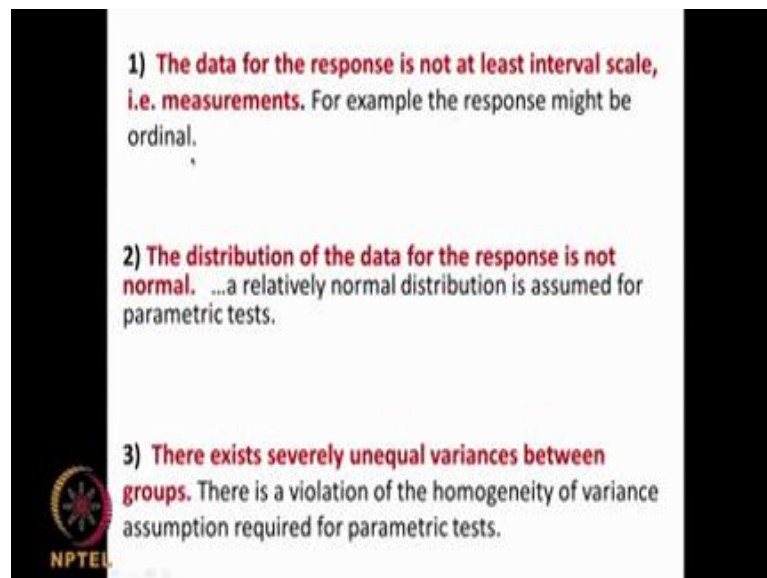


**Biostatistics and Design of Experiments**  
**Prof. Mukesh Doble**  
**Department of Biotechnology**  
**Indian Institute of Technology, Madras**

**Lecture - 29**  
**Nonparametric tests/Homogeneity of variance/Beta distribution**

Welcome to the course on a Biostatistics and Design of Experiments. We will continue on the topic of Nonparametric test. Yesterday I talked about Nonparametric test and generally these Nonparametric test is used when the data is ordinal, that means we do not have x axis.

(Refer Slide Time: 00:29)




Like for example, changes in a parameter as a function of time or change and so on actually. So, it is more of ordinal it is numbers and next one is if the distribution does not follow you say a Normal distribution or a **Chi square** distribution or a F distribution or T distribution, if the variances are unequal, so all these conditions we cannot use the parametric test. We have spent lot of time on this parametric test like your F test, T test and **Chi square** test, Z test and so on. So, we cannot use any one of these and we need to resort to nonparametric test.

(Refer Slide Time: 01:15)

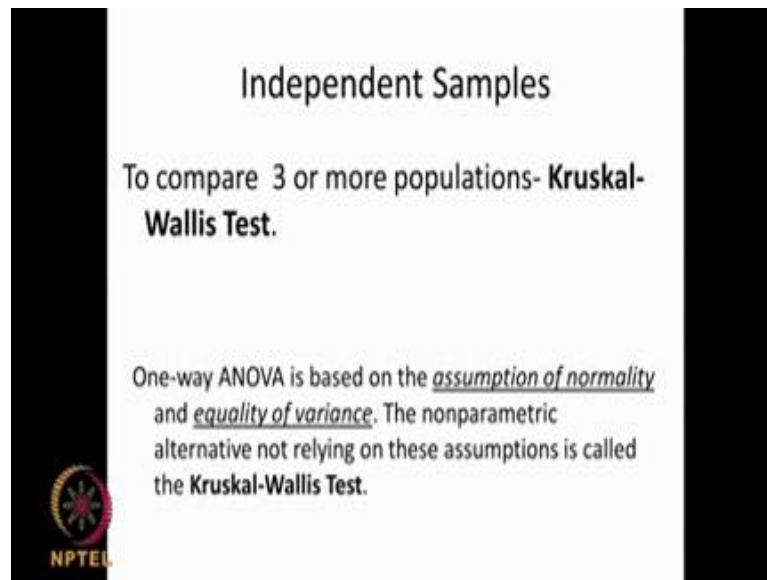
**Table of Parametric & Nonparametric Tests**

| Purpose of Test                 | Parametric Test                 | Nonparametric Test                     |
|---------------------------------|---------------------------------|--|
| Compare two independent samples | Two-Sample t-Test (either case) | Mann-Whitney/Wilcoxon Rank Sum Test    |
| Compare dependent samples       | Paired t-Test                   | Sign Test or Wilcoxon Signed-Rank Test |
| Compare k-independent samples   | One-way ANOVA                   | Kruskal-Wallis Test                    |



What are those Nonparametric tests? We have these on equivalent to a two-sample t-Test, if you are comparing 2 independent samples there is something called Mann-Whitney/Wilcoxon Rank Sum Test. Then if you are doing the Paired t-Test that means you are using the same subjects for say, control and test it is comparing dependent variables there is something called Sign Test or Wilcoxon Signed Rank Test. If you are doing a One-way ANOVA equivalent, that is comparing independent samples then there is something called Kruskal-Wallis Test. Yesterday in the previous class we looked at these two-rank sum test, sign test and signed rank test and so on. Now let us look at equivalent to your one-way ANOVA. What do we in One-way ANOVA? We have several sets of samples and we are trying to compare them. If you have the homogeneity of variance we can use ANOVA, but otherwise then we need to resort to something called Kruskal-Wallis test.


(Refer Slide Time: 02:19)



**Independent Samples**

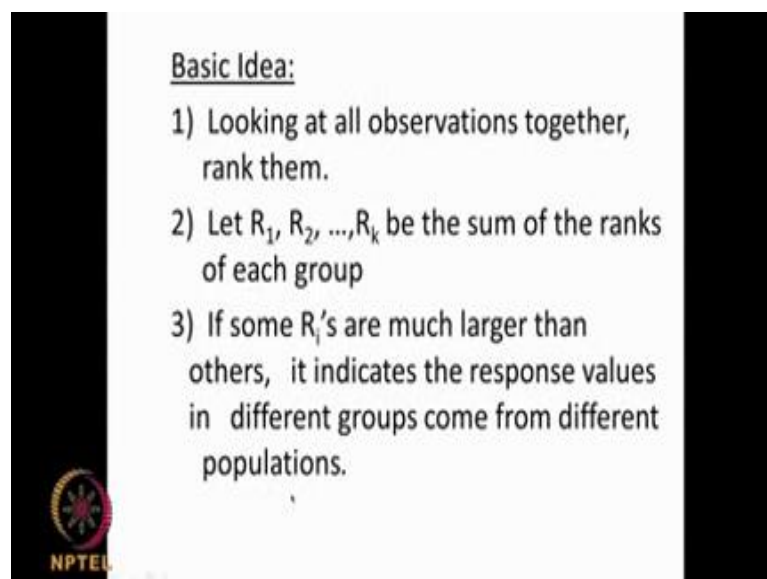
To compare 3 or more populations- **Kruskal-Wallis Test.**

One-way ANOVA is based on the assumption of normality and equality of variance. The nonparametric alternative not relying on these assumptions is called the **Kruskal-Wallis Test.**




So, if you want to compare 3 or more populations. If you have 2 or more, 2 population then of course we can use this Mann-Whitney/Wilcoxon Rank Sum Test or if it is Paired then we can use the Signed Test and so on. But if we are having 3 or more populations then we go resort to something called Kruskal Wallis Test. Because One-way ANOVA generally assumes normality or equivalent equality of variance or homogeneity of variance, so in a nonparametric situation we use this particular test.

(Refer Slide Time: 02:52)



Basic Idea:

- 1) Looking at all observations together, rank them.
- 2) Let  $R_1, R_2, \dots, R_k$  be the sum of the ranks of each group
- 3) If some  $R_i$ 's are much larger than others, it indicates the response values in different groups come from different populations.



What does it do? So, it looks at all the observations and then ranks them, and then once it ranks them we will sum up each of the group ranks, you understand?. We will sum up each of the group ranks and we will get say summation  $R_1, R_2, R_3$ . If some of these R's are larger than others, then it indicates the respond values in different groups come from different populations so that what it is. So, what we do is we put all the data together, we rank them in an ascending order and then for each group we sum up all the ranks and then we compare these ranks using some statistic, let us look at a problem.

(Refer Slide Time: 03:39)

The test statistic is

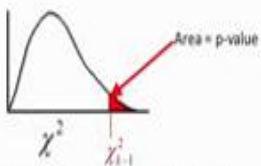
$$H = \frac{12}{N(N+1)} \sum_{i=1}^k n_i \left( \frac{R_i}{n_i} - \frac{N+1}{2} \right)^2 \sim \chi_{k-1}^2$$

where,

$N$  = total sample size =  $n_1 + n_2 + \dots + n_k$

$\frac{R_i}{n_i}$  = average rank for group  $i$

$\frac{N+1}{2}$  = average overall rank



Under the  $H_0$ , this has an approximate chi-square distribution with  $df = k - 1$ , i.e. the approximation is OK when each group contains at least 5 observations.

Before that this is the test statistic. So,

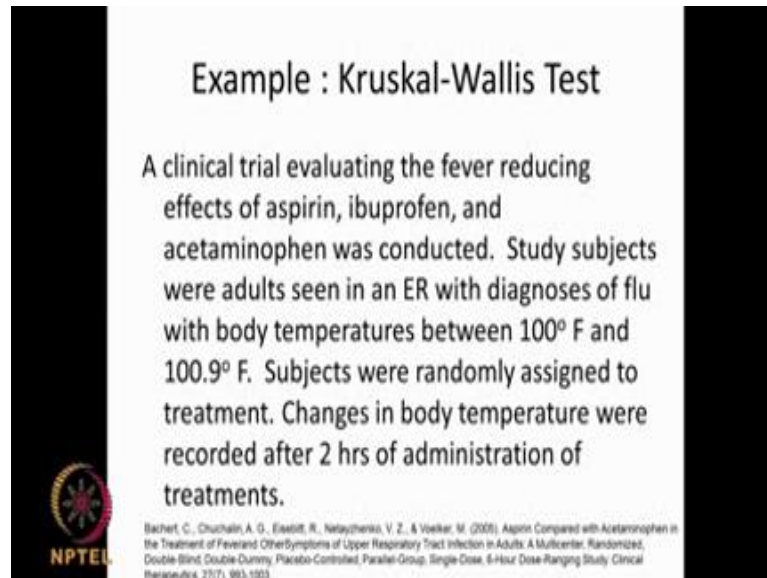
$$H = \frac{12}{N(N+1)}$$

where  $N$  is the total sample size, so if you have a say many groups each one having  $n_1, n_2, n_k$  sample size we add up all that gives you  $N$ . And then here we have a summation  $i = 1$  to  $k$ ,  $n_i$  that is the number of samples in that  $i^{\text{th}}$  data set,  $R_i / n_i$  is the average rank for group  $i - N + 1 / 2$  this is the test statistics. So, what we do under the null hypothesis, this is an approximate chi-square distribution with degrees of freedom

$$df = k - 1$$

, that is the approximation is ok when each group contains at least 5 observations. We compare this test statistics using the **chi square** distribution.

(Refer Slide Time: 04:44)



**Example : Kruskal-Wallis Test**

A clinical trial evaluating the fever reducing effects of aspirin, ibuprofen, and acetaminophen was conducted. Study subjects were adults seen in an ER with diagnoses of flu with body temperatures between 100° F and 100.9° F. Subjects were randomly assigned to treatment. Changes in body temperature were recorded after 2 hrs of administration of treatments.

Bachert, C., Chuchalin, A. G., Elisabitt, R., Nelaychenko, V. Z., & Voecker, M. (2008). Aspirin Compared with Acetaminophen in the Treatment of Fever and Other Symptoms of Upper Respiratory Tract Infection in Adults: A Multicenter, Randomized, Double-Blind, Double-Dummy, Placebo-Controlled, Parallel-Group, Single-Dose, 6-Hour Dose-Ranging Study. *Clinical Therapeutics*, 27(7), 993-1003.

Let us look at an example, this example was taken up from this particular paper which called Clinical Therapeutics. A clinical trial evaluating the fever reducing effect of 3 drugs are tested- aspirin, ibuprofen and acetaminophen. So, it was given to adults and they were tested and when they had a body temperature of 100 to 100.9. The subjects were randomly given either aspirin or ibuprofen or acetaminophen and after 2 hours, their body temperature was again recorded and then it is listed. 3 drugs so it is very random set of data.

(Refer Slide Time: 05:28)

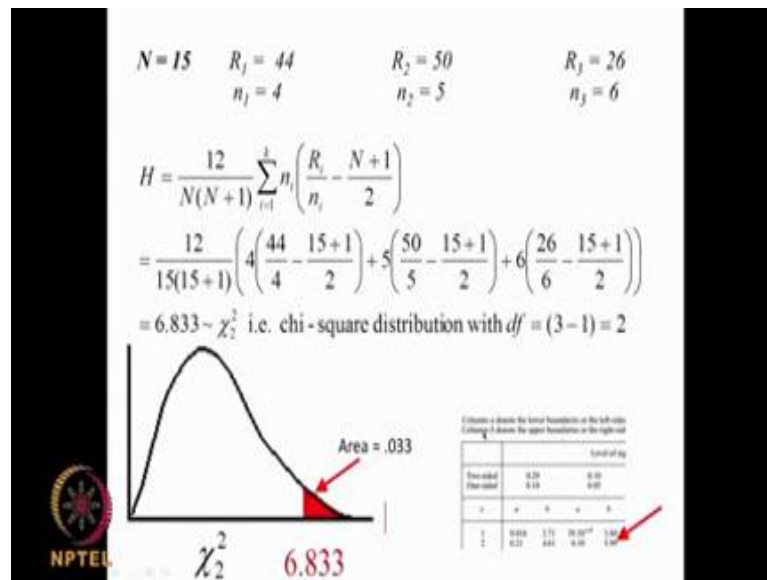
Data: Temperature Decrease (deg. F)

| Aspirin | Rank | Ibuprofen | Rank | Acetaminophen                | Rank |
|---------|------|-----------|------|------------------------------|------|
| .95     | 8    | .39       | 5    | .19                          | 4    |
| 1.48    | 14   | .44       | 6    | 1.02                         | 9    |
| 1.33    | 12   | 1.31      | 11   | .07                          | 3    |
| 1.28    | 10   | 2.48      | 15   | .01                          | 2    |
|         |      | 1.39      | 13   | .62                          | 7    |
|         |      |           |      | -.39<br>(i.e. temp increase) | 1    |

$N = 15$      $R_1 = 44$      $R_2 = 50$      $R_3 = 26$   
 $n_1 = 4$      $n_2 = 5$      $n_3 = 6$

The temperature decrease is shown here, of course in this particular case there is a temperature increase. So, Aspirin, we had 4 candidates they had temperature decrease of these numbers, ibuprofen there are 5 candidates who had a temperature decrease of these and acetaminophen we had 6 candidates, 5 of them temperature decreased, in one case temperature also increased. Now we add up all these and then rank them we start the smallest with rank 1 and then go upwards. When we do that, so these the 1 because this is the smallest number, then comes 2 then 3, 4, 5 like that it go on finally we end up with 15. Now, what do we do, we add up all these ranks together, these ranks together, these ranks together. The total data set is 15, 4 + 5 9, + 6 15 so capital N is 15. If you look at  $R_i$ 's that means, the summation if you add up all these you get 44 total number of data points is 4, if you add up all these you get 50 total number of data point is 5, you add up all these 6 total is 26. So obviously, there seems to be some large difference, you may expect it to be behaving differently. Now we go to these test statistics, ok?.

(Refer Slide Time: 06:57)




$H =$ , you remember this test statistics which I introduced. So  $12 \div 15 + 16$ , 3 terms are coming 4, 5, 6 because we have 3 sets of data 4, 5, 6 here. 4, 5, 6 then 44 is the sum for data set 1, 50 the sum for data set 2, 26 the sum of data set 3 so 4, 5, 6 is the terms that is coming in denominator. So, we do these addition we end up with 6.833, ok?. You look at the **chi-square** test for 6.833 so the area comes out to be 0.03. Obviously, it is significant so you reject the null hypothesis or you look at the **chi-square** table for this is a 3 data sets are there acetaminophen, ibuprofen and aspirin so the degrees of freedom is 2, we can look into this, you remember this table. Look into 2 for a 95 % 1 sided you get 5.99 and the H the statistics comes to be 6.833 so we reject the null hypothesis at 95 % confidence for 1 tail ok?, this is how you do this.

What we do is, we combine all the data, we rank them and then add up each one of the set of ranks separately and then we use this formula this is the statistic.  $12 \div$  the total number of data points capital N, this is the total sum of the ranks for each i, divided by n i that is the total number of points in that set - n is the total + 1  $\div 2$  and then you compare it with the chi-square table. Now the next question is which drug is statistically significantly different from the other? Now, again there is the different type of approach.

(Refer Slide Time: 09:07)

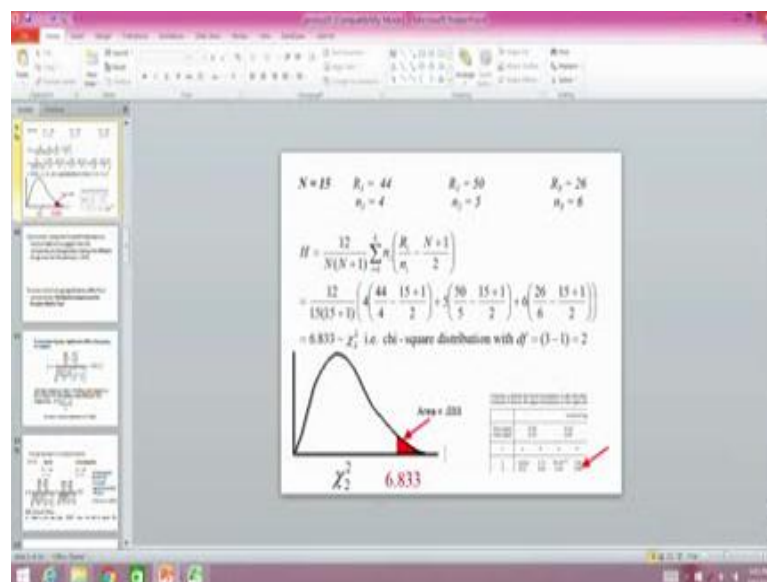
Conclusion: Using the Kruskal-Wallis test we have evidence to suggest that the temperature changes after taking the different drugs are not the same ( $p = .033$ ).

To know which drugs significantly differ from one another- **Multiple Comparisons for Kruskal-Wallis Test**



So, the conclusion using the Kruskal-Wallis test the evidence to suggest that the temperature changes after taking the different drugs are not the same because,  $p$  comes out to be 0.033 area we can get it using excel.

(Refer Slide Time: 09:21)



The image shows an Excel spreadsheet with the following data and calculations:

| $N$ | $R_1$     | $R_2$     | $R_3$     |
|-----|-----------|-----------|-----------|
| 15  | 44        | 50        | 26        |
|     | $n_1 = 4$ | $n_2 = 5$ | $n_3 = 6$ |

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k n_i \left( \frac{R_i}{n_i} - \frac{N+1}{2} \right)^2$$
$$= \frac{12}{15(15+1)} \left( 4 \left( \frac{44}{4} - \frac{15+1}{2} \right)^2 + 5 \left( \frac{50}{5} - \frac{15+1}{2} \right)^2 + 6 \left( \frac{26}{6} - \frac{15+1}{2} \right)^2 \right)$$
$$= 6.833 = \chi^2_2 \text{ i.e. chi-square distribution with } df = (3-1) = 2$$

The graph shows a chi-square distribution curve with a shaded area to the right of  $\chi^2_2 = 6.833$ , labeled "Area = .033".

Excel has this command called CHI DIST so we can put this CHI DIST i had explain this command. CHI DIST with it **6.833 comma 2**  $df$ , it gives you 0.03287 that is the probability or that is the area under this curve. We can use either the command CHI DIST from excel and get the probability value or we can use this table or we can use this



table for 0.05 one sided 2 degrees of freedom 5.99 and you can say this statistics calculated is larger than 5.99 so we reject the null hypothesis.

(Refer Slide Time: 10:20)

Data: Temperature Decrease (deg. F)

| Aspirin | Rank | Ibuprofen | Rank | Acetaminophen                | Rank |
|---------|------|-----------|------|------------------------------|------|
| .95     |      | .39       |      | .19                          |      |
| 1.48    |      | .44       |      | 1.02                         |      |
| 1.33    |      | 1.31      |      | .07                          |      |
| 1.28    |      | 2.48      |      | .01                          |      |
|         |      | 1.39      |      | .62                          |      |
|         |      |           |      | -.39<br>(i.e. temp increase) |      |

Is there a statistically significant difference between these 3 drugs, how do you go about doing that, there is another test, which drug significantly differ from another there is something called Multiple Comparisons for Kruskal-Wallis Test. So, from the Kruskal-Wallis test we concluded, there is a statistically significant difference between these 3 drugs. Now how do we find out which drug is statistically different?

(Refer Slide Time: 10:52)

To determine if group  $i$  significantly differs from group  $j$  we compute

$$z_{ij} = \frac{\left| \frac{R_i}{n_i} - \frac{R_j}{n_j} \right|}{\sqrt{\frac{N(N+1)}{12} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}} \sim N(0, 1)$$

and then compute p-value =  $P(Z \geq z_{ij})$  and compare to  $\alpha/2m$  where  $m$  is the number of possible pair-wise comparisons,  $m = \frac{k(k-1)}{2}$

Bonferroni corrected significance level =  $\alpha/2m$

To determine that, there is another formula like this, if group i is significantly different from group j that means, if you take one drug say aspirin and then you call this ibuprofen so you put in the aspirin data, you put in the ibuprofen data these number this is total and

$$\alpha/2m$$

then these the 15 is the number and then compute p value and compare to where, m is the number of possible pair-wise comparison that is m is given by

$$\frac{k(k-1)}{2}$$

. So, you need to compare this results, with the this particular m value and see whether this results is less than this value. What is k?

(Refer Slide Time: 11:44)

Comparing Aspirin to Acetaminophen

|          |                |                      |   |
|----------|----------------|----------------------|---|
| $N = 15$ | <i>Aspirin</i> | <i>Acetaminophen</i> |   |
|          | $R_i = 44$     | $R_j = 26$           |   |
|          | $n_i = 4$      | $n_j = 6$            | Computing the Bonferroni corrected significance level we have |

$$z_p = \frac{\left| \frac{R_i}{n_i} - \frac{R_j}{n_j} \right|}{\sqrt{\frac{N(N+1)}{12} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}} = \frac{\left| \frac{44}{4} - \frac{26}{6} \right|}{\sqrt{\frac{15(16)}{12} \left( \frac{1}{4} + \frac{1}{6} \right)}} = 2.31$$

$P(Z > 2.31) = 0.01044 \leftarrow$   
p - value is not less than .00833 thus we fail to reject  $H_0$ .

NPTEL

Let us look into this problem, so the statistics so let us compare aspirin, acetaminophen so  $R_i$  minus  $R_j$  here you have  $n_i$  you have  $n_j$  then N here  $n_i, n_j$ . As you know for aspirin  $n_i$  is 4, for acetaminophen  $n_j$  is 6 and then the total n is 15 so we substitute all these into this, ok?. We get the p z, that is once you calculate p z the probability of 0.0104 is obtained from your p z table. And then we compare it, then we look at  $0.05 \div 2.3$  that is k, that is  $\alpha$  is  $0.05 \div 2 m$ , where 2 m is given by

$$\frac{k(k-1)}{2}$$

, so you get this as 0.00833. And this is much less than this particular probability value so obviously, the  $p \text{ value} < 0.00833$  value so we failed to reject the  $H_0$ . That is  $H_0$  in this particular case is aspirin and acetaminophen behave in the same way, there is no reason for you to reject the null hypothesis.

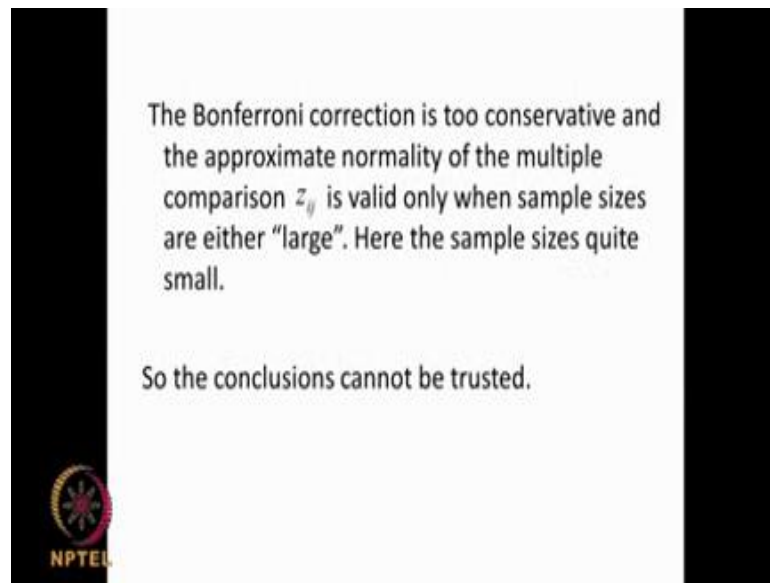
So similarly, we can take aspirin and ibuprofen and do the similar study and then we can take ibuprofen and acetaminophen and do the similar study and so on actually. It is very straight forward to that do you understand? How to do this problem. So, basically what you do is? First step is to look at this equation, substitute all the data and then get your statistic compare it with the table statistic and then you conclude in this particular case there, drugs differ differently. Once you do that you go into Multiple Comparisons for Kruskal-Wallis Test which makes use of this. In that Multiple Comparison, we calculate something called the  $z_{ij}$  that is comparing 2 cases, i and j and then you calculate the probability from this and after that there is something called Bonferroni corrected significance level where  $\alpha$  if you take it as  $0.05 / m$ , m is given by

$$\frac{k(k-1)}{2}$$

, where k here we have 3 drugs so k becomes 3, k -1 becomes 2. So, if you put them together what do you get. The Bonferroni corrected significance level is 0.00833 and the statistic you get 0.01044. So obviously, this is much larger than this, there is no reason for you to reject the null hypothesis do you understand how this is done?

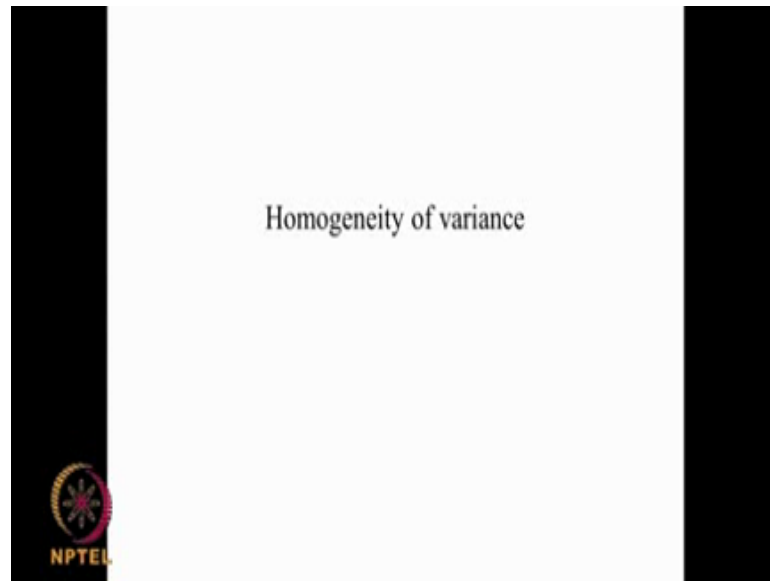
This is useful for comparing more than 2 sets of data, where as we used to use ANOVA where as in this particular situation when the data is ordinal or non-normal then we can use this type of Krus Wallis comparison, Krus Wallis Test ok?.

(Refer Slide Time: 15:39)



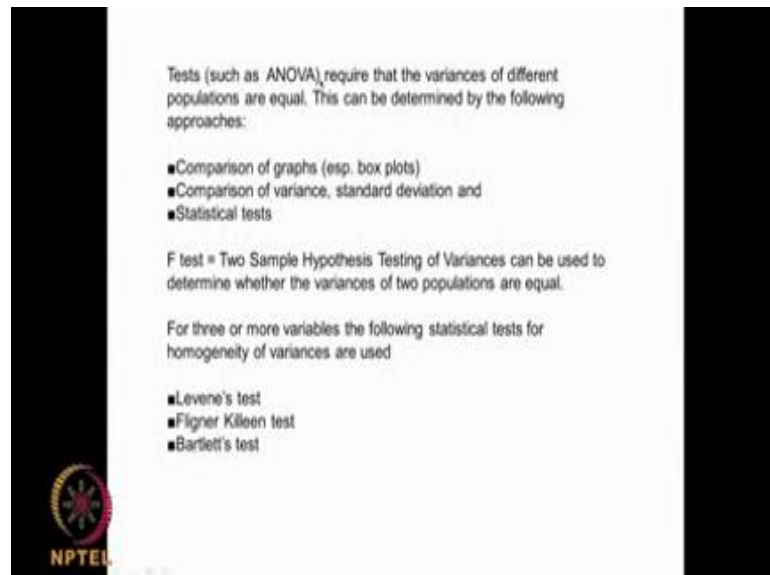
Now, we looked at large number of approaches for nonparametric test, looked at large number of approaches for nonparametric equivalent to your two sample and Paired t-test and One-way ANOVA we have all these rank based test, As you can see many of them makes use of the rank, if you have a large data set we rank them from an ascending order and see how the ranks are so that is how this type of nonparametric tests are conducted and then you have some cases like Wilcoxon Rank Test we have tables we compare it to the tables where as, Kruskal-Wallis we compare it with the **chi square**, ok?. Now there is another situation where you are talking, where you are having non-normal type of distribution.

(Refer Slide Time: 16:27)



In any normal if you are comparing 2 sets of data, multiple sets of data you consider something called Homogeneity of variance. We need check whether your data set has Homogeneity of variance that is very important, if they does not satisfy the Homogeneity of variance then we need to use some other type of test, ok?.

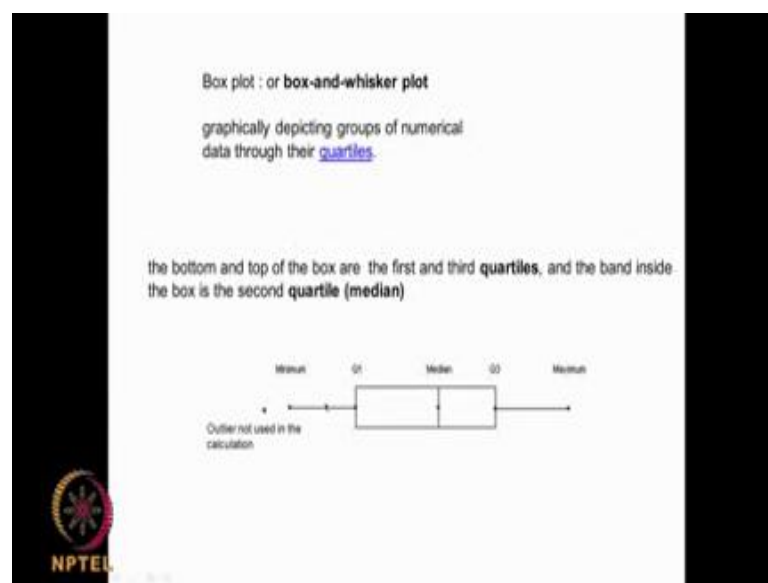
(Refer Slide Time: 16:54)



So, for example, ANOVA requires the variances of different populations are equal, this can be determined by the following approach. How do we do that? we can compare graphically there is something called Box Plots, I will show you how we can compare

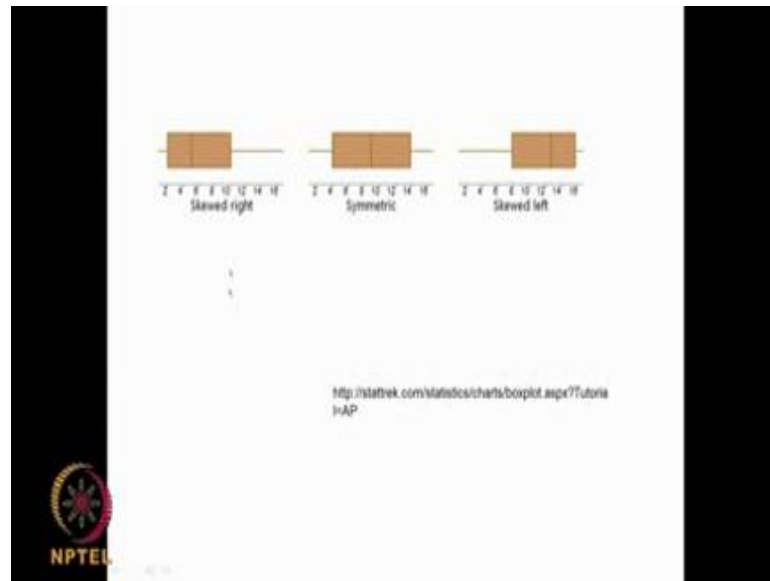
Variance, Standard deviation, we can even do this statistical tests like F test right?. Suppose we have 2 sets of samples we look at the variance of set 1, the variance of set 2 divide one and by another that is called F test. And if we have many, then we perform something called ANOVA. But then, there are other tests for checking the Homogeneity of variances like Levene's test, Fligner Kileen test, Bartlett's tests, so all these tests are available. Let us look at 1 or 2, let us not spend too much time on remaining test because, we can easily find out about the Homogeneity of variance by using even these approaches also, ok?.

(Refer Slide Time: 17:49)



What's a Box Plot? It is also called a box-and-whisker plot, this is a graphical depicting groups of numerical data through their quartiles. So, generally data is represented like this, so this is called the quartile 1, this is called the quartile 3 and this is the median or it is also called as 2nd quartile. This a maximum up to what the data goes up to, this is minimum up to what the data goes up to. Sometimes if you have outliers which you do not consider in your quartile calculation, put them as stars. This plot is very good because it tells you how the data is spread the quartile 1, quartile 3, median and what is a minimum value? what is a maximum value? So, it gives you nice picture and if there are any outliers also we can mark it. If I had 2, 3 data sets, we can draw this box-and-whisker plot and see how these rectangles are? are the rectangles very big? or the rectangles are very small? these are called Whiskers. So, how the whiskers are spread? And so on that is called a box-and-whisker plot, ok?.

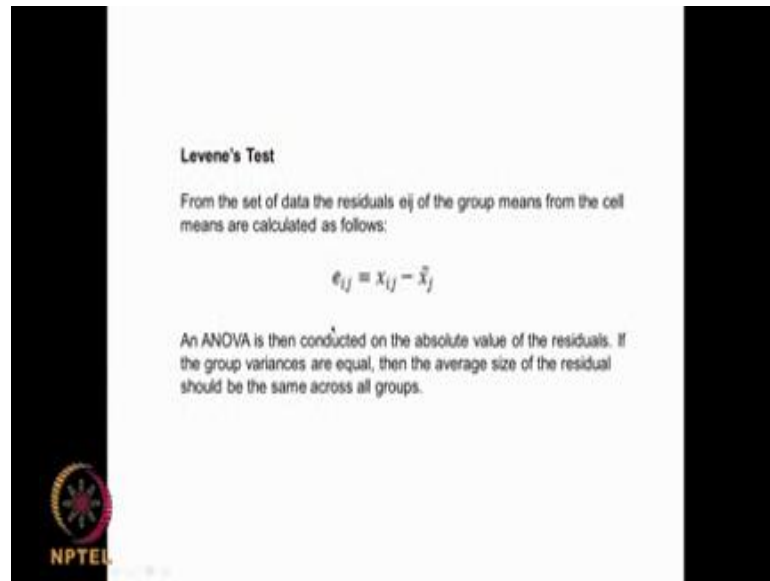
(Refer Slide Time: 19:05)



So, for example, if this is your data and your box-and-whisker looks like this obviously, it is skewed to the right. Symmetric means, it will be nice centered, it will be in the middle of the box, both sides equal like this, like this. Whereas, if it is skewed right we will have it like this, if it is skewed left you will have like this. By looking at it we can always say normality of the data. So, it is very good visual observation, this is called Box-and-whisker plot. This particular pictures are taken from these reference actually.

As you can see when it is skewed left, you will have lot of data as you can see the whisker very long, here the whisker is very short, if it is skewed right you will have very long whisker on right hand side , very short whisker on left hand side. If it symmetric you will try you will see both equal on both sides. So, it is a pic reference this is called the Box-and-whisker plot.

(Refer Slide Time: 20:11)




**Levene's Test**

From the set of data the residuals  $e_{ij}$  of the group means from the cell means are calculated as follows:

$$e_{ij} = x_{ij} - \bar{x}_j$$

An ANOVA is then conducted on the absolute value of the residuals. If the group variances are equal, then the average size of the residual should be the same across all groups.



Then you have the Levene's Test, so what you do is? You have a set of data you calculate the residuals from the group means so if you have the group mean then, we subtract each one of the cell you get the residual. So, once you get the residual, perform an ANOVA on the residuals, of course you take the absolute value of the residuals. If the group variances are equal then the average size of the residual should be the same across all groups, do you understand? It is very simple. So, what you do is, we take the mean and then we take the difference absolute value so those are the, that is called the Residuals absolute value and then you perform an ANOVA. We can perform it for 2 data sets or 3 data sets ANOVA and then we can show the  $H_0$ . Whether it satisfies the  $H_0$  or not satisfies the  $H_0$ . So, it is very simple that is called the Levene's Test.



(Refer Slide Time: 21:14)

**Fligner Killeen test**

The Fligner Killeen test is a non-parametric test for homogeneity of group variances based on ranks. It is useful when the data is non-normal or where there are outliers.

median-centering version of the [Levene's Test](#) by calculating the absolute values of the residuals from the group medians

all these residuals are ranked.. normalize these rankings

$$FK = \frac{\sum_{j=1}^k n_j (\bar{a}_j - \bar{a})^2}{s^2}$$

k = the number of groups, n<sub>j</sub> = the size of the jth group, a<sub>j</sub>-bar is the mean of the normalization values for the jth group, a-bar is the mean of all the normalization values and s<sup>2</sup> is the variance of all the normalization values.

CHIDIST(x,degrees\_freedom) ..... one-tailed probability of the chi-squared distribution

NPTEL

Then we have the Fligner Killeen test, what is this? This is also a nonparametric test for homogeneity of group variances based on ranks. Now you have heard quite a lot about ranks right, this is useful when the data is non-normal or when there are outliers. So, what do you do is? We calculate the absolute values of the residuals from the group medians and then all these residuals are ranked. First calculate the absolute values then put them together and then calculate the ranks of each one of them. And then you calculate the statistics called Fligner Kiileen where, this is summation of  $j = 1$  to  $k$   $n_j$  the size of the group  $j$ ,  $\bar{a}_j$  -  $\bar{a}$  square by  $s^2$ .  $\bar{a}_j$  is a mean of the normalization value for the  $j$ th group and  $\bar{a}$  is the mean of all the normalization values and  $s^2$  is the variance and  $k$  is the number of groups,  $n$  is the size of the group. You can do this and after that we can check it using a chi square distribution, 1 tailed probability of the chi square distribution to see whether they follow the chi square distribution, as simple as this, understand?.

You have the Levene's test which is looking at the residuals from the average values and then perform an ANOVA or we have the Fligner Kiileen test where we are looking at the FK which is the statistic, this is based on the absolute values of the residuals and then you perform chi square distribution test to see whether the  $H_0$  is agreed or it is rejected, ok?.

(Refer Slide Time: 23:21)

Bartlett's test for homogeneity of variances

Bartlett's test statistic B, is approximately chi-square:

$$B = \frac{(n-k) \ln s^2 - \sum_{j=1}^k (n_j - 1) \ln s_j^2}{c} \sim \chi^2(k-1)$$

where  $s^2$  is the pooled variance and


$$c = 1 + \frac{1}{3(k-1)} \left( \sum_{j=1}^k \frac{1}{n_j - 1} - \frac{1}{n-k} \right)$$

$s_j^2$  is the variance of the  $i$ th group,  $n$  is the total sample size,  $n_i$  is the sample size of the  $i$ th group,  $k$  is the number of groups, and  $s^2$  is the pooled variance. The pooled variance is a weighted average of the group variances and is defined as

$$s^2 = \sum_{i=1}^k (n_i - 1) s_i^2 / (n - k)$$

The null hypothesis that all the group variances are equal is rejected if  $p\text{-value} < \alpha$  where  $p\text{-value} = \text{CHIDIST}(B, k-1)$ .

B is only approximately chi-square, but the approximation should be good enough if there are at least 3 observations in each sample.

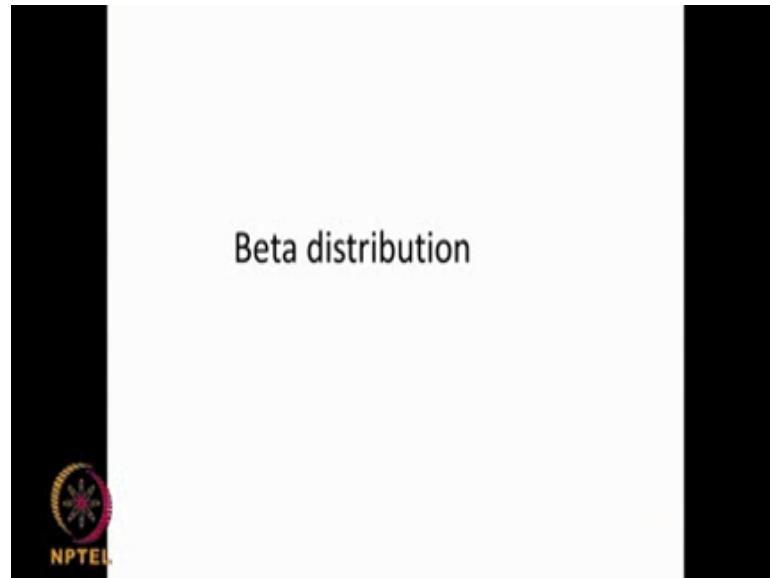


Then you have another test, that is called Bartlett's test for Homogeneity of variances. So you have here, again a Bartlett's test statistics it looks quite big. This also makes use of the **chi square** distribution. Equation looks big where your  $n$  is the sample size of the  $i$ th group,  $k$  is the number of groups,  $s$  is given like this,  $s^2$  is a total variance and  $s_i^2$  is the variance of the  $i$ th group as you can see here. The  $c$  here is given like in this formula so what you do is you calculate this test statistics substituting the  $c$  here, substituting the  $s$  here and then use the **chi square** distribution. So what is the null hypothesis? that all group variances are equal, it is rejected if the  **$p < \alpha$** . And your **chi square** distribution  $b$  will come here,  $k$  is the number of group so obviously  **$k - 1$**  is a degrees of freedom. Here, the Bartlett's test we have statistics and then we apply **chi square** distribution in the Fligner Killeen test we have a statistics which make use of residuals, then again we apply **chi square** distribution. In the Levene's test, we calculate the residuals with respect to the average from each group and then there are sets containing residuals we apply the One-way ANOVA.

All these different type of tests help you to determine the Homogeneity of variance and it tells you whether the variances are equal or they are largely different and of course, you have also have the box whisker plot which pictorially depicts how the variances are of each set of groups samples. We have been looking at nonparametric distribution where the data set could be non-normal, data set could be ordinal, the data set could have large difference in the variances. In such situations, what type of tests we can use? equivalent

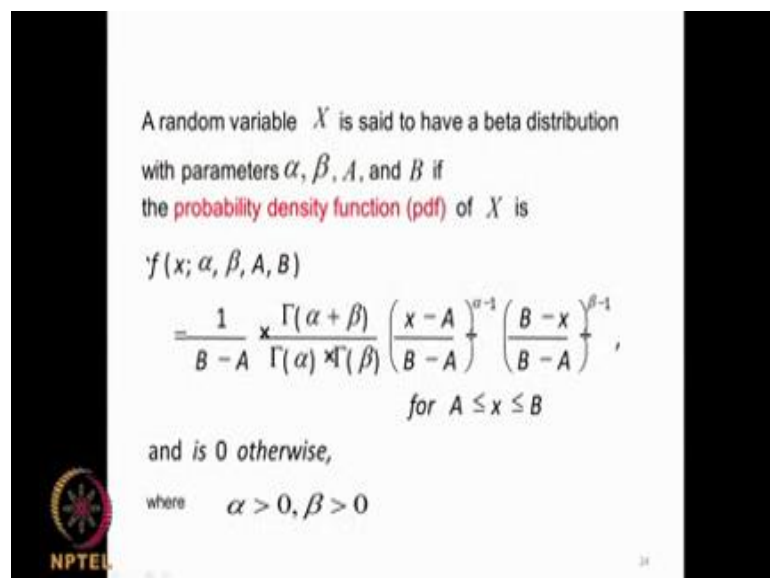
to your t test, Two sample t test, Paired t test, ANOVA. We have the different types of signed test, rank test and so on actually. That sort of will complete the various types of test,ok?.

(Refer Slide Time: 25:52)



And let us look at some other some more distributions we looked at Z distribution, then we looked at t distribution, then we looked at Chi square distribution, F distribution, let us look at one more distribution that is called the Beta distribution, this is also very useful distribution to talk about, ok?.

(Refer Slide Time: 26:17)




Beta distribution, is a random variable X is said to have a beta distribution with parameter. It is got 4 parameters  $\alpha, \beta, A$  and  $B$ , if the probability density function follows this type of relation. It is got 4 parameters  $\alpha, \beta, A$  and  $B$  and  $x$  here,  $x$  will be lying between  $A$  and  $B$  and  $\alpha, \beta$  are always  $> 0$ , understand?. So this is how the relationship will look like and  $\alpha, \beta$  are  $> 0$ . These are called  $\gamma$  functions. I talked about long time back, so  $\gamma \alpha$  is  $\alpha - 1$  factorial,  $\gamma \beta$  will be  $\beta - 1$  factorial,  $\gamma \alpha + \beta$  will be  $\alpha + \beta - 1$  factorial.

(Refer Slide Time: 27:16)

Beta distribution is a continuous probability distribution defined by four parameters:

| Parameter          | Description   | Characteristics            |
|--------------------|---------------|----------------------------|
| Min                | Minimum Value | Any number $= \text{low}$  |
| Max                | Maximum Value | Any number $= \text{high}$ |
| Alpha ( $\alpha$ ) | ShapeFactor   | Must be $> 0$              |
| Beta ( $\beta$ )   | Shape factor  | Must be $> 0$              |



You have 4 parameters here minimum, that is the minimum value it can be anything, maximum that is a maximum value anything  $\alpha > 0, \beta > 0$  both are called the Shape factors.

(Refer Slide Time: 27:30)


Standard Beta Distribution

If  $X \sim B(\alpha, \beta, A, B)$ ,  $A=0$  and  $B=1$ , then  $X$  is said to have a **standard beta distribution** with probability density function

$$f(x) = I_x(p, q) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

for  $0 \leq x \leq 1$

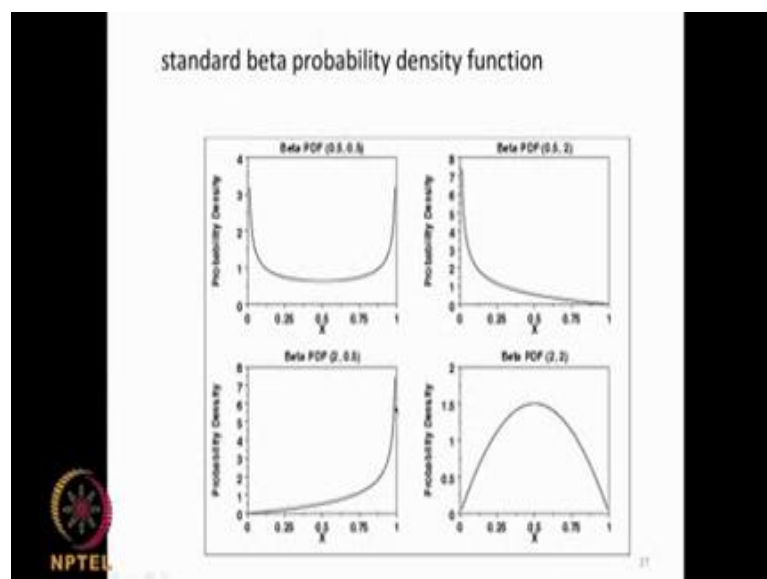
and 0 otherwise



26

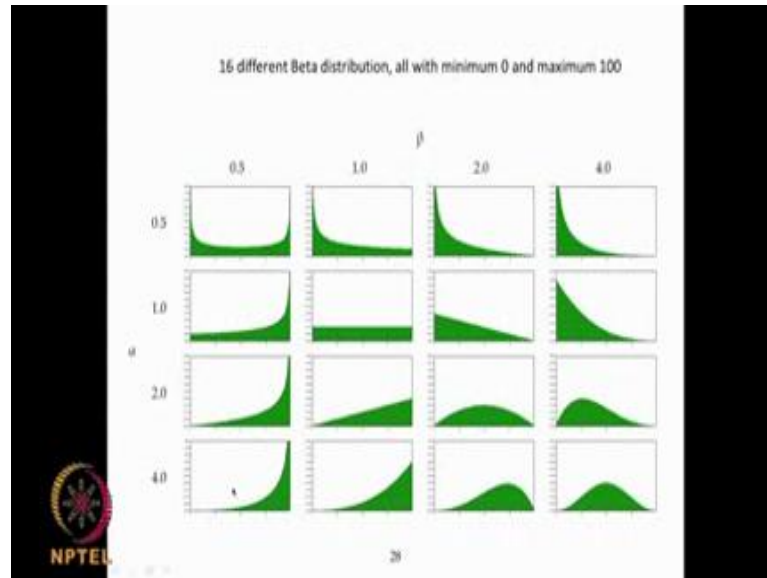
So, you can have a special case this called Standard Beta Distribution when  **$A = 0$  and  $B = 1$** . If you substitute that in the previous equation, in this equation  **$A = 0$  and  $B = 1$** . What do you have? You have something like this, quite simple looking equations for  $x$  lying between 0 to 1. So 0 to 1 it will have a distribution, if it is not between these it will be always 0 otherwise. For different values of  **$\alpha$  and  $\beta$**  we can have different types of shapes.

(Refer Slide Time: 28:08)



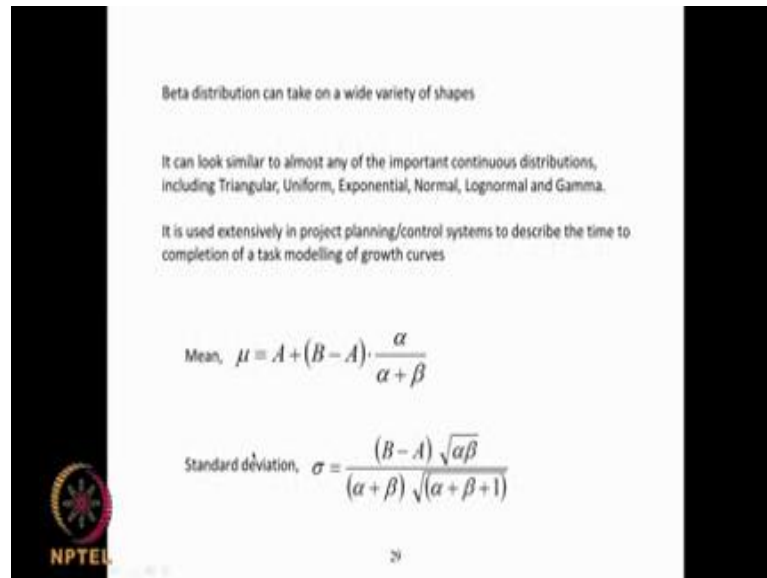
As you can see here, Beta distribution with  $\alpha$  0.5 and  $\beta$  0.5. Beta distribution that is probability density function,  $\alpha$  of 0.5 and  $\beta$  of 2. So, if you have both 2, 2 you get like this, so if you have 2, 1.5 you have get like this.

(Refer Slide Time: 28:31)



You can see that we can get a large number of shapes, with these  $\alpha$  and  $\beta$ . So, this very, very useful especially for simulations and modeling. We can generate any type of function as you can see, you can get maxima going down, you can get exponentially sort of falling down, you can have exponentially rising, you can a linear rising and then you can get bath tub type of curves. You can get any type of curve by manipulating your  $\alpha$  and manipulating your  $\beta$ . So, it is very useful for simulation purposes. So, if you want to generate any shape we take this beta distribution function, put A and B as 0 and 1 respectively, and we can use whatever  $\alpha$  and  $\beta$  we want to use to achieve the shape, ok?.

(Refer Slide Time: 29:29)



Beta distribution can take on a wide variety of shapes

It can look similar to almost any of the important continuous distributions, including Triangular, Uniform, Exponential, Normal, Lognormal and Gamma.

It is used extensively in project planning/control systems to describe the time to completion of a task modelling of growth curves

Mean,  $\mu = A + (B - A) \cdot \frac{\alpha}{\alpha + \beta}$

Standard deviation,  $\sigma = \frac{(B - A) \sqrt{\alpha\beta}}{(\alpha + \beta) \sqrt{(\alpha + \beta + 1)}}$

NPTEL

Let us go forward, it can take a wide variety of shapes, it can look like Triangular, Uniform, Exponential, Normal, Lognormal, Gamma so it is fantastic actually. It extends extensively in project planning controls, growth curves in your bacterial systems like as you can see we can get different types of growth curves, different types of growth patterns you know, using modifying your  $\alpha$  and  $\beta$ . So, the mean of this will be like this

$$A + (B - A) \cdot \frac{\alpha}{\alpha + \beta}$$

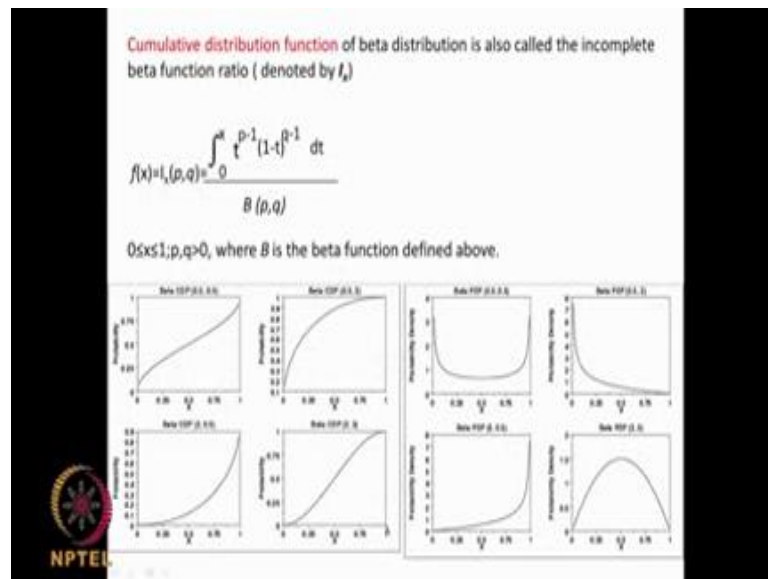
. So, if A is 0 B is 1, what happens? You will get

$$\frac{\alpha}{\alpha + \beta}$$

your standard deviation is given like this. If B is 1 this will go away, so your standard deviation be

$$\sigma = \frac{(B - A) \sqrt{\alpha\beta}}{(\alpha + \beta) \sqrt{(\alpha + \beta + 1)}}$$

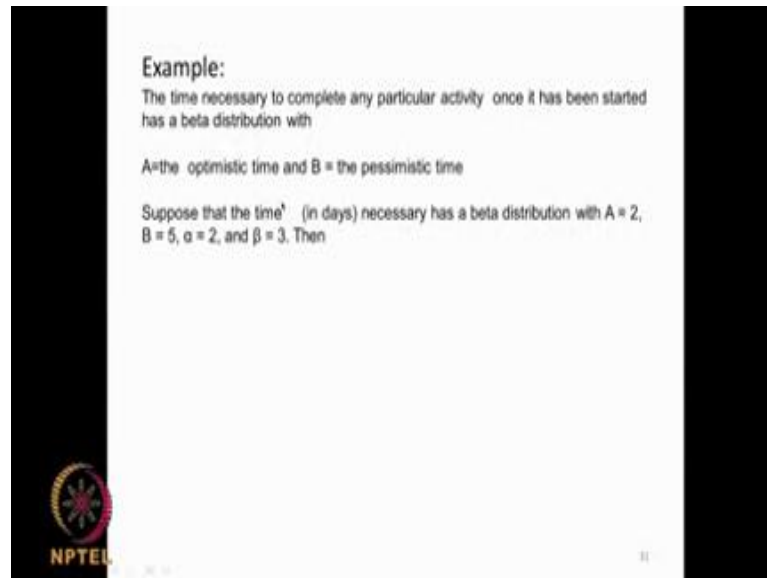
(Refer Slide Time: 30:20)



There is something called Cumulative distribution function. So, you just need to integrate between 0 to  $x$ , of the beta distribution and it is given like this  $B$  is your beta function, beta function is nothing but this is your beta function, sorry this is your beta function. So, you can integrate that to get your Cumulative distribution function. So, for example, if your probability looks like this, sorry probability density function looks like this your cumulative will look like this, if your probability density function will look like this cumulative will keep on increasing. Cumulative generally keep on increasing. If your probability density function looks like this for 2 and 2 your cumulative will go like this, if your probability density function looks like this your cumulative will go like that. So, this equation is also useful because it gives you an idea about the Cumulative density function.



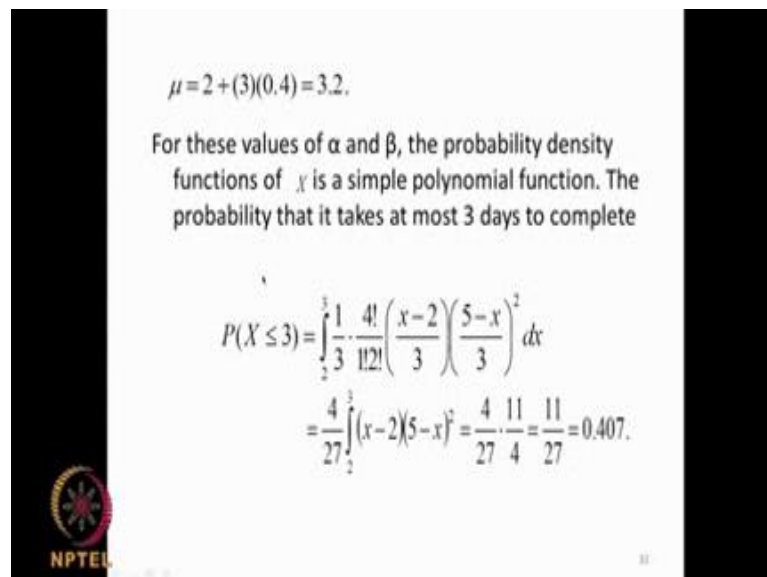
(Refer Slide Time: 31:19)



**Example:**  
The time necessary to complete any particular activity once it has been started has a beta distribution with  
A=the optimistic time and B = the pessimistic time  
Suppose that the time<sup>1</sup> (in days) necessary has a beta distribution with A = 2, B = 5,  $\alpha = 2$ , and  $\beta = 3$ . Then

Let us look at an example, time necessary to complete any particular activity once it has been started follows a beta distribution, suppose A is optimistic time is given in 2 days, B is the pessimistic time given 5,  $\alpha = 2$ , and  $\beta = 3$ , then the mean will be 3.2 that means, on an average it will take you 3.2 days to complete the task.

(Refer Slide Time: 31:36)



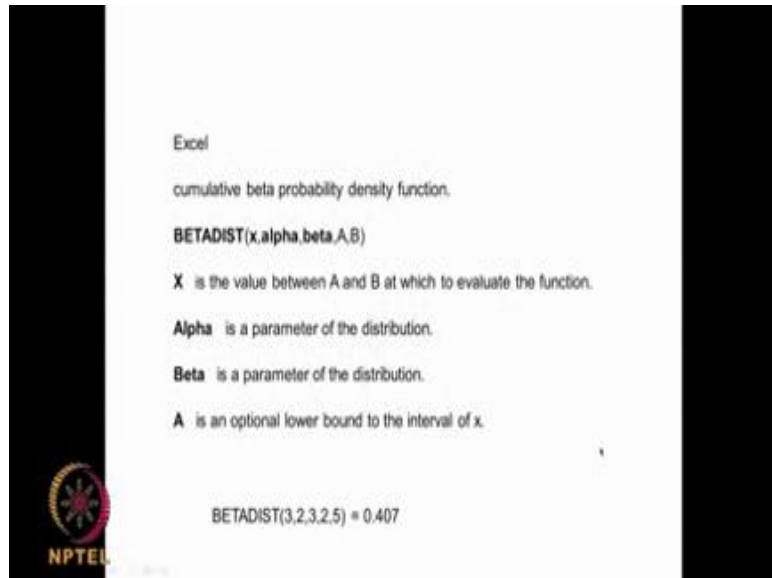
$\mu = 2 + (3)(0.4) = 3.2.$   
For these values of  $\alpha$  and  $\beta$ , the probability density functions of  $x$  is a simple polynomial function. The probability that it takes at most 3 days to complete

$$P(X \leq 3) = \int_2^3 \frac{1}{3} \cdot \frac{4!}{12!} \left( \frac{x-2}{3} \right) \left( \frac{5-x}{3} \right)^2 dx$$
$$= \frac{4}{27} \int_2^3 (x-2)(5-x)^2 dx = \frac{4}{27} \cdot \frac{11}{4} = \frac{11}{27} = 0.407.$$

Now for these values of  $\alpha$  and  $\beta$ , the probability density function of  $x$  is a simple polynomial because if we substitute the cumulative distribution function like this, so we put your **alphas**, your  $\alpha$  as 2,  $\beta$  is 3,  $\alpha$  is 2,  $\beta$  is 3 and then A as 2, B as 5. So substitute

these into the equation and integrate them, end up with probability of 0.407. The probability of completing the task at most 3 days that means maximum of 3 days is given as a probability of 0.47.

(Refer Slide Time: 32:09)



And also in your excel we have BETADIST, which gives you x,  $\alpha$ ,  $\beta$ , A, B. So, let me look at it, excel, so we have the excel function which is given by BETADIST, equal to BETADIST and you want to calculate at, see, it is given as x,  $\alpha$ ,  $\beta$ , A, B so in your case  $\alpha$ ,  $\beta$  is 2 and 3 respectively A and B is 2 and 5 respectively. If you want to finish it in 3 days, what happens? 2,5 sorry, want to finish it in 3 days as here at most 3 days, what it should be there? So, we need to put the values sorry, 3, 2, 3 because  $\alpha$ , is 2,  $\beta$  is 3, then A is 2 that is the optimistic time, then the pessimistic time is 3, 5 so you get 0.407. So, the probability of finishing it in 3 at most in 3 days is 0.407, that is same as what is got by integrating this Cumulative distribution function. So, we can use the excel to do the same thing and excel has this BETADIST command you want to know the probability at any value of x and you know the  $\alpha$ , you know the  $\beta$ , you know these A and you know the B so it gives you at 0.407. So, we will continue on this distribution in the next class also.

Thank you very much for your time.

**Key Words: cumulative distribution function, Beta distribution, Kruskal wallis test, multiple comparisons, homogeneity of variance, cumulative distribution function, test statistics, probability, chi square, probability density function**