Introduction To Reliability Engineering
Professor Neeraj Kumar Goyal
Subir Chowdhury School Of Quality And Reliability
Indian Institute of Technology, Kharagpur
Lecture 35
Goodness of Fit (GoF) Tests (Continued)

Hello everyone. So, we have been discussing about failure data analysis, we initially discussed non-parametric method where we looked into various types of data complete data or group data then sensor data.

And we have seen that how we can find out the based on the ranking how do we find out the FTI then from FTI we can get reliability, we can get a quality density function, we can also get the failure rate or hazard rate function, then we start discussing about the LSE method least squares estimation where we saw that if we have different distributions then we can convert the cumulative distribution function into a linear form and using that linear form then we can fit the data and you can find out that what is the parameter.

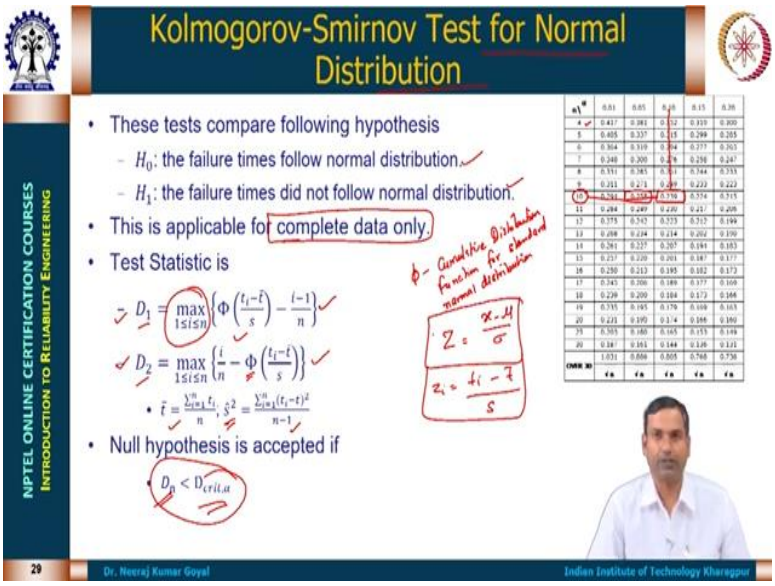
However, as we discussed earlier that many times MLE is considered a better approach than LSE for finding out or estimating the parameters of the distribution. So, then we discuss the MLE we discussed that how MLE can be used for estimating the parameters of exponential distribution, viable distribution and also normal and log normal distribution.

For normal and log normal distribution the sensor data problem becomes little bit more complex. So, that to solve those kinds of problems, we may need to use some software packages for that, but then we also discussed that if you are using LSE then how can we perform the goodness of fit test, we initially discussed a chi-square method. So, chi-square test we did which is general in nature and can be used for any distribution function to see whether the distribution is fitting to the data or not, that is based on the observed number of failures which is compared versus the expected number of failures which is coming after the distribution fitting.

So, based on that, we were able to know whether our distribution fitting is good or not. Then we also discuss the specific tests which we can do for exponential distribution, then for viable distribution.

Today, we will discuss the test which we one test which is used for finding out whether the distribution is normal distribution or not a normal distribution. Similarly, it can also be used for the log normal distribution or (now) whether it is a log normal distribution or not a log normal distribution.

(Refer Slide Time: 03:12)



Test Statistic is

$$D_1 \quad = \max_{1 \le i \le n} \left\{ \Phi\left(\frac{t_i - \bar{t}}{s}\right) - \frac{i-1}{n} \right\}$$

$$D_2 \quad = \max_{1 \le i \le n} \left\{ \frac{i}{n} - \Phi\left(\frac{t_i - \bar{t}}{s}\right) \right\}$$

- $\bar{t} = \frac{\sum_{i=1}^{n} t_i}{n}; \hat{s}^2 = \frac{\sum_{i=1}^{n} (t_i - \bar{t})^2}{n-1}$

Null hypothesis is accepted if

- $D_n < D_{crit,\alpha}$

So, going for that this test is called as KS test or Kolmogorov-Smirnov Smirnov test. So, this test is can be useful normal and log normal, I am showing it for normal distribution, as we discussed earlier that if the data is following log normal distribution then you just need to take the log of the data and then it will fit to the normal distributions. So, whatever we do for normal distribution same is applicable for log normal distribution, the only change is that rather than taking the data we will take the log of the data and then the same process can be carried out.

So, KS test works on the hypothesis for KS test is that failure times show normal distribution and alternative hypothesis is that failure times do not follow the normal distribution. So, as we see that this is also can be used as a test for the specific distribution and but, again the problem is that we can only use it for complete data, so sensor data is not considered here.

So, how does it work? For this we calculate the two statistic here D1 and D2 and from D1 and D2 which is the maximum value if you see that for all the sample points which you are consider, we try to find out the max value which is the difference between the cumulative or the cumulative distribution of ti minus t bar divided by s, minus i minus 1 divided by n and another D2 is the reverse of this, that is maximum value of same for i by n minus phi cumulative distribution, phi is the, what is phi? Phi is cumulative distribution function for standard normal distribution.

That means so, how do we get a standard normal distribution? From normal distribution has two parameters mu and sigma, but standard normal distribution does not have any parameter or we can get the standard normal distribution from normal distribution by keeping mean as zero and variance as one.

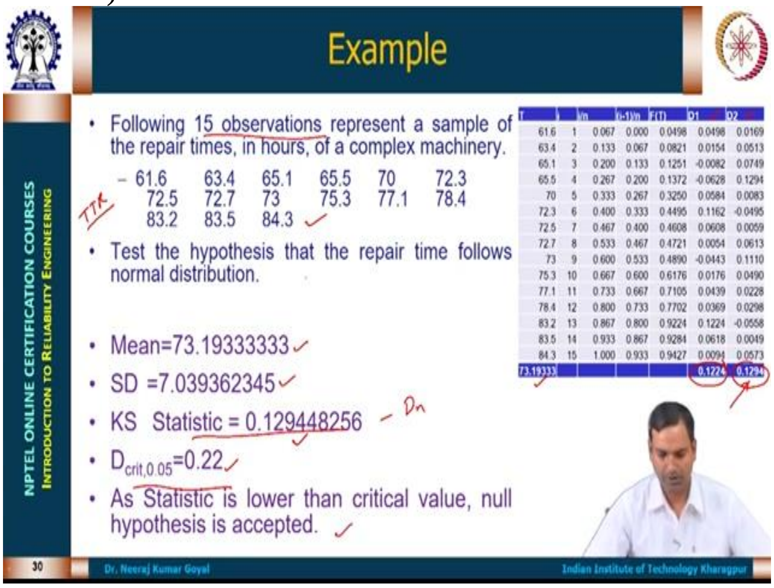
So, and that we get it like by transforming Z is equal to x minus mu divided by sigma. Now here, mu and sigma are not known to us which is to be which is calculated from this data. So, this actually becomes t, so if I say Zi, Zi will be equal to ti minus mu is mean value of ti divided by C rather than sigma we are using the notification as because this is this we are calculating from the samples and i minus 1 divided by n is usual we know the rank and from the rank we can subtract 1 divided by n.

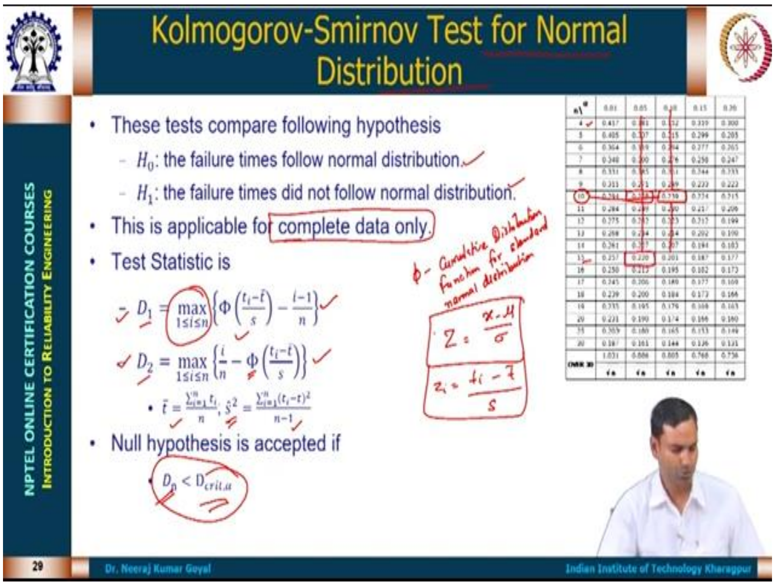
Similarly same thing we calculate from here. So, these two quantities when we calculate then where what is t? t is the average value of ti, and what is s square? s square is summation i equal to 1 to n, ti minus t bar whole square divided by n minus 1.

Here, we will see, we will calculate D1 D2 for all the sample points. So, I will show it with an example and then we find out which is the maximum value and out of the D1 D2 then again we will see which one is the maximum value and the maximum value which you are selecting we are calling as Dn this maximum value should be less than the our critical value and what is this critical value? Critical value we get from here.

So, let us see if sample sizes 10 and we want to have the significance level 0.1. So, 0.1 and 10 so, my critical value is 0.239 if I use 0.05 5 percent significance level then same thing would be here 0.258, so, we use this table to get this values.

(Refer Slide Time: 07:52)



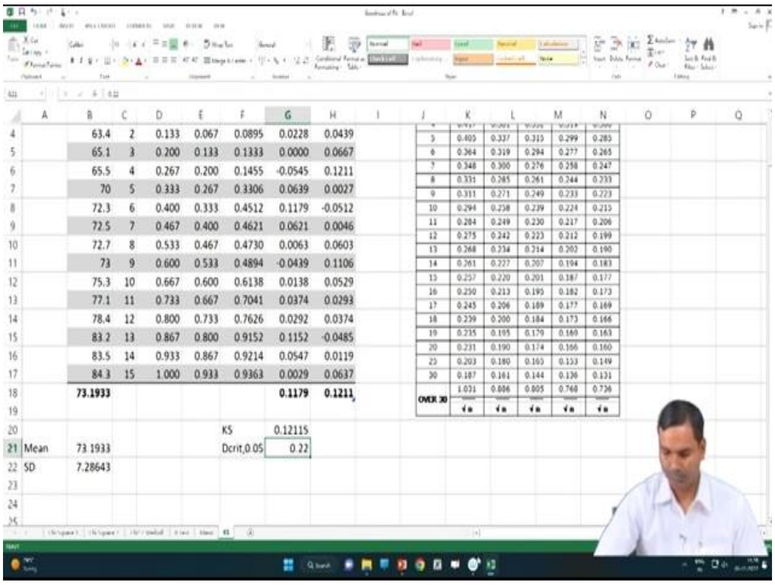


Now, let us say, let us take an example that we have 15 observations and these observations are given here I have solved this using the excel sheet I will show this again in excel sheet and we want to test the hypothesis. So, these are the repair time TTR. As you discuss you all the methods which you discuss can be applied to any random variable whether it is TTF TTR or any other of interest.

Now, to do that we will calculate the mean, mean is coming out to be 73.1991933 and standard deviation is 7.039, this we have already calculated and I will show it in the sheet also. Now, KS statistic which we calculate here like this is D1 value we have calculated using the formula which we discussed and D2 value. So, D1 maximum values 0.1224 and D2 maximum values 0.1294 out of the two this is the larger one.

So 0.1294 is used as the Dn value, this becomes our Dn value. And how much is the critical value for 0.05 because here our sample size is 15, so, n equal to 15 and critical we are taking the 0.05 so, this becomes our value 0.220 is the value, so critical value is 0.220. Since my KS statistic value is less than 0.22 my hypothesis null hypothesis is accepted. Now this example which I am showing, I will show it the same in excel sheet also.

(Refer Slide Time: 10:17)



This is the same sheet which we were discussing earlier. Now, let us go to the KS test. So, this data which was there I have already copied it here, our time was if you see here I copied all this data 61.6 to 84.3. And how do I calculate mean? Mean is the average, so I will calculate the average. And how much is standard deviation? Standard deviation is standard deviation dot I think we have to use the S here S and so, that will be 7.28 so, that is little bit correction here. And what is i? i is 1 to 15 and i by n is i divided by 15 same thing will come. What i minus 1 by n? Simple whatever I have taken I, so that will be i minus 1 divided by n, n is 15 here, so, I will use that 15 directly.

I made some error here, did not use the bracket, fine and what is capital F T? Capital F T calculations when we are doing that is nothing but the value based on the time I have I because based on this data t and i, I have already calculated mean and standard deviation. So, for mean standard deviation which I have calculated I can get the cumulative distribution function using the norm distribution normal distribution, normal distribution and what is the cumulative distribution for a given value x? And what is the x value? Value x is time here.

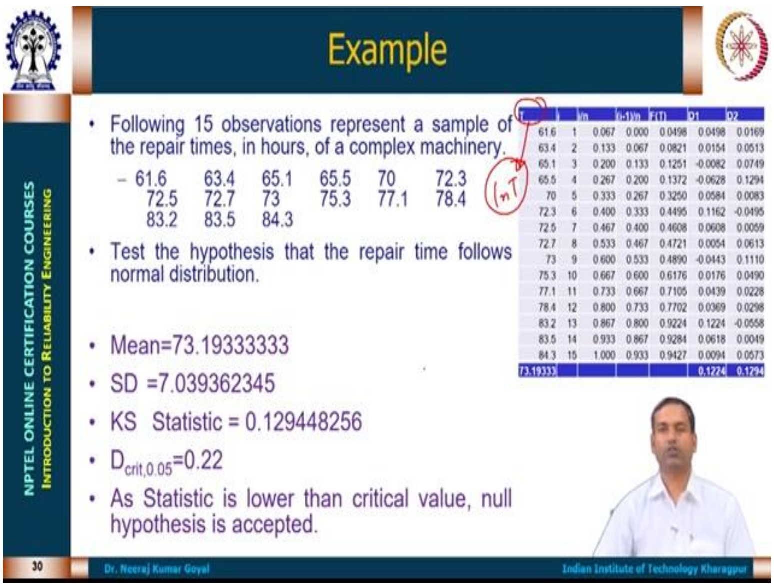
My concern value here is I will show this again, so that you can follow, this is equal to norm distribution. We can use the earlier one also we can use this one also. Now, normal distribution has to be calculated for this time. And what is the mean value? Mean value is this. And what is the standard deviation? Standard deviation is this, and this is the cumulative value I want, so I will use the true. For this b22, because I do not want it to change so, I will use the dollars here, so this is how I get the F T.

Now, let us see how do we calculate the D1? As we see here D1 is maximum value so rather than saying D1 I am calculating Di, what is Di? Di is this value phi of ti minus t divided by s which is nothing but the my F value which I have calculated and minus i minus 1 divided by n.

So, this becomes nothing but because that value is nothing but F T, F T minus i minus 1 divided by n. This becomes my D1 first set and second set is n by i by n minus F T, so we get this value. Now, here I have taken the maximum value this is the max of this, this is the max of this.

So, my KS is max of both of these, so max becomes 1211 and my D critical value I have got from this table, this table, so for 15 and 0.05 0.22, so, this will be my critical value. Once I use the same thing here then my D critical value is higher than KS or I can see my statistic value is lesser than the critical value therefore my null hypothesis is accepted. So I can say that this data is following the normal distribution, so this is how we can solve.

(Refer Slide Time: 14:51)



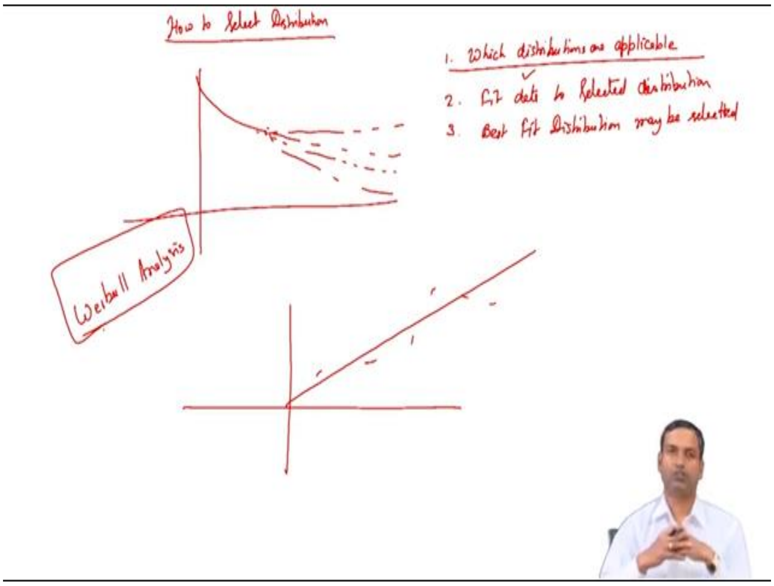
Same way if we do it for log normal distribution, for log normal distribution also we can do it by simply taking the rather than just in the same first column or we can make another column where rather than T we will take ln of T then ln of t then we will do the same thing i l, i f i by n and everything. And same procedure everything we will do and then we will see, the only change will be that this in this first column, whether I am using this column which is T here, rather than I will use the ln of T log values, because log values are expected to follow the normal distribution, rest remains same.

So, here like in these lectures in this week, we have seen that how we can have when we have the data, how we can use the data for estimating the parameters of interest which can be reliability failure rate, repair rate, if the time to failure rate is there, we will get the failure rate, if we have that time to repair data then using the same procedure we will get the repair rate. Similarly, we will get the maintainability if we have the repair data we will get MTTR if we have the repair data.

So, same procedures which we discuss can be used. These procedures which we have discussed in these classes, these procedures have been simpler one comparatively and which can be solved using hand or even excel but to whenever we want to do larger analysis, it is proposed that or it is suggested generally to use some statistical software, statistical software will allow you that for the same data, it can show you the results for multiple distributions all together that will be able to show you that which distribution is fitting better based on a goodness of fit test.

So, the goodness of fit test value which test statistic which is coming the lowest one based on that it will rank you that and suggest you that which distributions are fitting best. For distribution fitting whenever we are going it is better that we follow like there are certain questions like many people, many times people face this issue that what is the thing to do?

(Refer Slide Time: 17:33)



So, in that case they want to know that how to select a distribution. So, this problem you may face that how to select right distribution. Many times what happens, the commonly understood way of doing this is that you fit the data to multiple distributions and whichever distribution is giving you best fit that means the least error or the least value of the statistic that you consider to be the selected distribution.

However, that may many times lead to the wrong assessment, why? Because the problem is that when you are selecting the distribution, fitting of distribution is not just the enough criteria, because many times it may happen the distribution which is having more parameter or more flexible is able to fit your data better, but practically because it is a sample, because you are fitting the where distribution based on the sample.

So, it may happen that based on the distribution when you are making a prediction for the future, because let us say if you have the sensor data, then based on the distribution, you know reliability up to here, but by assuming that you follow the same distribution, you are assuming the pattern that it will follow this pattern, but if you have not carefully selected the distribution, the actual pattern would might have been like this, but it might have selected like this or it may have been like this, or it may have been like this, it may have happened that the distribution which are fitted, though it is fitting good but it may not be able to capture the real distribution.

The reason being that data is the sample data, so sample data has its own variability, all data is not going to fall on the line, so there is inherent variability in the data that is why we are doing the distribution fitting. So that may be confusing. So, (how to do) how we can do this step wisely? To do that, first we should ask a question that which distribution is applicable?

So rather than simply saying that best fit distribution you will choose, we first let us see find out which distributions are applicable for this. That means in general historically or with previous experience of with your overall experience, as you have seen with a similar system data, similar failure data, what kind of distribution has been fitting when you have the large number of data available.

So, that means that will guide you, that will suggest you that which are the distributions which can be the candidate for the selections. In general for failure data analysis Weibull is considered to be the unique distribution which is fitting most of the time, so that many times people consider either Weibull many times log normal also fits to most of the cases. So, many times when you fit the data you will find that log normal distribution is coming at the first priority.

However, that may not be enough because it is fitting to the data, but it may be taking us in the right, wrong direction also there is a possibility. So, how to correct that? That rather than depending on only on the fitting value, because if error is little less a little more that does not make sure that your estimation or your prediction would also be good. Therefore, it is essential that you first identify which distribution are applicable. So, you make a list of them then you fit that data then you fit data to select a distribution, once you do that then you will be able to know the parameters and then you use the best fit distribution.

This problem is much higher in reliability, the reason being in reliability we are dealing with the failure data. So, failure data is constrained especially if we are getting it from test etcetera, only few failures we observe, while number of devices which are working are much-much higher.

So, failure percentage may be somewhere around 1 percent 2 percent that is also for means when companies are having a good repu. So, even for bad repu nobody would be having the failures which is in range of 5 percent 10 percent. So, generally for good companies to survive, they want to bring their failure percentages below 1 percent. So, but they mostly work around these kind of percentages.

So, therefore, the failure data is only that much and that failure data is then also there is a inconsistency in the data there is a reporting time problems, there is a assessment problem there is a change in the applications.

So, there are many factors which affect the variability therefore, it is it will be wise that before selecting the distribution before saying that which is a best fit distribution, you first say and identify whether these distributions are applicable or not applicable, so, the distributions which are applicable to the situations only those distributions should be considered for the fitting and once you fit them then you can say that okay my this distribution is able to get the trend better or able to give the error better.

Similarly, we may also look into let us say if we have the data that certain data is here, then if it aligned to this. So, we can see that which portion our data is fitting better whether it is the initial phase it is (filling) fitting better, whether it is the later phase it is fitting better, sometimes that can also indicate and can help you to decide that which distribution is better, sometimes it may happen that initially it is fitting better but later on it is deviating from the points or the points are getting away from the line that is an indication that your points or the line which are fitting maybe having a different pattern then your data is actually having.

So, we can look into so, we need to try a few multiple distributions and based on that we can select but as we see that failure data analysis of distribution fitting, the Weibull has become so popular for failure data analysis like you will find that there is a complete site complete books which are written as the Weibull analysis. The failure data analysis or fitting the data is called Weibull analysis.

The reason being Weibull is a flexible distribution, which can take multiple shapes. So because of that it can capture multiple failure patterns. So because of that Weibull has been very popular so it can capture exponential also when the beta is one, it can capture normal also or log normal also when beta is four five or higher.

So, different pattern it is it can capture in increasing failure rate also, it can also capture the decreasing failure rate distribution, it can also capture the constant failure of distribution. So, because of this versatility and it is having two parameters only, there has been many times tendency to use three parameter distribution.

But before using the three parameter distribution you have to ensure that whether the third parameter which is saying that there is a minimum life to the equipment, whether that kind of scenario really exists or not. If that exists then only you try to use, if that does not exist then do not force it because otherwise you are forcing you know in your analysis that for a certain period of time there is no chance of failure that may be incorrect many times.

Similar case for the two parameters exponential distribution, therefore when we are doing all this analysis we have to properly follow that which distributions are the right candidate and once we find out the candidate distributions then we try to find out the best fit distribution and most of the time in reliability analysis practically also as I have observed when I have worked with various companies that they try to fit that data to the viable distribution, which is also acceptable approach and which has been successfully used by many companies, this makes their processes simpler.

They do not have to be statistician to have multiple distribution than fitting etcetera, you can develop a single tool single macro in excel sheet or simple way of doing the analysis, you can just make a simple code one code line of number small amount of code.

So, with that, you may not need this specific or highly these statistical software's and you can do all this analysis by using your excel sheet or using your C simple C program or some other programming language. However, to do the detailed statistical analysis, we have a significant number of tools available. Right now we have the Minitab, Minitab has a function which has which is can be used for the time to failure data analysis so, reliability analysis is in a way represented there.

We have a MATLAB also MATLAB data fitting is there curve fitting is there we can use them. Many other software's are their s is there, sorry, r is there and we have the statistical software we also have Python, Python can also be used. So, all these software's we can use for this purpose of doing the data fitting and finding out the parameters of the distribution, so, we find out which distribution is good that is the model selection and then we find out the parameters of the distribution that is the parameter estimation for the model.

Once we have these parameters, we can use these for the system study that we discussed in earlier weeks that we can find other based on the reliability we can find out the reliability of the components we can find out the reliability systems and we can also decide our maintenance policies we can also decide what to do about it.

So, with this we will now this our most of the discussion till now has been focused on reliability or time to failure. We will briefly discuss in next classes about the repair that systems which are repairable we discussed briefly about them in the during the first week where we discuss maintainability and availability, we will try to put little bit more discussion on the same aspect that how maintainability can be measured and how this can be useful. Similarly, how availability can be measured and that can be used, so that we will do in next classes. So, we will stop it here today. Thank you.