of \bar{x} however, is σ^2 by N .

So, the variance of \bar{x} is actually lower than the variance of the individual observations. The important point here to be noted is, if I have N repeats from the same distribution and if I take the average of them, the average will be less noisy than the original observations.

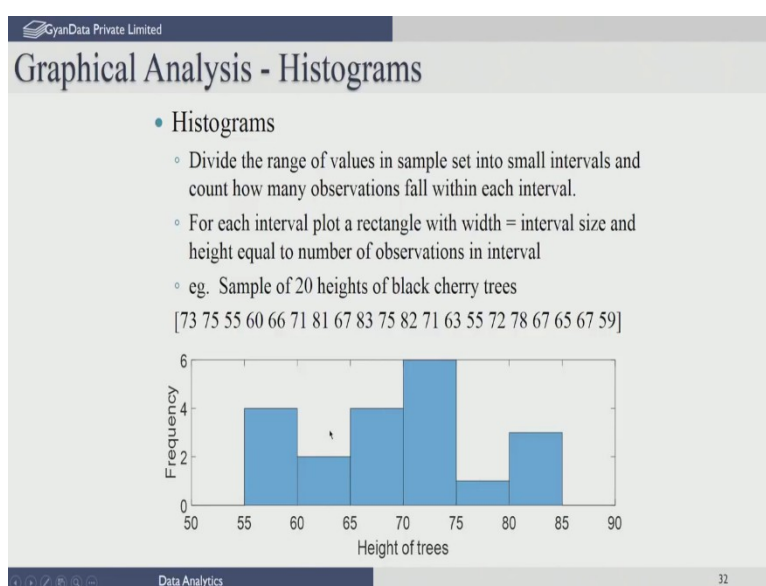
So, one simple way of dealing with noise and reducing the noise content in observations is to take n observations at the same experimental condition and average them. The average will contain less variability or less noise and it will reduce the variance of this

average will be 1 by N times the variance in your individual observations. So, what we call the noise will be reduced by square root of N where N is the number of samples.

Now, if you look at the sample variance and want to characterize the distribution of the sample variance, we can show again if you draw samples from the normal distribution with some mean μ and variance σ^2 and these observations I am going to assume are mutually independent.

Then if you take N - 1 times the sample variance divided by the population variance, we can show that this particular measure is a χ squared distribution with N - 1 degrees of freedom. We already saw the χ squared is a distribution of a random variable which varies between 0 and ∞ . And that distribution can be used to characterize s squared and we can later on do hypothesis testing, whether the σ^2 is some value and so on, using these distributions we will see.

(Refer Slide Time: 22:08)



Now, those are numerical methods of actually doing analysis of a sample data. What we want to do is also graphical analysis this is something that you should always do when you are given a data set. The first and foremost that you should do is to do some plotting to get a visual appeal because the mind is capable of inferring things what numbers do not tell you.

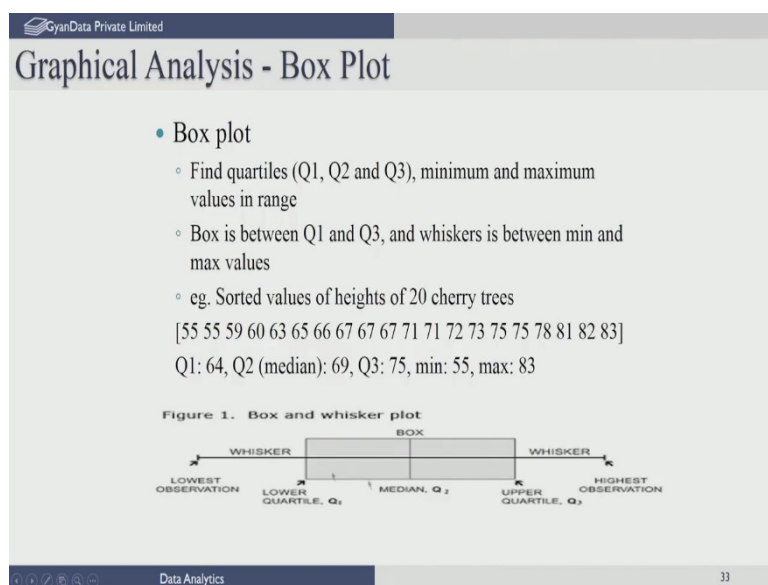
So, my suggestion is always when you have a data set, if you can plot and visualize it please do. So, let us see some of the standard plots again. Some of it you might have already encountered in your high school days. We will start with what is known as the histogram. Here I am given a sample set and what we do is first divide this sample set into small ranges; we define a small range and count how many

observations fall within that range or within each interval. And then we plot the width of the interval or the interval size of the x axis and the number of data points we see in that interval as the y axis, we call it the frequency and that is on the y axis.

So, let us take this example of the cherry trees. We have 20 data points. What I did was divided into small intervals of 5 feet which means I asked what are the number of cherry tree heights which are falling in the range 50 to 55, 55 to 60, 60 to 65 and so on, so forth. And I find between 50 and 55 there are no trees within that height, we find 4 trees with the height between 55 and 60 which we can easily see there is one data point here, there is second data point here, there is a third data point here and fourth data point is 60; so, the edge.

So, the 4 data points lying between 55 and 60 and similarly we find there are two data points lying just above 60 and up to 65 and so on, so forth. And that is what we plotted as a rectangle for each interval and this is known as a histogram. In fact, if I take 100 such examples and I plot, you will find this standard bell shaped curve. And that is because I drew these samples from the normal distribution. In this case because it is 20, you are not able to clearly see its bell shaped. You can see that the most of the data points are clustered around the middle point which is around 70 and you can see highest number 6 there. So, at least that is borne out.

(Refer Slide Time: 28:48)



You have other kinds of plots. One is called the box plot, which is used most often in sometimes in visualizing stock prices. Here you will compute quantities called quartiles Q1, Q2 and Q3 and the minimum and maximum values in the range. What are quartiles? Quartiles are

basically an extension of the idea of median. Q2 is exactly the median which means half the number of points fall below the value of Q2 and half the number of points are exactly about Q2.

Similarly, Q1 represents the 25 percent value which means 25 percent of the observations fall below Q1. 75 percent above Q1 and Q3 implies that 75 percent of the data points fall below Q3 and 25 percent above Q3. And once you have these values, the median, the quartiles and the minimum maximum, you can plot what is called the box and whisker plot in the box the median is the center value and the lower quartile Q1 and the upper quartile is also plotted.

And the box is drawn between Q1 and Q3 clearly showing where the median is. In this case its shown as symmetric, but generally need not be, either Q2 might be closer to Q1 or Q3. We also show the minimum and maximum values; here in this case the lowest observation the highest observation and those are called the whiskers. This gives you an idea better idea of the spread of the data than just giving you standard deviation or the mean absolute deviation and so on. This gives you a little more information about the spread of the data.

So, as an example if you take the 20 cherry trees and sort them out, sort it from the lowest to highest value, and we try to compute the median it turns out the median is the average of the tenth and eleventh point which is 69. Then the quartile one can actually be computed by just taking the first half which is the 10 points and computing the median of the first 10 points which turns out to be 64. And the Q3 can be computed as the median of the other half from 60 from 71 to 83 and that turns out to be around 75. So, the min and max of course, is 50 and 83 and therefore, you can perform this plot.

(Refer Slide Time: 27:21)

Graphical Analysis – Probability Plot

- Probability plot (p-p or q-q plot)
 - Determine different quantile values from sample set. Plot computed quantiles vs theoretical quantile values from chosen distribution
 - Same example (standardized and sorted values)

[-1.697 -1.697 -1.1016 -0.7443 -0.5061 -0.3870 -0.2679
-0.2679 -0.2679 0.2084 0.2084 0.3275 0.4466 0.6848 0.6848
1.0420 1.3993 1.5184 1.6374]

Normal Probability Plot

Observed Quantiles

Normal Quantiles

Normal Probability Plot

The third kind of plot which is very useful is to know about the distribution of the data and this is called the probability plot the p-p plot or the q-q plot. Here instead of determining just Q1, Q2, Q3 you compute several quantiles. And then plot these quantiles against the distribution which you think this data might follow. And if the data falls on the 45 degree line, then you can conclude that the sample data has been drawn from the appropriate distribution you are testing it against.

So, this is useful for visually figuring out from which distribution did the data come from. So, I have taken this example of these 20 cherry trees. I first standardized them, standardization means we remove the mean and divide by the standard deviation and we get the values. The 20 values as these are called the standardized values I have sorted them from the lowest to highest.

Now if you look at the 10 percent quantile, I can say that out of 20 points two points first two points fall below 1.679 - 1.679. So, - 1.679 represents the 10 percent quantile similarly - 1.1016 represents to 20 percent quantile and so on, so forth. For example, the 50 percent quantile or the median quantile, in this case the standardized measure, will be and which is around these two points around let us say - point or around 0 close to 0.

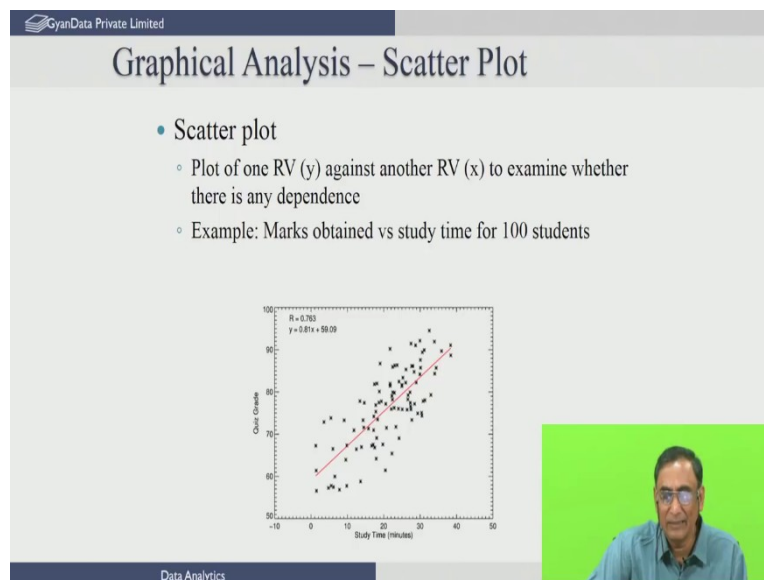
Notice that for a normal distribution, 50 percent of the data will lie below 0 and that is what this also seem to indicate. Now if you go to the standard normal distribution and try to compute the value below which the probability is 0.1 which is the lower tail probability we

talked about. Then you will find the value is around let us say - 1.7-1.5 whatever that value turns out to be. So, that is the 10 percent quantile. Similarly you ask what is the value below which 20 percent of the data lies or what is the value below which 20 percent of a standard normal distribution values will have a probability of area under the curve of 0.2 and that value you take that is the second 20 percent quantile and so on, so forth.

Then you plot the actual value obtained for the sample which is - 1.67 against the standard normal quantile and that is what is called the probability plot. So now, if this data has been drawn from the normal distribution then you should find a curve like this. I did not plug the normal probability plot for these 20 points, but for some other set of data. But typically if you find if you think that this data comes from the normal distribution, then you will find that in the normal probability plot the data will align itself on the 45 degree line and then you can conclude yes that the data has come from the distribution.

You can test this against any distribution. In this case, I have shown you how to test it against the normal distribution. You can take the quantiles from a uniform distribution or from the χ squared distribution what have you and the plot these sample quantiles against the expected population quantiles. And if they fall on the 45 degree line then you know that it comes from the appropriate distribution.

(Refer Slide Time: 30:57)



So, this is useful for determining visually the distribution from which the data have been drawn. Now the last kind of plot which is useful in data analysis is what is called the scatter plot. The scatter plot plots one random variable against another. So, if you have two random

variables, let us say y and x and I want to know whether there is any relationship between y and x , then one way of visually verifying this dependency or interdependency is to plot y versus x .

So, in this case we have taken some data corresponding to 100 students for which I mean students have spent time preparing for a quiz and they have obtained marks in that quiz. So, you should find typically if you spend more time study you should obtain typically more marks. And that is what this is trying to show on the x axis is a time in minutes that we have plotted. And the y axis we have plotted the marks obtained by the student and you can see there is an alignment.

The marks obtained seem to be dependent on the time spent and in fact, in this particular case you find that it looks like a linear dependency. So, you can plot a line approximate line through these data points we will show how to fit such lines using regression and how to obtain the parameters of this line that we have actually indicated here.

But more importantly if the random variable y , in this case the quiz marks has a dependency on the study time, then you will see an alignment of the data. On the other hand if there is no dependency you will find a random spread, this data will be spread all around with no clear pattern, in which case you will say that there are these two variables are more or less independent and you do not have to discover a relationship between these variables.

So, this is a plot which we will do in order to assess dependency between two variables and then proceed for further for analysis. In the next lecture we will take you through some decision making using hypothesis testing and how to perform estimation of parameters using sample data. See you in the next lecture.