

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

Lecture - 25
 χ^2 test/Weibull distribution

Hello everyone. Welcome to the course on Biostatistics and Design of Experiments. We will talk little bit more on χ^2 test and later on we will take up another new distribution called a Weibull distribution. As I said, χ^2 test looks at goodness of fit, it can see whether there is an association between a group and another group. So, it is quite very useful type of test.

(Refer Slide Time: 00:42)

The χ^2 -test for goodness of fit

To determine significance of the differences between observed data and the theoretical.


Test statistics is given by:
$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

where O_i and E_i are the observed and theoretical

$H_0 : O_i = E_i$
 $H_1 : O_i \neq E_i$

The test statistic is compared with the critical value from χ^2 tables with ν DF. Where, $\nu = k - 1$.

If the test statistics, χ^2 is $>$ critical value we reject the null hypothesis that the observed and theoretical distributions agree.

 NPTEL

For example, if I have some observed data and I expect something else then χ^2 test can be used to determine whether the observed data is very, very different from the expected data. So, here the null hypothesis will be observed will be same as expected and the alternate hypothesis will be observed \neq expected. The test statistics will be

$$\frac{(O_i - E_i)^2}{E_i}$$

, you sum it up over all the data points and that will give your χ^2 . Then you compare this χ^2 with the table χ^2 for $K - 1$ degrees of freedom and as usual if the test statistics is less than table, there is no reason for you to reject the null hypothesis, if the test statistics is greater then you need to reject the null hypothesis and accept the alternate hypothesis. Of course, there is something called Yate's correction. Especially it is very valid when you are doing a 2 by 2 type of contingency table.

(Refer Slide Time: 01:48)

Problems involving two categories, a correction is used (called Yate's correction) Because χ^2 distribution is continuous in two categories the data becomes integers or discontinuous) ...used in 2 x 2 contingency table

So $[E-O]_c = |E-O| - 0.5$

Subtract 0.5 from the absolute of $[E-O]$
If it becomes <0 , then equal it to 0

NPTEL

Because when you are doing a 2 / 2, you are converting a continuous distribution. χ^2 is a continuous distribution into sort of an integer you know, almost like a binomial, yes, no, live, dead and so on. So, in order to take care of this discontinuity especially when you are hand handling 2 by 2 contingency table there is something called the Yate's correction that has to be done. Some softwares might not do that actually. What does Yate's correction do? $[E-O]$, you take the absolute value and subtract from, - 0.5 from that, that means subtract 0.5 from that. You need to subtract 0.5 from the absolute value. Sometimes you will get $[E-O]$ negative but that does not mean that you will add another

0.5 negative to that. Then, if the subtraction reduces to below 0, then you just equate it to 0. So these points you need to remember especially when you are using Yate's correction.

(Refer Slide Time: 03:04)

Critical Values of the χ^2 distribution

- For upper-tail one-sided tests, the test statistic is compared with a value from the column of upper boundaries critical values.
- For two-sided tests, the test statistic is compared with values from both the table for the upper- boundaries critical values and the table for the lower- boundaries critical values.
- If the test statistic is $>$ than the upper-tail critical value or $<$ the lower-tail critical value, we reject the null hypothesis.

Columns 1 shows the lower boundaries or the left-tailed critical values.
Columns 2 shows the upper boundaries or the right-tailed critical values.

Degrees of Freedom	Level of significance α			
	0.10	0.05	0.025	0.01
χ^2	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000
2	0.010	0.004	0.001	0.000
3	0.051	0.078	0.016	0.004
4	0.708	0.711	0.485	0.316
5	1.610	1.236	0.831	0.554
6	1.237	0.924	0.676	0.475
7	1.219	0.872	0.637	0.445
8	1.345	0.833	0.609	0.420
9	1.380	0.808	0.589	0.400
10	1.372	0.788	0.569	0.380
11	1.357	0.770	0.551	0.362
12	1.350	0.755	0.535	0.347
13	1.345	0.742	0.521	0.334
14	1.341	0.730	0.508	0.323
15	1.338	0.719	0.496	0.313
16	1.336	0.709	0.485	0.304
17	1.334	0.700	0.475	0.296
18	1.333	0.692	0.466	0.289
19	1.332	0.685	0.458	0.283
20	1.331	0.678	0.451	0.277
21	1.330	0.672	0.445	0.272
22	1.329	0.666	0.439	0.267
23	1.328	0.661	0.434	0.263
24	1.327	0.656	0.429	0.259
25	1.326	0.651	0.425	0.255
26	1.325	0.646	0.421	0.252
27	1.324	0.642	0.417	0.249
28	1.323	0.638	0.414	0.246
29	1.322	0.634	0.411	0.243
30	1.321	0.630	0.408	0.241
31	1.320	0.626	0.405	0.238
32	1.319	0.623	0.402	0.236
33	1.318	0.620	0.400	0.234
34	1.317	0.617	0.397	0.232
35	1.316	0.614	0.395	0.230
36	1.315	0.611	0.392	0.228
37	1.314	0.608	0.390	0.226
38	1.313	0.605	0.387	0.224
39	1.312	0.602	0.385	0.222
40	1.311	0.600	0.383	0.221
41	1.310	0.597	0.381	0.219
42	1.309	0.595	0.379	0.218
43	1.308	0.593	0.377	0.216
44	1.307	0.591	0.375	0.215
45	1.306	0.589	0.373	0.214
46	1.305	0.587	0.371	0.213
47	1.304	0.585	0.369	0.212
48	1.303	0.583	0.367	0.211
49	1.302	0.581	0.365	0.210
50	1.301	0.579	0.363	0.209


Now, I also mentioned there is a table for χ^2 , just like a table for F test, t test, z test and so on. In that table as you can see here, here you have the degrees of freedom, here you have the probability, here you see the two-sided and one-sided probability, generally one-sided 95 % is here. This gives you the lower boundary and this gives you the upper

boundary. Generally, 95 % χ^2 we may end up using this particular column. Let us look at one more problem. I am trying to do as many problems as possible so that you gain lot of confidence in organizing and solving this type of problems.

(Refer Slide Time: 03:51)


Five different antiseptic formulations were given to rural medical centres and the feedback on their acceptability was received. The table shows number of medical centres which accepted or not accepted each of these

Formulations	Accepted	Not accepted
A	14	22
B	12	26
C	20	16
D	13	24
E	17	17



We have 5 different antiseptic formulations. So, 5 different antiseptic formulations are prepared and they are given to rural medical centers and the feedback on their acceptability was received, so accepted not accepted, you get feedback. So, the beauty of it is, it need not be the total, need not be the same. As you can see here you know, total need not be the same you may get different sets of feedback, some rural hospitals may give you feedback, some might not give, so it does not matter. So, we have 5 formulations, accepted for A we get 14, not accepted we get 22 and so on actually. So, we know this is the observed data. Obviously, we need to do the expected data and then we do the χ^2 test statistics calculations. Let us go forward.

(Refer Slide Time: 04:48)



Formulations	Observed		Total
	Accepted	Not accepted	
A	14	22	36
B	12	26	38
C	20	16	36
D	13	24	37
E	17	17	34
Total	76	105	181

Formulations	Expected		Total
	Accepted	Not accepted	
A	15.12	20.88	36
B	15.96	22.04	38
C	15.12	20.88	36
D	15.54	21.46	37
E	14.28	19.72	34
Total	76	105	181

Formulation A, B, C, D, E accepted, this is the observed. I have just copied the whole thing here, not accepted. So, if you add up all this accepted you get 76, if you add up all these not accepted you get 105. If you go across these rows we get 36, this gives you 38, this gives you 36, this gives you 37, this gives you 34 so the grand total is 181. If I want to calculate what is the expected here, I taught you last time it is quite simple $76 \div 181 * 181 * 36$ that will give you 15.12. If you want to know what about this, what is the expected for formulation been accepted? $76 \div 181 * 38$. Here for C, we do $76 \div 181 * 36$ and then for D, we say $76 \div 181 * 37$ and for E, $76 \div 181 * 34$. So, if you want to go here, what do I do? I say $105 \div 181 * 36$, $105 \div 181 * 38$, $105 \div 181 * 36$, $105 \div 181 * 37$, $105 \div 181 * 34$ and finally they will all adapt row wise, column wise, grand total just to cross check whether your calculation are ok. We end up having, these are the observed and these are the expected. These are the observed and these are the expected.

We start doing the test statistics estimation. It is quite simple. What do you do? [E-O]. Do you understand? Divided by expected in the bottom it will come right?.

(Refer Slide Time: 06:47)

E-O	(E-O) ²	(E-O) ² /E
1.12	-1.12	1.25
3.96	-3.96	15.65
-4.88	4.88	23.85
2.54	-2.54	6.43
-2.72	2.72	7.42
		6.16

at DF=4, p=0.05, $\chi^2=9.49$

Cannot reject null hypothesis

We cannot conclude that there is a real difference between each of these formulations

NPTEL

Columns A shows the lower boundaries of the left-sided critical values.
Columns F shows the upper boundaries of the right-sided critical values.

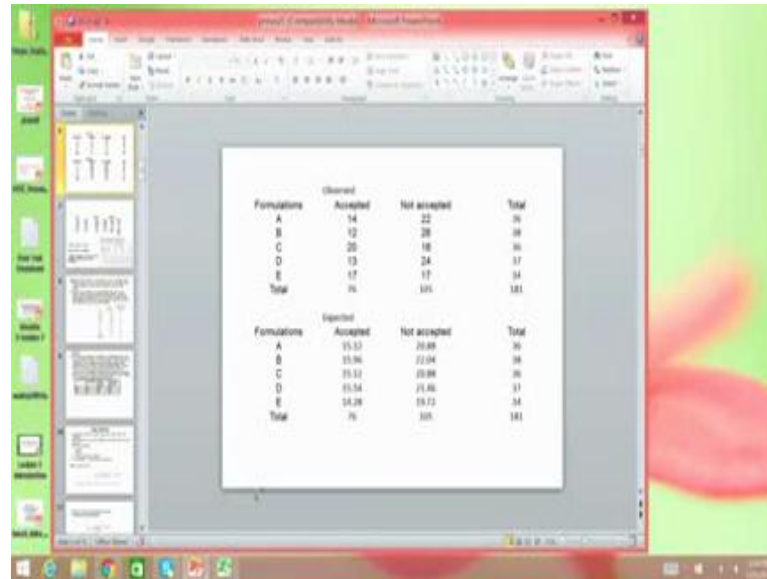
		Level of significance						
Tailed	Observed	0.10	0.05	0.01	0.001			
		0.10	0.05	0.025	0.01	0.005		
v		4	5	6	7	8		
1	0.000	2.71	30.81 ⁰¹	1.64	98.16 ⁰¹	5.02	38.89 ⁰¹	4.61
2	0.20	4.61	5.99	1.90	99.00	7.38	8.02	7.37
3	0.58	6.25	7.88	2.15	99.33	8.02	8.53	11.34
4	1.06	7.78	9.49	2.40	99.49	11.14	9.35	13.28
5	1.60	9.24	11.07	2.62	99.60	12.83	9.89	15.09

So, it can be **observed minus expected** or **expected minus observed** because we are any way going to square it but the bottom will be expected. These are the values; we are having how many? 5 here and another 5 here, totally 10 and then we do the squaring, we do the squaring, squaring, squaring, squaring divided by E, so

$$\chi^2 = \sum_{i=1}^K \frac{(O_i - E_i)^2}{E_i}$$

, it comes out to be 6.16. We have how many degrees of freedom? We have 5 antiseptics, obviously, we have 4 degrees of freedom, **p = 0.5**, 4 degrees of freedom. We go it down here we say 9.49 so obviously, the test statistics is less than the table, so cannot reject null hypothesis. We cannot conclude that there is a real difference between each of these formulations. We have to, as of now say that all the 5 formulations are the same. We cannot conclude because our test statistics is less than the table value. We can do the same problem using Excel also.

(Refer Slide Time: 08:14)



	Observed		Total
Formulations	Accepted	Not accepted	
A	14	22	36
B	12	28	40
C	20	18	38
D	13	24	37
E	17	17	34
Total	76	105	181

	Expected		Total
Formulations	Accepted	Not accepted	
A	15.52	20.48	36
B	15.96	24.04	40
C	15.12	22.88	38
D	15.34	21.66	37
E	14.28	19.72	34
Total	76	105	181

Let me show you how to do it in the Excel. So we do the Excel, copy these 5 items we put them together and then we copy these 5 items and put them together so that is the observed and then we copy these 5 items and put them here and then we copy these 5 items put them here. This gives you the expected, this gives the observed. What do you do? CHITEST, we have to put the **actual ÷ expected**. So, expected is this, this is the actual, we just do it comma, comma so the p value comes out to be 0.723. There is no reason for you to reject the null hypothesis. If you use the GraphPad, GraphPad cannot do a very detailed calculation as I showed you, originally GraphPad can estimate the p value from the **χ^2** or from the p value it can give you a **χ^2** that is all it can do. It cannot really process in detail like excel is able to do. So, what do we do? For a 95 %, for a 95 % and 05, how many degrees of freedom we have? If we look at our problem we have 5 formulations so obviously, we have 4 degrees of freedom. For 4 degrees of freedom it will give you what is the probability, it will give 9.4877.

And in fact, as I originally told you that is exactly like your table. So the GraphPad just gives you the **χ^2** value given your probability of 0.05. The CHITEXT command in Excel can do this problem and it can give you a probability and so we can conclude that there is no reason for us to reject the null hypothesis that means, we can conclude, so we cannot conclude that there is a real difference between each of these formulations, we cannot conclude that there is a difference. So obviously, that is because we have to


accept the null hypothesis.

(Refer Slide Time: 11:12)

• **Example:**

A study compared members of a medical clinic who filed complaints with a random sample of members who did not complain. The study divided the complainers into two subgroups: those who filed complaints about medical treatment and those who filed nonmedical complaints. Here are the data on the total number in each group and the number who voluntarily left the medical clinic. Set up a two-way table. Analyze these data to see if there is a relationship between complaint (no, yes - medical, yes - nonmedical) and leaving the clinic (yes or no)

	No complaint	Medical complaint	Non medical complaint
Total	743	199	440
Left	22	26	28



Now I want to leave it as a homework. Another problem here, I think you people should try to work it out. Here we have 3 situations No complaint, Medical complaints and Non medical complaints. So No compliant, Medical complaint, Non medical complaint and these are the total of people who stayed back and these are people who left the medical clinic. They are asking you to analyze this data to see if there is a relationship between complaint and Non medical that is Medical and Non medical and then leaving the clinic or not leaving the clinic. It is quite simple but only thing is, we need to make two different **2 / 2** type of table. One is we can look at only these Medical complaint, Non medical compliant other one is we can look at No compliant verses compliant. We can do that both actually. No complaint verses total of these complaint or Medical complaint verses Non medical complaint. So I am leaving this as homework for you people to look at actually.

(Refer Slide Time: 12:19)

Weibull distribution

- used in reliability engineering, medical research, quality control, finance, and climatology.
- model time-to-failure data, such as the probability that a part fails after one, two, or more years.
- described by 3 parameters:
 - shape (β)
 - Scale (η)
 - threshold parameters or location (γ).
- The three-parameter Weibull distribution expression, or:

Probability Density Function:

$$f(T) = \frac{\beta}{\eta} \left(\frac{T-\gamma}{\eta} \right)^{\beta-1} e^{-\left(\frac{T-\gamma}{\eta}\right)^\beta}$$

$f(T) \geq 0, T \geq 0 \text{ or } \gamma, \beta > 0, \eta > 0, -\infty < \gamma < \infty$

NPTEL

Now, let us change gears. We will look at another distribution. This is called the Weibull distribution. Weibull distribution is used quite a lot in reliability analysis. So, imagine we have a part, what is the time taken for the part to fail, time to failure that is very very important because nowadays lot of bio materials are implanted into the body one needs to know, how long it will last? Will it last for hours, weeks, months, years? What is the probability associated with that? So it is extremely useful, the Weibull distribution used in medical research, quality control, you are making millions of artificial valve. What is the failure rate with these artificial valves? How **much** percentage of this fails in say 2 months? How much percentage of this fails in 1 year and so on actually?

Finance, of course is quite a lot used. It tells you how reliable a particular share market is and so on actually. Climate, what will be reliability of the climate? So basically, it is looking at time to failure data such as the probability that a part fails after 1, 2 or more years. Like I said, it is extremely useful in bio material research, when you are making various types of bio materials like artificial valves or tendons or some of the knee joints. One is interested to know the reliability of these materials. Will it last for 10 years, given a 95 % probability? Will it last for life long and so on actually? So, the Weibull distribution plays a very important role in identifying such situations. It has got three-parameters: **β** it is called Shape, Scale it is called **η** and **γ** gamma threshold parameter or location actually. So this is generally is called a three-parameter

Weibull distribution.

Shape, Scale, threshold or location it is. They saw the probability density function for a Weibull distribution looks like this,

$$f(T) = \frac{\beta}{\eta} \left(\frac{T - \gamma}{\eta} \right)^{\beta-1} e^{-\left(\frac{T - \gamma}{\eta} \right)^\beta}$$

You do not have to get scared about it but the softwares can determine these know, there are softwares which can give this. So,

$$f(T) \geq 0, T \geq 0 \text{ or } \gamma, \beta > 0, \eta > 0, -\infty < \gamma < \infty$$

f t is always greater than 0 as t go becomes 0 and the gamma that is the threshold parameter location lies between minus infinity to infinity, whereas beta and eta are always greater than 0, understand? This is how the distribution function looks like.

(Refer Slide Time: 15:27)

• Generally, the threshold parameter is set to zero

• 2-parameter Weibull distribution.

$$f(t) = \left(\frac{\beta}{\eta} \right) \left(\frac{t}{\eta} \right)^{\beta-1} e^{-\left(\frac{t}{\eta} \right)^\beta}$$

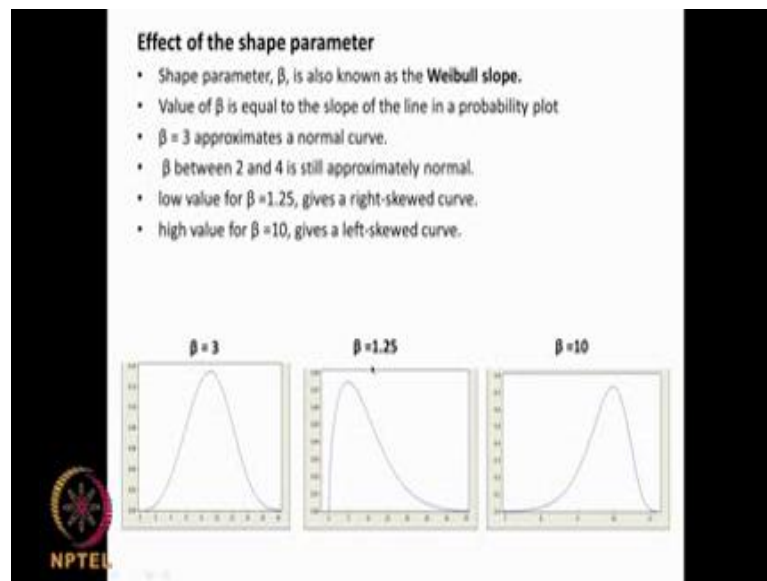
• defined only for nonnegative variables.

NPTEL

Generally the threshold parameter is set to 0 that means the γ is set to 0 then you end up having a 2-parameter Weibull distribution. Do you understand? So here β and

η will always be > 0 . So non negative variables β and η , it is easy look at. Note t is your time, probability, we have 2 different parameters β indicates shape and η indicates scale.

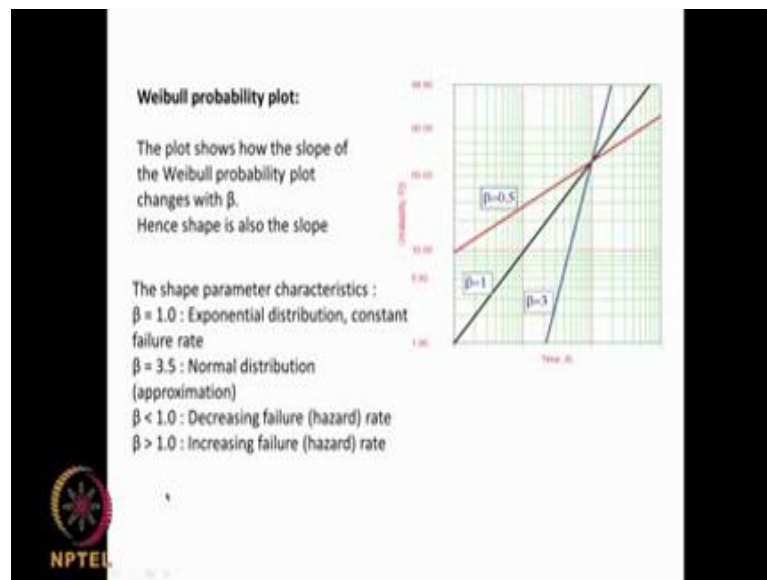
(Refer Slide Time: 16:00)



How does it look like? Let us look at the shape parameter β and let us look at the other parameter the scale parameter η later. Shape parameter, if β is 3, beta is also known as the Weibull slope because you can see here you know it is a multiplication number, so obviously it is called a slope actually. β is equal to the slope of the line in a probability plot. When $\beta = 3$ it looks like a normal distribution, you see this and β between 2 and 4 still approximately normal. So β lower, now you can see this it is a right skewed curve and $\beta > 10$, so you get a left skewed curve that means, data is missing on the left, here you have the data missing on the right. By varying the β we can arrive at different skewed distributions and between 2 to 4 values of β you end up having here approximately a normal distribution. As you can see here when

• β is 3 it is approximately normal distribution.

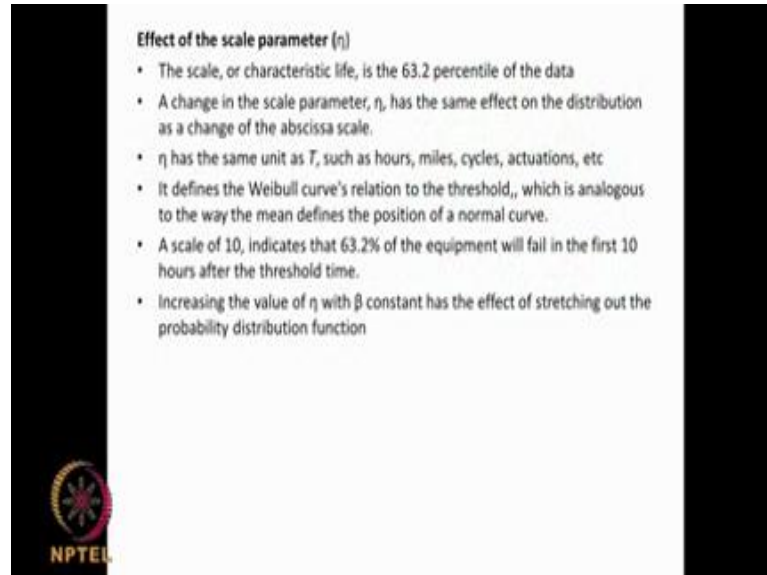
(Refer Slide Time: 17:08)



Let us look at the probability plot. Here we have the time here and then we have the unreliability here that is the f_T here, capital f t as against small f t , please note. So β is the slope, $\beta = 0.05$, $\beta = 1$, $\beta = 3$ and so on actually and generally around 3.5, it is normal distribution actually. $\beta = 1$ is exponential distribution, constant failure rate that means parts will fail in a very constant manner, there is no relationship with respect to time at all, it will keep on failing regularly. $\beta < 1$, decreasing failure rate or decreasing hazard rate, so as time goes up the failure rate goes down. $\beta > 1$, increasing failure rate know some of these, for example electronic components as time goes up the % of failure will also go up actually, so $\beta > 1$. $\beta < 1$, we will say it is an exponential distribution, β about 3.5 is a normal distribution, $\beta < 1$ increasing failure rate, $\beta < 1$ decreasing failure rate actually. Let us look at the scale parameter η . So we looked at β , η is the shape parameter like this actually. It is also an indication of the failure rate. If $\beta > 1$, it is increasing failure rate as time goes up, the % of failure rate also increases, many parts may follow this and $\beta < 1$, it is a

decreasing failure rate that means as time keeps increasing the failure rate goes down that is $\beta < 1$.

(Refer Slide Time: 19:12)



Effect of the scale parameter (η)

- The scale, or characteristic life, is the 63.2 percentile of the data
- A change in the scale parameter, η , has the same effect on the distribution as a change of the abscissa scale.
- η has the same unit as T , such as hours, miles, cycles, actuations, etc
- It defines the Weibull curve's relation to the threshold,, which is analogous to the way the mean defines the position of a normal curve.
- A scale of 10, indicates that 63.2% of the equipment will fail in the first 10 hours after the threshold time.
- Increasing the value of η with β constant has the effect of stretching out the probability distribution function

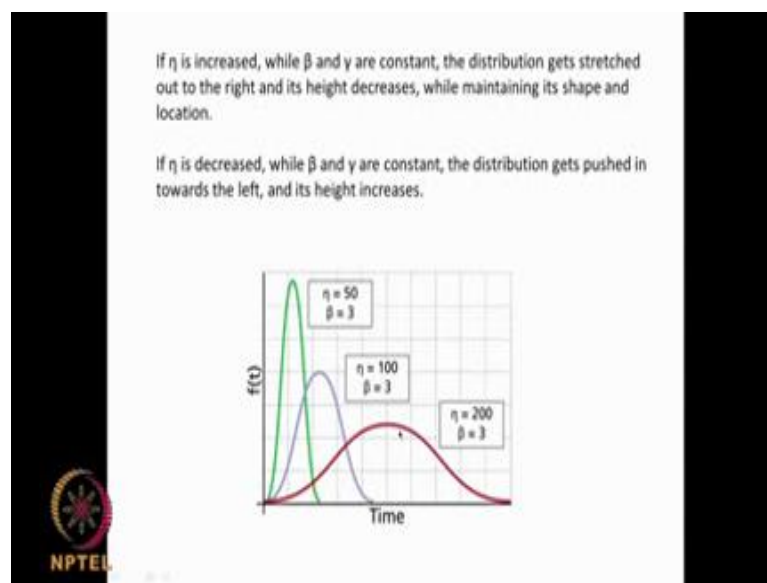
NPTEL

The next one is called the scale parameter, η . It gives you characteristic life, the scale or characteristic life and it is the 63.2 percentile of the data. Now, this η has a same unit as time that is hours, miles, cycles, actuations and so on actually. So when I change η , it has the same effect on the distribution change of the abscissa, it is like the x axis. It defines a Weibull curve's relation to the threshold, which is analogues to the way the mean defines the position of a normal curve. Do you understand? If I have a mean of 20 so the normal distribution will get shifted more in the 20, if I have the mean of say 5 then normal distribution will be closer to 5, do you understand? So it gets moved this way, that way and that is called the position of the normal curve. So η is an indication of that, a scale of 10 indicates that 63.2 % of the equipment will fail in the first 10 hours after the threshold time. So when I say $\eta = 10$, it is extremely useful because we can say 63.2 of % of the equipment will fail in the first 10 hours after the threshold time.

Ideally, we would like to have larger, larger η . Increasing the value of η with

β constant has the effect of stretching out the probability distribution function, you understand? Ideally, I should have a very large β so that the failure can be pushed more and more, more to larger times, understand? So both are very logical the first one is called the Shape, it tells you how it looks like whether it is normal, whether skewed. If it is < 1 , it is decreasing failure, > 1 it is increasing failure rate. $\beta = 1$, constant failure rate. So irrespective of time, failure happens regularly, regularly, regularly and so on actually. Whereas, η is like stretching your moving your curved to the right, so larger the η , longer is it for failure. If $\eta = 10$, it indicates 63.2 % of the equipment will fail in the first 10 hours after the threshold time.

(Refer Slide Time: 21:49)



If η is increased while β and γ are kept constant, the distribution gets stretched out to the right, as you can see in this picture, right?. We have $\beta = 3$, $\eta = 50$ that means, 63.2 % of equipment will fail in the 50 time units whereas, if I put $\eta = 100$, so the 63.2 will fail in 100 time minutes. The time is hours so in 100 hours 63.2 of minute % of the material will fail. If $\eta = 200$ then 63.2 % of the

material will fail within 200 hours. If η is decreased while β and γ are constant, the distribution gets pushed towards the left and its height increases. Whereas, if η is increased the height will decrease but the graph will get pushed more and more to the right. Do you understand? It plays a very important role.

(Refer Slide Time: 23:00)

Effect of the threshold parameter

- The threshold is a shift of the distribution away from 0.
- A negative threshold shifts the distribution to the left of 0, whereas, a positive threshold shifts the distribution to the right of 0.
- All data must be greater than the threshold.
- A 2-parameter Weibull (4,100) distribution is exactly the same as a 3-parameter Weibull(4,100,30) except for the 3-parameter Weibull is shifted 30 units to the right of 0.
- Weibull distribution is an alternative to the normal distribution in the case of skewed data.
- The exponential distribution is a special case of Weibull distribution ...
..... frequently used to study the scattering of radiation or wind speed.
- γ has the same units as t

The slide includes a graph showing two probability density function curves on a grid. The x-axis is labeled 'Time, (t)' and the y-axis is labeled 'f(t)'. One curve starts at the origin (0,0) and rises to a peak. The second curve starts at a positive value on the x-axis (representing a threshold γ) and rises to a peak that is lower than the first curve's peak. This illustrates how a positive threshold shifts the distribution to the right and reduces its maximum height.

The threshold parameter γ , we generally said we do not have to bother about it. So threshold parameter generally indicates the shift of the entire distribution to the right or to the left. If it is 0, we can say the graph will start from the origin, if we give some value the whole graph gets shifted here as you can see in this picture. A negative threshold shifts the distribution to the left of 0 whereas, a positive threshold shift it to the right. All data will be greater than that the threshold. So a 2 parameter Weibull distribution like say 4,100 exactly same as a three-parameter Weibull distribution 4,100,30 except the whole thing has been shifted by 30 units to the right of 0 actually. So if this is on 10, 20, 30 a 2-parameter 4, 100 is same as a three-parameter 4, 100, 30 but only thing is the graph will look like this. Weibull distribution has alternate to the normal distribution in the case of skewed data, if the data is skewed we can use Weibull distribution. The exponential distribution is a special case of Weibull. It can be used frequently used to study the scattering of radiation or wind speed because I said if $\beta = 1$ constant failure rate. So irrespective of time, a constantly, mid parts fail actually. So, wind speed follows may be

exponential distribution. Now γ also has the same units as t because it is shifting the graph to the right or to the left, so it should have the time unit. So γ and η will have the units of time, understand?. We will continue further on this Weibull distribution in the next class.

Thank you very much.

Key Words: Weibull distribution, reliability analysis, shape, scale, threshold or location parameter, skewed distribution, failure rate, exponential distribution, normal distribution.