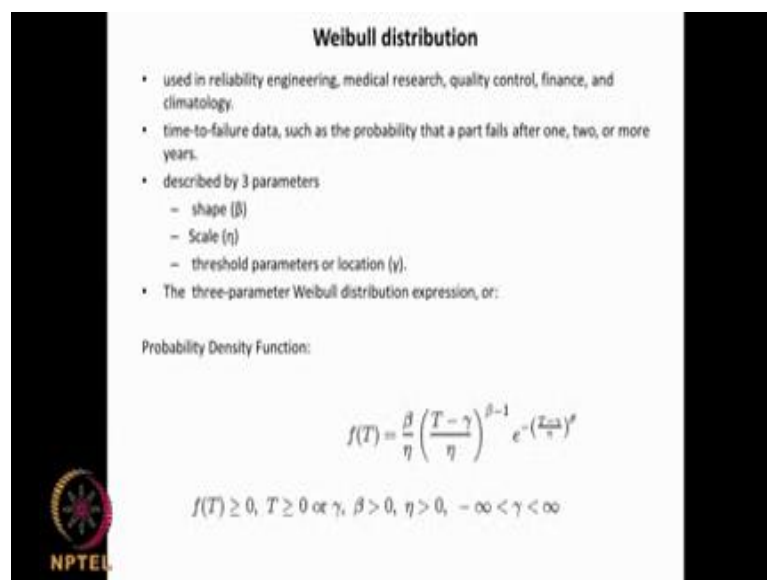


Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 27
Weibull distribution

Welcome to the course on Biostatistics and Design of Experiments. We will continue on this Weibull distribution and as I said Weibull distribution is very important in reliability of parts, various bio materials in plants, devices, machine parts, bolts and so on actually.

(Refer Slide Time: 00:22)




Weibull distribution

- used in reliability engineering, medical research, quality control, finance, and climatology.
- 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:

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

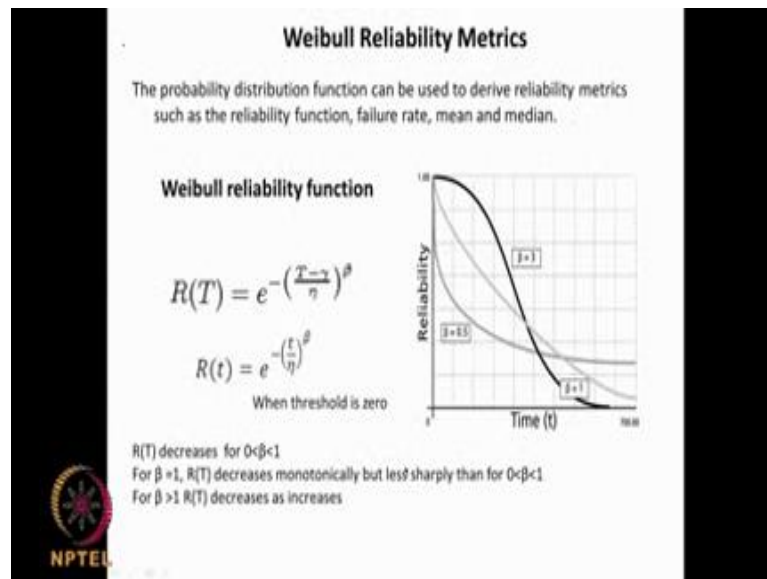
$f(T) \geq 0, T \geq 0 \text{ or } \gamma, \beta > 0, \eta > 0, -\infty < \gamma < \infty$



Even your tube lights, even your lights, your bulbs, your laptops. Anything can have a reliability and Weibull distribution can be used there actually. It determines time to failure, how many days it may take to fail or how many hours it may take to fail, how many weeks or years it may take to fail. So, it is got three parameters, shape parameter called β , scale parameter called η , threshold parameter called γ . $\beta, \eta > 0, \gamma$ can be < 0 also or > 0 . γ is helping you to shift the entire equation to the right it depending upon the γ value. β is called the shape, so as the β goes up to 3, you end up having a normal distribution. β is lower or higher, you can have a skewed distribution left skewed, right skewed and so on and η is a scale parameter, it sort of tells you where the maximum occurs. So, lower η , the maximum occurs at lower value, higher η maximum occurs at

higher value.

(Refer Slide Time: 01:51)



Then there are other terms, they are called Weibull Reliability Metrics. So, **1 - reliability** can be unreliability, it is given by this formula

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

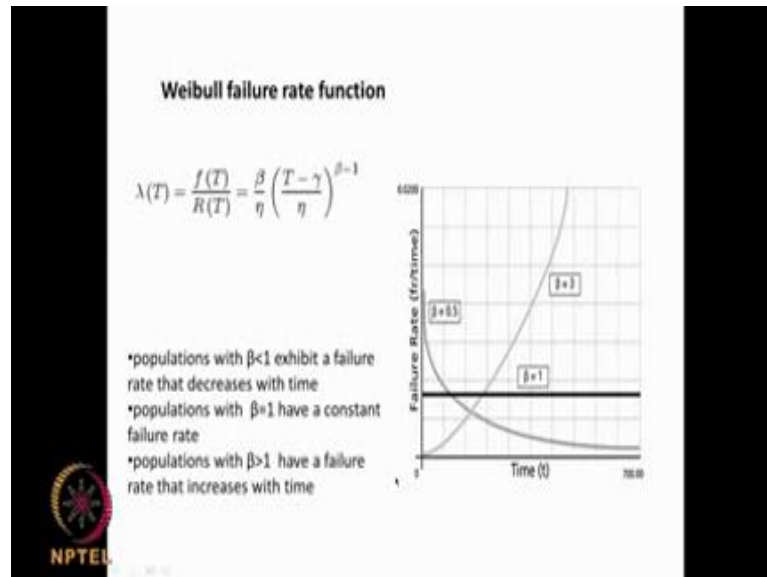
. If **your γ** is 0, that means if the threshold is 0, then you can call it

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

. As you can see depending upon the **β** , you can have different types of curves here. You can have different types of curves depending upon the **β** value here, then we call the Weibull failure rate function, which is given by this,

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

(Refer Slide Time: 02:33)



If you remove γ so it will become

$$\frac{\beta}{\eta} \left(\frac{T}{\eta} \right)^{\beta-1}$$

So, if $\beta < 1$, their failure rate decrease at the time, if $\beta = 1$, failure rate is constant with respect to time, if $\beta > 1$, failure rate increases with time.

(Refer Slide Time: 03:07)

Weibull mean life, or mean time to failure MTF

$$\bar{T} = \gamma + \eta \cdot \Gamma\left(\frac{1}{\beta} + 1\right)$$


where $\Gamma(\cdot)$ is the gamma function

$$\Gamma(n) = \int_0^{\infty} e^{-x} x^{n-1} dx$$

Median life, or B_{50} life, for the Weibull distribution

$$\check{T} = \gamma + \eta (\ln 2)^{\frac{1}{\beta}}$$

$\Gamma(-1) = (-2)! = \infty$
$\Gamma(0) = (-1)! = \infty$
$\Gamma(1) = 0! = 1$
$\Gamma(2) = 1! = 1$
$\Gamma(3) = 2! = 2$
$\Gamma(4) = 3! = 6$
$\Gamma(-\frac{1}{2}) = \frac{1}{\sqrt{\pi}} \approx 2.363271801207$
$\Gamma(-\frac{3}{2}) = -\frac{2}{\sqrt{\pi}} \approx -3.544907701811$
$\Gamma(\frac{1}{2}) = \sqrt{\pi} \approx 1.772453850905$
$\Gamma(\frac{3}{2}) = \frac{1}{2}\sqrt{\pi} \approx 0.88622692545$
$\Gamma(\frac{5}{2}) = \frac{3}{4}\sqrt{\pi} \approx 1.32934038818$
$\Gamma(\frac{7}{2}) = \frac{15}{8}\sqrt{\pi} \approx 1.32333907045$



Then we introduced another term that is called mean time to failure. That is mean, what is the average time for failure. It is given by this relation,

$$\bar{T} = \gamma + \eta$$

$$\bar{T} = \gamma + \eta \cdot \Gamma\left(\frac{1}{\beta} + 1\right)$$

This is a γ function

is your shape factor, γ is your moving factor, η is your scale factor. This γ function is given by, so please do not confuse with this γ and this γ . This is a γ function and this is a γ parameter. So this $\Gamma(\cdot)$ is given by 0 to ∞ ,

$$e^{-x} x^{n-1} dx$$

, so you do not get scared by this, so depending upon that the term inside, we will know

what is a gamma function value using this table.

For example, $\Gamma(1)$ is 1, $\Gamma(2)$ is 1, $\Gamma(3)$ is 2, $\Gamma(4)$ is 6 and so on. The median life or B_{50} life for the Weibull distribution is given by

$$\check{T} = \gamma + \eta (\ln 2)^{\frac{1}{\beta}}$$

. We are going to use these at some point. We did some problem in the previous case; we can use it especially in bio material body parts. How long? What is the probability that the bio material will last for so many hours? So that sort of problems are attempted. Now I need to know how to get these parameters? How do I get β, η, γ from the data? And that is what we are going to talk about in today's class. So if you take the unreliability function and you get rid of this γ so

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

and then $1 -$ that gives you the unreliability function.

(Refer Slide Time: 05:04)


In the case of the 2-parameter Weibull, the cdf (also the unreliability $Q(t)$) is given by:

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

This can be linearized by:

$$Q(t) = 1 - e^{-\left(\frac{t}{\eta}\right)^\beta}$$
$$\ln(1 - Q(t)) = \ln\left[e^{-\left(\frac{t}{\eta}\right)^\beta}\right]$$
$$\ln(1 - Q(t)) = -\left(\frac{t}{\eta}\right)^\beta$$
$$\ln(-\ln(1 - Q(t))) = \beta \left(\ln\left(\frac{t}{\eta}\right)\right)$$
$$\ln\left(\ln\left(\frac{1}{1 - Q(t)}\right)\right) = \beta \ln t - \beta \ln(\eta)$$

This forms $y = mx + b$

$$y = \ln\left(\ln\left(\frac{1}{1 - Q(t)}\right)\right) \quad \text{and} \quad x = \ln(t)$$


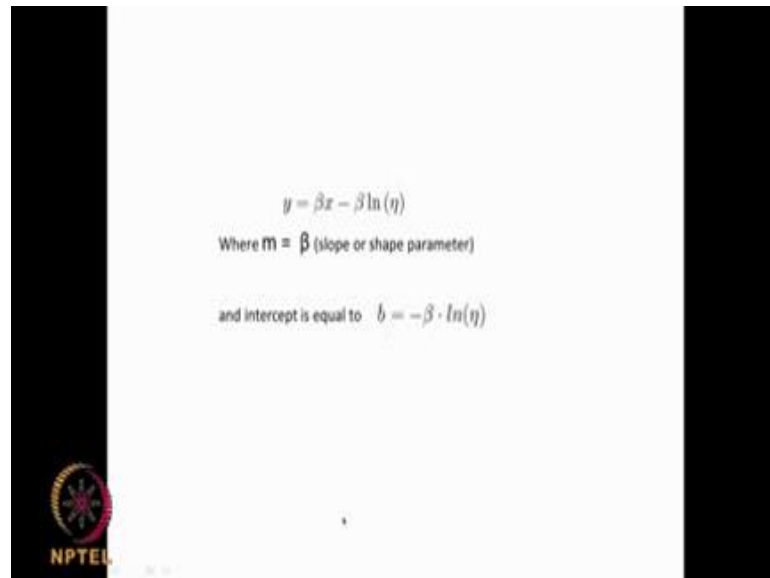
This can be linearized, last time I mentioned this can be linearized. If you keep linearizing this, how do you linearize? You take the logarithm on both sides then again you take logarithm and so on. You end up having of

$$\left(\ln\left(\frac{1}{1 - Q(t)}\right)\right) = \beta \ln t - \beta(\eta)$$

so this looks like a linear regression relation. This is why, it is on the left hand side t is your time, $\ln t$ is your x so β is your m slope and this whole thing could be the intercept. If I plot this whole thing as a function of $\ln t$, I will get the slope as β and the intercept as

$$b = -\beta \cdot \ln(\eta)$$

(Refer Slide Time: 05:56)



$$y = \beta x - \beta \ln(\eta)$$

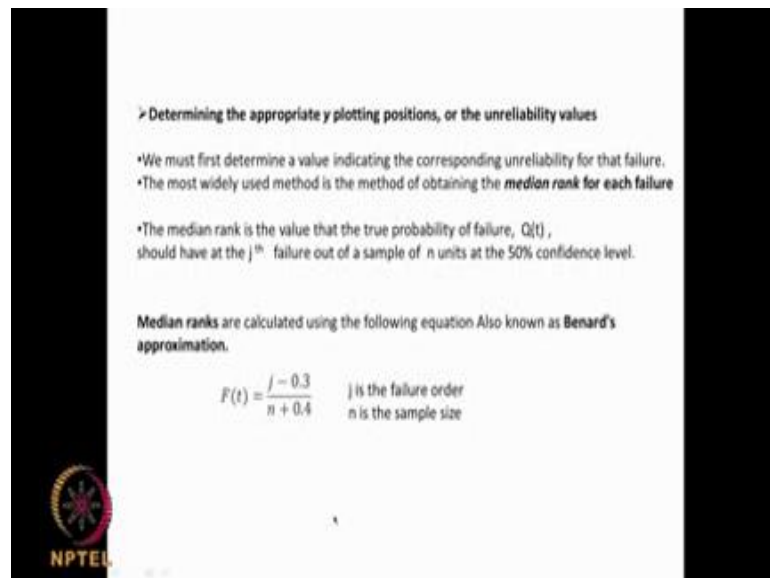
Where $m = \beta$ (slope or shape parameter)

and intercept is equal to $b = -\beta \cdot \ln(\eta)$

NPTEL

It is very easy to do that, but it is not very easy to do that because we need to look at many many things as we go along.

(Refer Slide Time: 06:06)



➤ Determining the appropriate y plotting positions, or the unreliability values

- We must first determine a value indicating the corresponding unreliability for that failure.
- The most widely used method is the method of obtaining the **median rank for each failure**
- The median rank is the value that the true probability of failure, $Q(t)$, should have at the j^{th} failure out of a sample of n units at the 50% confidence level.

Median ranks are calculated using the following equation Also known as Benard's approximation.

$$F(t) = \frac{j - 0.3}{n + 0.4}$$

j is the failure order
 n is the sample size

NPTEL

There is something called median ranks which you need to calculate and then you need to plot that actually. We need to determine the appropriate y plotting positions, because I need to know how to do this, this is the y plotting position, so I need to know this. So in order do that it is bit tricky, first we need to determine a value indicating the corresponding unreliability for that failure and general method that is used, it is called

the median rank for each failure. The median rank is the value that the true probability of failure $Q(t)$ should have at the j th failure out of a sample of n units at the 50 % confidence level.

How do you calculate median rank? We use this formula

$$F(t) = \frac{j - 0.3}{n + 0.4}$$

j is the failure order and n is the sample order. So, if I have 10 sample and this particular part fails at the third position, failure order is 3 then $3 - 0.3 \div 10 + 0.4$. That will be the median rank of that particular part, understand? So the whole problem here is; How do I estimate this q ? In order to estimate the q , there is an approach called median rank for each failure. So we are trying to determine that, it can be calculated by using this particular formula. You think it is of no use but then look at this problem.

(Refer Slide Time: 07:45)

Example:
 A company wants to compare the dependability or reliability of two proposed designs of heart diaphragm valves. The desired reliability at 400,000 cycles is 0.90, i.e. 90 % of the valves to survive at least 400,000 cycles. It can be mathematically expressed as $R(400,000) = 0.90$. Ten samples from each of the two designs (Design A and Design B) were tested until their action failed.

Design A		Design B	
Sample	Cycles	Sample	Cycles
1	726,044	11	529,082
2	615,432	12	729,957
3	508,077	13	650,570
4	807,863	14	445,834
5	735,223	15	343,280
6	848,953	16	959,903
7	384,558	17	730,049
8	666,686	18	730,640
9	515,201	19	973,224
10	483,331	20	258,006

A company has two designs for manufacturing heart diaphragm valves. They call it design A, they call it design B. The desired reliability at 400,000 cycles is 0.9 that is 90 % of the valves should survive at least 400,000 cycles. So you take 10 samples from each of these design and they were tested and at which cycle they failed is plotted here. For example, this sample 1 failed at 726,044 th cycle. Sample 2 failed at 615,432 nd

cycle. They are all from the design 1 or design A and if you look at these samples the sample 11 which was taken from design B, failed at 529,082nd cycle. Now the question is, How do I calculate this $Q(t)$, the unreliability? So what do we do is? First we need to organize this data.

(Refer Slide Time: 08:59)

Answer:

1. First consider the Design A. Type the data in excel sheet and sort them in ascending order of failure cycles and then rank them 1 to 10 as given the table
2. Using the Linearizing the Unreliability Function
3. Now, estimate of the proportion of the population that will fail by the number of cycles listed. This can be done by determining the median rank.

$$F(t) = \frac{j - 0.3}{n + 0.4}$$

j is the failure cycles and $n = 10$

In excel:
 *enter the formula:
 $=((82-0.3)/(10+0.4))$

*Then determine
 $1/(1 - \text{Median Rank})$

*Followed by
 $\ln(\ln(1/(1 - \text{Median Rank})))$

Also determine $\ln(\text{Design A Cycles})$.

Design A Cycles	Rank
384.558	1
483.321	2
508.077	3
515.201	4
615.452	5
668.888	6
728.544	7
759.223	8
807.863	9
848.853	10

Design A Cycles	Rank	Median Rank	1/(1-Median Rank)	ln(1/(1-Median Rank))	ln(Design A Cycles)
384.558	1	0.07207292	1.07218498	-2.892842085	12.8586499
483.321	2	0.18181818	1.195402299	-1.72326719	13.086417
508.077	3	0.29090909	1.388648181	-1.202023118	13.1383629
515.201	4	0.35454545	1.532238808	-0.821986118	13.15221238
615.452	5	0.41818182	1.64981454	-0.58588304	13.33207974
668.888	6	0.48181818	1.74298987	-0.285905445	13.41307443
728.544	7	0.54545455	1.81381041	-0.13204982	13.4953619
759.223	8	0.60909091	1.86181818	0.28632032	13.53478638
807.863	9	0.67272727	1.89772727	0.58872217	13.60214777
848.853	10	0.73636364	1.92272727	0.94288929	13.6117191

First take design A, let us do design a put this data in excel for example, and then sort them in ascending order. So, 384.558 this sample fails the earliest, right in within very short cycle this will come first and then this will come second like that you know you put them in that order. So rank them 1, 2, 3, 4, 5, 6 up to 10. Now the question is how do I calculate these q? In order to calculate this q, as I said we need to use this Benard's approximation,

$$F(t) = \frac{j - 0.3}{n + 0.4}$$

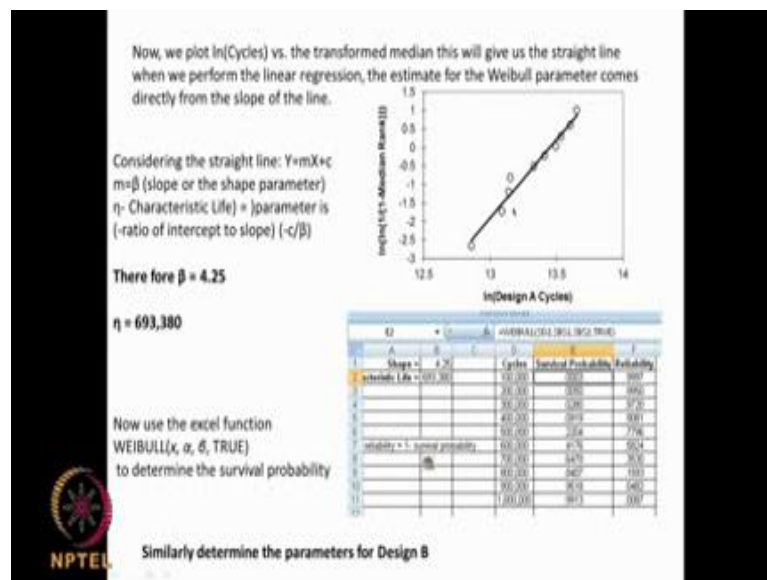
j is a failure order, n is the number of data points here n is 10. So this particular has a failure order of 1. So same thing I am putting it in excel for example, like this, so this has a failure order of 1. So $1 - 0.3$ is 0.7, $0.7 \div 10$, 0.4 will be 0.067.

Like that for each one of these items, you find out a median rank. Then you calculate $1 \div 1$ minus median rank. Then you take logarithm of that, then again you take logarithm of

that. Why do you need to do all these? Because if you remember,

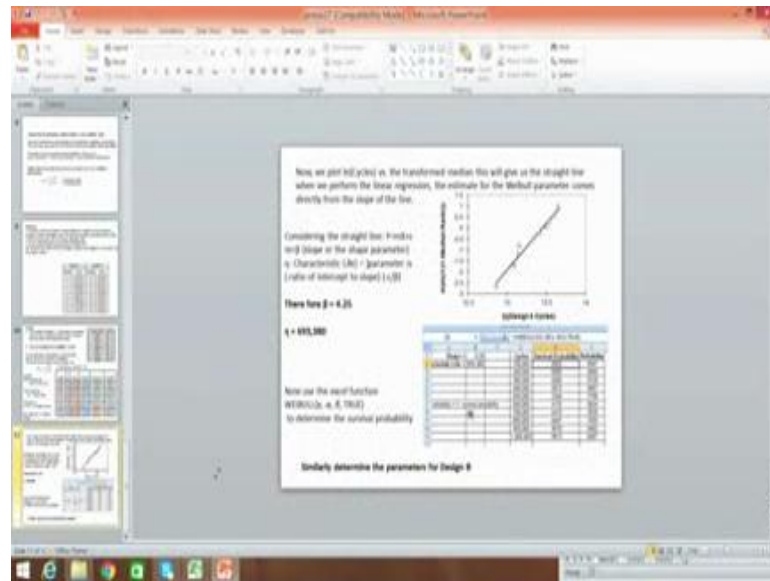
logarithm of logarithm of $1 \div 1 - Q(t)$ versus logarithm of t is what we are going to plot. So in order to estimate $Q(t)$, we use this type of approximation, the median rank approach and median ranks can be calculated using Benard's approximation. Do you understand? So we are doing logarithm of logarithm of this, and then other side is do the logarithm of design a cycles, that is logarithm of each one of this. You are going to plot this in the y axis, this in the x axis, so when you plot these 2, as you can see from our figure, slope will give you the β and the intercept will give $-\beta \ln \eta$. If I divide this by this, I should be able to get my η .

(Refer Slide Time: 11:23)



I am plotting logarithm of logarithm of 1 divided by 1 minus median rank because median rank is considered to be your $Q(t)$. Then this the design cycle, logarithm of design cycle data falls here, you get the slope, slope is your β . β and then the other parameter, η is calculated by, in ratio of intercept by slope. Because that is given by this β , so if you divide you will get the η ; η is given by this. So β is 4.25, that is the shape factor, η is 693,380. What is the question? The desirability at 400,000 cycle is 0.9. Now, we can calculate the survival probability using the Weibull function available in excel, we know 400,000 as our data x and we know β as 4 point, sorry the α as 4.25, β as 693,380 when we do that we can calculate. Let get the survival problem.

(Refer Slide Time: 12:56)



Let me show how to do that, so we have a Weibull; 400,000 comma 4.25 comma 693380 true. We get 0.092 that is the survival probability for 400,000 cycles; 0.092. Similarly we can do it for different numbers also instead of 400,000 we can do it for 300,000, we can do it for 500,000 and so that gives you survival probability for different cycles. We get 0.028 and similarly we can do it for 500,000 also.

So that gives a point survival probability of 0.22055. For different values of these cycles like 300,000, 400,000, 500,000 we can calculate the survival probability using this Weibull function. The other one, the design B same type of calculation and we can determine the survival probability. And then from these two we can decide which design to accept. This is how you need to do, it is looks very complicated but then it is a very useful type of approach, let me explain once more for the convenience. We are looking at the reliability, reliability function is given by this,

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

. So the unreliability will be 1 - that, $1 - e^{-t/\eta}$, raise to the power β . Now, I want to estimate the η and β from the experimental data and then use that β and η for doing some performing some calculations. How do I do that? Now if I take logarithm of this data, this equation this is the unreliability equation $Q(t) = 1 - e^{-t/\eta}$, take a one logarithm

then again I take logarithm, the equation ends up like this. So this looks like y, this is your x, t is your time. Slope will give you β and intercept will give this whole thing. Right? So we can calculate β and η , from the experimental data.

But it is not so simple we will come across that later actually, but basically if I am able to get a q that is if I am able to get a unreliability value, then if I plot, keep that as my y axis and then $\ln(t)$ as my x then, from the slope and intercept I should be able to get β and η . Then I can use it further for other type of calculation later. How do I do this? That is where the median rank for each failure comes into picture. That means we need to calculate the median ranks for the failure. What do you do? Median rank is the value that the true probability of failure should have at the j th failure out of a sample of n units at the 50 th % confidence level. How do you calculate median ranks? It is given by this particular relationship called Benard's approximation. **This**

$$F(t) = \frac{j - 0.3}{n + 0.4}$$

; j is the failure order, n is the sample size.

So if I have 20 samples and if a particular sample failure order is say 4, then j will be 4 and n will be 20 that will give me my Q(t). It is not so simple as we come along later, a company wants to compare the dependability or reliability of 2 proposed designs of heart diaphragm valves. So they want to have at least 90 % of the valve should survive 400,000 cycles, so they have 2 designs. They have 10 samples from design 1 and check the number of cycles it takes for it to fail, so you can see sample 7 fails at this particular cycle the lowest. The sample 4 survives up to 807,000 th cycle, the design B you have sample 15 failing at the earliest and sample 16 lasting for, no, sample 19 lasting for much longer time, sorry, sample 20 is quickest and sample 19 lasting the longest. So we use calculate the median ranks, so you put them first of all you arrange them in the order of increasing, so this sample fails at the earliest, this sample fails at the last.

Then what do you do? You put $j - 0.3$, so for this sample j is 1, n is 10; for this sample j is 2, n is 10 and so on, so we calculate the median rank. Then $1 \div 1 - \text{median rank}$ is what you do, that means $1 \div 1 - 0.06$ that will give you this and so on actually; like that you calculate. Then you take a logarithm of this and then again take a logarithm, 2 times you

take the logarithm. Why? Because if you remember this, you have to take 2 times logarithm and you plot keeping this as y axis and logarithm of t as your x axis, that is the design cycles is your t. So you take the logarithm, you plot this versus this. You get a beautiful straight line, logarithm of designs cycles here, here you plot the $\ln \ln$ of $1 \div 1 -$ **medium rank**. So the slope will give you the β and intercept divided by slope will give you the η . Because see as you can see here, this intercept this the slope, so divide this by that we end up getting the η . β is there, η is there so I can use the Weibull distribution for any value of x. For example, I want to see, I want last it for 400,000 cycles I substitute here and then I put true and I can calculate the probability. So it is very simple, you have done it for design A, now you go about doing it for design B. Repeat the whole process. So you arrange them in the ascending order so this will come as 1, then this will come as 2, this will come as 3 and so on and this will come as 10.

Once you arrange them, then you calculate the rank, median ranks so this will remain the same, so $1 \div 1 -$ **median rank** also will remain the same, this also will remain the same only thing is the \ln of design cycles will have different sets of numbers. When you plot that as the x axis and this as the y axis, so you may get a different slope and an intercept, correspondingly the slope will give you the β for the design B and corresponding η will get for design B. So you will have β and η for design A. Similarly, we will calculate the β and η for design B and then we can do a Weibull for a design B, using this Weibull command available in excel and then we can compare the results with that we got for design A.

So it is very simple to do, but it is quite an involved problem. This is a very useful type of approach especially if I am taking samples and then measuring their failure rates. If you already I know the β and η it is very very simple. I can use the Weibull function to get the probability of survival and so on. But this is more like you collect the samples from their failure rates; you estimate the β and η that means it is curve fitting and parameter estimation and then use that for further calculation, that way this is very useful approach.

As I said Weibull distribution is important especially in bio materials type of research, like I showed you this problem. Two designs are available for a heart diaphragm valve and you want to decide which design to take up. I have some body implant and I want to know, what is the failure rate? How long it will survive? Will it survive for years or

months or days? If I have different designs? So which design will last much longer? So in that sort of situation Weibull distribution is very useful. Weibull distribution has 3 parameters like I showed you the β and η and γ . β is called the shape, shape means whether it is like a normal or whether it is skewed left or right. The scale is the η it basically moves the maximum to the right or to the left. The maximum can occur later or earlier, so obviously if I want to increase the reliability, I want to push it more to the right.

The γ is the threshold parameter location if γ is 0 that means your result data starts from the origin. Whereas if $\gamma > 0$, that means the data is starting after the; it is getting shifted the whole curve get shifted to the right. Whereas the $\gamma < 0$ the whole curve gets shifted to the left. So basically γ is what it does? It shifts the curve either to the right or to the left completely. So these are the three parameters available in Weibull distribution.

We can use Weibull distribution for determining reliability function, like we talk discussed about recently and then we can use it for failure rate function; that means what is the rate of failure as a function of time, where as in the reliability what is the reliability as a function of time. After 1 year how reliable it is? After 2 years how reliable it is? After 10 years how reliable it is? Like that. Then we can calculate the mean time to failure, What is the mean time for it to fail? Here the equation looks like this, this is a $R(t)$ function which is given by this relationship and then we can also calculate median life, median life of this of a data set. Then we can also calculate unreliability function $Q(t)$, which is given by $1 - \text{the reliability function}$, 1 minus of this will give the unreliability function.

So there are different functions available in Weibull which have their own usefulness and significance in many problems which involve reliability. As I said reliability is even important whether you take a light bulb, whether you take a laptop, whether you take switches, whether you take a heart valve or whether you consider an implant and so on actually. So we will continue on some other topic in the next class.

Thank you very much for your time.

Key Words: Reliability, reliability function, mean time to failure, failure rate, probability, median life, median rank, unreliability function