Inferences and Notes about Weibull Distributions and their Applications

ABU ELGASIM ABBAS ABOW MOHAMMED
College of Business and Economics
Qassim University Kingdom Saudi Arabic
College of Economics and Political Science
Omdurman Islamic University Sudan

Abstract

This paper used a Weibull distribution to study the breast cancer data taken from the Ministry of Health in Khartoum State. The paper found that the linear regression model can also be used to numerically assess good ness of fit and estimate parameters of the Weibull distribution. The gradient informs one directly about the shape parameter β and the scale parameter a can also be inferred. Inside that the failure rate decreases over time.

Key words: Weibull distribution, estimate parameters, fit model, failure rate, breast cancer.

0.   INTRODUCTION

In probability theory and statistics, the Weibull distribution is a continuous probability distribution. Itis named after Swedish mathematician Waloddi Weibull, who described it in detail in 1951, although it was first identified by Fréchet (1927) and first applied by Rosin & Rammler (1933) to describe a particle size distribution [1]. Weibull distribution provides a useful description for many different kinds of data , especially in emerging areas such as electrical engineering to represent overvoltage occurring in an electrical system and industrial engineering to represent manufacturing and delivery times, reliability engineering and failure analysis and in the weather

forecasting and the wind power industry to describe wind speed distribution often matches the Weibull shape, and finance such stock prices and actuarial data, in general insurance to model the size of reinsurance claims, and the cumulative development of asbestosis and a general model forecasting technological change[2] The Weibull can also fit a wide range of data from many other fields, including: biology, economics, engineering sciences, and hydrology Rinne [3].

1.   Problem Statement

The Weibull distribution is important in the scientific field in general and statistics in particular the problem of the paper is that the distribution of Weibull has several features varies according to parameters, this paper try to introduce model of Weibull distribution and the usefulness of using this distribution in data of breast cancer.

2.   Objective of Study

The objectives of paper represented the following:

a.   Estimate the Weibull distribution using breast cancer data
b.   data modeling under study
c.   diagnosis of the model of the breast cancer data.

3.   Importance Study

The importance of this paper stems from the role played by the distribution of the Weibull in many different fields and its use of the study data.

4.   Limitation of Study

The limitations of this study are breast cancer data from the ministry of Health in Khartoum state for year 2016.

5. Study Hypotheses

H_0 : Weibull distribution is not a good fit to the breast cancer data.

H_1 : Weibull distribution is a good fit to the breast cancer data.

6.   Literature Review

some method for estimating Weibull parameters (2000), by Al-Fawzan in this study estimated shape parameter and scale parameter and classified method in two categories: graphical and analytical method [4]. A comparison of method for the estimation of Weibull distribution

parameters (2014), by Nwobi and Ugomma the study provided the different methods for estimation of the parameter of Weibull distribution. The result is showed that the maximum likelihood estimation method (MLE) signicaltly out performed other method [5] This study presents method for estimating Weibull distribution parameters as well as the possibility of constructing a predictable model.

7. Methodology

The paper was based on theoretical approach that dealt with Weibull distribution and supported the practical side that depends on the data of breast cancer from Ministry Heath of Khartoum state and used the SPSS and Excel for analyzing data.

8. Theoretical Formulation

8.1 Weibull Distribution PDF

Two versions of the Weibull probability density function (pdf) are in common use: the two parameter pdf and the three parameter pdf this explained in the text below.

8.2 Three parameter Weibull

The formula for the probability density function of three parameter general Weibull

$$f(x) = \frac{\beta}{\alpha} \left(\frac{x-\mu}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x-\mu}{\alpha}\right)^{\beta}} \tag{1}$$

Where $f(x) \geq 0, x \geq 0, or \mu \geq 0, \beta > 0, \alpha > 0, -\infty < \mu < \infty$

B is shape parameter (also known as Weibull slope or the threshold parameter) and α is scale parameter, also called the characteristic life parameter μ is the location parameter, also called the waiting time parameter or sometimes the shift parameter.

8.3 Two parameter Weibull

The formula is practically identical to the three parameter Weibull, except that μ isn't included:

$$f(x) = \begin{cases} \frac{\beta}{\alpha} \left(\frac{x}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x}{\alpha}\right)^{\beta}} & x \geq 0 \\ 0 & otherwise \end{cases} \tag{2}$$

The two parameter Weibull is often used in failure analysis, because no failure can happen before time zero. If know μ, the time where the failure happens, we can subtract it from x (i.e. Time x). Therefore,

when we move from the two-parameter to the three-parameter version, all you have to do is replace each instance of x with (x-μ).

8.4   Characteristics of the Weibull distribution

a) Weibull shape parameter β

The Weibull shape parameter, ß, is also known as the Weibull slope. This is because the value of ß is equal to the slope of the line in a probability plot. Different value of the shape parameter can have marked effects on the behavior of the distribution in fact, some values of the shape parameter will cause the distribution equations to reduce to those of the of the other distribution. For example, when ß= 1, the pdf of the three-parameter Weibull reduces to that of the two-parameter exponential distribution. R; P [6].

Looking at the same information on a Weibull probability plot, one can easily understand why the Weibull shape parameter is also known as the slope. Later we will show how the slope of the Weibull probability plot changes with ß. So we show that the models represented by the figures all have the same α.

Another characteristic of the distribution where the value of ß has a distinct effect is the failure rate. The following plot shows the effect of the value of ß on the Weibull failure rate.

This is one of most important aspect of the effect of ß on the Weibull distribution. As is indicated by the plot, Weibull distributions with $ß < 1$ have a failure rate, that decreases with time, also known as infantile or early-life failures. Weibull distribution with ß close to or equal to 1 have a fairly constant failure rate, indicative of useful life or random failures. Weibull distribution with $ß > 1$ have a failure rate that increases with time, also as wear-out failures. Mixed Weibull distribution with one subpopulation with $ß < 1$, one subpopulation with $\beta = 1$ and one subpopulation with $\beta > 1$ would have a failure rate plot that was identical to the bathtub curve.

b) Weibull scale parameter α

A change in the scale parameter α, has the same effect on the distribution as a change the abscissa scale. Increasing the value of α while holding ß constant has the effect of stretching out the pdf. Since the area under a pdf curve is a constant value of one, the "peak" of the pdf curve will also decrease with the increase of α , as indicated in the

1- If α increased, while ß and μ are kept the same, the distribution gets stretched out to the right and its

height decreases, while maintaining its shape and location.

2- If α is decreased, while ß and μ are kept the same, the distribution gets pushed in towards the left (i.e., towards its beginning or towards 0 or μ), and its height increases.

3- α has the same unit as X , such as hours , mile, cycles, actuations, etc.

8.5 Weibull plot

The fit of data to a Weibull distribution can be visually assessed using a Weibull plot [7]. The Weibull plot is a plot of the empirical cumulative distribution function $\hat{F}(x)$ of data on special axes in a type of Q Q plot the axes are $\ln(-\ln(1 - F(x)))$ versus $\ln x$. the reason for this change of variables is the cumulative distribution function can be linearized:

$$F(x) = 1 - e^{-\left(\frac{x}{\alpha}\right)^{\beta}}$$
$$-\ln(1 - F(x)) = \left(\frac{x}{\alpha}\right)^{\beta}$$
$$\ln(-\ln(1 - F(x))) = \beta \ln x - \beta \ln \alpha \text{ there}$$
$$y = \beta \ln x - \beta \ln \alpha \qquad (3)$$

Which can be seen to be in the standard from of a straight line. Therefore, if the data came from a Weibull distribution then straight line is expected on a Weibull plot. There are various approaches to obtaining the empirical distribution function from data: one method is to obtain the vertical coordinate for each point using $\hat{F} = \frac{i-0.3}{n+0.4}$ where i is the rank of the data point and n is the number of data point Wayne Nelson [8]

Linear regression can also be used to numerically asses' goodness of fit and estimate the parameters of the Weibull distribution. The gradient informs one directly about the shape parameter β and the scale parameter α can also be inferred.

8.6 Weibull Reliability

The pdf can be used derive commonly used reliability metrics such as the reliability function, failure rate, mean, and median. The equations for these functions of the Weibull distribution will be presented in the following section, without derivations for the sake of brevity and

simplicity. Note that in the rest of this section we will assume the most general form of the Weibull distribution, the three- parameter form. The appropriate substitutions to obtain the other form, such as the two-parameter form where μ=0, or the one-parameter form where ß is a constant, can easily be made.

The Weibull reliability function is given by:

$$R(X) = e^{-(\frac{X-\mu}{\alpha})^{\beta}} \qquad (4)$$

The Weibull failure rate function is given by:

$$h(X) = \frac{f(X)}{R(X)} = \frac{\beta}{\alpha}(\frac{X-\mu}{\alpha})^{\beta-1} \quad (5)$$

The Weibull mean life is given by:

$$\bar{X} = \mu + \lambda.\Gamma\left(\frac{1}{\beta}+1\right) \qquad (6)$$

Where $\Gamma\,(*)$ is gamma function The gamma function is defined as:

$$\Gamma\,(n) = \int_0^\infty e^{-x} x^{n-1} dx \qquad (7)$$

The equation for the median life for Weibull distribution is given by:

$$\check{X} = \mu + \alpha\,(\ln 2)^{\frac{1}{\beta}} \qquad (8)$$

9. Estimation of Weibull Distribution parameter's
9.1 Graphical procedure

$$F(x_i) = \begin{cases} 1 - exp\left\{-\left(\frac{x_i}{\alpha}\right)^{\beta}\right\} & x \geq 0; \ \alpha \, \beta > 0 \\ 0 & other\ wise \end{cases}$$

$$F(x_i) = 1 - exp\left\{-\left(\frac{x_i}{\alpha}\right)^{\beta}\right\}$$

$$1 - F(x_i) = exp\left\{-\left(\frac{x_i}{\alpha}\right)^{\beta}\right\}$$

$$\ln\left(\frac{1}{1-F(x_i)}\right) = \left(\frac{x_i}{\alpha}\right)^{\beta}$$

$$\ln\left[\ln\left(\frac{1}{1-F(x_i)}\right)\right] \beta \ln x_i - \beta \ln \alpha$$

x_i actually represented the order statistics $x_{(1)} < x_{(2)} < \cdots.< x_{(n)}$

If we let $y = \ln\left[\ln\frac{1}{1-F(x_i)}\right]$ ⬜ $X = \ln x_i$ ⬜$a = -\beta \ln \alpha$ ⬜ then represent a simple linear regression corresponding to:

$$y = \beta X + a \qquad (9)$$

The unbiased of estimate of α the scale parameter is calculate:

$$\hat{\alpha} = exp\left[-\frac{a}{\beta}\right] \qquad (10)$$

Where a is intercept of the linear equation we perform the estimation of α and β using the following methods of estimation .

9.2 Maximum Likelihood Estimation(MLE)

The method of maximum likelihood estimation used for estimating parameters Cohen and Harter and Moore (1965).

Let $x_1.x_2.....x_n$ be a random sample of size n drawn from a population with probability density function is defined by

$$l = f(\propto.\beta) = \prod_{i=1}^{n} f(x_t.\lambda) \tag{11}$$

The maximum likelihood of $\underline{\lambda} = (\beta.\propto).$ maximize L or equivalently , the logarithm of L when

$$\frac{\partial \ln L}{\partial \ln \underline{\lambda}} = 0 \tag{12}$$

for example, Mood et al (1974). Consider the Weibull pdf given in equation its likelihood function is given as:

$$L(x_1.x_2.....x_n; \beta.\alpha) = \prod_{i=1}^{n} \left(\frac{\beta}{\alpha}\right)\left(\frac{x_t}{\alpha}\right)^{\beta-1} exp\left[-\left(\frac{x_t}{\alpha}\right)^{\beta}\right] =$$
$$\left(\frac{\beta}{\alpha}\right)\left(\frac{x_t}{\alpha}\right)^{n\beta-n} \sum_{i=1}^{n} x_t^{(\beta-1)} exp\left[-\sum_{t=1}^{n}\left(\frac{x_t}{\alpha}\right)^{\beta}\right] \tag{13}$$

Taking a natural logarithm of both sides yields:

$$\ln L = n\ln\left(\frac{\beta}{\alpha}\right) + (\beta-1)\sum_{t=1}^{n} x_t - \ln(\alpha)^{\beta-1} - \sum_{t=1}^{n}\left(\frac{x_t}{\alpha}\right)^{\beta} \tag{14}$$

And differentiating equation (14) partially w.r.t β and $\Box$ α in turn and equating to zero obtain the estimating equation as follows :

$$\frac{\partial L}{\partial \beta} = \frac{n}{\beta} + \sum_{t=1}^{n} \ln x_t - \frac{1}{\alpha}\sum_{t=1}^{n} x_t^{\beta} \ln x_t = 0 \tag{15}$$

$$\text{And } \frac{\partial \ln L}{\partial \propto} = \frac{-n}{\propto} + \frac{1}{\propto^2}\sum_{t=1}^{n} x_t^{\beta} = 0 \tag{16}$$

From (16) we obtain an estimator of α as

$$\hat{\propto} = \frac{1}{n}\sum_{i=1}^{n} x_t^{\hat{\beta}} \tag{17}$$

And on substitution of (17) in (15) obtain

$$\frac{1}{\beta} + \frac{1}{n}\sum_{t=1}^{n} \ln x_t - \frac{\sum_{t=1}^{n} x_t^{\beta} \ln x}{\ln x_t^{\beta}} = 0 \tag{18}$$

Which may be solved obtain the estimate of β using Newton-Raphson method or any other numerical procedure because (18) does not have a closed from solution. When $\hat{\beta}$ is obtained, the value of $\hat{\propto}$ follows from (17)

9.3 Least squares method (LSM)

The least squares method assume that there is a linear relationship between two variables assume a data set that constitute a pair $(x_t \square y_t) = (x_1 \square y_1) \square (x_2 \square y_2) \ldots \square (x_n \square y_n)$ were obtained and polled the least square principle minimizes the vertical distance between the data points and straight line fitted to the data $\square$ the best fitting line to this data is the straight line : $y_t = \alpha + \beta x_t$ such that

$$Q(X \square \alpha \square \beta) = \sum_{i=1}^{n}(y_t - \propto -\beta x_t)^2 \tag{19}$$

To obtain the estimators of $\alpha \square$ and β we differentiate Q w.r.t $\alpha \square \beta$. Equating to zero
Subsequently yields the following system of equation:

$$\frac{\partial Q}{\partial \alpha} = -2\sum_{i=1}^{n}(y_t - \propto -\beta x_t)^2 = 0 \tag{20}$$

$$\frac{\partial Q}{\partial \beta} = -2\sum_{i=1}^{n}(y_t - \propto -\beta x_t)^2 x_t = 0 \tag{21}$$

Expanding and solving equation (20), (21) simultaneously, we have

$$\hat{\beta} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \tag{22}$$

And $\hat{\propto} = \bar{y} - \hat{\beta}\bar{x}$

$$\hat{\propto} = \exp\left(-\frac{a}{\hat{\beta}}\right) \tag{23}$$

Where $\hat{\propto}$ and $\hat{\beta}$ are the unbiased estimators of the α and β respectively.

9.4 Weibull Regression

Another approach to finding parameters for a Weibull distribution is based on linear regression first note that the cumulative distribution function of a Weibull distribution can be expressed as

$$F(x) = 1 - e^{-\left(\frac{x}{\propto}\right)^\beta}$$

And so $\quad\quad 1 - F(x) = e^{-\left(\frac{x}{\propto}\right)^\beta}$

Taking the natural log of both sides of the equation yields the equation

$$\ln(1 - F(x)) = -\left(\frac{x}{\propto}\right)^\beta$$

Multiplying both sides of equation by -1 and taking the log again yields' the equation

$$\ln(-\ln(1 - F(x)) = \beta \ln x - \beta \ln \alpha$$

This can be expressed as the linear equation

$$y = \beta \acute{x} + a$$

Where $y = \ln(-\ln(1 - F(x)))$. $\acute{x} = \ln x$. and $a = -\beta \ln \propto$

Thus if the sample has a Weibull distribution then we should be able to find the coefficient linear regression [9]

10- Goodness-of fit test

A goodness-of fit test is appropriate when one wishes to decide if an observed distribution of frequencies is incompatible with some preconceived or hypothesized distribution [10] by other the test procedures are intended to defect the existence of a significant difference between the observed (empirical) frequency of occurrence of an item and the theoretical(hypothesized) patter of occurrence of data here we assume that the Weibull distribution is a good fit to the given data set ; otherwise, this assumption is nullified.

A goodness- of fit test is

11. Application

In this section we apply to a data which is the time of breast cancer patients in moths we focus on analysis using Weibull distribution of two parameters model. The sample data obtained from ministry of health in Khartoum state in year 2016- 2017 and sample data in table (1) bellow.

Table 1. sample patient times pear month of breast cancer

2	2	10	3	2	9	6	10	6	1
3	13	8	4	4	4	4	8	5	10
4	5	4	6	1	4	4	6	6	8
6	2	4	6	4	6	2	4	5	6
6	4	4	6	10	8	5	4	6	4
5	1	6	4	8	2	13	6	4	9
6	10	8	4	6	4	5	8	4	4
4	8	2	5	4	6	2	5	2	4
4	6	4	5	4	6	4	2	5	6
2	4	6	13	7	5	1	3	13	8
5	4	6	5	6	6	10	3	5	5
13	9	5	2	8	4	8	6	2	2
5	6	6	4	5	5	6	6	4	4
2	8	4	1	2	13	4	5	1	6

Sources data: military medical hospital

We calculate from data the estimated value of $F(x)$, using the formula

$F(x) = \frac{i - 0.3}{n + 0.4}$ so as to estimate ß the shape parameter and α the scale

parameter, whereas Xi represent times patient of breast cancer and Oi represent observed frequency looking in table (2) bellow

**Table 2. Case processing**

Xi	Oi	LnXi	F(x)	y
1	6	0	3.801582	-23.8342
2	17	11.7835	11.03802	-38.5525
3	4	4.394448	1.975392	-6.48534
4	36	49.90658	23.93856	-33.873
5	19	30.57932	12.72983	-6.02668
6	30	53.75277	20.22288	5.427269
7	1	1.94591	0.677563	0.484315
8	12	24.9533	8.166744	8.235932
9	3	6.591675	2.049615	2.535178
10	6	13.81551	4.113402	5.804146
≥11	6	15.38969	4.148622	9.317563
SUM	140			

Using the equation (9)& (10) we found that $\beta = 0 \cdot 093896$ & $\alpha = 0 \cdot 061584$ we notice that value of β is less than one indicates that the Weibull failure rate decreases over lime which given by equation(5) the formula of Weibull with two parameter which often use in failure analysis, because no failure can happen before time zero. Therefore, μ=0, when we move from three- parameter to the two-parameter we have to do is replace each instance of $x - \mu$ with x , the formula for the pdf reduces to:

$$h(X) = \frac{f(X)}{R(X)} = 15 \cdot 25 \left(\frac{x}{0 \cdot 061584}\right)^{-0 \cdot 0611}$$

looking that in figure (1) below
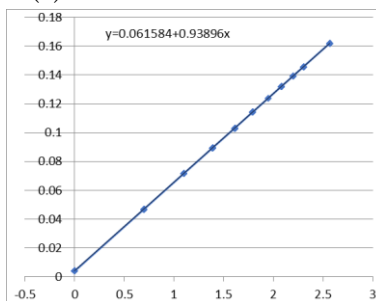
Figures 1: Weibull plot

Based on these results we find that the parameters of Weibull distribution estimated by the linear regression. Using the equations

(9), (22), (23) and the calculations in above table and the linear regression model represented with equation:

$$y = 0 \cdot 061584 + 0 \cdot 936896x$$

to check the hypotheses uses a goodness- of fit test to find the statistics using table (3) bellow

Table 3. test of goodness fit

i	Oi	Ei	(Oi-Ei)2 / Ei
1	6	19.03046	8.922161
2	17	26.93018	3.661637
3	4	4.217913	0.011258
4	36	28.42864	2.016472
5	19	11.98655	4.103647
6	30	15.75144	12.88908
7	1	0.449501	0.67419
8	12	4.714465	11.25876
9	3	1.046555	3.646198
10	6	1.881927	9.011256
>=11	6	25.56238	14.9707
SUM			71.16535

$$\chi^2 = 71 \cdot 16535 > \chi^2(0 \cdot 05 \; \square 10) = 18 \cdot 31$$

For this test the computed value is greater than the defined critical value then acceptance the null hypothesis. Looking at the same information on a Weibull probability plot, one can easily understand why the Weibull shape parameter is also known as the slope. The following plot shows how the slope of the Weibull probability plot changes with ß. Note that the models represented by lines all have the same α. The following figure (2-6) shows that the effect of the different values of the shape parameter, ß on the shape of the of the pdf (while keepingα constant).one can see that the shape of the pdf can take on a variety of forms based on the value of ß
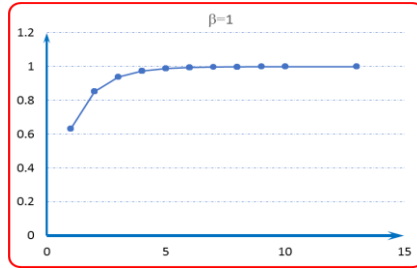
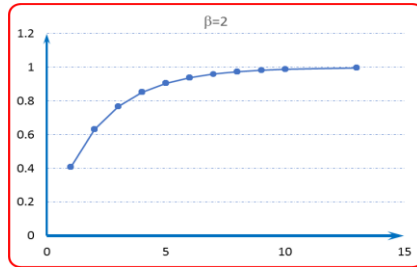**Figure 2: effect of Weibull shape parameter on probability**

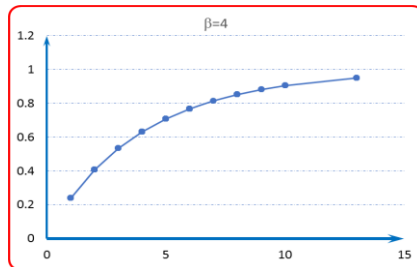**Figure 3: effect of Weibull shape parameter on probability**

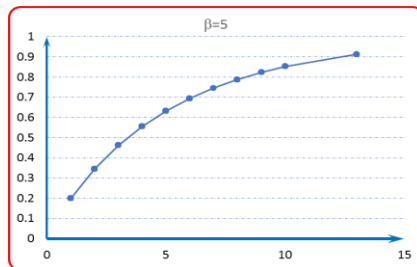**Figure 4: effect of Weibull shape parameter on probability**

Figure 5: effect of Weibull shape parameter on probability
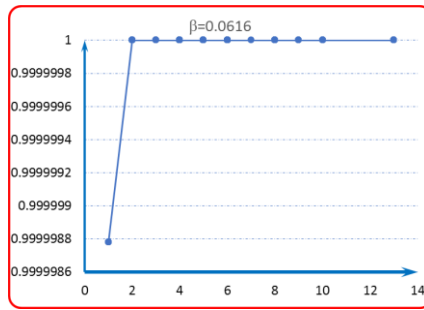
Figure 6: effect of Weibull shape parameter on probability

12. CONCLUSION

In this paper find that results as follows:

1- The data of the breast cancer distributed with linear model
2- The linear regression can also be used to numerically assess goodness of fit to estimate the parameters of Weibull distribution
3- The failure rate decreases over time.

13. RECOMMENDATION

Based on the results the paper recommends the following:

1- Weibull recommended because it has flexibility to process difference data
2- using linear model if other methods cannot be a good fit to estimate model parameters.

REFERENCES

[1] Papoulis, Athanasios Papoulis; Pillai, S. Unnikrishna (2002). Probability, Random Variables, and Stochastic Processes (4th ed). Boston: McGraw-Hill. ISBN 0-07-366011-6

[2] Sharif, M.Nawaz; (1980), "The Weibull distribution as a general model for forecasting technological change".

[3] Rinne.H, (2008), "the Weibull distribution"

[4] Al-Fawzan. M (2000). "Methods for Estimating the Parameter of the Weibull Distribution" King Abduaaziz City for Science Technology. Saudi Arabia

[5] Nwbi.F.N, Ugomma.Ch.A, (2104),"Comparison of Methods for the Estimation of Weibull Distribution Parameters,

[6] Jiang, R.; Murthy, D.N.P (2011)." A study of Weibull shape parameter: Properties and significance". Reliability Engineering and system Safety. 96(12): 1619-26.

[7] Weibull Plot " (http://www.itl.nist.gov/div898/handbook/eda/section3/welibplot.htm)www.itl.nist.gov.

[8] Wayne Nelson (2004) "Applied Life Data Analysis. Wiley-Blackwell ISBN0-471-64462-5

[9] Law less J.F (2003),"Statistical Models and Methods for Life Time Data" Edition, John Wiley and Sons, New York

[10] Daniel, Wayne (1999) " Biostatistics A foundation for Analysis in the Health Sciences" 7 th ed . John Wiley & Sons, Ney York, USA