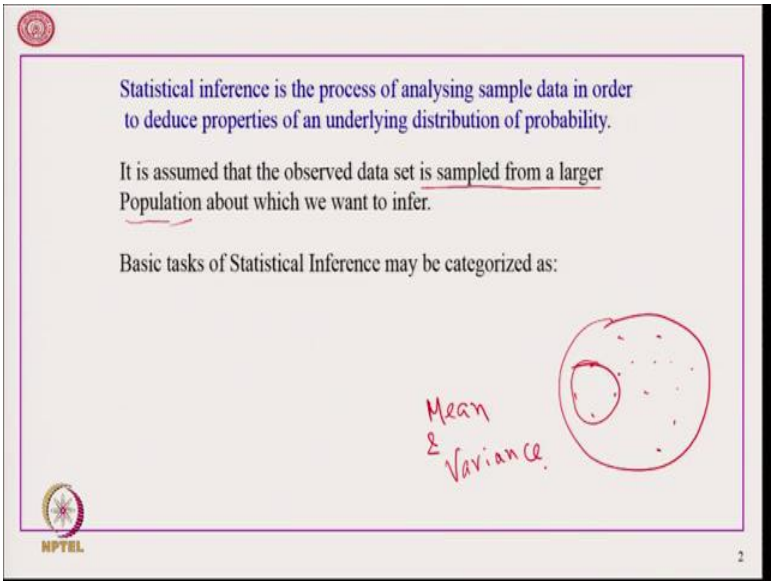
Nonparametric Statistical Inference
Professor Niladri Chatterjee
Department of Mathematics
Indian Institute of Technology, Delhi
Lecture 1
Nonparametric Statistical Inference

Welcome students to the MOOC's lecture series on Nonparametric Statistical Inference, this is lecture number 1. In this lecture I shall first explain to you what Statistical Inference is. Basically there are two types of statistical inference; one is parametric and the other is nonparametric.

I assume that most of you know at least the basics of parametric inference therefore in this lecture I shall not discuss much of parametric. Mostly my focus will be on nonparametric statistical inference. However, I will often draw comparison between parametric and nonparametric so that you will understand where these techniques stand in comparison with your background knowledge of parametric statistical inference.

(Refer Slide Time: 01:25)



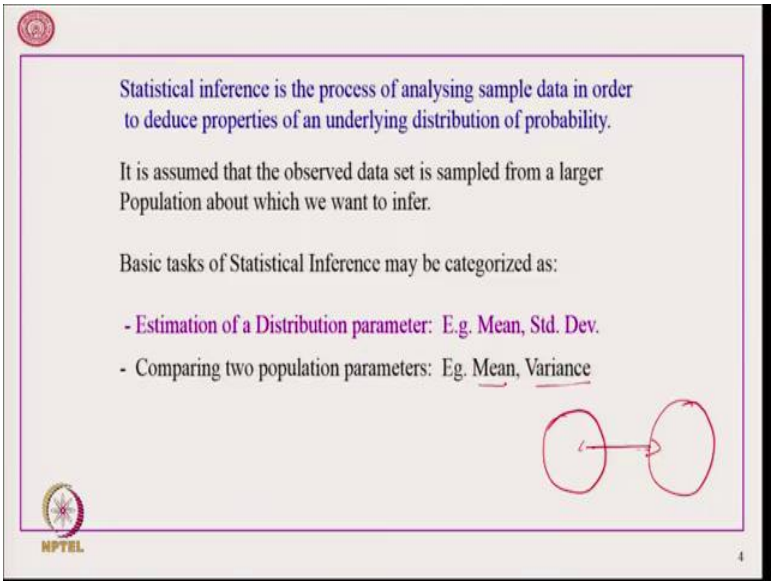
So let us first understand what statistical inference is, this is the process of analyzing a sample data in order to deduce properties of an underlying distribution of probability, say for example, here is a population; there are individuals, the population may be very large

and our aim is to understand certain properties of the population, most important ones are the mean and variance.

We cannot study the entire population if it is very large, therefore what we do?

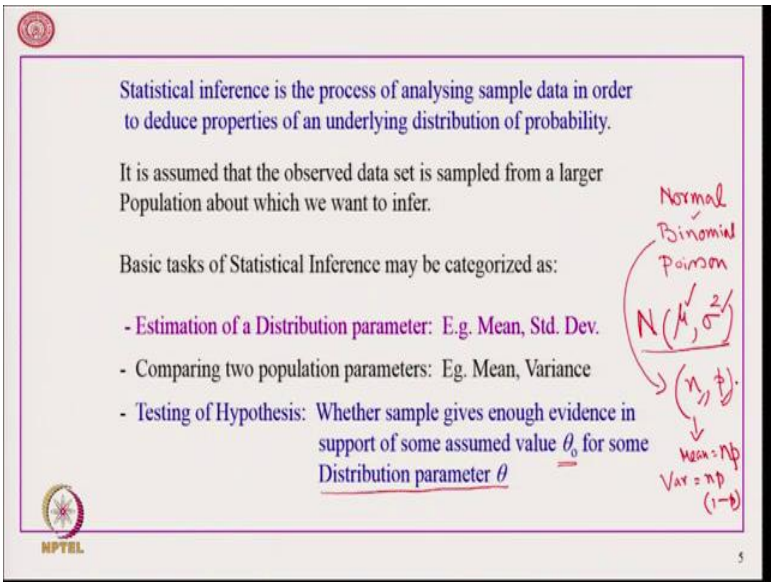
We sample some data from the larger population, for which we want to infer, therefore what is being done, suppose there is a huge set of data, we take a small sample typically randomly chosen and based on that our aim is to infer about this population or sometimes we shall see that we compare two populations.

(Refer Slide Time: 02:58)



Therefore the basic tasks of statistical inference can be said, one is that estimation of distribution parameter; it can be comparing two population parameters namely mean and variance, say for example, here is a group of students and here is another group of students, we want to compare whether they have the same mean or they have the same variance or if they are not same, then what is the relationship between them, we try to get a feeling about that.

(Refer Slide Time: 03:37)

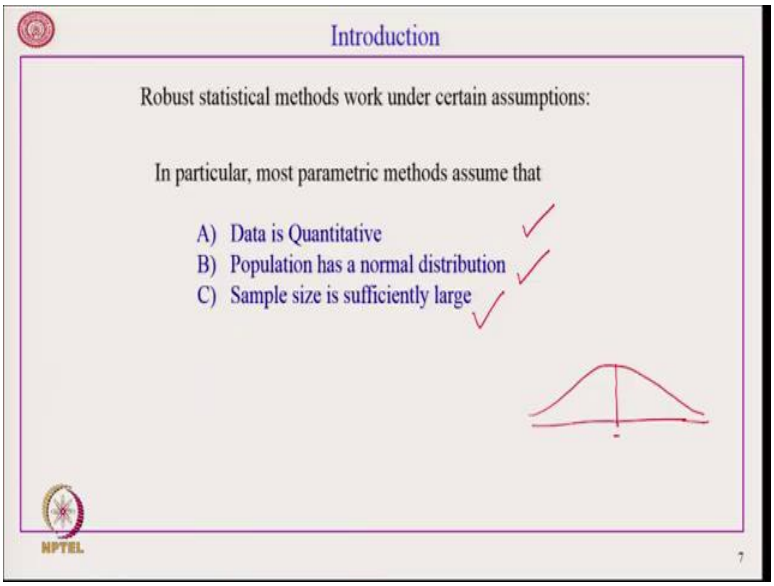


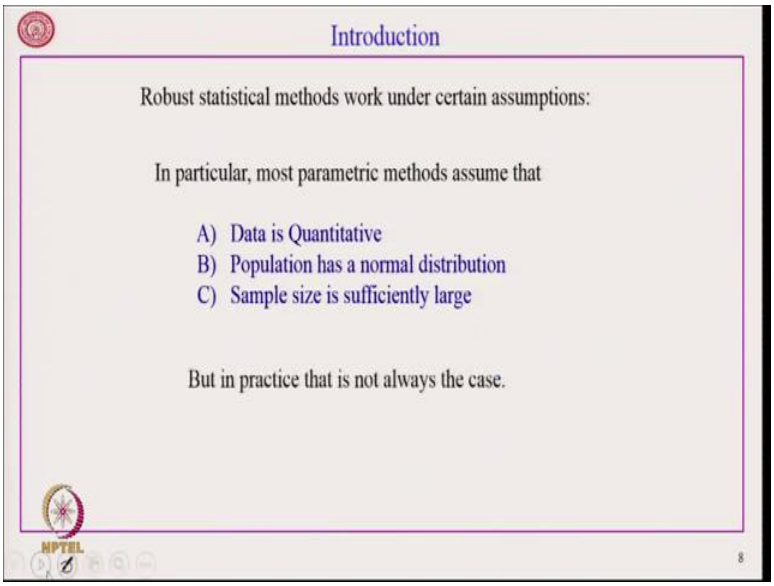
Other thing is testing of hypothesis, which means that whether a sample gives enough evidence in support of some assumed value θ_0 for some distribution parameter θ. Now, many of you may not know what I am mean by distribution parameter. All of us I hope are familiar with certain distributions, for example normal, binomial, Poisson etcetera.

Whenever we talk about a certain distribution, we associate with them certain parameters for example, normal has two parameters μ and σ^2, binomial has two parameters, the number of trials n and the probability of success p. Now, these are important parameters for a distribution because if you look at them carefully you will see that the mean and variance come from the distribution parameters.

For example, for normal, the mean is μ and the variance is σ^2 and they are involved in the pdf of normal distribution as parameters. For binomial, the mean is np and variance is equal to $np(1-p)$, again you see that we can derive them from the parameters given corresponding to a binomial distribution. Hence, if we estimate the parameters of a distribution then we can get a feel of the overall population.

(Refer Slide Time: 05:54)





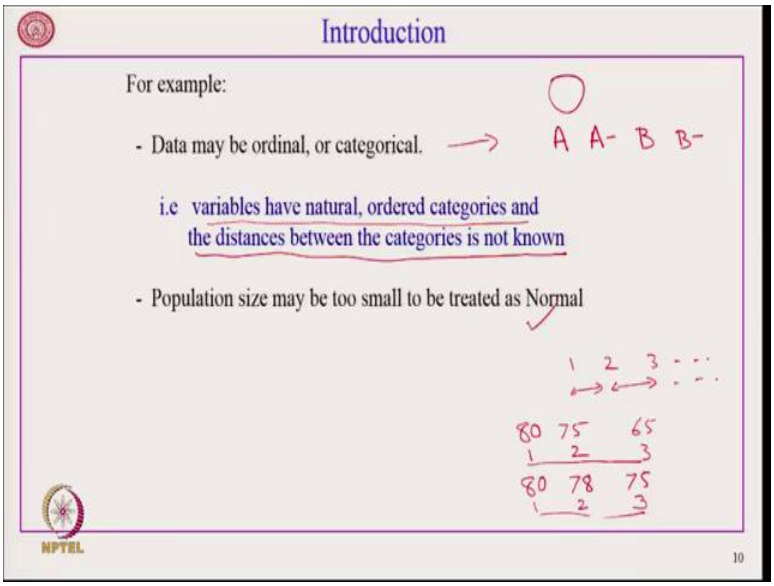
Now, robust statistical methods work under certain assumptions, by robust we mean typically the statistical methods that we use, namely, say many of you know about Z test, T test, chi square test, F test etcetera. These are some robust statistical methods but there are certain assumptions for them.

In particular most parametric methods assume that the data is quantitative, if the data is not quantitative the statistical methods that we know will not work. Population has a normal distribution; it is an underlying assumption that the population with respect to certain

properties will have a normal distribution that is, it is central around the mean and it is more or less symmetrically spread around the mean. The sample size is sufficiently large; that means in order to infer for the population we need quite a large sample size.

But in practice it always does not happen.

(Refer Slide Time: 07:17)



For example, data may be ordinal or categorical for example, when we talk about marks of the students we may give them A, A-, B, B-, etcetera., in our IIT system.

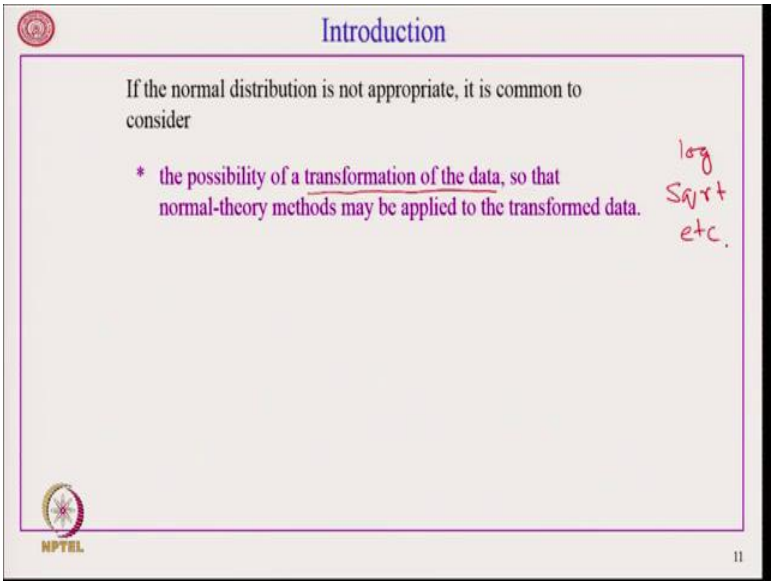
Therefore in some sense, all students getting a grade *A* are somewhat equivalent but they do not have the same marks, therefore the data is no more numeric and therefore we cannot process it using the standard statistical techniques.

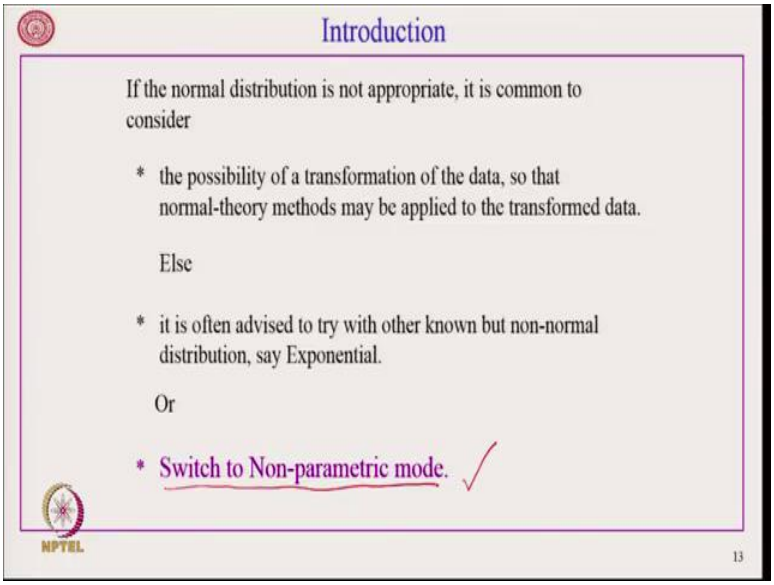
Data may be ordinal that means we are ranking them as first, second, third but we do not know the distance between the second and third and the distance between the first and second etcetera.

And that means that, here I am saying that variables have natural ordered categories and the distances between the categories is not known. I hope you understand that. Suppose the first boy got 80, second one got 75 and third one got 65. So I will give them the rank 1, 2, 3 but if the first one is 80, second one is 78 and third one is 75, again they get the same

rank 1, 2, 3. Therefore, when I have the data in this ordinal form I lose a lot of information. Also, another problem is that the population size may be too small to be treated as normal.

(Refer Slide Time: 09:12)

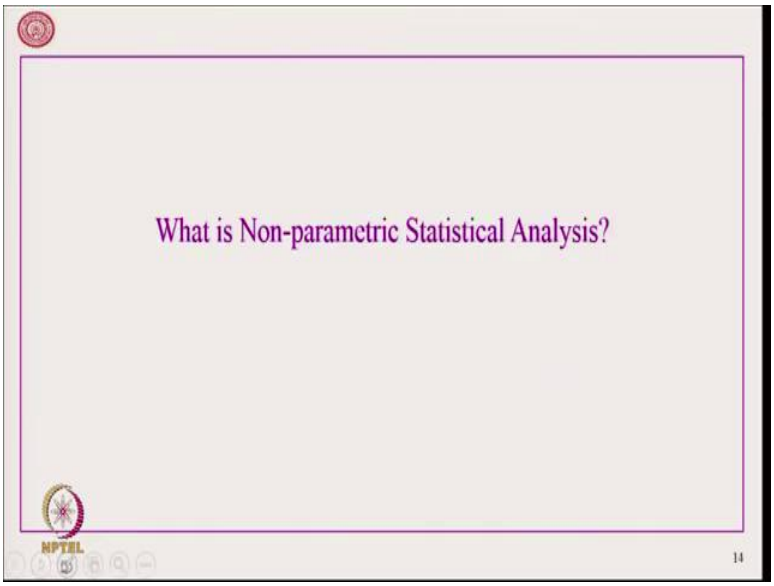




Therefore the question comes if the normality assumption is not there then what we should do? There is one common technique of transformation of data, in transformation what we do? If we take a function of the data then we may get the normality assumption, some standard transformations are log, square root, etcetera.

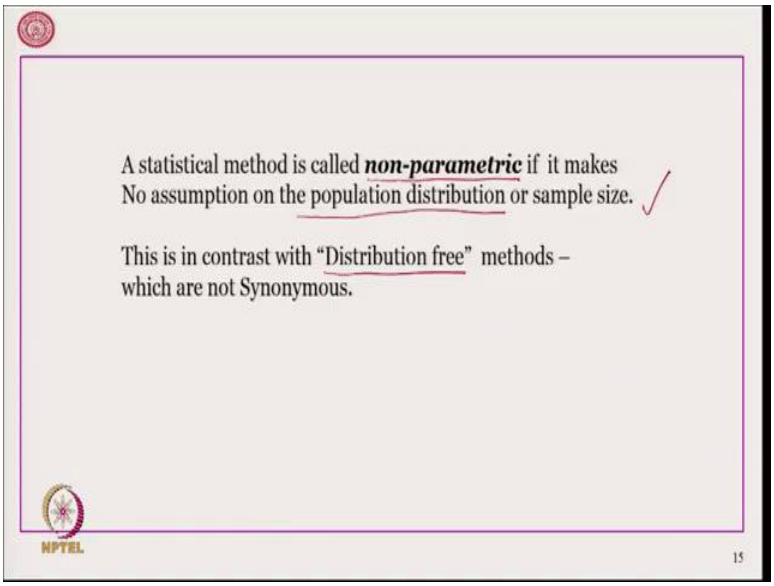
Else what we can do? Else it is often advised to try with other known but non-normal distribution say exponential distribution, that is another important thing that can be done or the third option is that we go to nonparametric mode, that means now we are switching over to a nonparametric analysis of the data and this is the focus of this series of lectures that I am going to present to you.

(Refer Slide Time: 10:29)



Therefore the basic question comes, what is nonparametric statistical analysis?

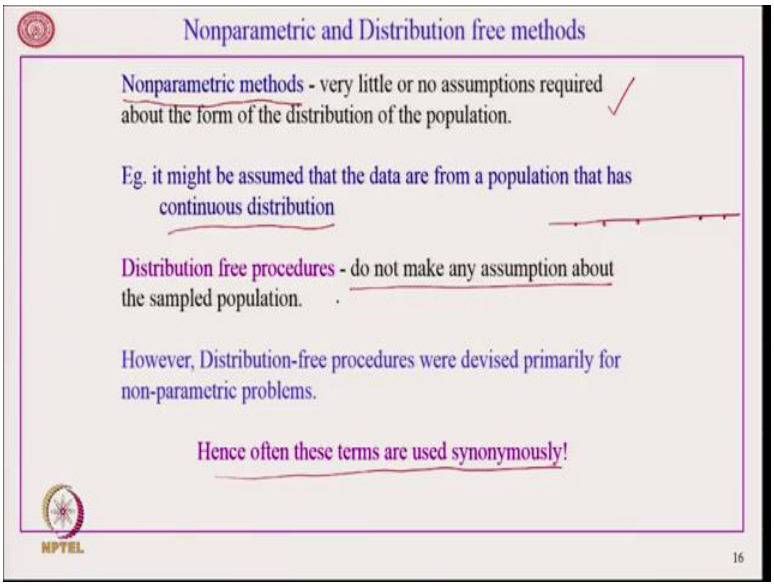
(Refer Slide Time: 10:37)



A statistical method is called nonparametric if it makes no assumption on the population distribution or the sample size. Therefore we are not having any restriction on sample size. You will see later that there are statistical tests on nonparametric way where sample size may be 2, 3 but still we can work with that small size data.

And generally there is no assumption on the population distribution but this is not quite correct, sometimes we may have a very general assumption as I am explaining later but this is in contrast with distribution free, so these two terms distribution free and nonparametric are not quite synonymous, they have a subtle difference.

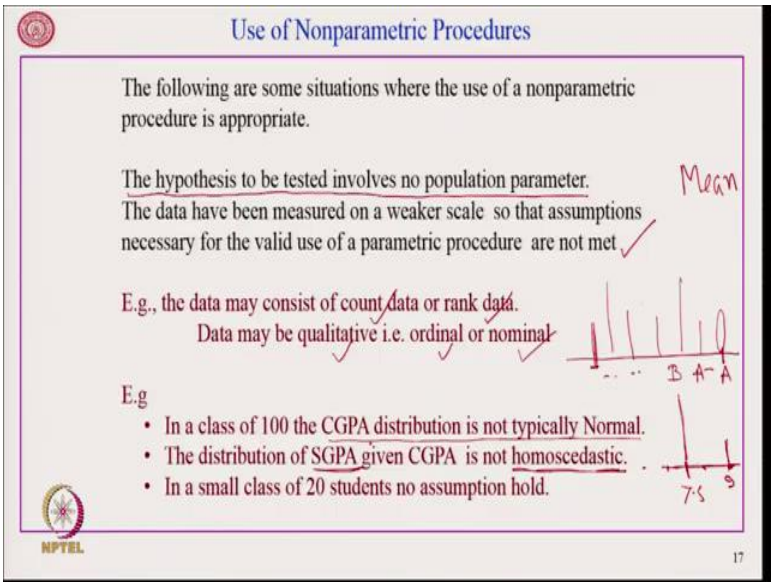
(Refer Slide Time: 11:34)



As I said here earlier that nonparametric methods have very little or no assumptions required but sometimes we shall assume that it is from a population with a continuous distribution, so that is the only assumption, we are not talking about anything else, what does it mean that the sampling that we are taking that is coming from a continuous set of data but the samples that we have taken as you can understand they may be discrete.

Distribution free properties do not make any assumption about the sampled population, so absolutely no assumption is there, so there is a subtle difference. However distribution free procedures were devised primarily for nonparametric problems, therefore often these two terms are used synonymously. But for this class we shall stick to nonparametric methods and we are not going to call them distribution free methods which you may often find in some literature.

(Refer Slide Time: 12:50)



So, question comes where do we use non-parametric methods? Very valid question, so the following are some situations where the use of a nonparametric procedure is appropriate. The hypothesis to be tested involves no population parameter, as we have discussed sometime back, that when we are talking about mean of a distribution this actually can be found from the parameter of the underlying distribution or underlying pdf or pmf like say normal distribution function or binomial distribution, Poisson distribution, etcetera.

So, that is first thing, second thing is that the data have been measured on a weaker scale so that assumptions necessary for the valid use of parametric procedure are not met. For example, the data may consist of count data and rank data, data may be qualitative, ordinal or nominal, I have in some sense explained this thing, so I am not going into details. We will come across them in course of time but let me give you an example.

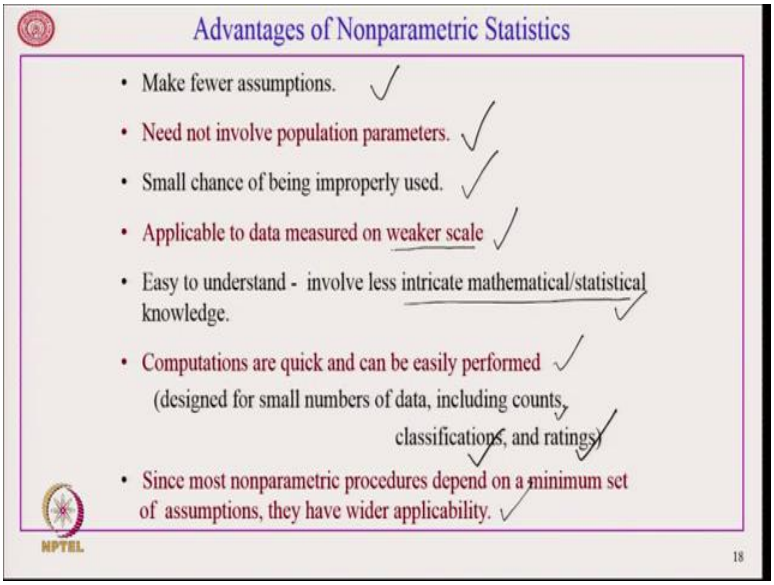
In a class of 100, the CGPA distribution is not typically normal. If we look at the distribution of marks in a class of say 100 we will find that it is not normal, there will be a set of boys who are very good. Thus at the right end of the data there will be high frequency, similarly somewhere near the pass and fail boundary there will be high frequency.

In between it is not necessary to be symmetric, we have often seen distribution like this with respect to different grades like A, A-, B, etcetera. The distribution of SGPA given

CGPA is not homoscedastic, that means even if we assume that the CGPA is given then it is not mandatory that the semester grade point average will have equal variance along the or for the same CGPA.

Say for example, when the CGPA is something like say 9 out of 10 we will find very low variance among the SGPA. But if the CGPA is something like say 7.5 we will find a wide range of variation in the SGPA or the semi-structured point average. So homoscedasticity means that the variance will be same with respect to different points for SGPA but that is not a very valid assumption with respect to the CGPA distribution and if we consider a small class of 20 students then such assumptions do not hold.

(Refer Slide Time: 16:30)



Therefore, the question comes what are the advantages of nonparametric statistics? It makes fewer assumptions in comparison with the robust statistical procedures, it need not involve population parameters, we are not bothered about the distribution parameters as I said that we are not even assuming any distribution for the data which has come from the population.

Small chance of being improperly used because nonparametric methods have been developed for very-very specific purposes, so it is very unlikely that you are going to use it in a wrong way. Applicable to data measured on weaker scale as I have been telling that

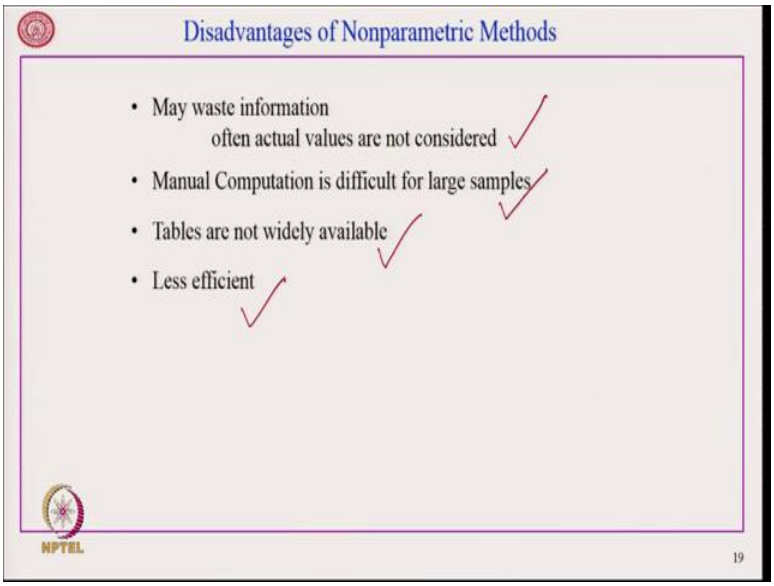
marks are a much stronger scale in comparison with grades because it merges several marks into the same grade, therefore that is a weaker scale.

It is easy to understand and involves less intricate mathematical statistical knowledge. We will be doing lot of mathematics in establishing or improving certain nonparametric results but we will see that the mathematics is pretty simple, although laborious, let me warn you at the very beginning that this is going to be little bit laborious but simple.

Computations are quick and can be easily performed because less mathematical details and design for small numbers of data including counts, classifications, and ratings. Since these are working on a small set of data, overall computational requirement is less and since most nonparametric procedures depend on a minimum set of assumptions they have wider applicability.

So they can work on various types of data in comparison with normal parametric procedures which work only under certain restricted cases.

(Refer Slide Time: 18:46)



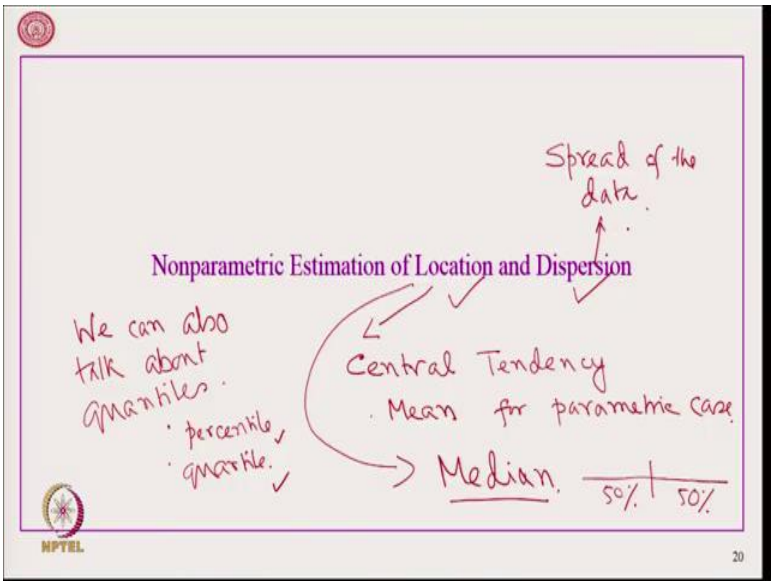
However there are certain disadvantages also, we may waste some information of an actual values are not considered. For example, given the marks we have converted them into ranks and that way we lose a lot of information.

Therefore that is one drawback. Manual computation is difficult for large samples because as I said the computation is laborious if the sample size is large then it takes quite some time to finish the computation. Tables are not widely available, say normal table, chi square table, t, f all these distribution tables, all those who are familiar with parametric inference know very well.

But similar tables are there with respect to many nonparametric techniques, but those tables are less freely available. As we go through the series of lectures I will refer to many different tables, I will give you the links and I shall also give examples of how those tables are to be read when you want to use them for your inferencing purpose and definitely these are less efficient.

If you can do something with a T test or Z test that is going to be much more efficient to achieve the same efficiency level with respect to nonparametric cases, you may need a prohibitively large sample size.

(Refer Slide Time: 20:34)



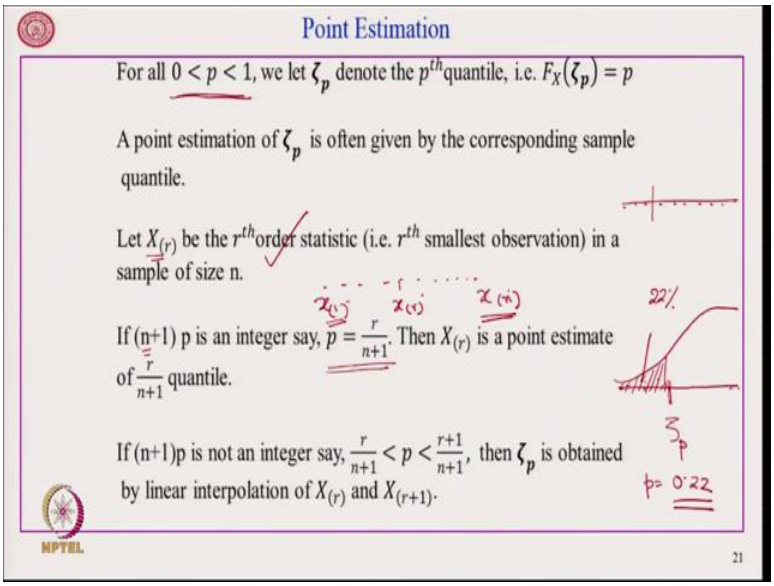
So let us look at the most common problem, nonparametric estimation of location and dispersion, so what do we mean by location? By location essentially we mean the central tendency, when it is numeric case we talk about mean but mean as I said has come as a distribution parameter and we are not assuming any distribution, so when we are talking about nonparametric for us this is going to be median.

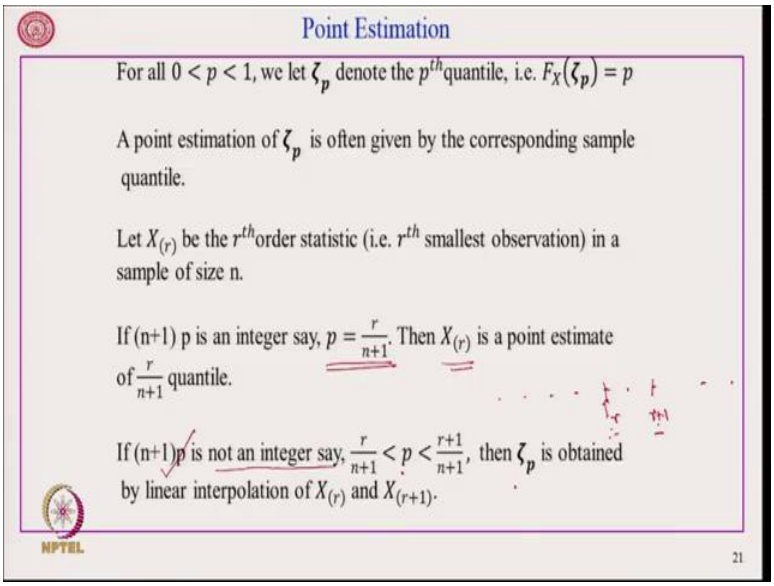
I hope all of you know what is median that means it is the value such that 50% observations are will be on left side of it that is lesser than that and 50% of the observations will be greater than this. Not only median we can also talk about quantiles that is, many of you know percentile, quartile, etcetera.

So we can even try to estimate this type of quantiles also there is no problem that we can do using nonparametric methods and dispersion, by dispersion basically we mean the spread of the data. With respect to parametric case we use variance or range, we shall see how we use the data given for measuring the spread. So, this is the basic background.

Let us go into some estimation first.

(Refer Slide Time: 23:04)





For all $0 < p < 1$, let ζ_p denote the p^{th} quantile that is $F_X(\zeta_p) = p$.

So suppose, this is a distribution, this is the data, and suppose this is the $F(x)$ which you know is the cumulative distribution function, so this point is called as ζ_p if the proportion of observations less than this is p.

Therefore if p is equal to say 0.22 basically what we try to identify the point such that the proportion of observations below that is going to be 22% and that is true for all p between

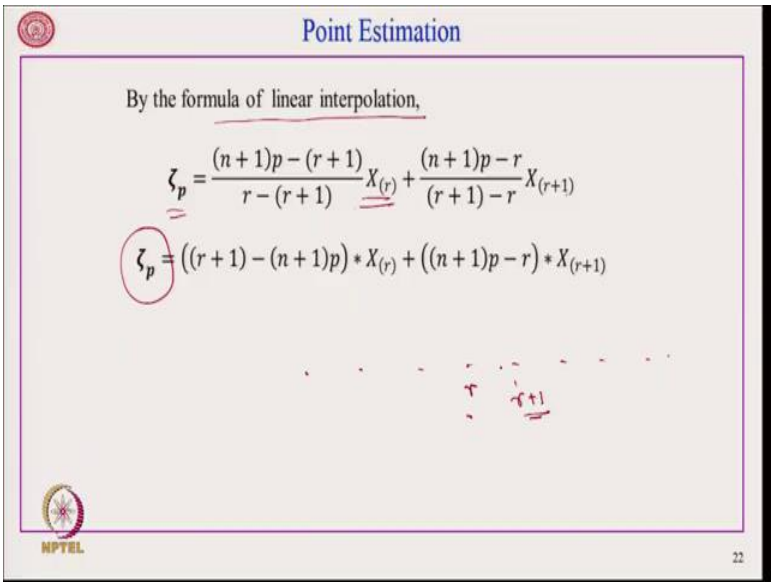
0 to 1, a point estimation of ζ_p is often given by the corresponding sample quantile, that means, if this is the sample that is given, we will try to see which values are there from which I can get the ζ_p.

So in order to do that what we shall have to use is the r^{th} order statistic, I hope all of you know what is an order statistic but basically this is that, suppose we have a sample then we order them in terms of magnitude, the smallest one is $X_{(1)}$ and the largest one is $X_{(n)}$, it is the n^{th} order statistic, it is the first order statistic and the r^{th} one in number that will be called the r^{th} order statistic.

Therefore this is a random variable, we put it with capital X when we are looking at small x this is basically the observation. Now n is the sample size, if $(n+1)p$ is an integer, that is $p = \frac{r}{n+1}$, then $X_{(r)}$ is the point estimate of $\frac{r}{n+1}$th quantile.

However if $(n+1)p$ is not an integer, that is the most likely situation, then p will be lying between two observations the r^{th} and $(r+1)^{th}$, say p lies between this is the r^{th} observation in terms of order and this is $(r+1)^{th}$ observation, then p is suppose lying somewhere here, then we have to do an linear interpolation of these two observed values to get the estimate of ζ_p.

(Refer Slide Time: 26:35)

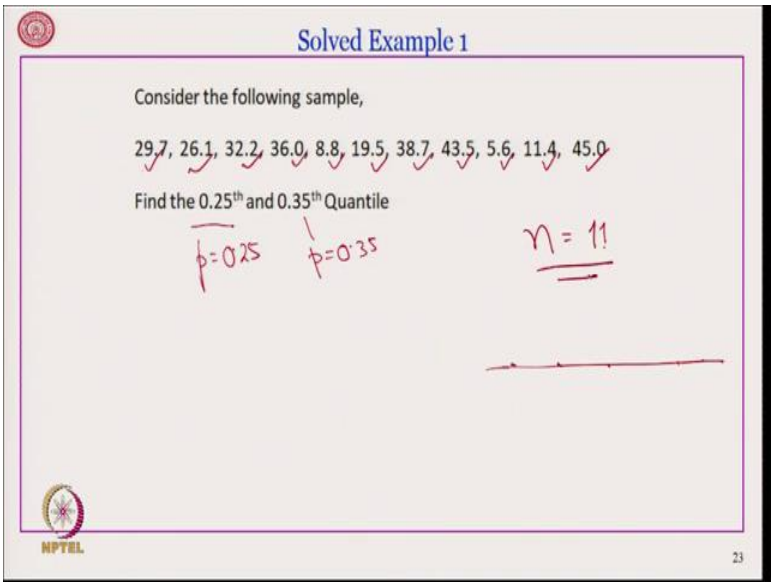


So if we use the linear interpolation then,

$$\zeta_p = \frac{(n+1)p-(r+1)}{r-(r+1)} X_{(r)} + \frac{(n+1)p-r}{(r+1)-r} X_{(r+1)}$$

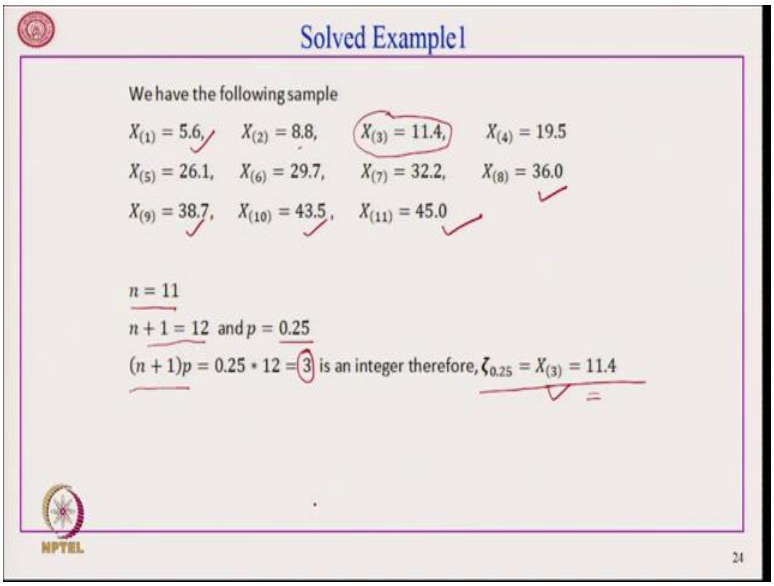
$X_{(r)}$ is the r^{th} order statistic, that is in the all the observations the one that is r^{th} smallest that is going to be called $X_{(r)}$ and this is therefore going to be $(r+1)^{th}$ smallest and I am going to interpolate between them and from there we are going to get a value for ζ_p. That is $\frac{r}{n+1} < p < \frac{r+1}{n+1}$.

(Refer Slide Time: 27:23)



So, let me give you an example, suppose I have observed this set of data, there are 11 information 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 what does it mean? That suppose this is a set of possible values we have considered 11 observations from this line we need to find out what is the 0.25 quantile that is $p = 0.25$ and what is the p^{th} quantile when $p = 0.35$, when $n = 11$ and these are the observations. So how will we proceed?

(Refer Slide Time: 28:17)

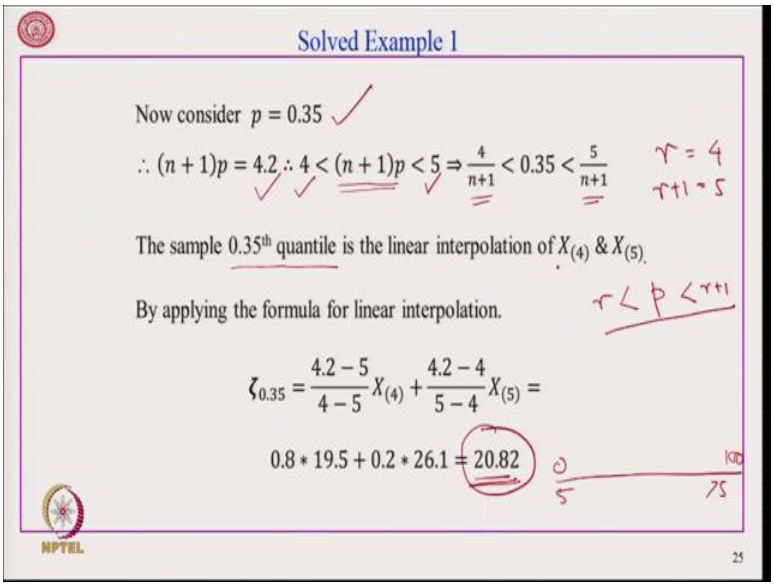


We first compute the order statistic, that means we have sorted the values from the smallest to the largest, the smallest observation has been 5.6, second smallest is 8.8 like that eighth one is 36, ninth one is 38.7, tenth one is 43.5. So this is the order statistics of the observed data, do not forget that that is the order in which we have sampled.

Sample might have come in this order as given there but when I sort the observations I get this sequence, so $n = 11$, therefore $n + 1 = 12$ and we are looking at the ζ_p, when $p = 0.25$, therefore $(n + 1)p = 3$, therefore $\boldsymbol{\zeta}_{0.25}$we can replace it with $X_{(3)}$, the third order statistic which is 11.4, what does it mean?

This inference says that the population is such that 25% of the observations are going to be less than equal to 11.4, very-very simplistic thing but that is how practically it works.

(Refer Slide Time: 29:50)

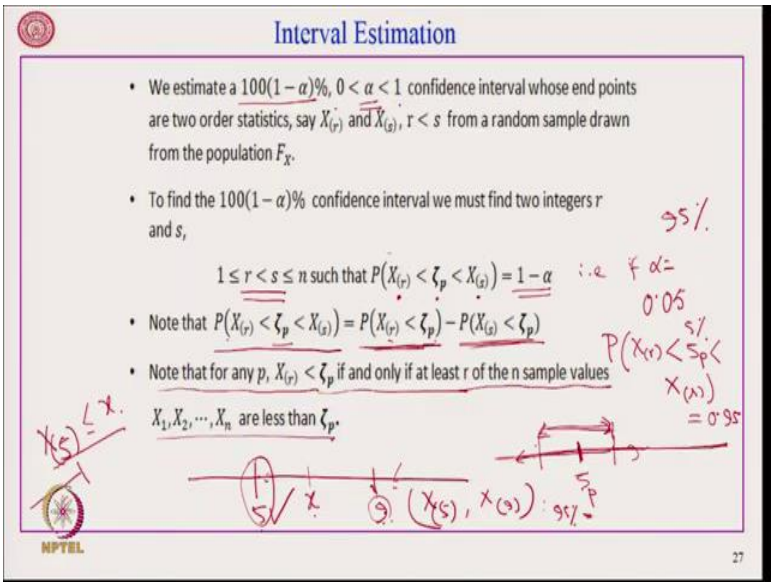


Now let us consider $p = 0.35$, therefore $(n+1)p = 12 * 0.35$ so which is coming out to be say 4.2 because $n + 1 = 12$ it is 4.2, therefore we see $(n+1)p$ lies between 4 and 5, which implies that $\frac{4}{n+1} < 0.35 < \frac{5}{n+1}$

Therefore we have got $r = 4$, $r + 1 = 5$, therefore p is lying between these two, therefore what we can understand that the 0.35 quantile is the linear interpolation of the fourth order of statistic and the fifth order statistic. Therefore using the interpolation we get that the value is coming out to be 20.82, what does it mean?

It means that the 35% of the observations are going to be less than equal to 20.82 therefore without knowing anything about the scale whether it is from 0 to 100 or say from 5 to 75, we get a feel of what is going to be the 35th quantile of that one. So that is the simplicity for nonparametric estimations but you should understand that it is the simplest type of inference that we are going to do in a nonparametric way.

(Refer Slide Time: 31:52)



Now question comes about interval estimation, say suppose this is the observed data, this is my ζ_p, this is done on the basis of the sample that we have but perhaps most of you will say that based on one sample how can we say that this is the actual p^{th} quantile which is very obvious. Therefore what we try to do, we try to build a confidence interval around the point that we have calculated such that we can be confident enough that ζ_p will lie within this interval.

Now this size of the interval will depend upon the level of confidence which you callα, therefore if alpha is equal to 0.05 that is 5 percent which is between 0 to 1, therefore effectively we are looking at $100(1-\alpha)$ % which is equal to 95% confidence interval, I hope most of you have the idea about this but still I am explaining it, what does it mean?

It means that although we have done it on the basis of one sample, 95% confident that the ζ_p, the p^{th} quantile with be will be within this interval and that chance is only 5 percent that actual ζ_p will lie outside this interval.

So this is what is called interval estimation, so when we are going for nonparametric we are looking for two values r and s such that the probability of $X_{(r)} < \boldsymbol{\zeta_p} < X_{(s)}$ is going to be $1-\alpha$, that is if $\alpha = 0.05$ we are looking for probability, that $X_{(r)} < \boldsymbol{\zeta_p} < X_{(s)}$ is equal to 0.95

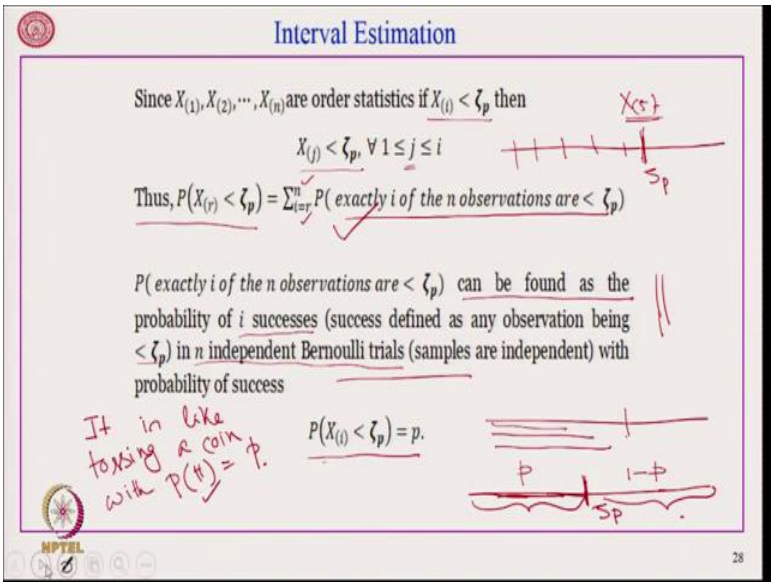
So our aim is to identify this r and s, that is we can say that, suppose r is 5 and s is say 9, so basically we are saying that $X_{(5)} < \boldsymbol{\zeta_p} < X_{(9)}$ with probability 95 percent or 0.95. Now how to obtain this? Note that probability, $P(X_{(r)} < \boldsymbol{\zeta_p} < X_{(s)})$ is equal to what is the probability that $X_{(r)} < \boldsymbol{\zeta_p}$ minus probability $X_{(s)} < \boldsymbol{\zeta_p}$, right.

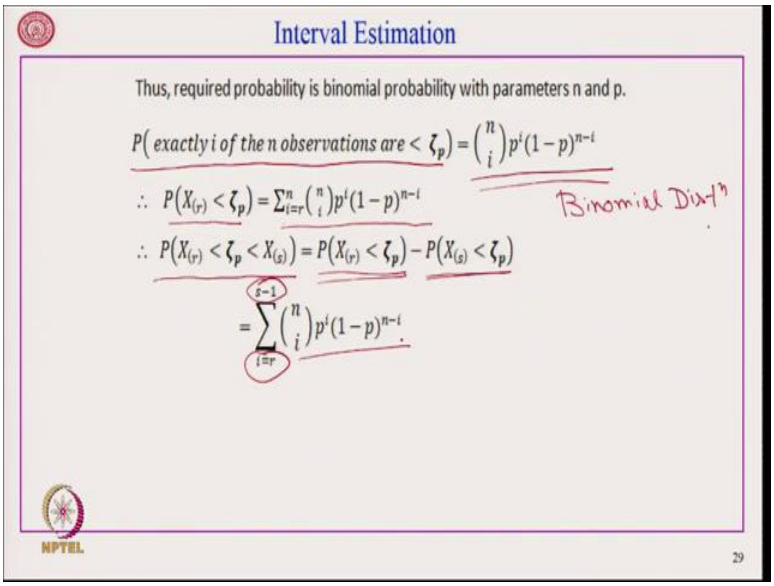
Why it is? So consider the same $r = 5$ and $s = 9$, therefore $X_{(5)} \leq x$ when first 5 observations are less than equal to x but it is also true for first 6 observations are less than equal to x, it is also true if 7 observations are less than equal to x.

Therefore probability $X_5 \leq x$ where $X_{(5)}$ is the fifth order statistic is greater than the probability that $X_{(9)} \leq x$ because whenever this happens automatically this is going to be less than equal to x but reverse is not true there can be a situation when fifth order statistic is less than equal to x but the ninth order statistic is more than x.

Therefore probability that $X_{(r)} < \boldsymbol{\zeta_p} < X_{(s)}$ will, what we will do? We shall subtract the probability $X_{(s)}$ is less than equal to ζ_p from probability $X_{(r)}$ is less than ζ_p. This is what is written here that for any p, if and only if at least r of the n sample values are less than equal to ζ_p.

(Refer Slide Time: 37:22)





Therefore since $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ are order statistics if $X_{(i)} < \boldsymbol{\zeta_p}$ then $X_{(j)} < \boldsymbol{\zeta_p}$, $\forall\, 1 \le j \le i$ that means suppose this is my ζ_p and I am saying that the fifth order statistic is less than equal to ζ_p that means that $X_{(4)}, X_{(3)}, X_{(2)}, X_{(1)}$ they are all less than ζ_p.

Therefore , $P(X_{(r)} < \boldsymbol{\zeta_p}) = \sum_{i=r}^{n} P(\text{exactly } i \text{ of the } n \text{ observations are} < \boldsymbol{\zeta_p})$ as I have explained earlier that $X_{(5)}$ the fifth order statistic is less than ζ_p, if below ζ_p there are five observations, six observations, seven observations in fact all the n observations are less than ζ_p, in all these cases $X_{(5)}$ is going to be less than ζ_p.

Therefore this is clear, therefore probability exactly i of the n observations are less than ζ_p can be found as the probability of i success, success defined as any observation being less than ζ_p in n independent Bernoulli trials with probability of success is $P(X_{(i)} < \boldsymbol{\zeta_p}) = p$, I hope you understood it but let me explain again.

So suppose this is my ζ_p, that means probability an observation is less than this value is equal to p and probability an observation is greater than this value is $1 - p$. Now, therefore when I am choosing a sample it has a probability p to remain on the left side of ζ_p and probability $1 - p$ to be on the right side of the ζ_p.

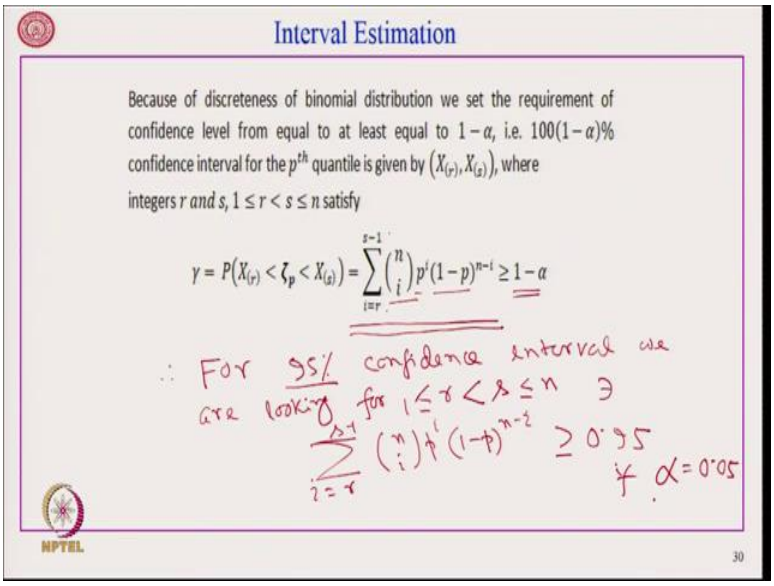
Therefore it is like tossing a coin with probability of head equal to p, right, because if we take a random sample there is a small p probability that it is on this side which we are calling it a success and if it is on this side it is going to be a failure and you know that in Bernoulli by head we mean the success, therefore we are trying to compare this situation with a Bernoulli tossing situation.

Thus required probability is a binomial probability that out of n tosses r of them are less than equal to p, therefore probability exactly i of the n observations are less than ζ_p that probability is coming out to be $\binom{n}{i}p^i(1-p)^{n-i}$, this is coming from binomial distribution, therefore probability $P(X_{(r)} < \boldsymbol{\zeta_p}) = \sum_{i=r}^{n}\binom{n}{i}p^i(1-p)^{n-i}$

Therefore,

$$P(X_{(r)} < \boldsymbol{\zeta_p} < X_{(s)}) = P(X_{(r)} < \boldsymbol{\zeta_p}) - P(X_{(s)} < \boldsymbol{\zeta_p}) = \sum_{i=r}^{s-1}\binom{n}{i}p^i(1-p)^{n-i}$$

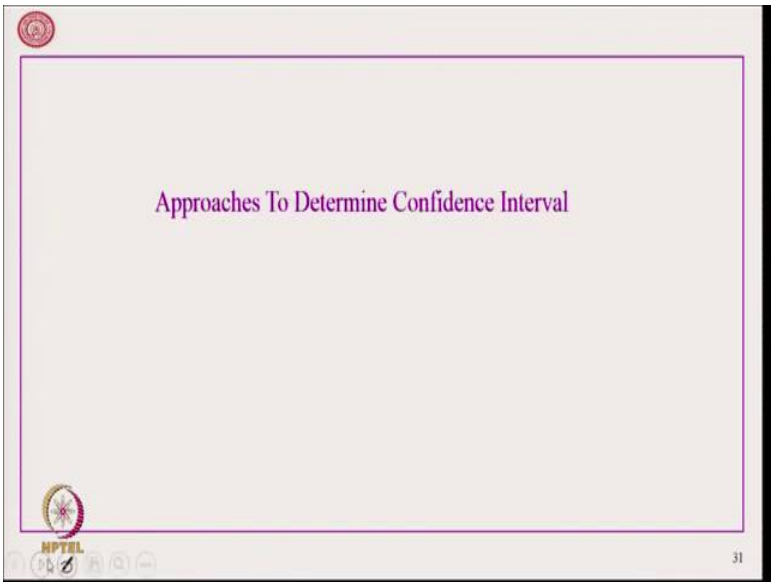
(Refer Slide Time: 41:53)



Now when we are going to use this for interval estimation, therefore if I want a 95% confidence level, therefore we will look at that I need a s and r such that this probability, $\sum_{i=r}^{s-1}\binom{n}{i}p^i(1-p)^{n-i} \geq 1-0.05 = 0.95$.

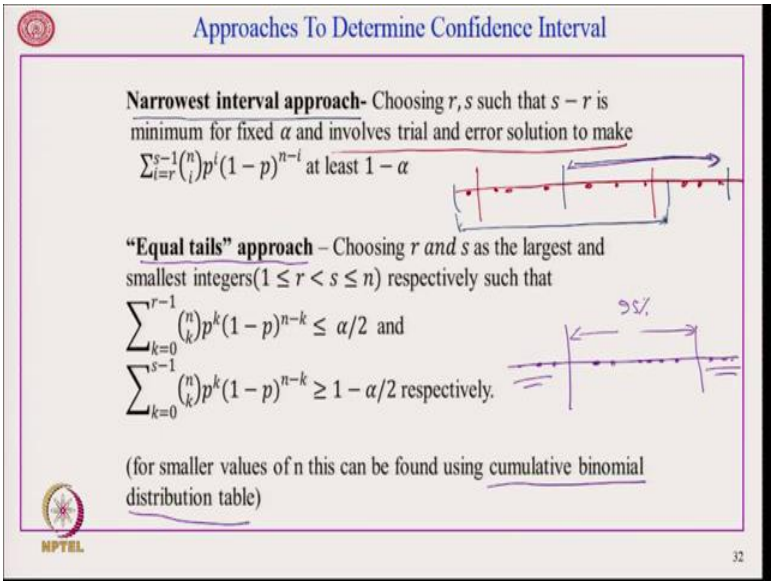
Therefore for 95% confidence interval we are looking for $1 \leq r < s \leq n$, such that $\sum_{i=r}^{s-1}\binom{n}{i}p^i(1-p)^{n-i} \geq 0.95$, for 95% confidence interval, if $\alpha = 0.05$. Similarly for 90 % confidence interval value of α is going to be 0.1.

(Refer Slide Time: 43:37)



Therefore what we do for determining the confidence interval?

(Refer Slide Time: 43:41)



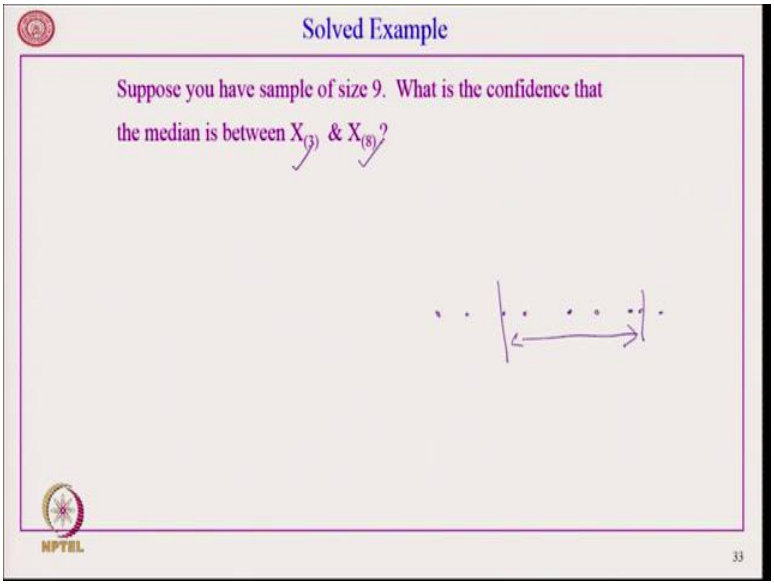
There are different ways of choosing. Narrowest interval approach: choosing r and s such that $s - r$ is the minimum for a fixed α and involves trial and error solution to make this summation $\sum_{i=r}^{s-1}\binom{n}{i}p^i(1-p)^{n-i}$ to be greater than equal to $1 - \alpha$. So what does it mean?

Suppose the values are along this line $X_{(1)}, X_{(2)}$, again I tell you these are order statistic, this is not the actually first, second, third sample. Suppose this is how the values are ordered I am looking at the 95 percent confidence level. Now it can be that between these there are 95% observations, similarly it can be that between this and this there are 95% observations or it can be that between this and this there are 95% observations.

Therefore question comes which r and s we should choose? The narrowest interval approach tells us that we will take that to be the 95% confidence interval which is of smallest length. So in this case most probably this is going to be the 95% confidence level, another way of looking at this is the following, this is called Equal Tails approach.

So if we have the data like this, then we will partition it in such a way that on both the sides there are same number of observations, so this is the 95 percent confidence interval we are going to consider. So these are the two basic approaches for considering the confidence interval. So this we obtain typically using cumulative binomial distribution table given n and p we will try to see how to get the confidence interval.

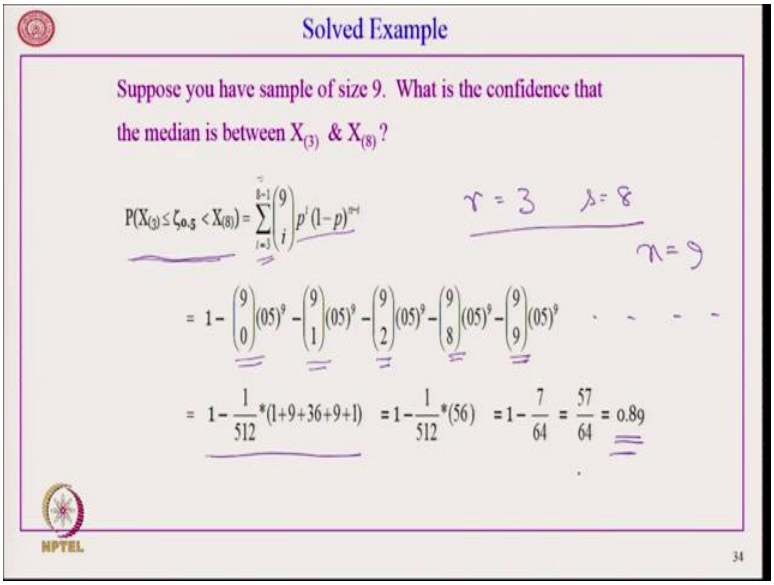
(Refer Slide Time: 46:21)



So let me solve one example, suppose you have sample of size 9 what is the confidence that the median is between $X_{(3)}$ and $X_{(8)}$? The third order statistic and the eighth order statistic, that means we have 9 samples and we are looking at that the median will lie in

this interval, what is the confidence? That is what we are trying to figure out. So, we go as follows.

(Refer Slide Time: 46:57)



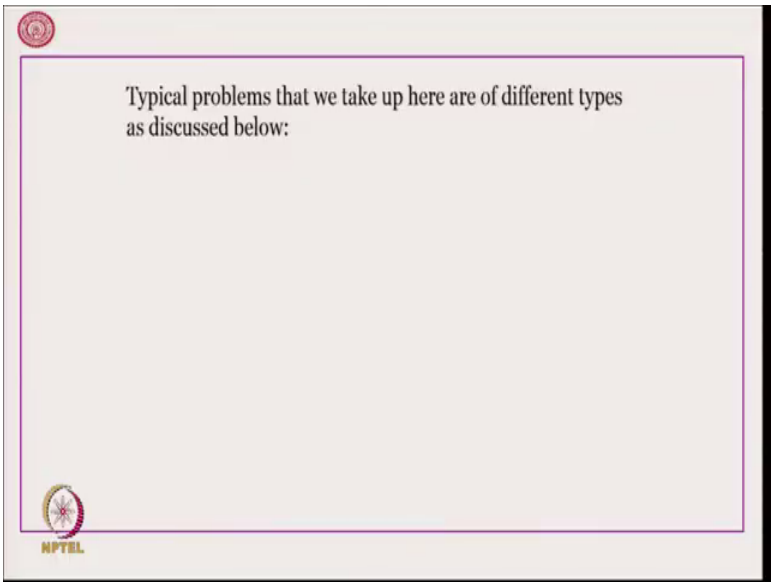
$$P\left(X_{(3)} < \boldsymbol{\zeta_{0.5}} < X_{(8)}\right) = \sum_{i=3}^{8-1} \binom{9}{i} p^i (1-p)^{n-i}$$

therefore $r = 3$ and $s = 8$ and as we have summed earlier for 3 to $8-1$, $\binom{9}{i} p^i (1-p)^{n-i}$. Here $n = 9$, therefore this is equal to,

$$1 - \binom{9}{0}(0.5)^9 - \binom{9}{1}(0.5)^9 - \binom{9}{2}(0.5)^9 - \binom{9}{8}(0.5)^9 - \binom{9}{9}(0.5)^9$$

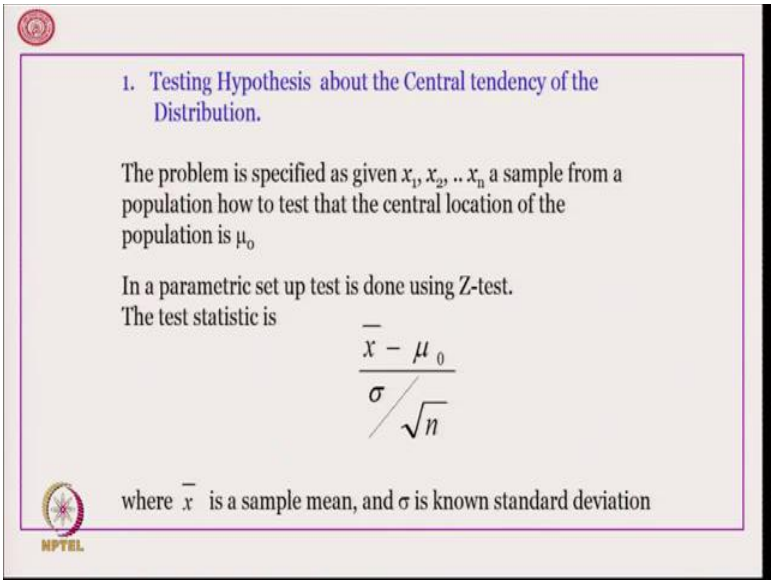
So I am just taking out all those things which are not included and that comes out to be after this arithmetic 0.89 therefore the confidence that even if you choose any random sample of size 9, then the median of the population will lie between the third and eighth order statistic is confidence is 0.89 that is nearly 90% confidence interval. So this is the basics of point and interval estimation and this is the very simple of the problems that we will be dealing with for nonparametric tests.

(Refer Slide Time: 48:40)



So under nonparametric setup the type of problems that we take up are of different types, so let me discuss the problems that we are going to take up in this series of lectures.

(Refer Slide Time: 48:57)



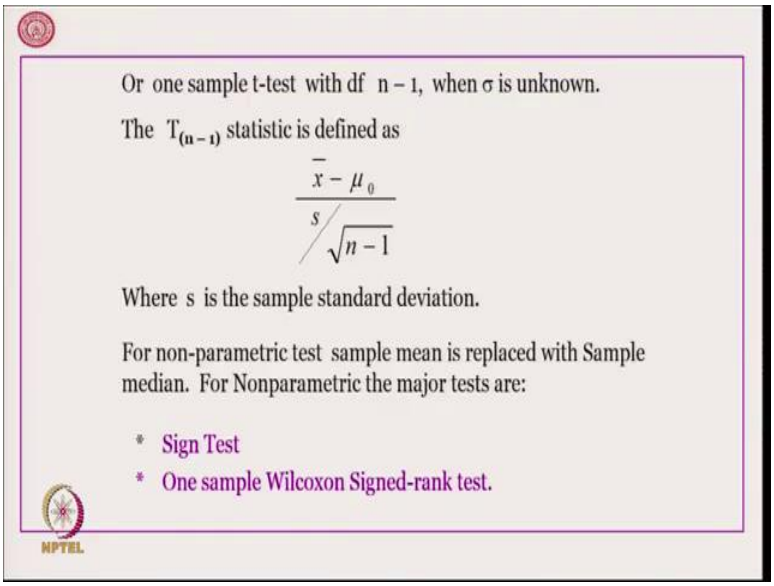
So first one is of testing of hypothesis about the central tendency of the distribution, so the problem is specified as, suppose we have taken a sample $x_1, x_2, \cdots, x_n$ from the population, question is how to test the central location of the population is actually μ_0?

So μ_0 is a given value and we want to test whether this sample gives us enough evidence that the sample has come from a population with central location μ_0. In parametric setup the test is done using z test or the normal test where the test statistic is

$$\frac{\bar{x} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

where $\bar{x}$ is the sample mean and σ is the known standard deviation and n is the sample size.

(Refer Slide Time: 50:06)



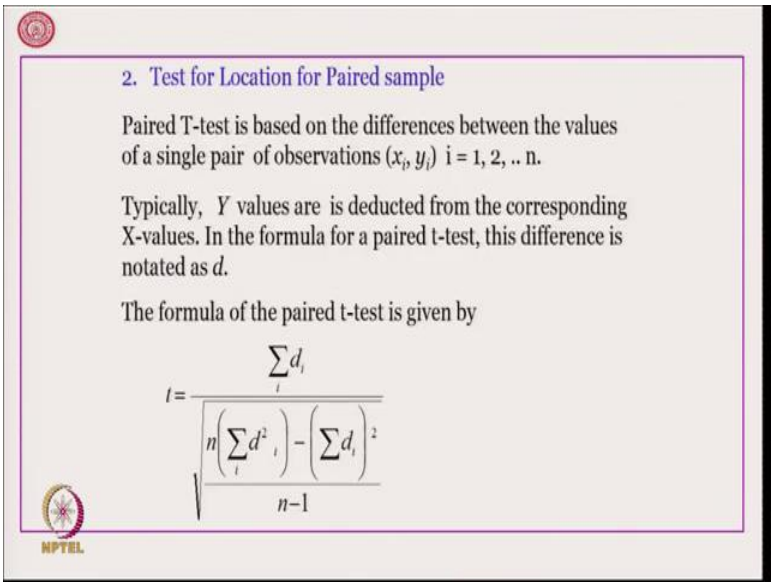
Alternatively one can use one sample t-test with $n-1$ degrees of freedom when σ is unknown, the corresponding statistic is therefore going to be T with $n-1$ degrees of freedom which we write as a subscript, $T_{(n-1)}$ and the statistic is defined by,

$$\frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n-1}}\right)}$$

I am sure most of you are familiar with these tests, but I am providing here for some recapitulation.

For nonparametric test sample mean is replaced with sample median and the corresponding important tests are Sign Test and one sample Wilcoxon Signed rank test. In this series of lectures we shall study these two tests in detail.

(Refer Slide Time: 51:15)

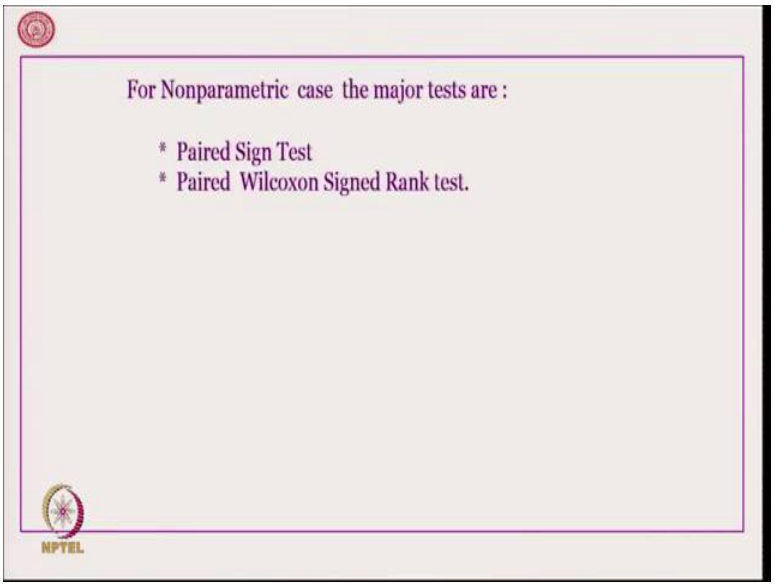


Another important problem is test for location for a paired sample for a parametric case paired T test is done here and it is based on the difference between two values of a pair of observations $(x_i, y_i), i = 1,2, \cdots, n$ where n is the sample size. We shall discuss paired test in detail later but basically on the same sample we take two observations x_i and y_i and the statistic is computed on the basis of their difference which typically we denote by d_i.

Therefore corresponding t statistic again with $n - 1$ degrees of freedom comes out to be as given below sigma di divided by square root of n into sigma di square minus sigma di whole square divided by n minus 1.

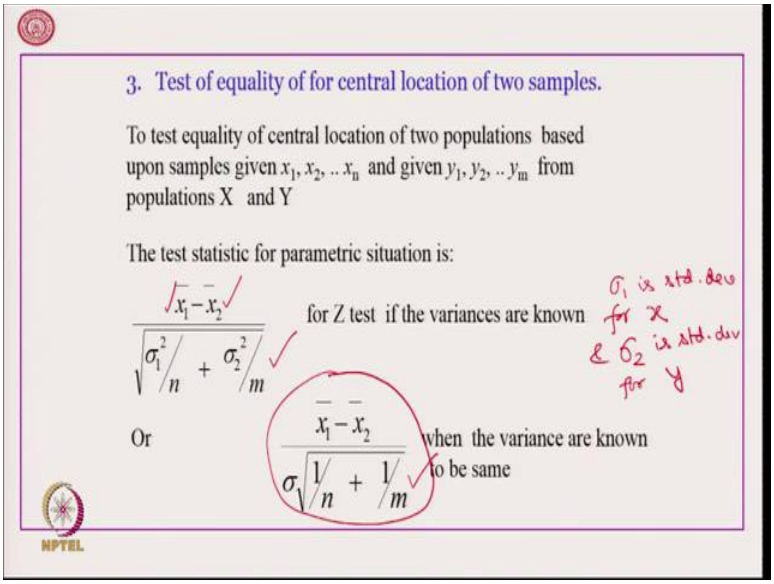
$$\frac{\sum_i d_i}{\sqrt{\frac{(n(\sum_i d_i^2) - (\sum_i d_i)^2)}{n-1}}}$$

(Refer Slide Time: 52:24)



So for nonparametric case the major tests are Paired sign test and Paired Wilcoxon signed rank test.

(Refer Slide Time: 52:34)



Another type of problem is equality for central location of two samples. So far the two tests that we said it is one sample but now the motivation is to compare the central location of the two populations where from X and Y have come. So in order to do that what we do?

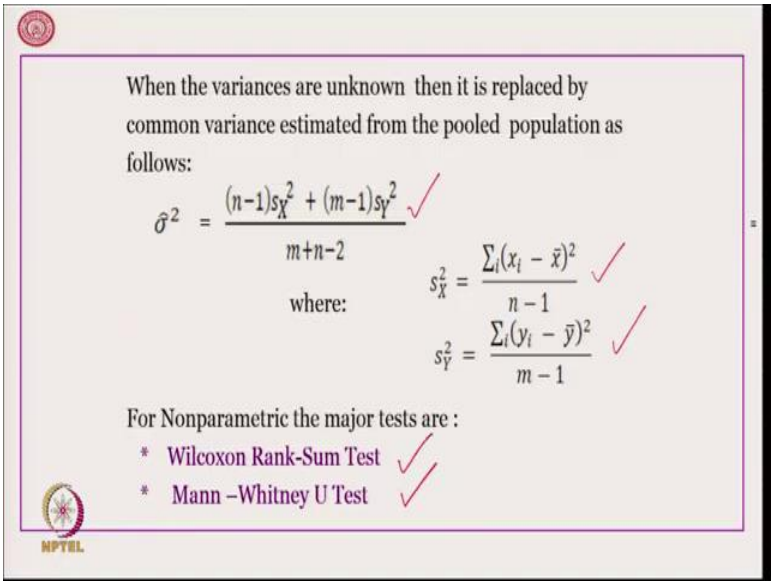
We take samples $x_1, x_2, \cdots, x_n$ and $y_1, y_2, \cdots, y_m$, say from X and Y population if we want to do Z test, then corresponding statistic is,

$$\frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

where σ_1 is standard deviation for X and σ_2 is standard deviation for Y and n and m are corresponding sample size $\overline{x_1}$ is the sample mean for X population and $\overline{x_2}$ is the sample mean of Y population.

If we know that σ_1 and σ_2 are same then the statistic becomes slightly simplified as you can see we just take out σ from the square root and therefore this is the statistic when we are knowing that σ is same for both the populations.

(Refer Slide Time: 54:20)

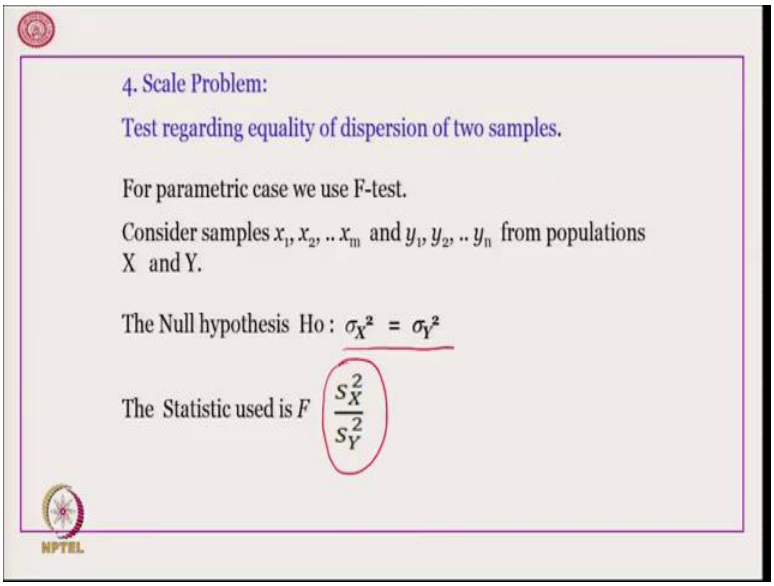


However when the variances are unknown then it is replaced by a common variance and that is typically estimated by,

$$\frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}$$

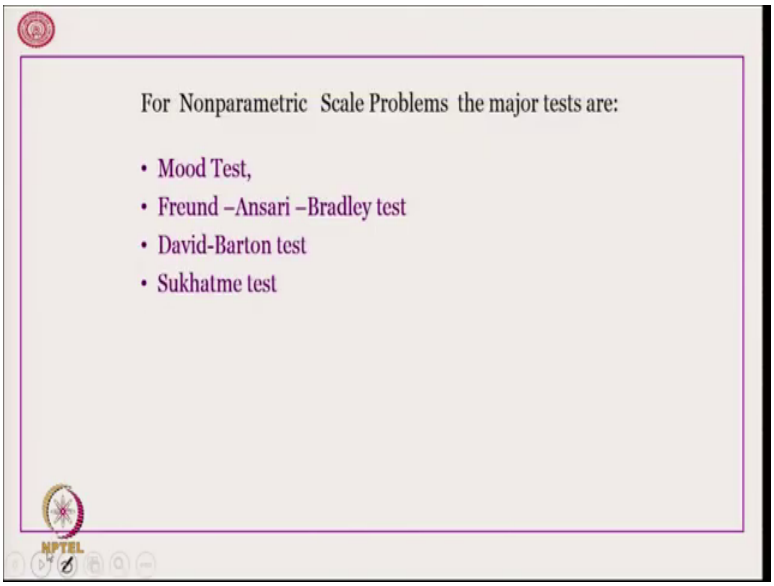
, where s_X^2 is the sample variance for X, s_Y^2 is the sample variance for Y. So this is what we do for parametric case, for nonparametric case two major tests that we are going to study are Wilcoxon Rank Sum Test and Mann-Whitney U Test.

(Refer Slide Time: 55:01)



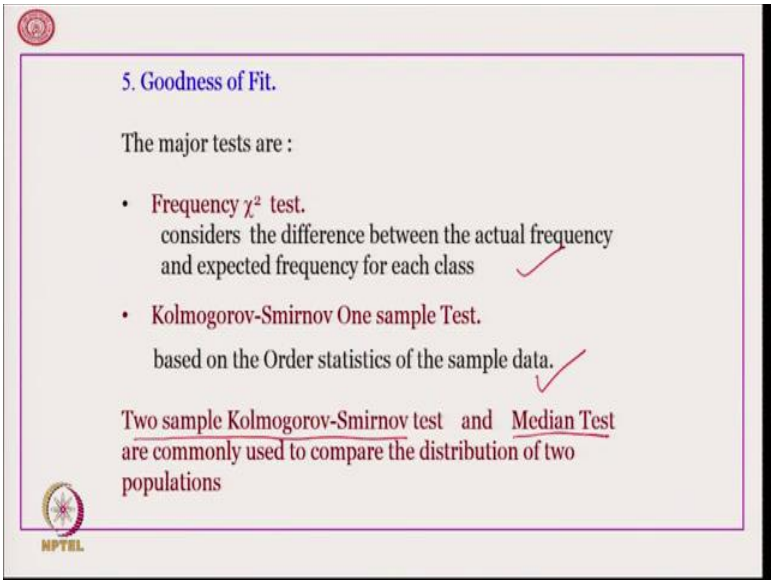
Scale Problem, this is another type of problem that we will be handling, here you want to test the dispersion of two samples. For parametric case, suppose $x_1, x_2, \cdots, x_m$ is a sample from X population $y_1, y_2, \cdots, y_n$ is a sample from the Y population and we are trying to test that whether they have the same variance, the corresponding statistic is F statistic which is computed by the ratio of the two sample variance, we shall talk about it later when we come to the scale problem.

(Refer Slide Time: 55:48)



But for nonparametric scale problem there are many different tests in particular we shall study Mood Test, Freund- Ansari- Bradley test, David-Barton test and Sukhatme test.

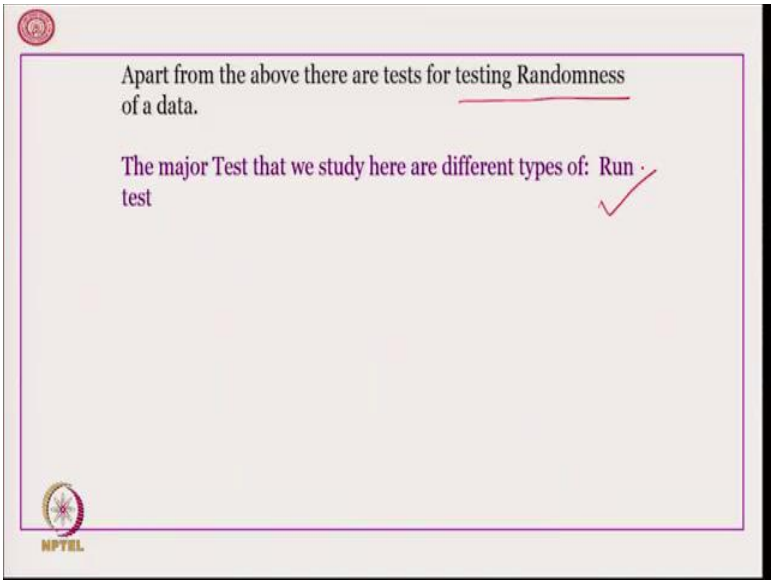
(Refer Slide Time: 56:04)



Goodness of Fit is another type of problem that we shall handle in the nonparametric statistical inference. Here the two major tests are Frequency chi square test, it considers the difference between the actual frequency and the expected frequency for each class and the statistic computed is following chi square distribution.

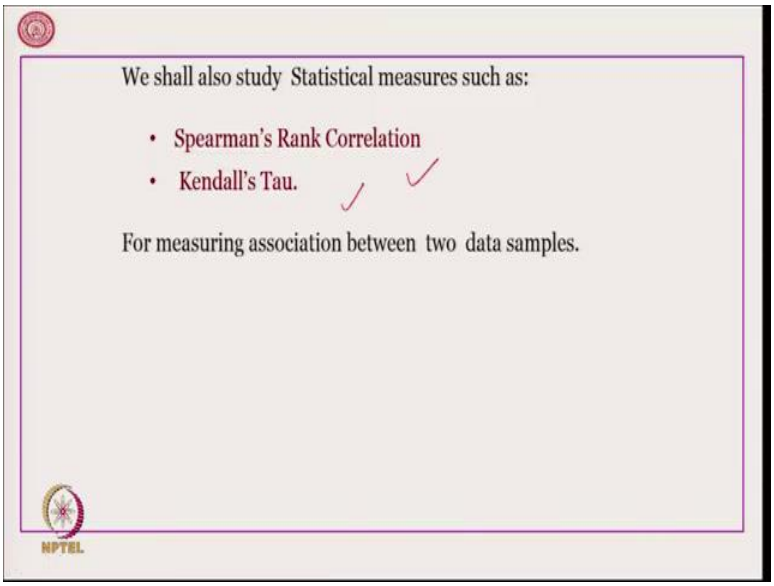
Hence it is called chi square test and another important test is Kolmogorov-Smirnov One sample test which is based on the order statistic of the sample data. In a similar way slightly more extended version of Kolmogorov-Smirnov One sample test is called the Two Sample Kolmogorov-Smirnov test and also there is something called Median test which are commonly used to compare the distribution of two different populations. We shall study them as an extension from Kolmogorov-Smirnov One sample test.

(Refer Slide Time: 57:13)



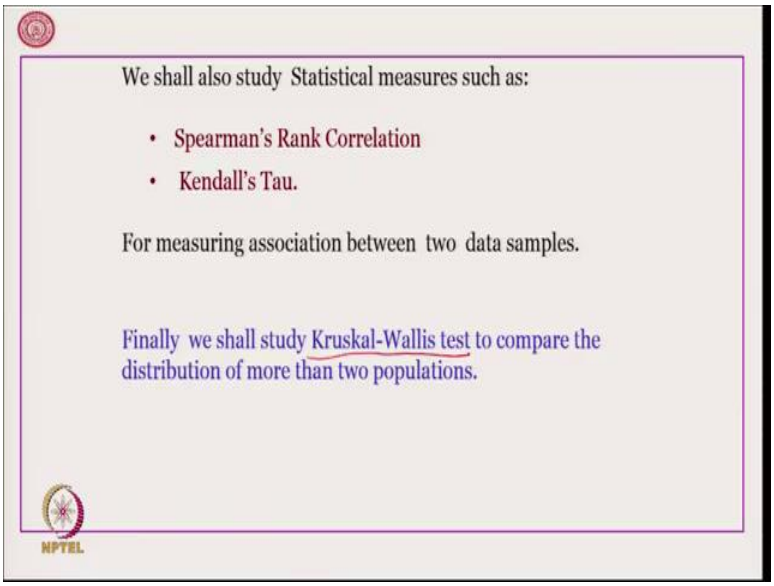
Apart from the above there are also tests for testing randomness of a data, that means the sample that we have taken whether it is actually collected randomly or not that itself is necessary to judge before actually putting it for some inference mechanism, therefore we need to first check if the data is actually random and there are several run test that allows us to see the randomness of a given data.

(Refer Slide Time: 57:48)



Furthermore we shall study statistical measures such as Spearman's Rank Correlation and Kendall's Tau, these two are important for measuring association between two data samples, in case of parametric inference we use covariance or correlation coefficient for measuring the association between data, for nonparametric case we shall look at these two tests in detail.

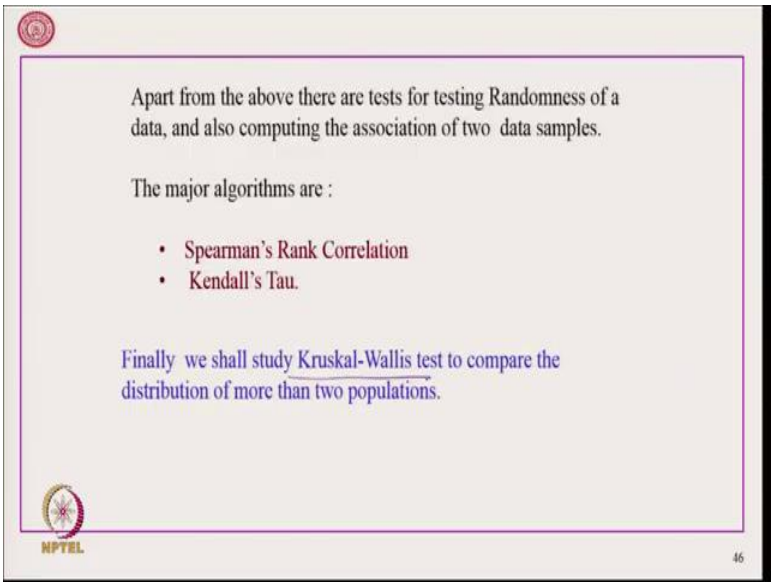
(Refer Slide Time: 58:20)



Finally we shall study Kruskal-Wallis test which is to compare the distribution of more than two populations. So we have mentioned about one sample, then two sample, what happens if you want to test the equality of distribution of more than two samples say k samples where k is equal to 3, 4, 5, etcetera, the corresponding test is called Kruskal-Wallis test and we shall study it towards the end of this lecture series.

So these are the major nonparametric tests that we shall study in course of time during the lecture series.

(Refer Slide Time: 59:06)



So over the next 9 lectures we shall learn these major nonparametric tests with great details, with solved examples and also I shall show you how to check the test from the tables or how to see the tables to understand whether to accept or to reject a particular hypothesis, okay friends I stop here today from the next class I shall start with different nonparametric tests thank you.