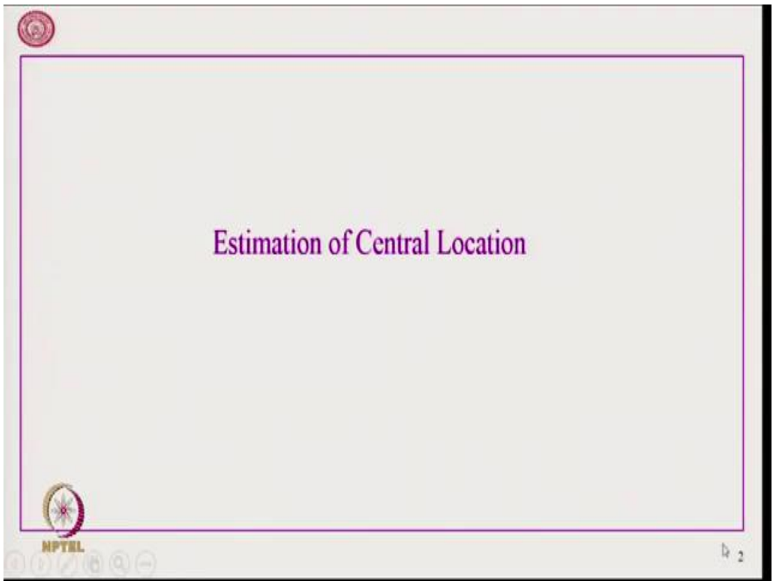
Nonparametric Statistical Inference
Professor Niladri Chatterjee
Department of Mathematics
Indian Institute of Technology, Delhi
Lecture 2
Nonparametric Statistical Inference

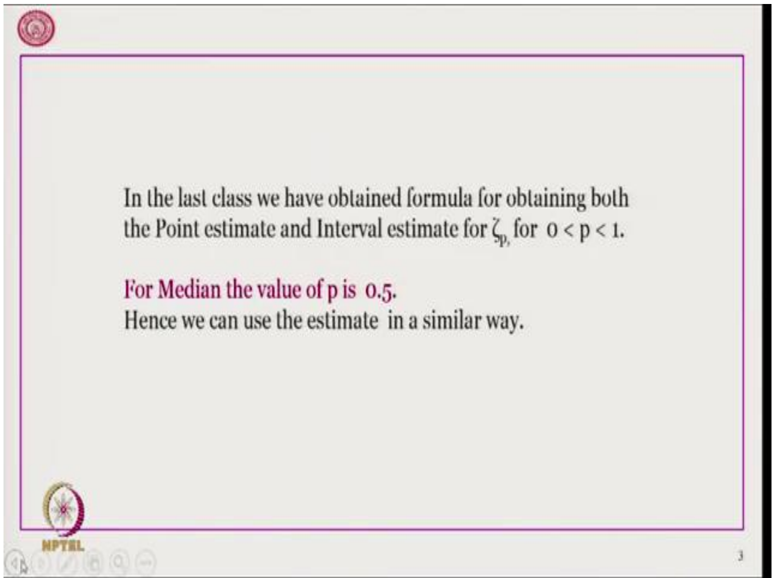
Welcome students to MOOCs series of lectures on Nonparametric Statistical Inference, this is lecture number 2. In the last class, we have seen how to do point estimate for different quantiles.

(Refer Slide Time: 00:38)



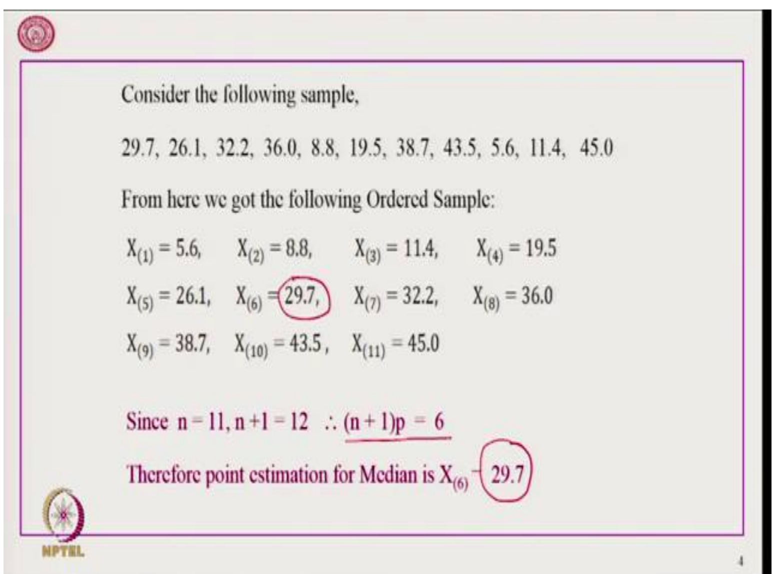
Today, we shall start with Estimation of Central Location. When we talk about parametric inference, by central location typically we think about mean. But in a nonparametric scenario, the major central location that we consider is median.

(Refer Slide Time: 01:08)



In the last class we have seen how to do ζ_p for $0 < p < 1$ and we have found out the point estimate and interval estimate for ζ_p for any given p. When we are looking at median, then value of p is 0.5. Therefore, effectively we can estimate median in a very similar way, the way we have done for any arbitrary p by considering the value of p is equal to 0.5.

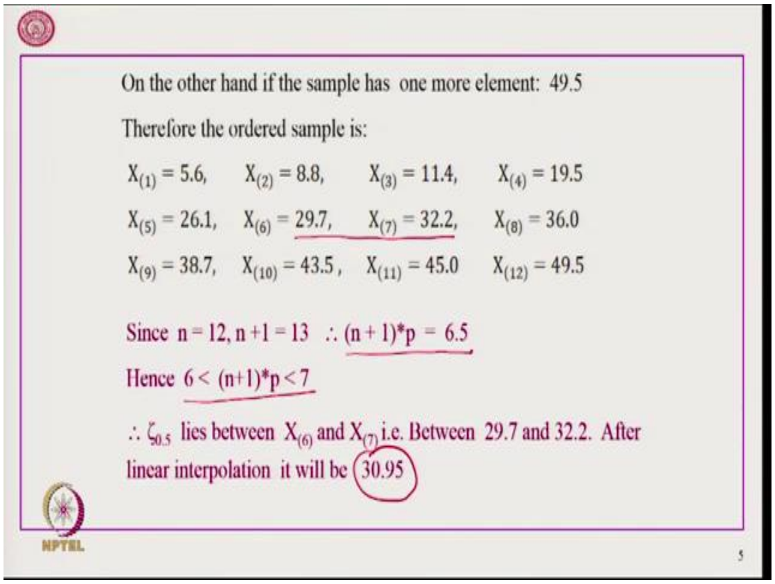
(Refer Slide Time: 01:49)



So, consider for example, this set of data there are 11 data points 29.7, 26.1 up to say 11.4 and 45. From here, what we do just to recollect that, we take the ordered sample that is, we sort them in increasing order, since $n = 11$ therefore, $n + 1 = 12$ and hence $(n + 1)p = 6$, where $p = 0.5$.

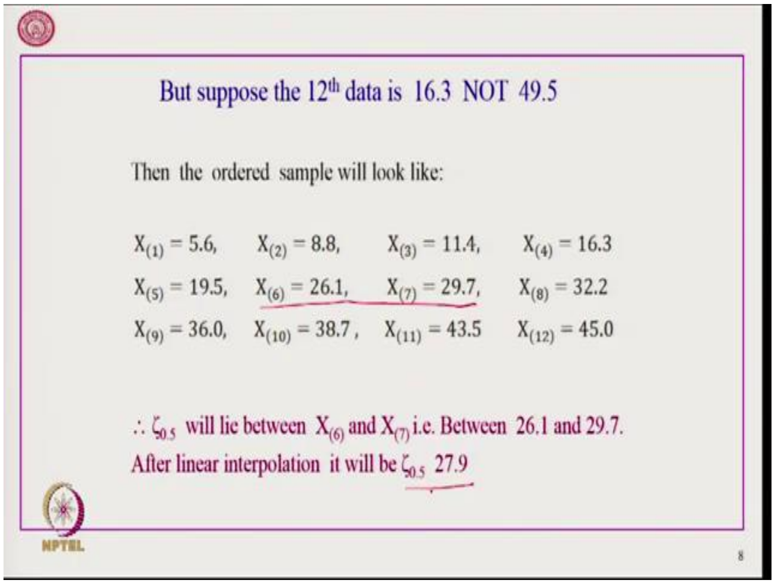
Hence, the point estimation for median is $X_{(6)}$ which is 29.7. So, that way we can get the value of the median or an estimate for the median.

(Refer Slide Time: 02:51)



On the other hand, suppose, there is one additional element namely 49.5 therefore, the ordered sample is now $X_{(12)}$ is 49.5 because that is the maximum of all the elements, but in this case $(n + 1)p = 6.5$ which is between 6 and 7. Therefore, by the concept of linear interpolation, we get an average of these two which is the value 30.95. So, this is going to be the estimate for median.

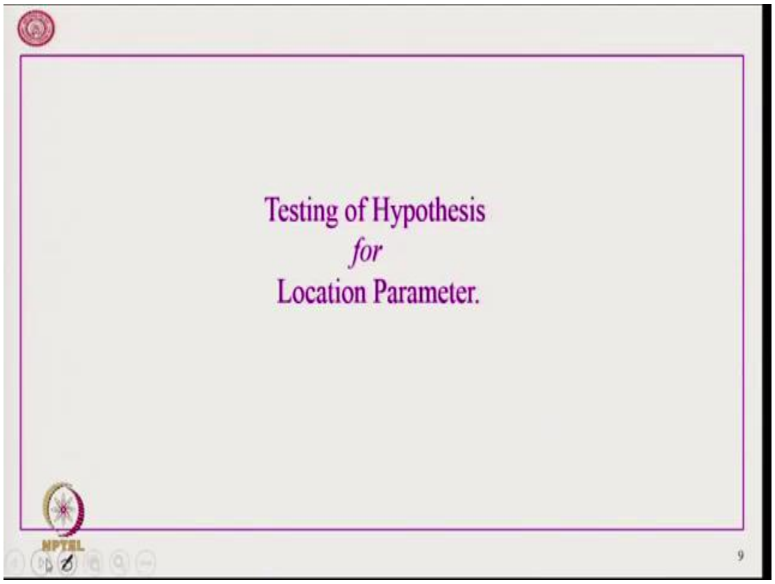
(Refer Slide Time: 03:40)

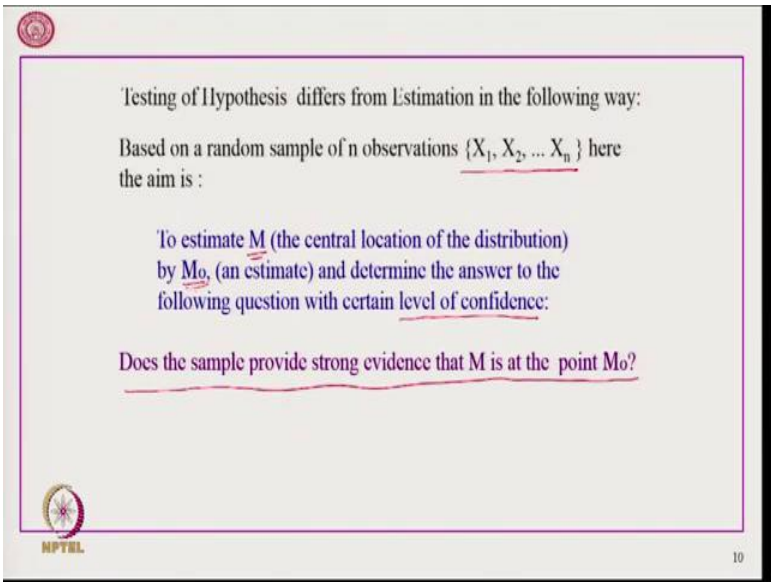


Instead, if the new data that we have, data would have been 16.3 then what would happen? Then we get the ordered sample is like that. So, we can see that 16.3 is coming here, but in this case the median is going to be the arithmetic average of these two, which is going to be 27.9. So, that is going to be the point estimate for the median.

Therefore, we can see that a single observation may change that estimate of the central location of the distribution, which is true even for parametric inference, because in that case the sample mean is going to be the estimate for population mean and therefore, one single value can change the estimate of the central location.

(Refer Slide Time: 04:48)

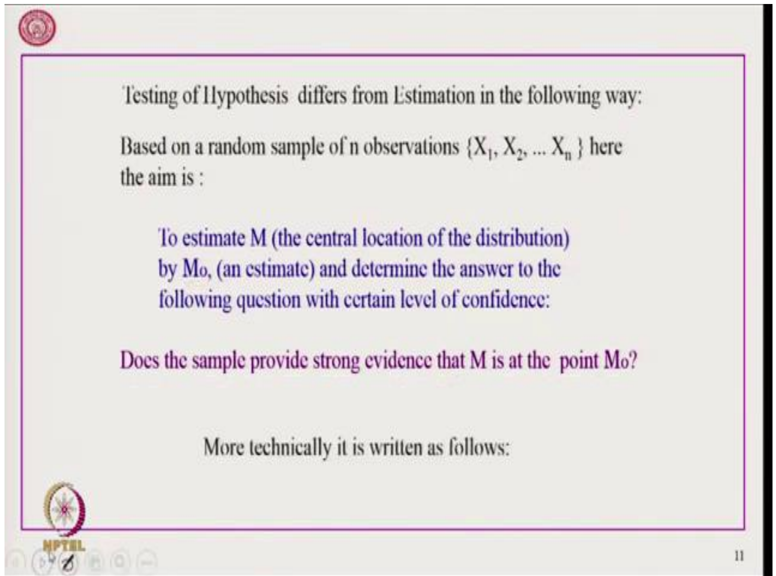


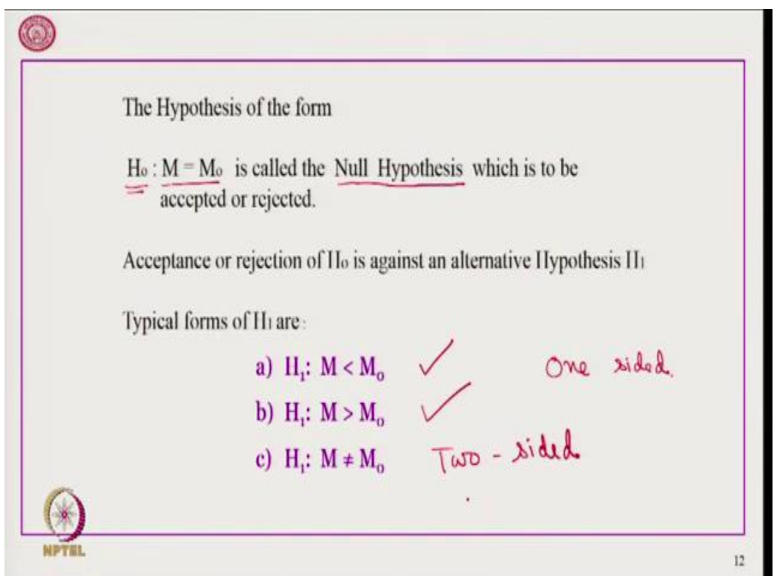


Therefore, the most important concept is testing of hypothesis. So, in testing of hypothesis, it differs from the concept of estimation in the following way. Here based on a random sample of n observations, let me call them $X_1, X_2, \cdots, X_n$. Our aim is to estimate M, the central location of the distribution by some value M_0. M_0 is an estimate which may be based on the sample or which may be theoretically one can think of and determine the answer to the following question with certain level of confidence.

Does this sample provide strong evidence that M is at the point M_0? So, those who are familiar with statistical inference, you understand what I mean and those who are not very familiar. Let me tell you, what we try to do here. We want to see that if a given value M_0 can be considered as the central location, and if we can, then what is going to be the level of confidence.

(Refer Slide Time: 06:12)

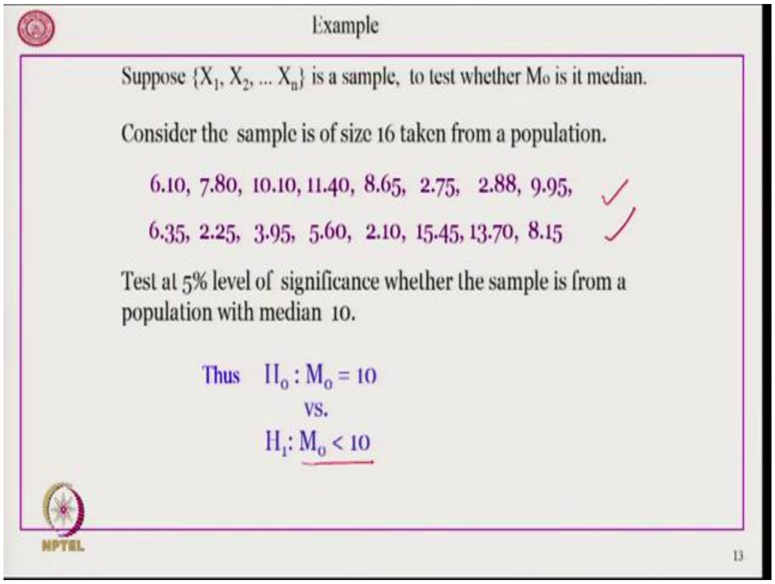




So, let us go and study. So, more theoretically we shall write it as follows. This H_0 is called the null hypothesis. As I have indicated H_0 is the notation for null hypothesis. So, the hypothesis that we want to test is whether M can be considered as the value M_0 and we are going to accept or reject this hypothesis viz-a-viz some alternative which can be of 3 types.

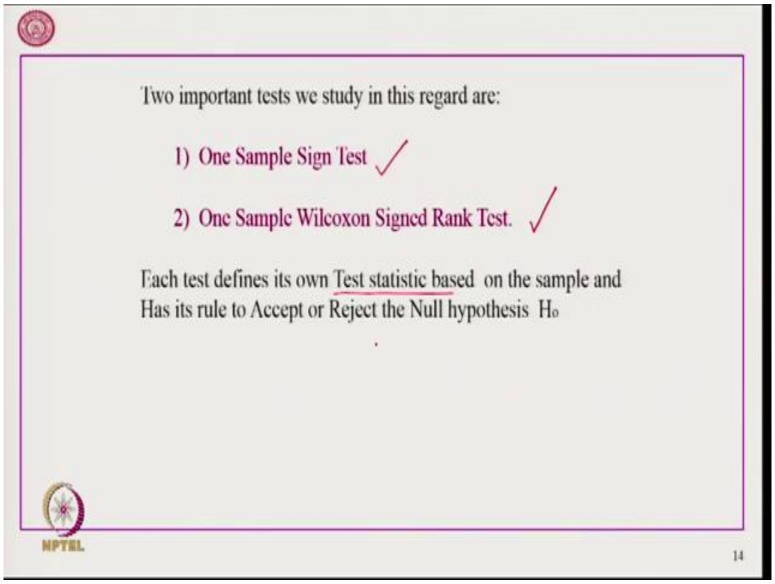
$M < M_0$, that is, the sample evidence is such that we cannot accept that the value of M is M_0 rather it can be that $M < M_0$ or it can be $M > M_0$. These two are called one-sided or we can think of $M \neq M_0$ which is a two-sided alternative. These alternative hypotheses are denoted by H_1

(Refer Slide Time: 07:33)



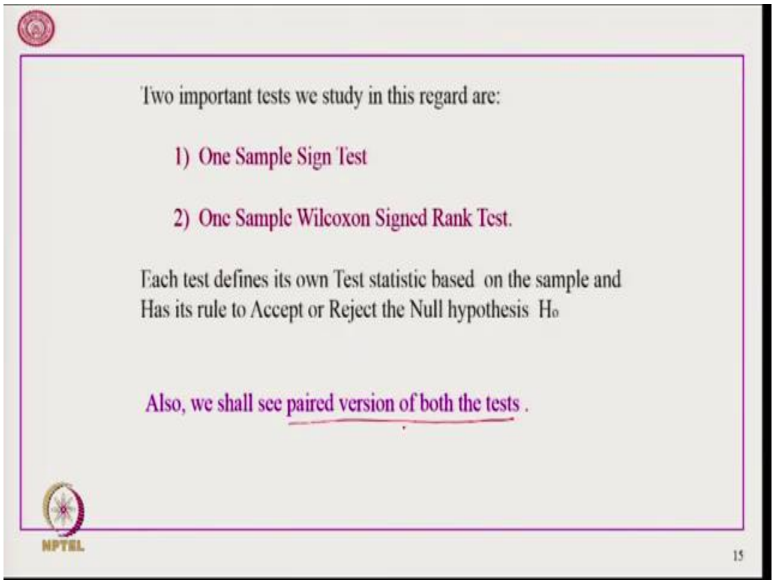
For example, suppose I have n samples $X_1, X_2, \cdots, X_n$ and we have taken a sample with $n = 16$ and I am giving you the values of the 16 observations. That is, you can see the 16 values here and we want to test at 5 percent level of significance whether we can accept the value $M_0 = 10$ against the alternative is $M_0 < 10$. That is, you are looking at one-sided alternative.

(Refer Slide Time: 08:15)



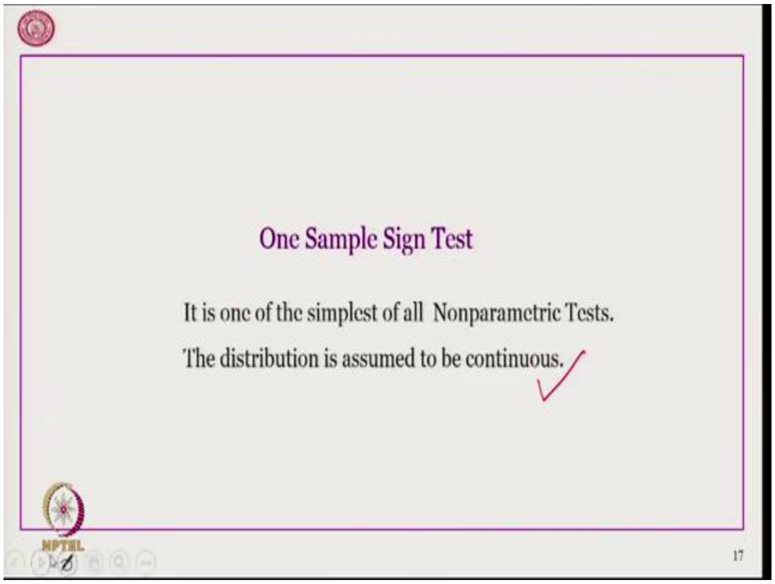
Two important tests in this regard are One Sample Sign Test and One Sample Wilcoxon Signed Rank Test. So, in this lecture, we shall be studying these two tests in detail. Each test defines its own test statistic based on the sample and has its rule to accept or reject the null hypothesis H_0 .

(Refer Slide Time: 08:47)



Also, we shall see a paired version of both the tests which perhaps we shall do in the next class but let us first see what these two tests are.

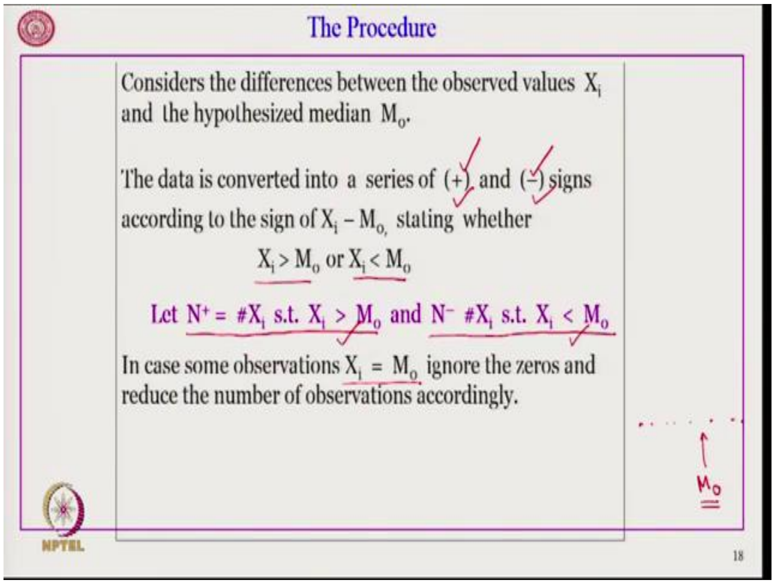
(Refer Slide Time: 09:01)



So, the first test we are looking at is One Sample Sign Test. By one sample, we mean that the sample is taken from one population. Later, we shall see that we will be learning tests, which are based on more than one sample. For example Two Sample Sign Test, when we will be comparing the central location of two different populations.

So, let us first understand what One Sample Sign Test is, it is one of the simplest of all nonparametric tests, the distribution is assumed to be continuous. So, this is the most basic assumption that we are not looking at a discrete population, rather it is a continuous population, but the test is actually very-very simple as we see now.

(Refer Slide Time: 09:59)



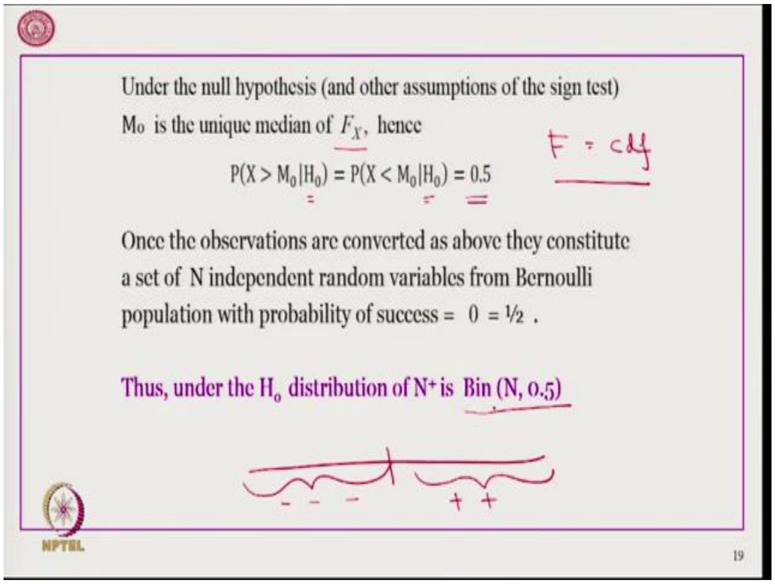
So what we do here in One Sample Sign Test is the following. It considers the differences between the observed values X_i and the hypothesized median M_0 what does it mean? So, suppose these are my observed values and suppose this is my hypothesized value M_0. So, we want to see with respect to M_0 how the observed values are distributed.

So, in order to test that, we do the following. The data is converted into a series of plus and minus signs according to the sign of $X_i - M_0$ and stating whether $X_i > M_0$, or $X_i < M_0$.

So, if the observed value is more than M_0 then $X_i - M_0$ is going to be a plus. And if it is less than M_0 t, then it is going to be a minus. And therefore, we reduce the observed sample into a sequence of plus and minus. Therefore, what happens? We can check what is the number of pluses that is the N^+ number of X_i such that $X_i > M_0$ and N^- that is number of X_i such that $X_i < M_0$.

So, here we are checking number of X_i which are greater than M_0. Here we are checking number of X_i which are less than M_0. In case some observations X_i is same as M_0, then we ignore that value, because in that case $X_i - M_0$ is going to be 0 and if it is a continuous distribution we can assume that the random variable taking a particular value has probability 0. So, that is the reason.

(Refer Slide Time: 12:12)

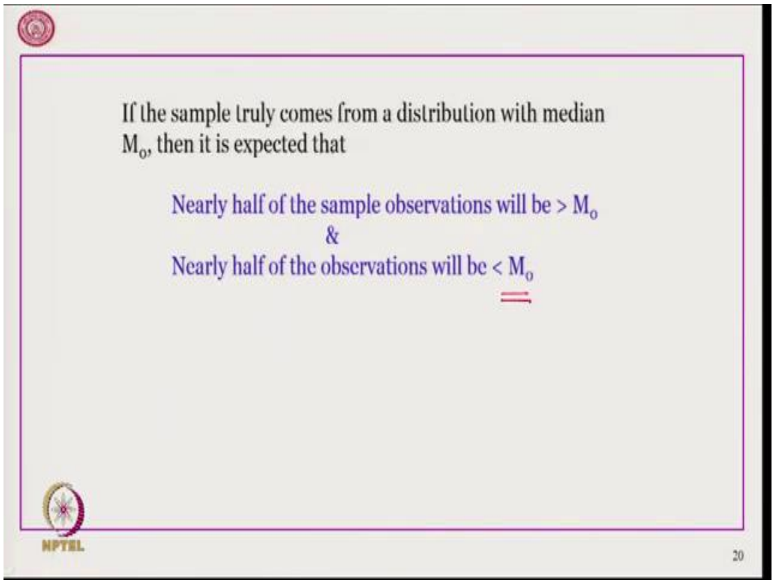


Now, under the null hypothesis, let us see what happens. M_0 is the unique median of a F_X. Let X be the underlying random variable and F is the cdf that is the cumulative distribution function. Therefore, probability $X > M_0$ under the null hypothesis is same as probability $X < M_0$ under the null hypothesis and that value is equal to half i.e. $P(X > M_0|H_0) = P(X < M_0|H_0) = 0.5$.

That is very easily understandable because if M_0 is the true median, then we would expect half of the observations will be here which are greater than M_0 and other half will be here, which are less than M_0. All these are going to give me plus signs and all these are going to give me minus signs, right.

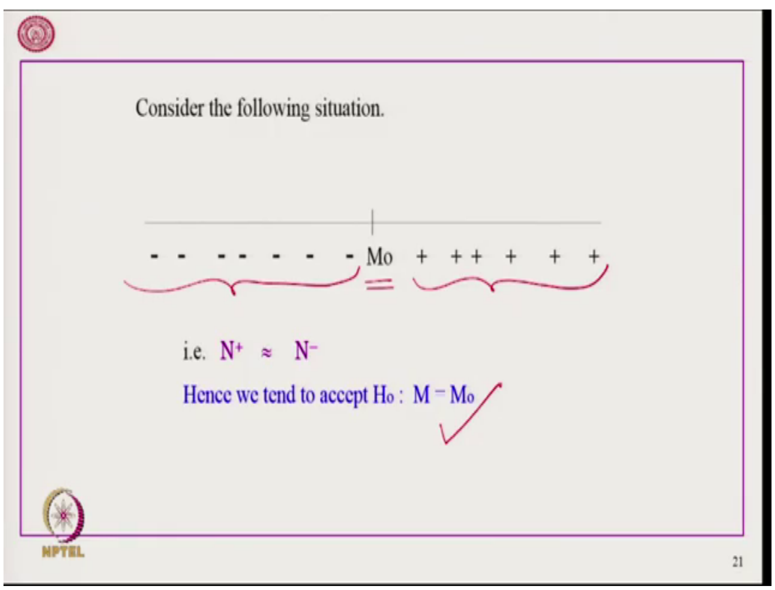
Therefore, under H_0, the distribution of N^+ which is the number of observations giving plus sign is a binomial distribution with parameter N and 0.5 is the probability of success, and where N is the number of observations.

(Refer Slide Time: 13:35)



So, if the sample is truly coming from that distribution with median is equal to M_0 then we would expect that nearly half of the sample observations are going to be greater than M_0. And nearly half of the observations are going to be less than M_0. That comes under the null hypothesis. That is the median is actually at that point M_0.

(Refer Slide Time: 14:04)

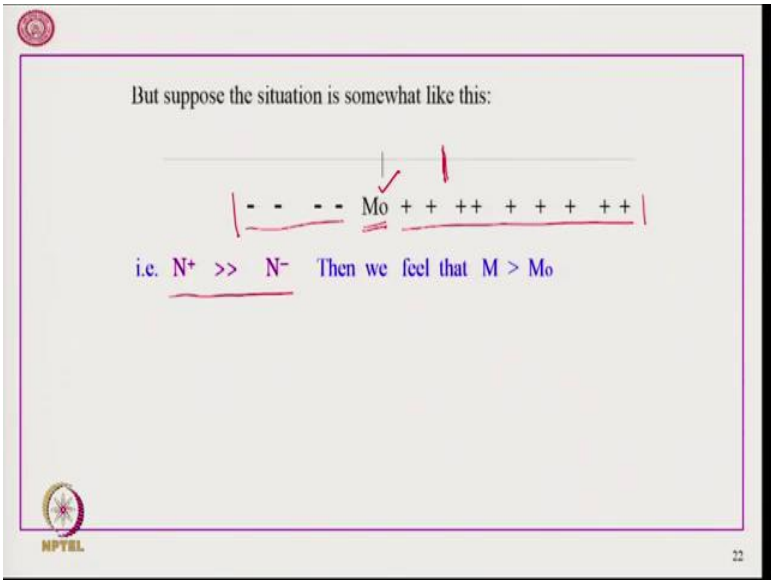


So, let me illustrate. Suppose this is the scenario, M_0 is the hypothesized median. So, these are the observations which are greater than M_0, these are the observations which are less than M_0 and if

M_0 is truly the median, then N^+ is equal to or very close to N^-, they need not be equal, because we know that statistically it cannot happen that both sides will have equal number of observations.

But there should be a limit that is what the test is all about whether the difference that is there in the number of pluses and minuses is acceptable under the null hypothesis. So, in this case it appears that they are more or less same. So, we are likely to accept that M is same as M_0.

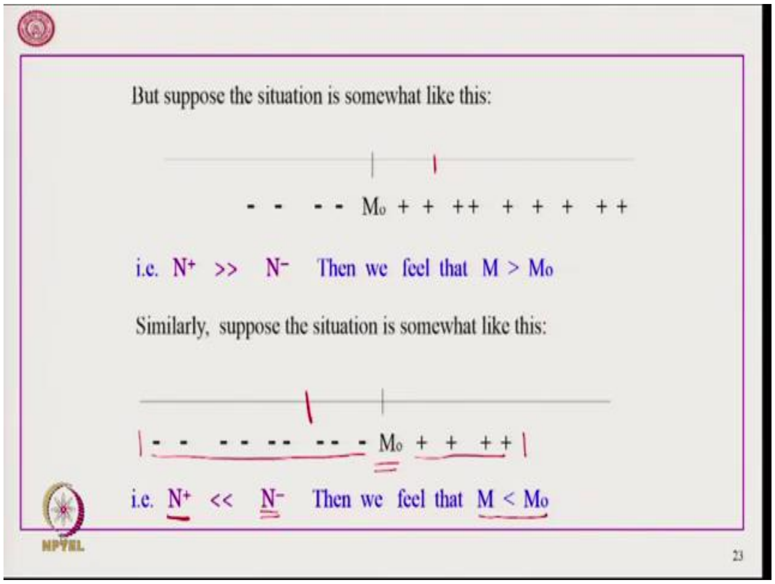
(Refer Slide Time: 15:04)

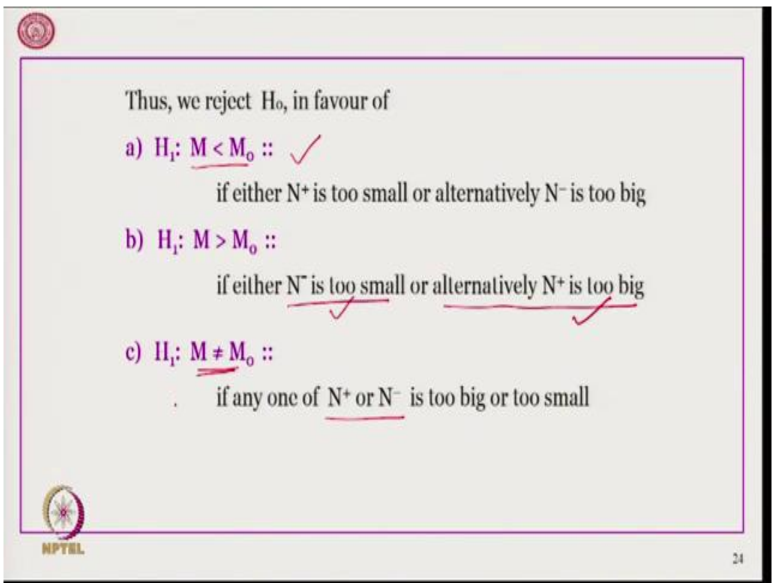


Now consider this scenario. Suppose this is my hypothesized M_0 and after we take the differences and convert the observed sample into pluses and minuses we can see that there are so many plus and so less number of minus. That is the number of plus signs is much-much greater than the number of minus signs, then what we feel because it is a continuous distribution, I can see that our observation has come in this range.

Therefore, we would expect the median is going to be somewhere here that is actual median of the population is greater than the hypothesized median M_0. I hope you understand this.

(Refer Slide Time: 15:57)





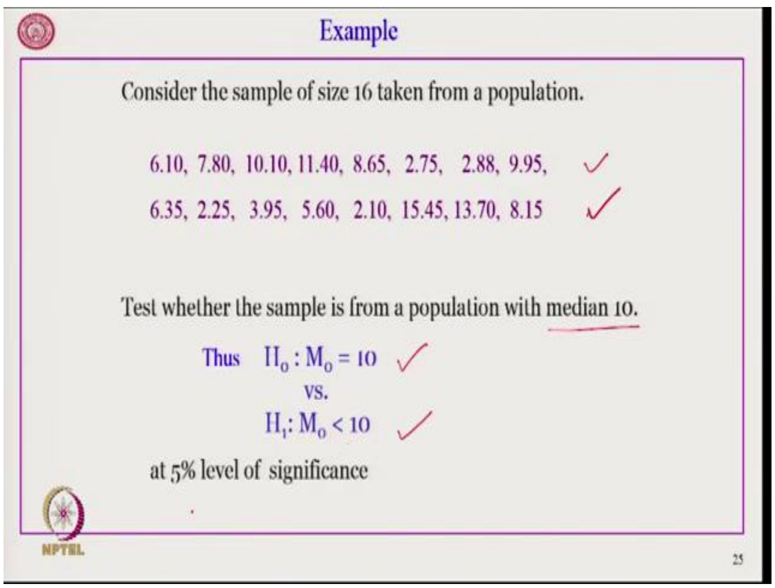
In a similar way, suppose the situation is somewhat like this, this is the hypothesized M_0 and then we can see that the number of plus signs of N is less than much less than number of minus signs. Therefore what we would expect that the median should come somewhere here because it is a continuous distribution in this range and therefore, we would feel that M is actually less than M_0 .

Therefore, we reject the H_0, that is the null hypothesis in favor of $M < M_0$ that is one-sided alternative if either N^+ is too small or alternatively, N^- is too big. Say for this example, if N^+ is too small then we would expect, that it is more likely that $M < M_0$ or in other words, if N^- is too big, then we would expect that M is somewhat less than M_0 and therefore, we would reject the null hypothesis in favor of $M < M_0$.

In a similar way, if N^- is too small or alternatively N^+ is too big, then we would rather think that $M > M_0$ say something like this, we would expect that M is somewhere greater than M_0 . Therefore, if the observed statistic is something like that, that N^+ is too big or N^- is too small, then you would expect that $M > M_0$.

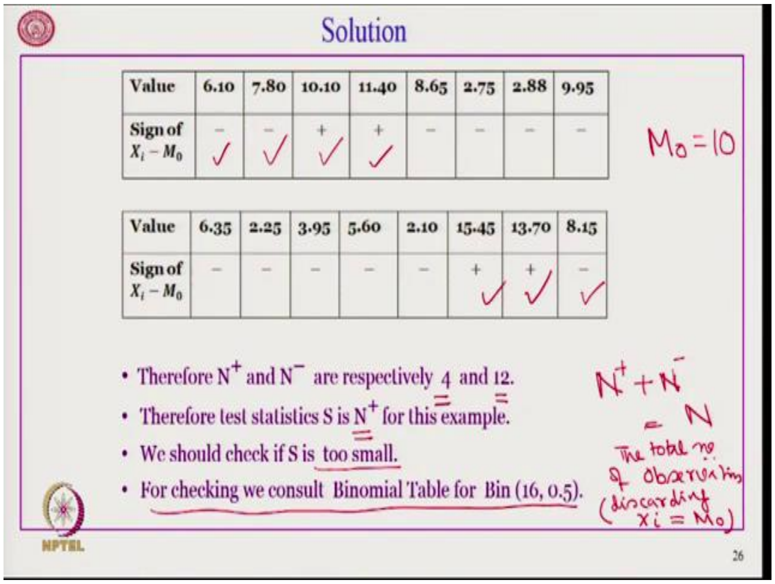
For a two sided alternative $M \neq M_0$, if anyone of them N^+ or N^- is too big or too small, we tend to reject the null hypothesis in favor of the alternative $M \neq M_0$.

(Refer Slide Time: 18:16)



So, let me illustrate with one example considered the sample of size 16 taken from a population. So, there are 16 observations as you can see enlisted here and we want to test whether this sample is from a population with median is equal to 10. Therefore, we were testing $M_0 = 10$ against say alternative $M_0 < 10$ and our level of significance is 5%.

(Refer Slide Time: 18:56)

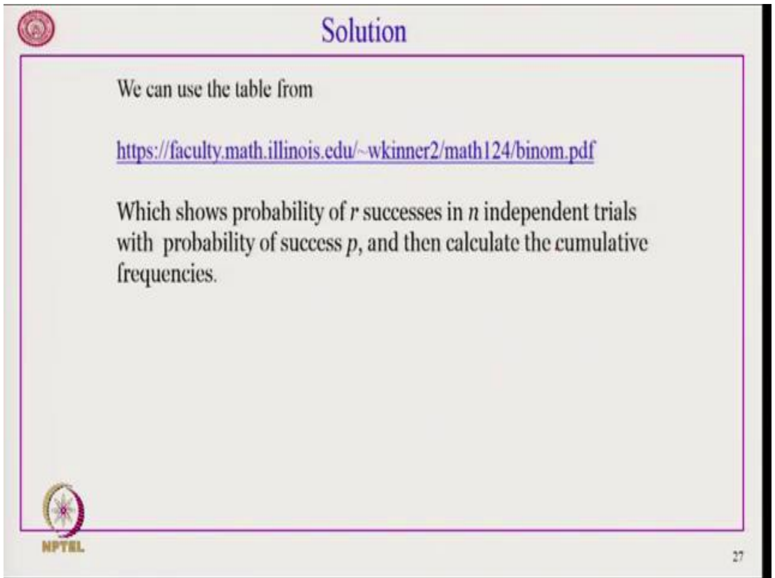


So, what do you do? We do the following we construct the table. We are considering the sign of $X_i < M_0$ our $M_0 = 10$. Therefore, what we do? 6.10 that gives a minus, 7.80 that gives a minus, 10.10 that gives a plus and like that, we compute it for all the 16 observations. Therefore what is the value of N^+? N^+ is 1, 2, 3, 4 and therefore, quite naturally N^- is going to be 12 because we have 16 observations and $N^+ + N^- = N$, the total number of observations.

Note that if some of the observations are same as M_0 we have to discard them. So, I write discarding $X_i = M_0$. So, all those observations which are equal to M_0 , we have to discard them. Therefore, the test statistic we can consider to be N^+ there is no hard and fast but typically the smaller of these we try to use as our statistic and we need to check if it is too small.

So, if it is too small that means, it is so small that it is very unlikely to happen under the null hypothesis, then we are going to reject the null hypothesis in favor of the alternative but how to do that? For checking that we consult a binomial table of 16 and p is equal to 0.5.

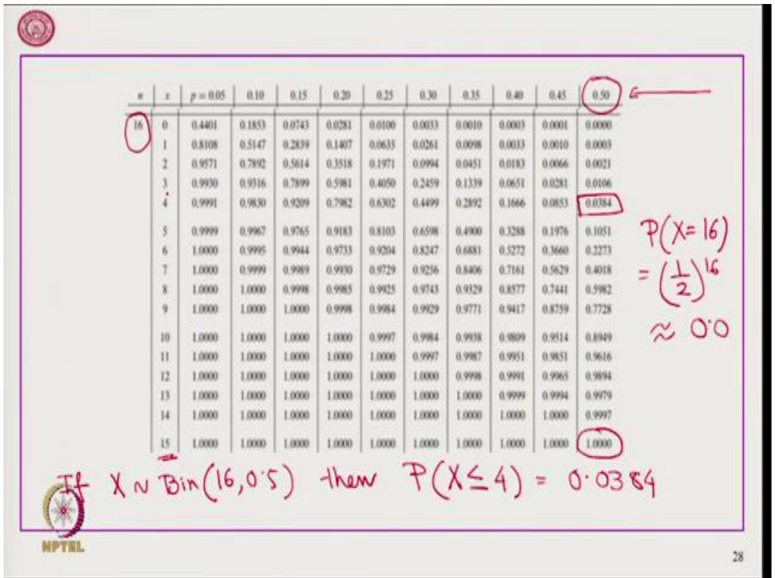
(Refer Slide Time: 21:07)



Now, all of you, I assume are familiar with different statistical tables, like normal table, chi square table, t, F, etc. For parametric, these four tables are most important and with that we can test most of the hypothesis that we are going to test. In case of nonparametric for different tests, we need to have corresponding tables, which will give us the probability.

So, for Sign test, we have to use the binomial table and which gives us the probability of r success in n independent trials with probability of success is equal to p. So, that is a general binomial table. For Sign test we need to look at p is equal to 0.5.

(Refer Slide Time: 22:00)



n	x	p = 0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
16	0	0.4401	0.1853	0.0743	0.0281	0.0100	0.0033	0.0010	0.0003	0.0001	0.0000
	1	0.8108	0.5147	0.2839	0.1407	0.0635	0.0261	0.0098	0.0033	0.0010	0.0003
	2	0.9571	0.7892	0.5614	0.3518	0.1971	0.0994	0.0451	0.0183	0.0066	0.0021
	3	0.9930	0.9316	0.7899	0.5981	0.4050	0.2459	0.1339	0.0651	0.0281	0.0106
	4	0.9991	0.9830	0.9209	0.7982	0.6302	0.4499	0.2892	0.1666	0.0853	0.0384
	5	0.9999	0.9967	0.9765	0.9183	0.8103	0.6598	0.4900	0.3288	0.1976	0.1051
	6	1.0000	0.9995	0.9944	0.9733	0.9204	0.8247	0.6881	0.5272	0.3660	0.2273
	7	1.0000	0.9999	0.9989	0.9930	0.9729	0.9256	0.8406	0.7161	0.5629	0.4018
	8	1.0000	1.0000	0.9998	0.9985	0.9925	0.9743	0.9329	0.8577	0.7441	0.5982
	9	1.0000	1.0000	1.0000	0.9998	0.9984	0.9929	0.9771	0.9417	0.8759	0.7728
	10	1.0000	1.0000	1.0000	1.0000	0.9997	0.9984	0.9938	0.9809	0.9514	0.8949
	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9951	0.9851	0.9616
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9991	0.9965	0.9894
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9979
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

So, consider this table, I am just giving you for $n = 16$. Similar tables are there for different values of n. On the header we can see different values of p starting from 0.05 to 0.50 for this example our consideration is 0.50,$n = 16$. So, we are looking at for different values of x, what is going to be the cumulative probability.

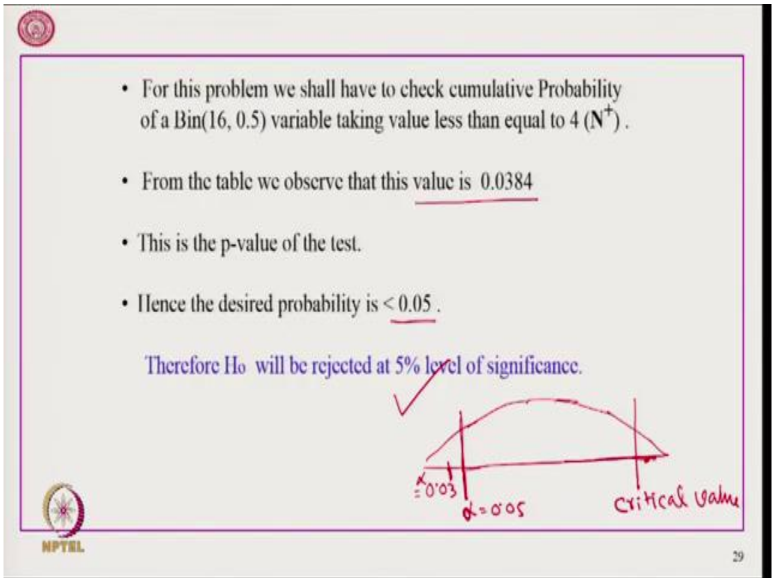
Say for example, for the value of 4, the cumulative probabilities 0.0384. What does it mean? This means that if X is a binomial 16 comma 0.5 then probability $P(X \leq 4) = 0.0384$. Just for the sake of completeness, let us look at what is the value at 15. It is showing that if X is binomial 16 comma 0.5 then cumulative probability $P(X \leq 15) = 1$.

Some of you may think how it can happen, because if X is binomial with n is equal to 16, then there is a positive probability that X will take the value 16, but we know that probability X is equal to 16 is equal to half to the power 16, which we consider to be almost close to 0.

That is why at this point, this is given to be 1, I would suggest that you go to the table and try to see the cumulative probabilities. And if you are trying to experiment, you can compute these values for some x, for some n, say n is less than 16, you can take it to be 5 and try to compute the values of probability for a given value of p. And for a given value of x, what is going to be the cumulative probability for that one.

That will be a good exercise and that will tell you how one has computed such large tables for different n and different x for different values of p, okay. So, let us go back to our test.

(Refer Slide Time: 25:11)



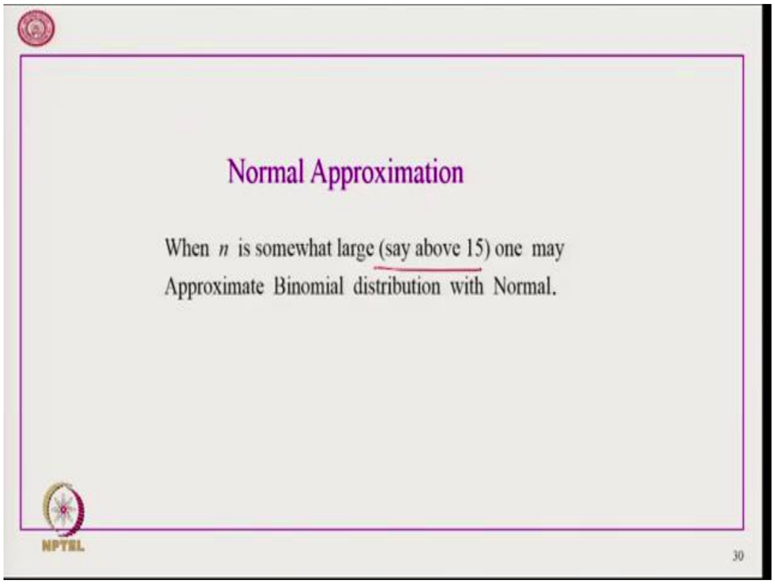
So, for this problem, we have to check the cumulative probability of binomial 16 comma 0.5 is taking a value less than equal to 4, as I have already explained, and we observed that the value is equal to 0.0384. So, this is what is called the p value of the test.

So, that is the observed value of the probability corresponding to the statistic that is given from the sample. Since this value is less than 0.05, we will be rejecting the hypothesis at 5% level of significance. So, if you do not understand this, so, let me give you an illustration. So, if it is a binomial distribution with p is equal to 0.5 it is going to be very symmetric around the mean.

And we are looking at 5 percent level of significance. That means, if this is the cutoff, above which there should be 95% of observations, but, in this case we have found the value to be 0.03 which is less than for alpha is equal to 0.05 that is 5%. Therefore, because it is a one-sided test and this value is less than the cutoff point, we are going to reject the null hypothesis.

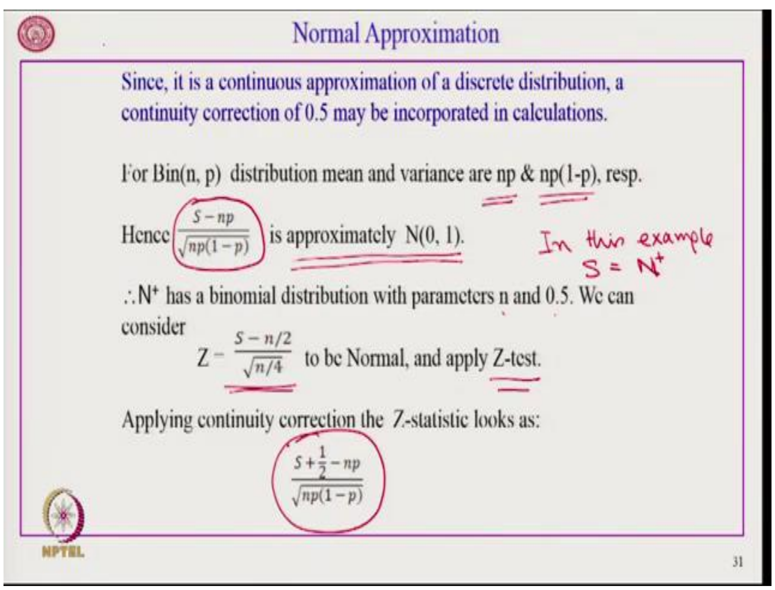
Later, we shall see some examples when we will be looking at the upper sided tests and we shall reject the null hypothesis, if the observed value falls somewhere higher than this value, this is called the critical value. I am sure most of you are familiar with this term. So I am giving you a brief idea about what we do, when we are going to test a particular hypothesis.

(Refer Slide Time: 27:27)



Now, we can do normal approximation, because we know that from Central Limit Theorem that if the number of observation is somewhat large, then we can approximate it with normal in this case, typically, if the value is greater than equal to 15. Then we try to approximate the binomial distribution with a normal distribution.

(Refer Slide Time: 27:54)



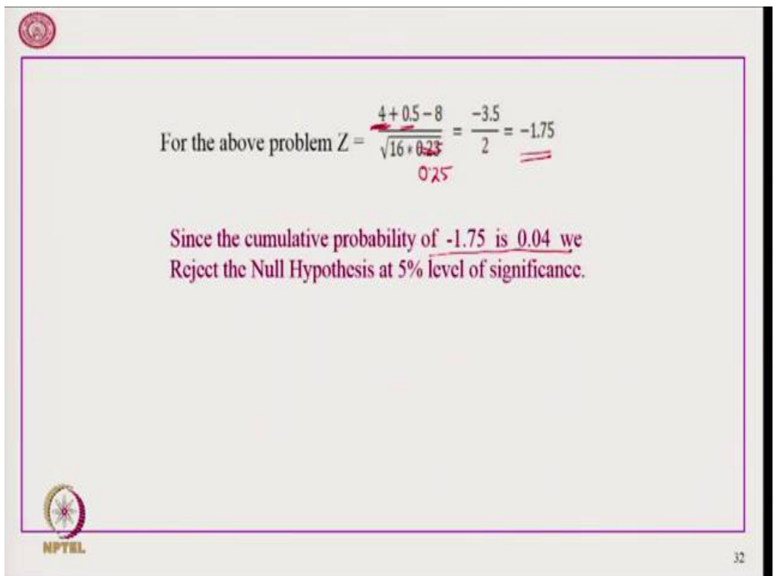
In order to approximate any distribution with a normal we need to understand what is going to be the mean and standard deviation of the random variable. In case of a binomial $Bin(n,p)$, we know that the expectation is equal to np and the variance is equal to npq or $np(1-p)$. Therefore, the standardized variable is S where S is the statistic. In this case we have takenN^+.

Therefore, we are looking at $\frac{S-np}{\sqrt{np(1-p)}}$ and this is going to be approximated by normal $N(0,1)$

Therefore, N^+ has a binomial distribution parameters if n and 0.5 then we can consider $Z = \frac{S-n/2}{\sqrt{n/4}}$

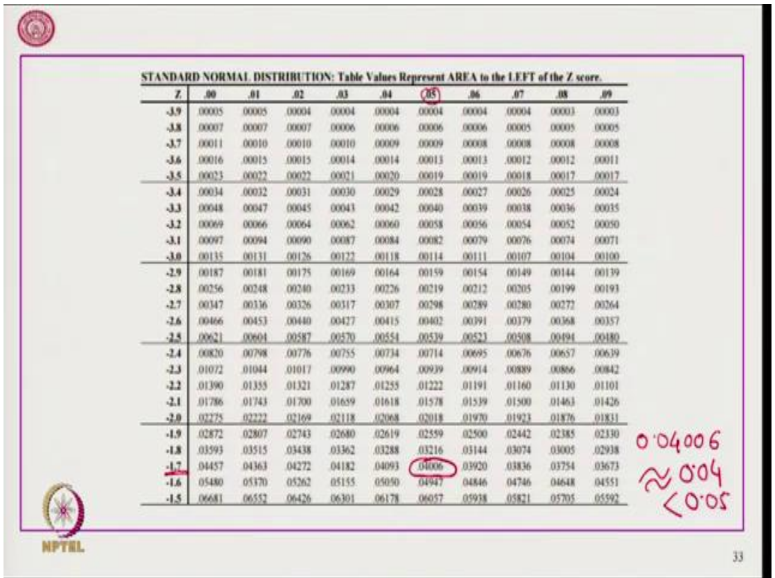
And therefore we can use Z test. But sometimes we use continuity correction, because it is a discrete density. Therefore, we use a factor 0.5. And under continuity correction, this is going to be the statistic $\frac{S+0.5-np}{\sqrt{np(1-p)}}$.

(Refer Slide Time: 29:45)



So, for the given problem, our S was 4, with a continuity correction 0.5, np is equal to 8 because n is equal to 16 divided by $\sqrt{16p(1-p)}$ and that gives us, the value this should be 25. And therefore, the obtained value of the statistic is equal to -1.75. So, since that cumulative probability of -1.75 is 0.04 we reject the null hypothesis at 5% level of significance. What does it mean? Those who are not familiar with this, for them let me explain.

(Refer Slide Time: 30:37)



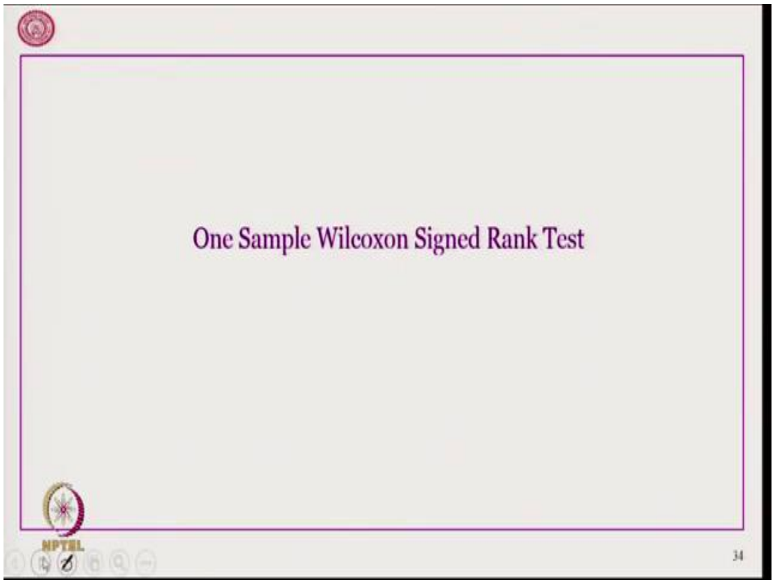
STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592

We go to the normal table and the observed value of the statistic if you look at it, it was minus 1.75. So, we go to the normal table, we look at -1.7 and 0.05. We come to that value which is equal to 0.04006 which is approximately 0.04 which is less than 0.05.

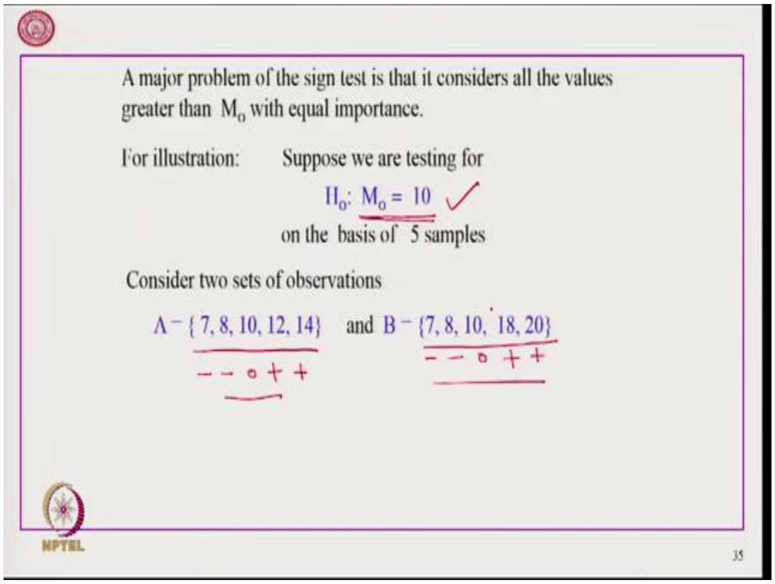
Therefore, we are going to reject the null hypothesis under normal approximation as well. So, as I said it was a very simple test, which is called Sign test and like that we can check whether we can accept the null hypothesis that $M = M_0$.

(Refer Slide Time: 31:43)



Now, let us look at the slightly improved test, which is called Wilcoxon Signed Rank Test. Sign test is very simple, but it has one problem.

(Refer Slide Time: 31:53)

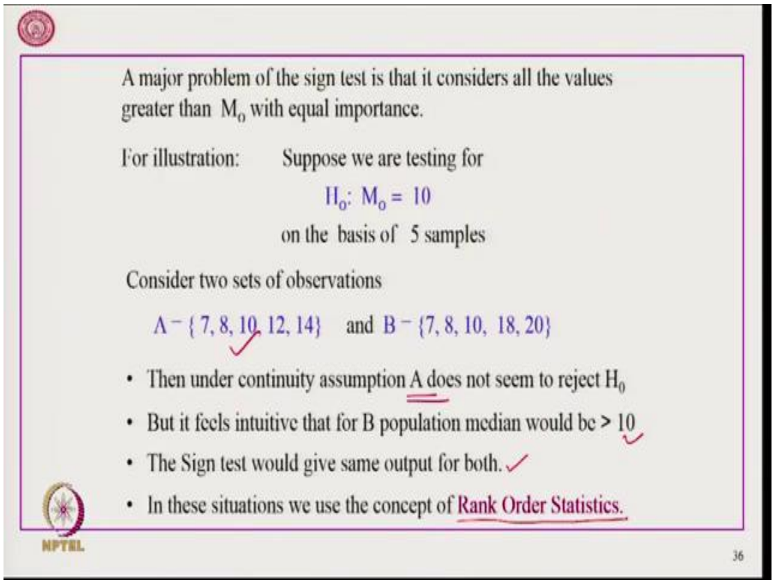


As I illustrate here, suppose with respect to some data, we are testing if $M_0 = 10$ that is the hypothesized median is equal to 10 and we have taken 5 observations. Suppose these observations are 7, 8, 10, 12 and 14. In this case, what is going to happen?

We will have minus, minus, this is 0 we can ignore, plus, plus. So, number of plus signs and number of minus signs are same. And we can feel that therefore we can accept 10 to be the median of the population. Suppose on the other hand my observations are 7, 8, 10, 18, and 20. If we use simple sign tests then again we will get minus, minus, this is 0, plus, plus.

Therefore by Sign test, it also has the same value as this. Hence, we are going to accept the null hypothesis that $M = 10$. However, If we look at this, that is a continuous distribution and observed values are from 7 to 20, then we can easily feel that the median should not be 10 it should be something higher than that, that is a basic intuition.

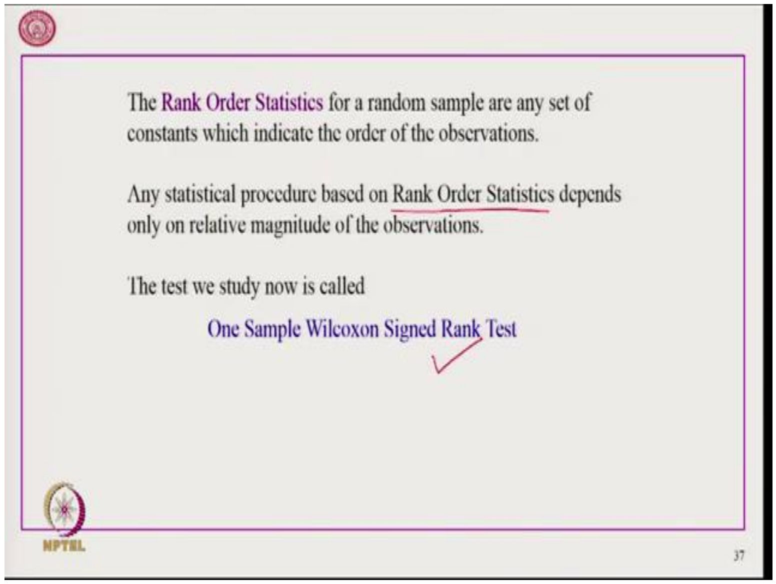
(Refer Slide Time: 33:28)



Hence, what I am saying that under the continuity assumption, it does not seem to reject H_0, because both of them are giving two minus and two plus therefore, A does not seem to reject H_0. This sample was not going to reject H_0, but it feels intuitive that the B population median would be greater than 10 as I have already explained.

But Sign test would give the same output for both as I have already explained, therefore, we need to do something better and we introduced that concept of rank order statistic. What is that?

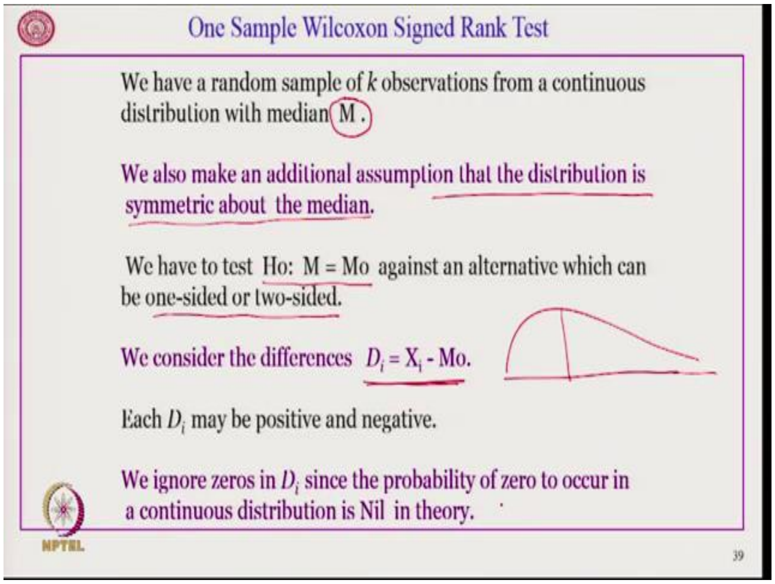
(Refer Slide Time: 34:13)



The rank order statistics for a random sample are any set of constants which indicate the order of the observations. So, we are not just looking at it is plus or minus, we are also looking at the value. That means in this case, whether it is not greater than or less than M_0, we are looking at how much it is greater than M_0 or how much it is less than M_0.

So, that is what we will be computing which is called a rank order statistic. And from this rank order statistic, the basic test that comes up is called Wilcoxon Signed Rank Test.

(Refer Slide Time: 35:00)



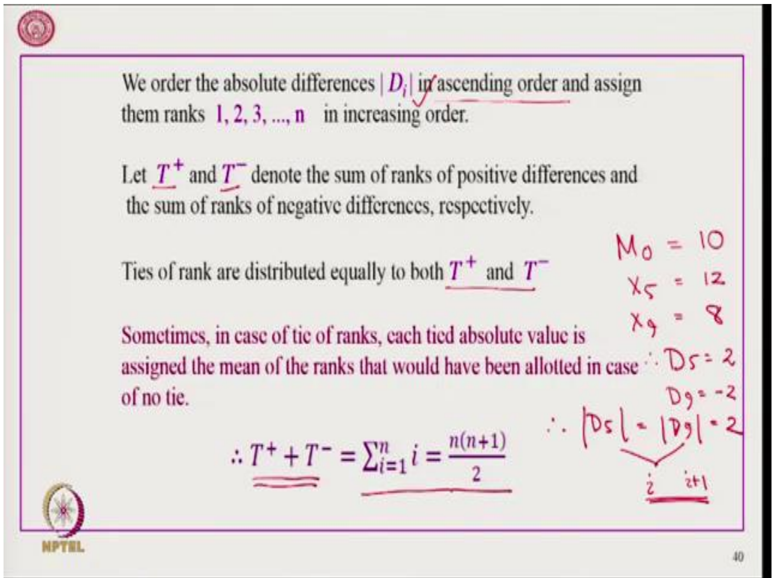
So, let us first study one sample Wilcoxon Signed Rank Test, what it is? So, as before, we have a random sample of k observations from a continuous distribution with a median is equal to M. So this is the population median, we do not know about it, this is what we are trying to estimate, or we are trying to test based on a sample.

But to apply Wilcoxon Signed Rank Test, we need another additional assumption that the distribution is symmetric around the median. It is not necessary that all distributions are symmetric about the median. For example, if it is a positively skewed distribution, then median may come somewhere here and we can see that it is not symmetric around that median, but that will not work for applying Wilcoxon Signed Rank Test.

So this assumption is needed. And we are testing $H_0: M = M_0$ against an alternative which can be one-sided or two-sided this I have already explained. So I am not going to repeat that here, what we do, we considered the difference $D_i = X_i - M_0$. Each D_i may be positive or negative.

Again we ignore zeroes in D_i since the probability of 0 to occur in a continuous distribution is nil in theory. Because it is a continuous distribution, the probability that X is taking the value of M_0 is 0 therefore, we ignore the 0 values when you are considering D_i.

(Refer Slide Time: 36:51)

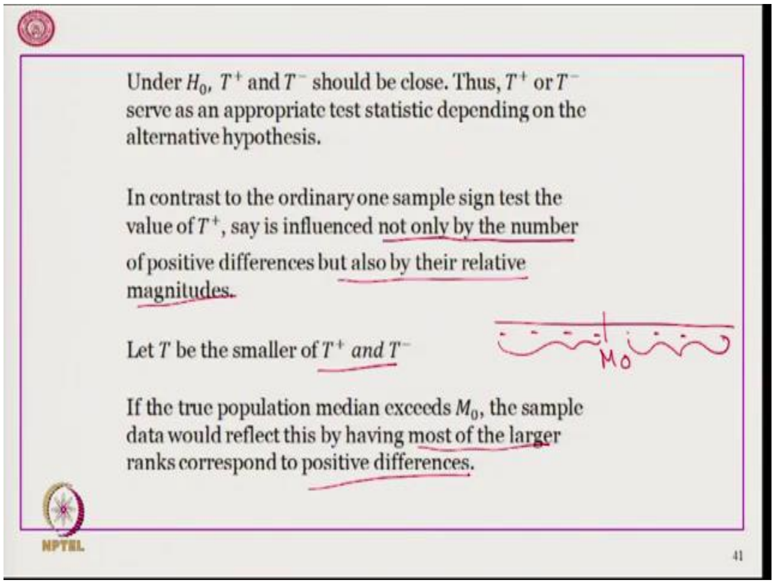


Now, what we do? We look at the absolute value of the difference D_i for all the n observations, $X_1, X_2, \cdots, X_n$. Let T^+ be the sum of ranks of the positive differences and T^- be the sum of ranks of the negative differences, when we arrange all these $D_i's$ in ascending order of magnitude. What will happen? If ties are there, T^+ and T^- say for example, my $M_0 = 10$ and suppose $X_5 = 12$ and suppose $X_9 = 8$.

Therefore, $D_5 = 2$ and $D_9 = -2$. Therefore,$|D_5| = |D_9| = 2$. Therefore, in the ascending order of the $D_i's$ these two will be tally. What to do? What we typically do is we assign the average values there, as I will explain later. So, sometimes in case of tie of ranks, each tied absolute value is assigned the mean of the ranks, that would have been allotted in case of no tie.

So, if there were no tie, perhaps one of them would have got the rank i other one who would have got the rank $i + 1$, then I will give the average of these 2 to both of them. So, that is the idea. Therefore, $T^+ + T^-$ assuming there is no tie or if there is a tie, we break it like that is equal to $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$

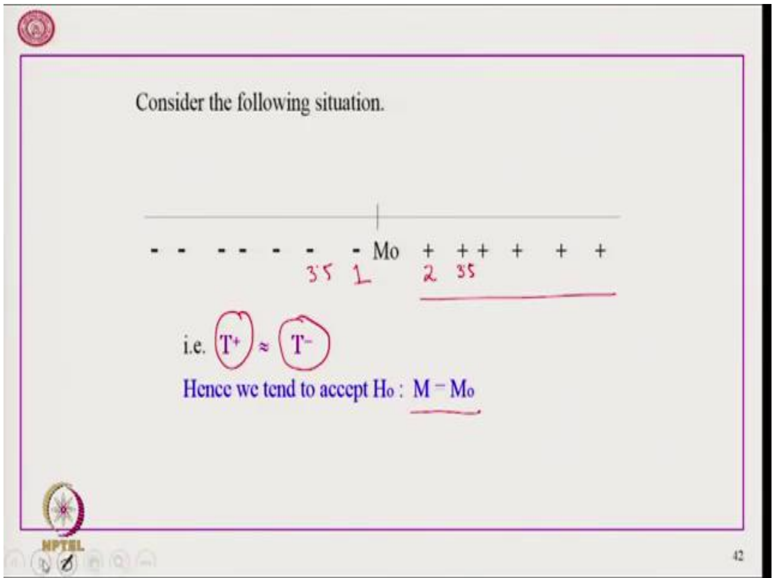
(Refer Slide Time: 38:57)



Now, let us see what happens under H_0 . Under H_0, T^+ and T^- should be close because if this is the M_0 and these are my observations, what I would expect the ranks of the differences is going to be equally distributed. Therefore, the sum of ranks of the negative ones and the sum of ranks of the positive ones should be very close.

So, in contest with ordinary One Sample Sign Tests, the value of T^+ is influenced not only by the number of positive differences, but also by their relative magnitudes. So, I hope you understand that otherwise I will explain later, let T be the smaller of T^+ and T^- . If true population median exceeds M_0 the sample data would reflect this by having most of the larger ranks corresponding to the positive difference.

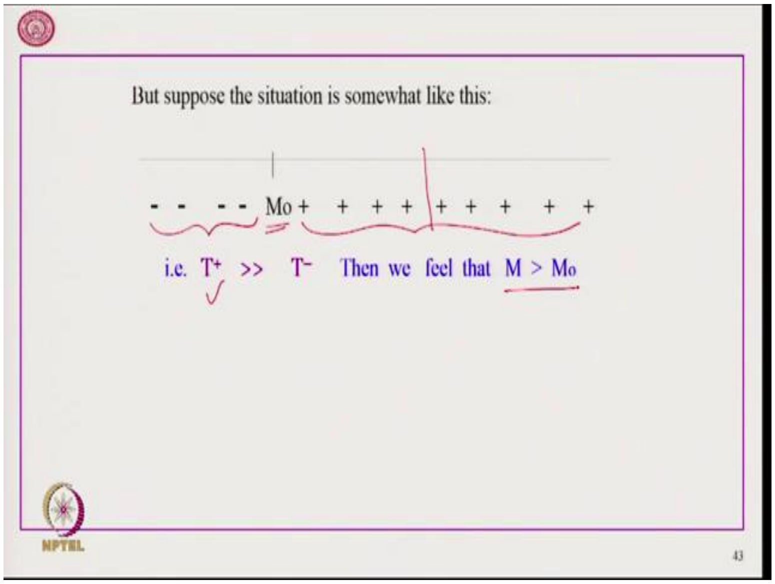
(Refer Slide Time: 39:59)



So let me explain with a very similar diagram. If M_0 is that true median, then in this scenario, perhaps this is going to get the rank 1, this is going to get the rank 2. If both of them are going to get 3 and 3, then I would rather give 3.5 and 3.5. Like that, you would rank all of them.

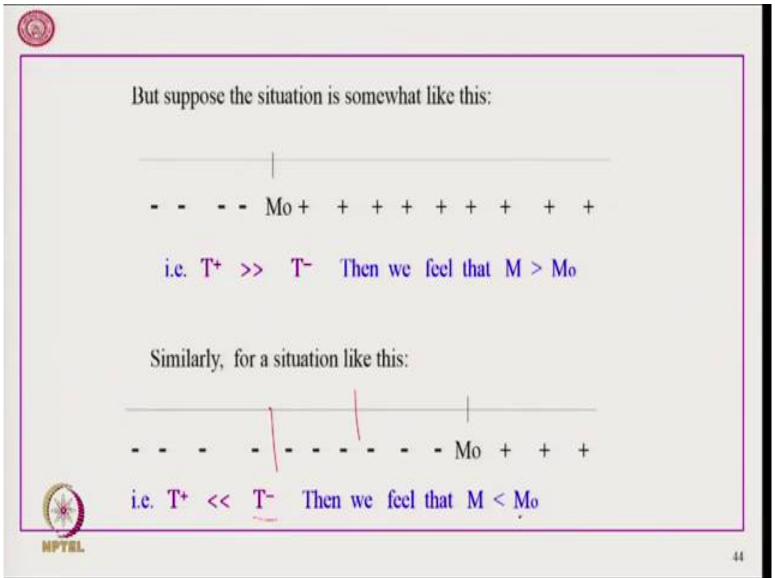
Therefore, as we can understand from this diagram, T plus which is the sum of ranks of the positive ones, is going to be approximately close to the sum of ranks of the negative ones under $M = M_0$

(Refer Slide Time: 40:41)



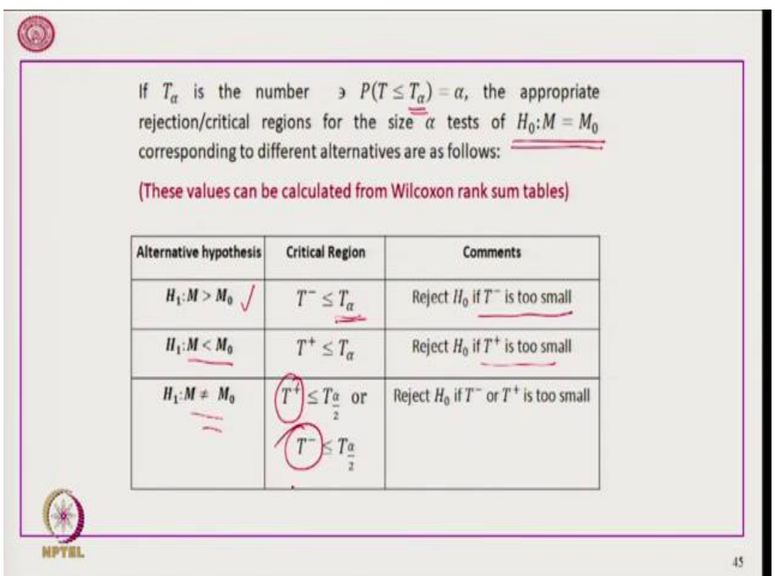
But suppose the situation is somewhat like this, we would expect the alternative is like this, but the observation is such that there are so many positive signs here and so few negative signs here, then what will happen, all the higher rank values are going to be to the positive side. Therefore T^+ is going to be much-much higher than T^+ and then when you would expect the actual median is greater than M_0 .

(Refer Slide Time: 41:15)



In a similar way, if we get a situation like this, then all higher rank values are coming with negative signs, therefore T^+ is going to be much-much less than T^-, and we would expect that actual median is less than M_0 .

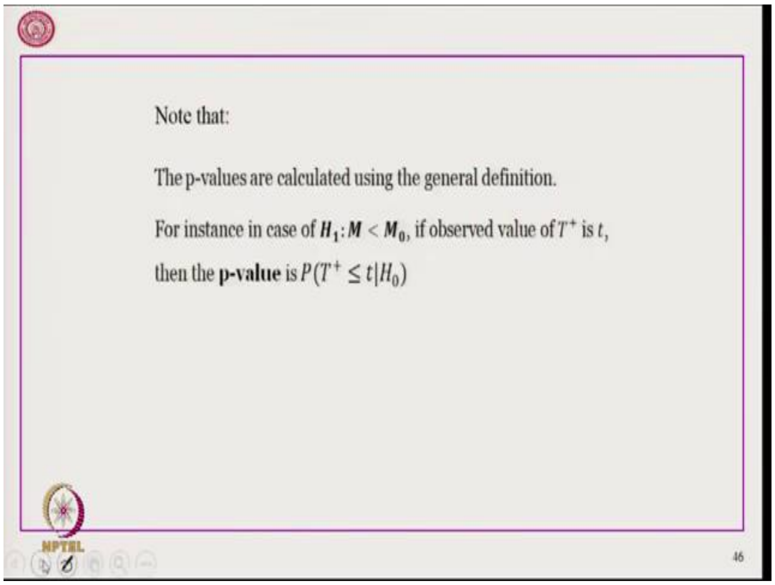
(Refer Slide Time: 41:36)



So, we will have very similar type of testing procedure. Let T_α be the critical point corresponding to α and we are testing $H_0: M = M_0$. So, if the alternative is $H_1: M > M_0$, then we will reject the null hypothesis. If T^- is too small that is T^- is less than the critical value T_α where α can be say 5% or 10% or 1% depending upon how you are going to test.

Similarly, we will reject null hypothesis in favor of $M < M_0$, if T^+ is too small. And for a two-sided test we will look at $T^+ \leq T_{\frac{\alpha}{2}}$ or $T^- \leq T_{\frac{\alpha}{2}}$, that is, if any one of T^+ or T^- is very small, then we are going to reject the null hypothesis.

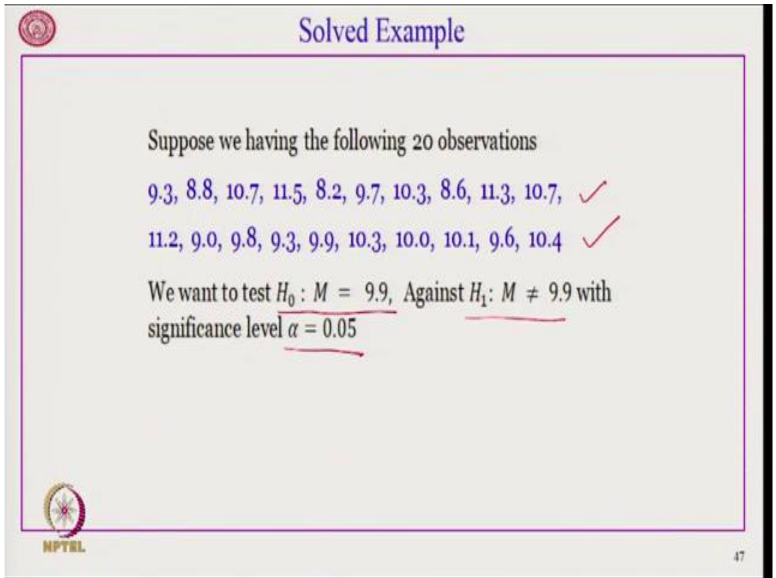
(Refer Slide Time: 42:39)



The p values are calculated using the general definition. So, some of you may not know what is the p value. The p value, if the observed value of the T^+ is t, is $P(T^+ \leq t|H_0)$ under the null hypothesis for instance if $H_1: M < M_0$.

So, let me give you an example.

(Refer Slide Time: 42:54)



Suppose, we have 20 observations, which are like this, we want to test if it is coming from a population with median is equal to 9.9. And we are looking for a two-sided test that H_1 is $M \neq 9.9$ and suppose the significance level is 5% that is $\alpha = 0.05$. How do you solve it?

(Refer Slide Time: 43:21)

X_i	$D_i = X_i - 9.9$	$\lvert D_i \rvert$	Sign (D_i)	Rank$(\lvert D_i \rvert)$
9.3	-0.6	0.6	-	9.5
8.8	-1.1	1.1	-	14
10.7	+0.8	0.8	+	11.5
11.5	+1.6	1.6	+	18
8.2	-1.7	1.7	-	19
9.7	-0.2	0.2	-	3.5
10.3	+0.4	0.4	+	6.5
8.6	-1.3	1.3	-	15.5
11.3	+1.4	1.4	+	17
10.7	+0.8	0.8	+	11.5

X_i	$D_i = X_i - 9.9$	$\lvert D_i \rvert$	Sign (D_i)	Rank$(\lvert D_i \rvert)$
11.2	+1.3	1.3	+	15.5
9.0	-0.9	0.9	-	13
9.8	-0.1	0.1	-	1.5
9.3	-0.6	0.6	-	9.5
9.9	0	0	ignored	0
10.3	+0.4	0.4	+	6.5
10.0	+0.1	0.1	+	1.5
10.1	+0.2	0.2	+	3.5
9.6	-0.3	0.3	-	5
10.4	+0.5	0.5	+	8

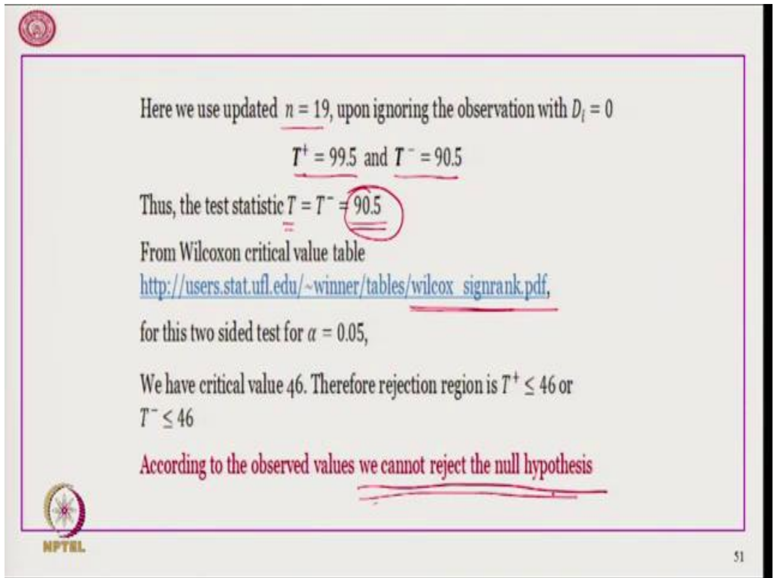
So, let me go to the solution what we do, we will prepare a table like this. So, first value is 9.3, M_0 is 9.9. So, $D_i = -0.6$ therefore what we understand that it is negative, but not only that the value of the absolute difference is 0.6.

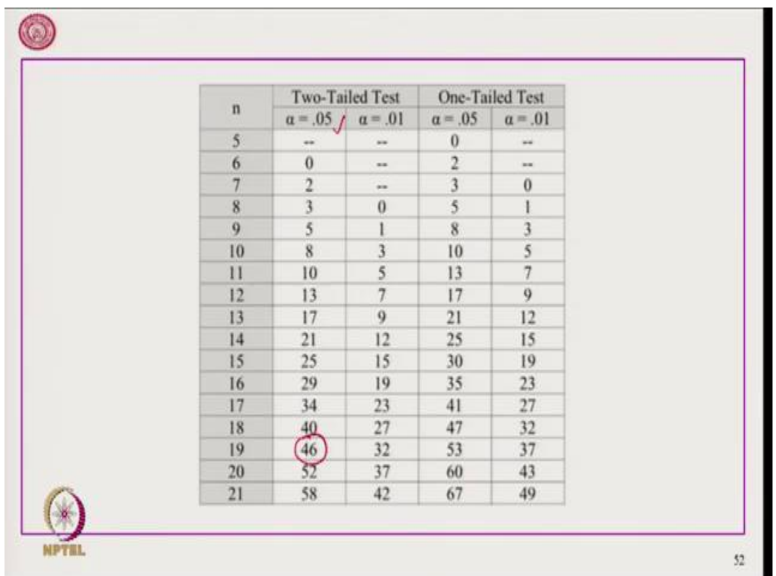
Similarly, for 8.8 the value of D_i is equal to -1.1, on the other hand, if it is 11.5 then the difference is 1.6 it is a plus and the absolute value the D_i and like that we calculate it for all the 20 values and we get the plus and minus sign also we get the absolute value of the D_i and then we rank them and we have broken the ties by assigning equal weight to them.

Say for example 9.8 and 10, one is giving me D_i is equal to -0.1 other is giving me the value to be $+0.1$ and that is the smallest. So one should get the value 1 other should get the value 2, we are equally distributing them between these two, so both of them get the rank 1.5.

Similarly, you see that if one observed value is 9.9 which is equal to the M_0 in that case we are ignoring this and we are giving a value 0. So that observation will have to go out of my consideration. Therefore, we will be working with only 19 observations.

(Refer Slide Time: 45:11)





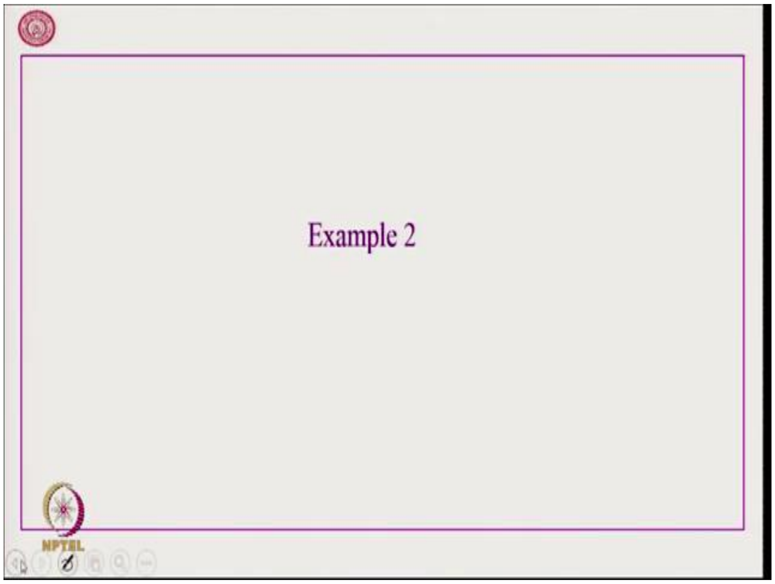
n	Two-Tailed Test		One-Tailed Test	
	α = .05	α = .01	α = .05	α = .01
5	--	--	0	--
6	0	--	2	--
7	2	--	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15
15	25	15	30	19
16	29	19	35	23
17	34	23	41	27
18	40	27	47	32
19	46	32	53	37
20	52	37	60	43
21	58	42	67	49

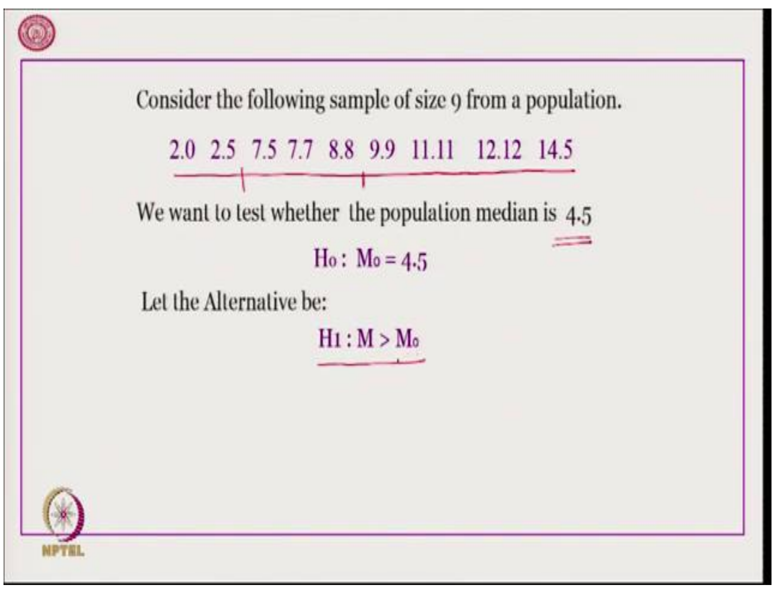
Now, what we do, the updated value of n is equal to 19. Now, we calculate the sum of ranks for positive ones which is 99.5 the sum of ranks of the negative ones, which is 90.5. So, typically we take minimum of these, so, in this case that test statistic T is equal to 90.5.

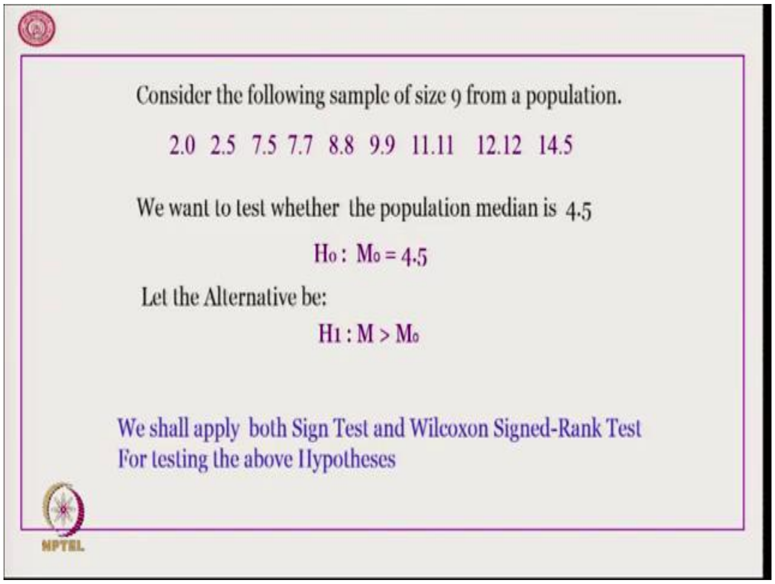
Now, the question comes whether we should accept the null hypothesis or reject therefore, we go to the table, in this case Wilcoxon signed rank table which is available in this location. This is available over the internet you can refer to that but for your understanding I am showing some portion of that table; n is equal to 19 at 5% level of significance it is a two-tailed test because we were testing.

Therefore, the critical value is coming out to be 46. Therefore, we shall reject the null hypothesis, if that statistic T comes out to be less than 46. But what we have got, we have got the value 90.5 which is greater than that, therefore, we cannot reject the null hypothesis. So, that is the conclusion after this test that we cannot reject this null hypothesis if that is the sample that we have got.

(Refer Slide Time: 46:46)



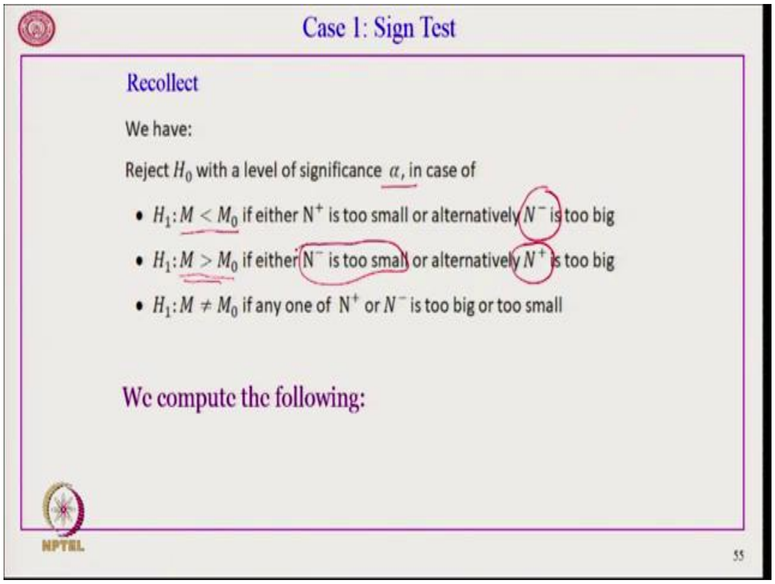




Let me now give you a very interesting example involving Sign test and Signed rank test consider, the following sample of size 9 from a population. So, the data points are 2, 2.5, 7.5 like that up to 14.5. And we want to test whether the population median is 4.5. That means, we are trying to check if this data is good enough to support the claim that the data has come from a population whose median is 4.5 that is somewhere here, what is going to be the alternative?

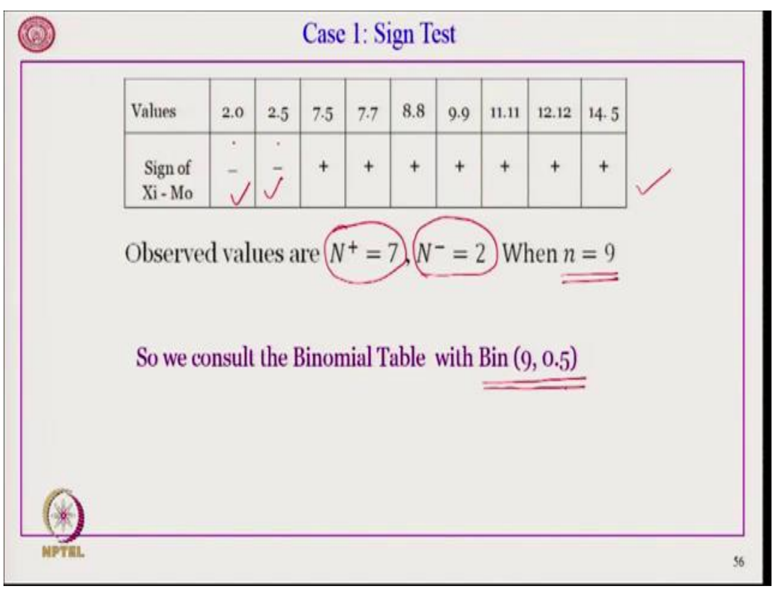
Looking at the data, it appears that it should not be at 4.5 rather, it should be somewhere bigger than that is 4.5 maybe somewhere like this. Therefore, quite naturally, the alternative is going to be $M > M_0$. So, we shall apply both sign test and Wilcoxon Signed Rank Test for testing the above hypothesis.

(Refer Slide Time: 48:05)



So, let us first consider sign test. We know that we reject H_0 with a level of significance α, if $M < M_0$ is my alternative hypothesis, then we shall look at N^- and that should not be too big. But if $M > M_0$ is the hypothesis, then we shall look at if N^+ is too big or N^- is too small. So, let us compute the following.

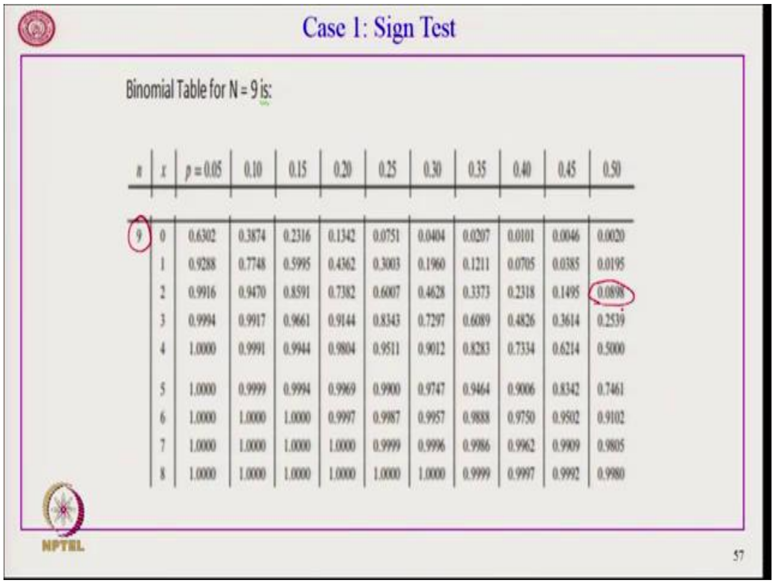
(Refer Slide Time: 48:42)

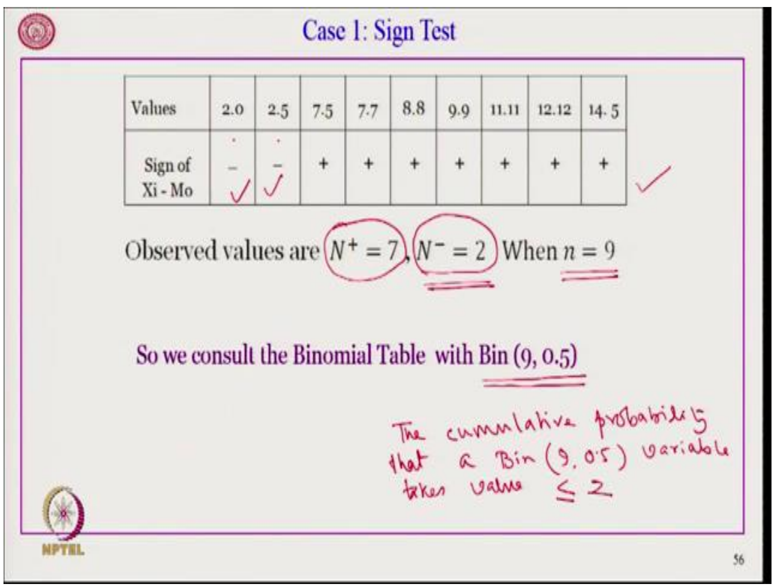


In Sign test we are just taking the difference from the observed value of the median and looking at the sign. Therefore, for 2.0 and 2.5 we have got negative sign and for all remaining ones, we have got positive sign because, number of observations is 9. Therefore, $N^+ = 7$ and N^- is equal to 2.

Therefore, what we have to do, we have to look at binomial table with parameters 9 and 0.5. And to see whether we can accept the null hypothesis or reject.

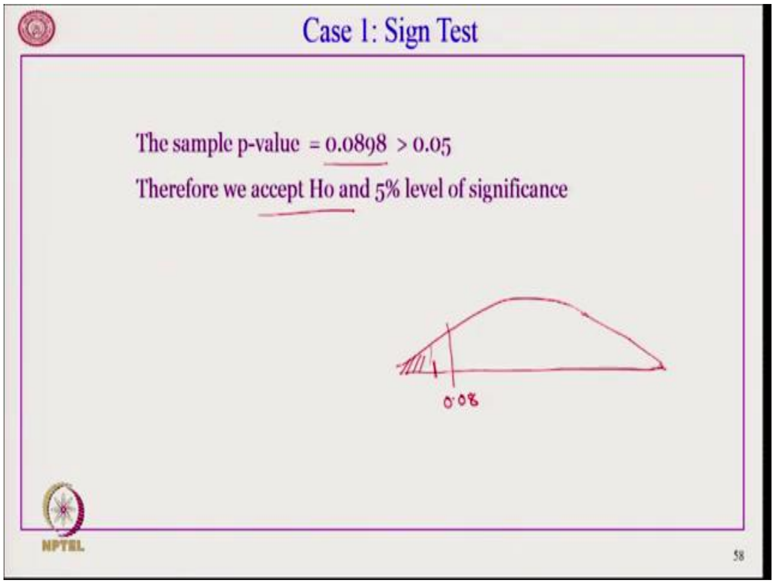
(Refer Slide Time: 49:33)





So, this is the binomial table and we have got N^- is equal to 2. So, we are looking at the cumulative probability that is, binomial variable $Bin\ (9,0.5)$ takes value less than equal to 2. From the tables we get when n is equal to 9, the cumulative probability for 2 is 0.0898.

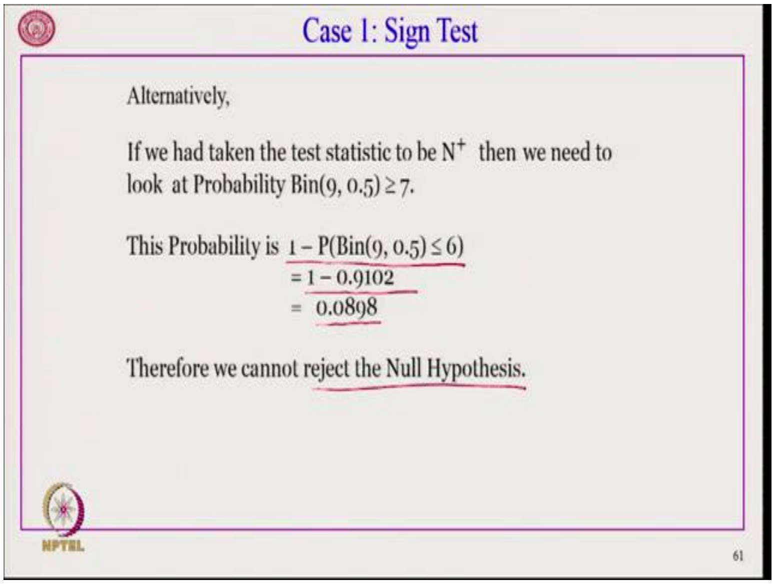
(Refer Slide Time: 50:27)

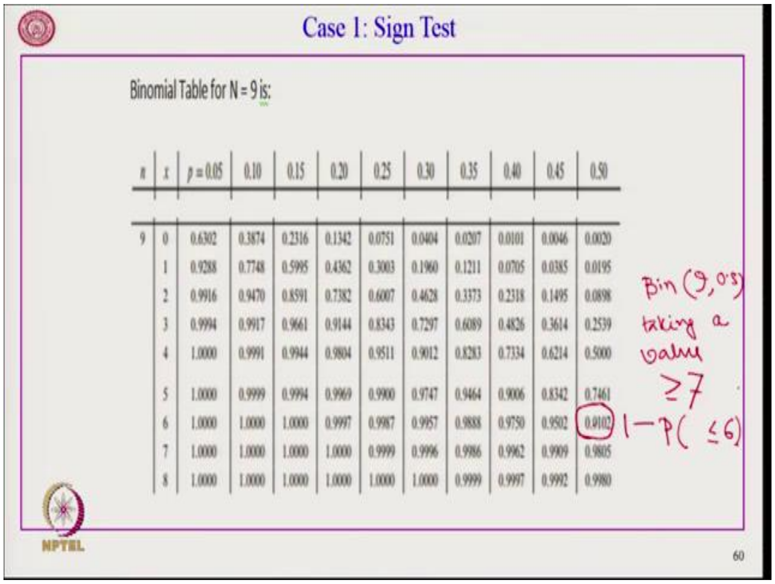


Since this sample p value, which is 0.0898 is greater than 0.05. Therefore, we accept H_0 at 5% level of significance. So, this I have explained to you, let me just illustrate it once more. So, suppose it is the binomial distribution with 0.5 it is going to be symmetric. So, p value is 0.08 but the critical value for 5% level is somewhere here.

So if the value is lesser than this then only we will reject the null hypothesis, but that did not happen in our case. Therefore, we accept the H_0 or we cannot reject the H_0 against the alternative H_1.

(Refer Slide Time: 51:25)



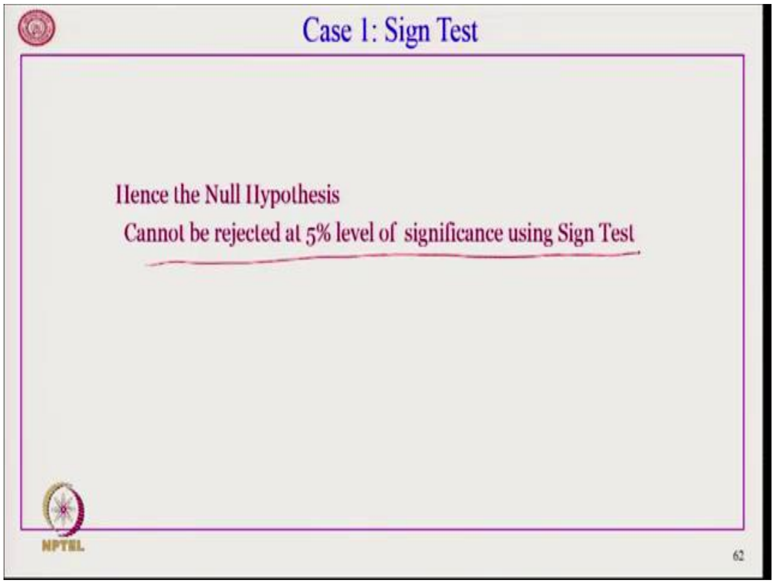


n	x	p = 0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
9	0	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
	1	0.9288	0.7748	0.5995	0.4362	0.3003	0.1960	0.1211	0.0705	0.0385	0.0195
	2	0.9916	0.9470	0.8591	0.7382	0.6007	0.4628	0.3373	0.2318	0.1495	0.0898
	3	0.9994	0.9917	0.9661	0.9144	0.8343	0.7297	0.6089	0.4826	0.3614	0.2539
	4	1.0000	0.9991	0.9944	0.9804	0.9511	0.9012	0.8283	0.7334	0.6214	0.5000
	5	1.0000	0.9999	0.9994	0.9969	0.9900	0.9747	0.9464	0.9006	0.8342	0.7461
	6	1.0000	1.0000	1.0000	0.9997	0.9987	0.9957	0.9888	0.9750	0.9502	0.9102
	7	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9986	0.9962	0.9909	0.9805
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9980

Alternatively, if we had taken that test statistic to be N^+, then we will have to look at the probability that a binomial variable $Bin(9,0.5)$ is taking value greater than equal to 7. So, let us look at the binomial table again till 7 we need to find out the probability that a binomial Bin (9,0.5) taking a value greater than or equal to 7 which is nothing but 1 minus the probability that the random variable is less than equal to 6.

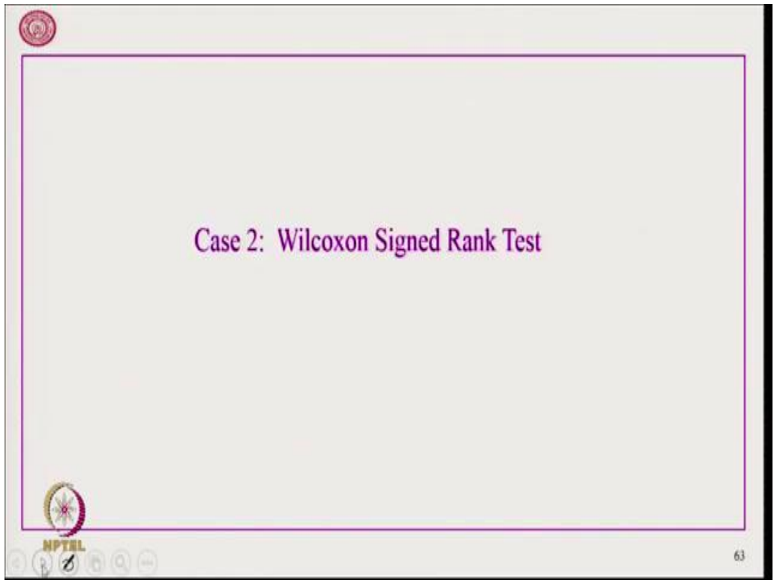
Therefore, we look at the cumulative probability that a binomial $Bin(9,0.5)$ variable is taking the value less than equal to 6. So, that value if we look at the table carefully, is given us 0.9102. Therefore, probability is 1 minus probability that a $Bin(9,0.5)$ variable is less than equal to 6 is equal to $1 - 0.9102$ which is coming out to the 0.0898. Therefore, again by the same logic, we cannot reject the null hypothesis.

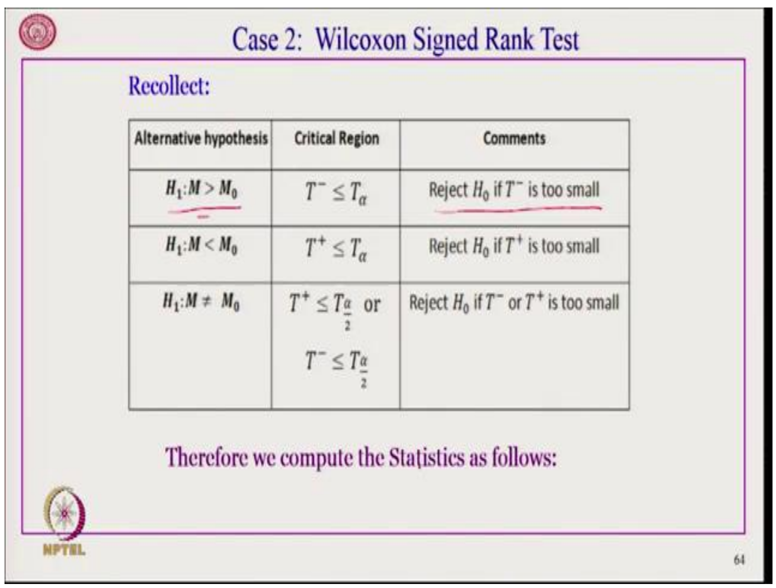
(Refer Slide Time: 53:20)



That means that if we are going to use Sign test, we are not able to reject the null hypothesis that is what we have written, the outcome of the testing is that the null hypothesis cannot be rejected at 5% level of significance using Sign test.

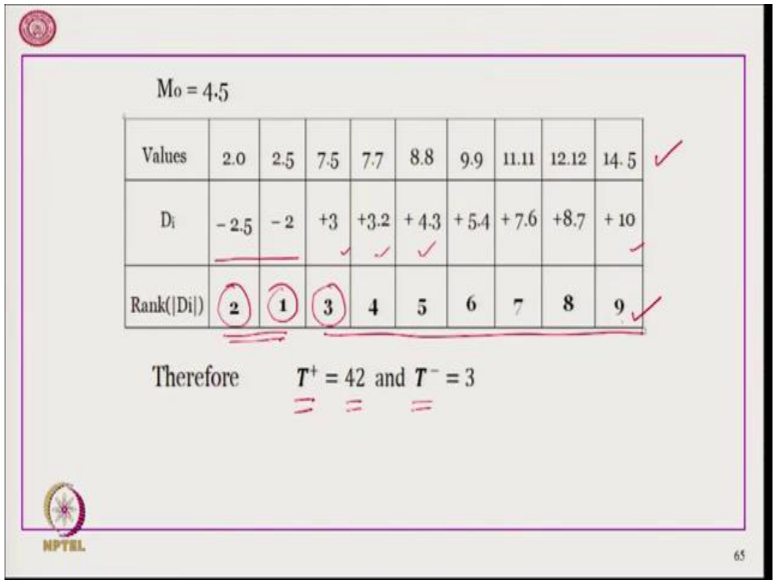
(Refer Slide Time: 53:31)





Now, let us look at what happens when we use Wilcoxon Signed Rank Test, we already know this thing, this you have already seen that the critical region for alternative $M > M_0$ is going to be that we reject H_0 if T^- is too small. So, T^- is the sum of ranks of the negative $D_i's$. Therefore, T^- is the sum of ranks of all those differences which are negative that means which are less than the median M_0 . Therefore, we compute the statistic as follows.

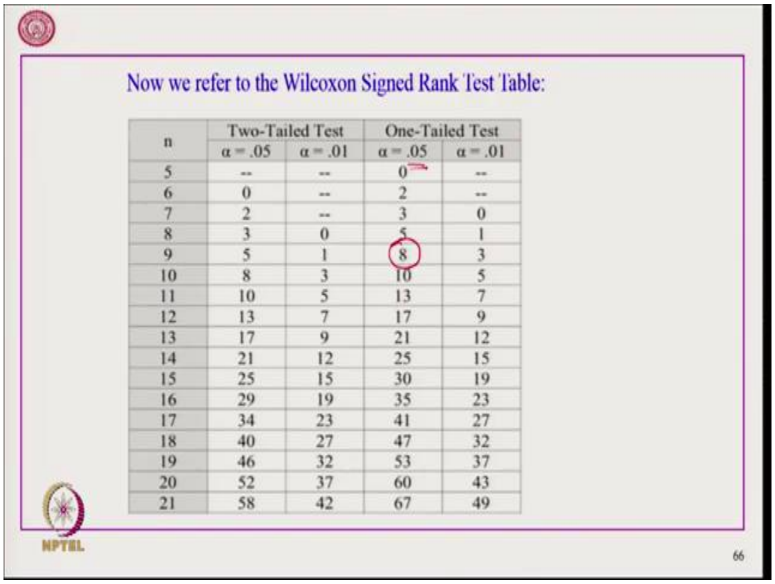
(Refer Slide Time: 54:16)



We have the values given like this, then we are subtracting 4.5 from all of them, therefore, we got these two are negative, all these are positive, but not only that, we have to look at their ranking in absolute value terms. So, 2 is the smallest therefore, this is the getting the value 1, 2.5 is the second smallest it is getting the value 2, 3 third smallest it is getting the value 3 and like that, till this point we have got 9.

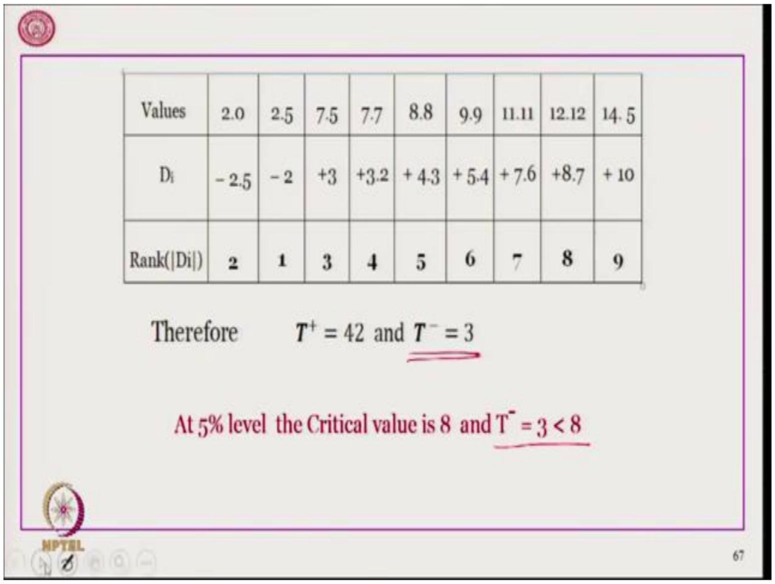
Therefore, T^+ is equal to sum of ranks of all those ones which are positive and that is there what is going to be 3 plus 4 plus 5 up to 9 which is coming out to be 42. On the other hand $T^- = 3$, because it is the sum of 2 and 1.

(Refer Slide Time: 55:25)



Now we refer to the Wilcoxon Signed Rank Test Table:

n	Two-Tailed Test		One-Tailed Test	
	α = .05	α = .01	α = .05	α = .01
5	--	--	0	--
6	0	--	2	--
7	2	--	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15
15	25	15	30	19
16	29	19	35	23
17	34	23	41	27
18	40	27	47	32
19	46	32	53	37
20	52	37	60	43
21	58	42	67	49



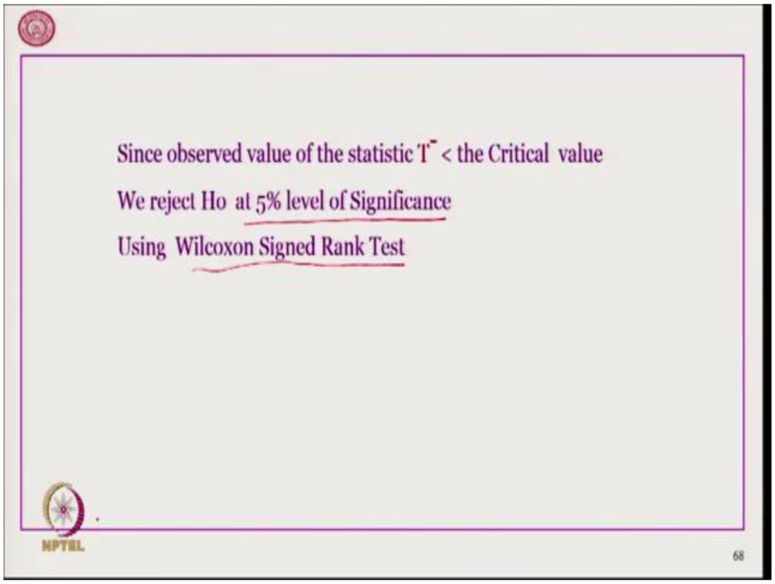
Values	2.0	2.5	7.5	7.7	8.8	9.9	11.11	12.12	14.5
D_i	-2.5	-2	+3	+3.2	+4.3	+5.4	+7.6	+8.7	+10
Rank($\lvert D_i \rvert$)	2	1	3	4	5	6	7	8	9

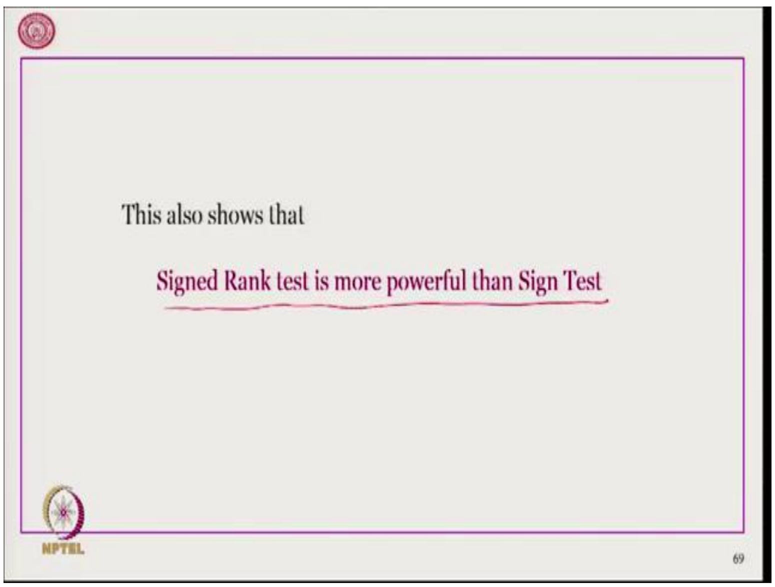
Therefore $T^+ = 42$ and $T^- = 3$

At 5% level the Critical value is 8 and $T^- = 3 < 8$

Therefore, we look at now, this one-tailed test with 9, and here the critical value is giving out to be 8 for alpha is equal to 0.05. Our observed value was $T^- = 3$. Therefore, at 5% level, the critical value is 8, but T^- is less than that.

(Refer Slide Time: 55:56)

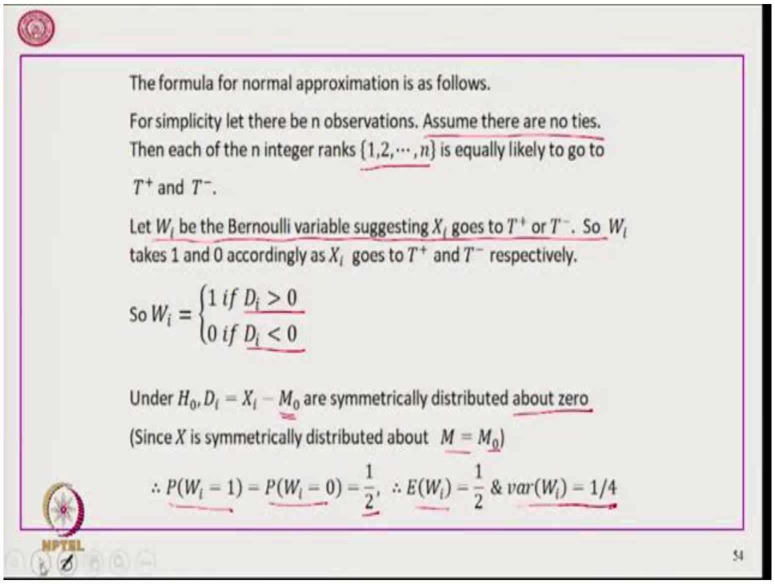


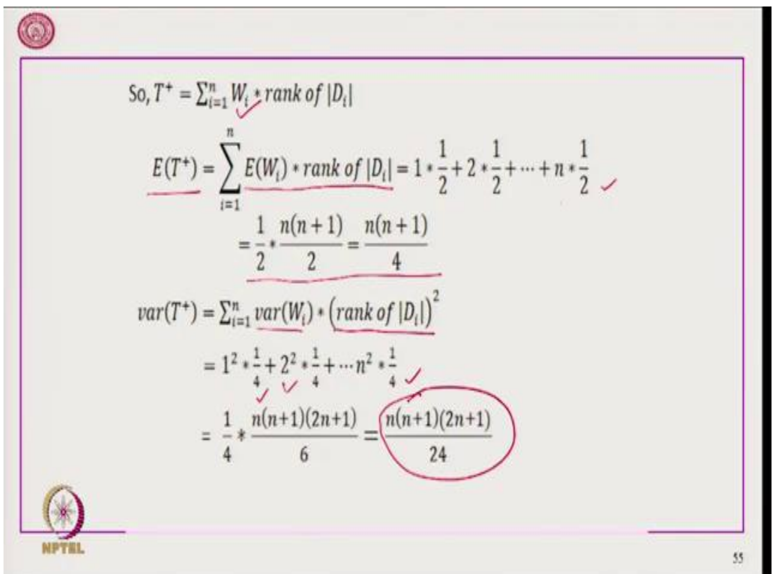


And therefore, what we do we have to reject the H_0 at 5% level of significance using Wilcoxon Signed Rank Test. That is why I said that this is a very interesting example, because when we are using sign test, we are going to accept the null hypothesis at 5% level of significance, but when we are using signed rank test, we are able to reject that one.

And if we had looked at the data, we found that it is natural intuition is that 4.5 cannot be the median of that test. Therefore, in that sense, signed rank is more powerful than signed test.

(Refer Slide Time: 56:44)





Normal approximations like the case of sign test, here also we can do normal approximation if the sample size is somewhat large. So the formula for normal approximation is as follows. So, let us assume that n observations are there. I assume that there are no ties. Therefore, all the n, D_i will have different ranks and they are going to be 1, 2, 3 up to n.

Now, each of those observation is equally likely to go to either the positive side or the negative side because it is the median that we are testing. So, with probability half it will be greater than median and with probability half, it is going to be less than median. So, let us consider the W_i corresponding to each observation.

So, if X_i is observation and W_i is something which is a Bernoulli random variable which is giving us the value if it is going to the positive side or it is going to be a negative side. So, $W_i = 1$ if $D_i > 0$ and $W_i = 0$ if $D_i < 0$, a typical Bernoulli distribution. So, under H_0 the median is equal to actually M_0 , $X_i - M_0$ is symmetrically distributed about 0. Since X is symmetrically distributed about $M = M_0$.

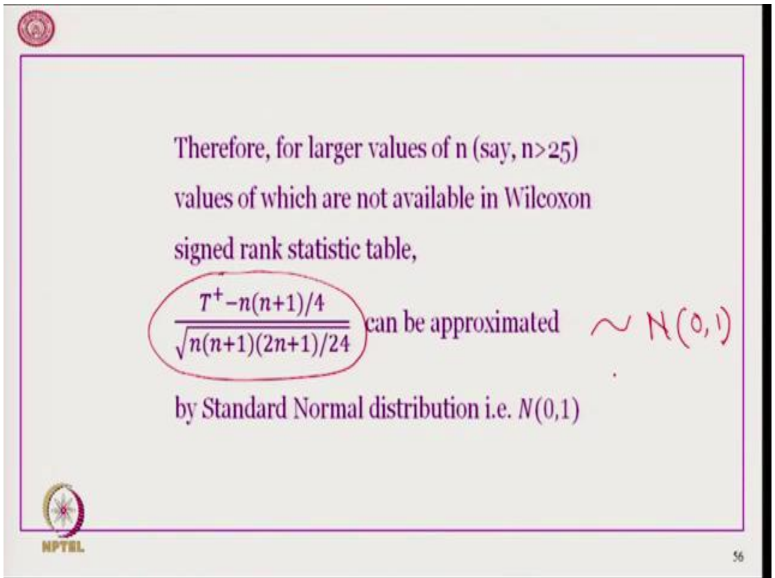
Therefore, $P(W_i = 1) = P(W_i = 0) = \frac{1}{2}$.

Therefore, the expected value of W_i is equal to $1/2$ and variance of W_i is equal to $1/4$ that we know from the definition of Bernoulli random variables. Therefore, $T^+ = \sum_{i=1}^{n} W_i * rank\ of\ |D_i|$, what it is going to do?

It is going to consider only those which are positive and adding up their ranks. If it is less than the median, the value of $W_i = 0$. So, therefore, it is not coming to the summation. Therefore $E(T^+) = \sum_{i=1}^{n} E(W_i) * rank\ of\ |D_i| = 1 * \frac{1}{2} + 2 * \frac{1}{2} + \cdots + n * \frac{1}{2}$, because of the linearity of the expectation which is equal to $\frac{n(n+1)}{4}$.

In a similar way, therefore, $var(T^+) = \sum_{i=1}^{n} var(W_i) * (rank\ of\ |D_i|)^2$ because the observations are independent. Therefore, we are not considering any covariance terms. Therefore, it is $1^2 * \frac{1}{4} + 2^2 * \frac{1}{4} + \cdots n^2 * \frac{1}{4}$. So, this is coming out to be $\frac{n(n+1)(2n+1)}{24}$.

(Refer Slide Time: 60:04)



Therefore, for a larger values of n say n is greater than 25, values of which are not available in the Wilcoxon signed rank statistic table, we have to approximate it with the following statistic $\frac{T^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$, where $n(n+1)/4$ is the mean and $\frac{n(n+1)(2n+1)}{24}$ is the variance so its square root is giving us the standard deviation. Therefore, this whole thing is distributed as standard normal i.e. $N(0,1)$.

Therefore, for a larger values of n we can convert the statistic to a normal $N(0,1)$ distribution. And from there by reference to the normal table, we can accept or reject the null hypothesis. Okay friends, I stop here today. I hope I could make the concept clear about how to apply Sign test and signed rank test for testing the central location of unknown distribution.

As I said at the beginning that similar tests can be used for paired data, but for that you have to understand what a paired sample is and what is the significance of a median there and how we can apply the sign test and signed rank test for the hypothesis with respect to paired sample. Okay students, in the next class, I shall start with that. Thank you.