VIDYASAGAR UNIVERSITY

PROJECT SUBMITTED FOR PARTIAL FULFILMENT OF BACHELOR'S DEGREE IN STATISTICS HONOURS

A STUDY OF WAITING TIME TO JUDGEMENT OF COURT CASES A SURVIVAL ANALYSIS APPROACH

REGISTRATION NO: 1160464

SESSION: 2020-2021

DEPARTMENT OF STATISTICS
HALDIA GOVERNMENT COLLEGE

ACKNOWLEDGEMENT :

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along with the completion of my project. All that I have done is only due to their supervision and assistance and I would not forget to thank them. I respect and thank, for providing me an opportunity to do the project work under the department of Statistics, Haldia Government College and giving you all the support and guidance that I required, which made me complete the project duly on time. I owe my deep gratitude to our project guides Dr. Shyamsundar Sahoo, Mr. Sibsankar Karan, Mr. Tanmay Kumar Maity and Mr. Bijitesh Halder, who took keen interest on my project work and guided me all along, till the completion of our project work by providing all the necessary information for developing a good system. I would not forget to remember my parents, and also my friends for their encouragement and more over for their timely support and guidance till the completion of our project work.

CONTENTS

A study of waiting time to judgement of court cases filed in Indian High Courts ;
A survival analysis approach

- **INTRODUCTION:**

The path to justice in the context of the Indian legal system is oftentimes a lengthy one, often stretching for a decade or two. The time to disposal of a case depends on many factors , such as whether it is a criminal or civil case, in which court the case is being filed, the judge or the bench, availability of evidence to continue hearing and so on. All these factors contribute to the waiting time to judgement of a particular case. If one looks at the status of pending cases in but High Courts , there will be some homogeneity regarding the material conditions as compared to that of the lower courts or when compared to the Supreme Court. To avoid such parity , I will focus my discourse only in the domain of cases filed in Indian High Courts. It is impossible to claim that there is no variation in material conditions across High Courts, but they are positionally equivocal. When a case is being filed , the waiting time to judgement is uncertain, even though lawyers do get a rough idea regarding the running time of the case by virtue of experience , the estimates are not scientifically valid. This is the problem I have chosen for this study; to estimate waiting time to judgement of cases filed in Indian High Court.

In my study , the time to judgement is the study variable. I treat the time to judgement as a continuous random variable , and correspondingly seek how the waiting time to judgement is distributed over different intervals of time. The challenge in doing so is the unavailability of plentiful information and the element of censoring present in the data I am working with. For instance, the variation between civil and criminal cases could not be compared in this study due to lack of sufficient information on the number of cases withdrawn in a given period of time. The ultimate intention of this study is to estimate an interval of time, in which the waiting time to judgement will lie with a prefixed 95% probability. However, it should not be expected that the estimated confidence interval applies uniformly to civil and criminal cases over all Indian courts alike. The unavailability of sufficient information is a pivotal limitation to the findings of this paper.

The mathematics involved in doing so will be discussed in the later sections of this dissertation.

- **OBJECTIVE:**

  The path to justice in context of the Indian legal system is usually a lengthy one, oftentimes a decade or more. The objective of this study is to estimate the survival function for the data on waiting time to disposal of court cases filed in Indian High Courts.

  Inferences about the results obtained by the two approaches mentioned will be drawn, leading to a comparison of the two approaches and their validity for the given problem.

  Furthermore, the mean and median waiting time to disposal of a case will be estimated , rendering some insight on the efficiency and shortcomings on the working of the legal institutions of the country .

- **SOURCE OF DATA:** The data is sourced from National Judiciary Data Grid's website.
  The data used is given below:

Particulars	Civil Cases Pending	Civil Cases Disposed	Criminal Cases Pending	Criminal Cases Disposed	Total Pending	Total Disposed	Total Cases Withdrawn
0 to 1 Years	869119	158549	418943	172433	1288062	355040	5897
1 to 3 Years	689971	124573	207748	123096	897719	257261	4634
3 to 5 Years	799885	44620	288642	18385	1088527	63175	1660
5 to 10 Years	1022405	70539	367314	19898	1389719	90658	2624
10 to 20 Years	781107	30279	341417	6056	1122524	40636	1126
Above 20 Years	211116	6534	80304	1565	291420	8103	243
Total					6077971	814873	16184

- **METHODOLOGY:**

  The data at hand is used to estimate the non-parametric survival function using a naïve estimator devised from the idea of the Kaplan-Meier estimator. From there, I estimate the cumulative hazard function non-parametrically. Using large sample method.
  The cumulative hazard function gives us an idea about the underlying probability distribution of the rv. From thereon, I fit an appropriate distribution to the given data and thus estimate the survival function parametrically. Here , I obtain a 95% confidence interval.
  Now, I will be heading towards a comparison of the results obtained from these two different approaches . The two estimated survival functions will be plotted alongside to show the parity between them.

THEORY:

The data at hand is an interval censored data. The interval censoring scheme can be described as follows. Suppose n identical items are put on a life test and let $T_1, \ldots, T_n$ be the lifetime of these items. For the i-th item, there is a random censoring interval (L_i , R_i), which follows some unknown bivariate distribution. Here Li and Ri denote the left and right random end point, respectively, of the censoring interval. The life time of the i-th item, T_i , is observable only if $T_i \notin [L_i , R_i]$, otherwise it is not observable. Define $\delta_i = I(T_i \notin [L_i, R_i])$, then $\delta_i = 1$ implies the observation is not censored. In that case the actual value of T_i is observed. When $\delta_i = 0$, only, the censoring interval $[L_i, R_i]$ is observed. For all the n items, the observe data is of the form (y_i , δ_i), i = 1, . . . , n, where

$$(y_i , \delta_i) = \begin{cases} (T_i, 1) & if\ T_i \notin [L_i, R_i] \\ ([L_i, R_i], 0) & otherwise \end{cases}$$

- The survival function S(.) at a given time point is defined as the probability of a case not facing the vital event (disposal of the case in this case) before the given time point,
  i.e. ***S(t) = Pr {τ > t }***

- **Non-parametric approach:**
- To construct the estimator, let 0 = τ0 < τ1 < τ2 < · · · < τm be a grid of time which includes all the points Li and Ui for i = 1, . . . , n. For the ith observation, define a weight αij to be 1 if the interval (τj−1, τj) is contained in the interval (Li , Ui] and 0, otherwise. The weight αij indicates whether the event which occurs in the interval (Li , Ui] could have occurred at τj . An initial guess at S(τj) is made and the Turnbull´s algorithm is as follows:.
- • Step 1: Compute the probability of an event occurring at time τj by pj = S(τj−1) − S(τj) j = 1, . . . , m;
- Step 2: Estimate the number of events which occurred at τj by $d_j = \sum_{i=1}^{n} \frac{\alpha_{ij} p_j}{\sum_{k=1}^{m} \alpha_{ik} p_k}$
- Step 3: Compute the estimated number at risk at time τj by Yj = $\sum_{i=1}^{j} d_j$
- • Step 4: Compute the updated Product-Limit estimator using the pseudo data found in Steps 2 and 3. If the updated estimate of S is close to the old version of S for all τj´s, stop the iterative process, otherwise repeat Steps 1-3, using the updated estimate of S.

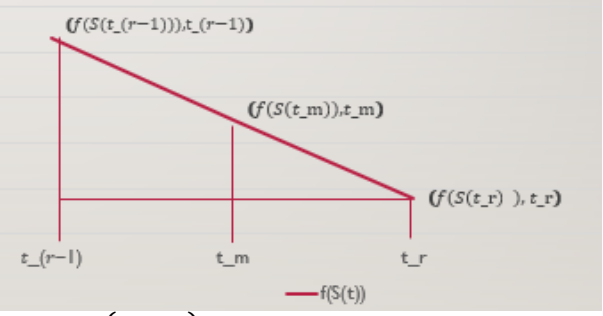
Let, $t_1 < t_2 < \cdots < t_k$ be k fixed time points not necessarily at intervals of one year. So, the Kaplan-Meier estimate of the survival function could be written as:

$$\widehat{S(t_r)} = q(t_r)q(t_{r-1}) \ldots q(t_1)q(0)$$

From these estimates, a suitable transformation is taken of the estimated survival function values and the values of S(.) at intermediate time points are obtained by linear interpolation.
i.e. if we have $\widehat{S(t_r)}$ and $\widehat{S(t_{r-1})}$ and we intend to find $\widehat{S(t_m)}$, where $t_{r-1} < t_m < t_r$.

First, some suitable transformation of $\widehat{S(t)}$ is taken, say $f(\widehat{S(t)})$, such that f is a

linear function of t and it is invertible.



Secondly, for linear interpolation, we have the following setup, as shown in the figure.
Thus, it could be written that;

$$\frac{f(S(t_m)) - f(S(t_{r-1}))}{(t_m - t_{r-1})} = \frac{f(S(t_r)) - f(S(t_m))}{(t_r - t_m)}$$

On simplification , this equation yields

$$f(S(t_m)) = \frac{(t_r - t_m)f(S(t_{r-1})) + (t_m - t_{r-1})f(S(t_r))}{(t_r - t_{r-1})}$$

Thirdly, we have the estimated values of $f(S(t))$. By taking the inverse of f, we obtain the estimated $S(t)$ at the intermediate timepoints.
i.e. $\widehat{S(t_m)} = f^{-1}f(S(t_m))$

Then, to estimate the cumulative hazard function, the estimator due to Nelson and Alen is being used.
Let, $I_j = [t_j, t_{j+1})$ be disjoint time intervals, with

$$n_j = number\ of\ cases\ at\ risk\ of\ vital\ event\ in\ I_j$$
$$d_j = number\ of\ cases\ disposed\ in\ the\ time\ interval\ I_j$$

The Nelson-Alen estimator for the cumulative hazard function is given as;

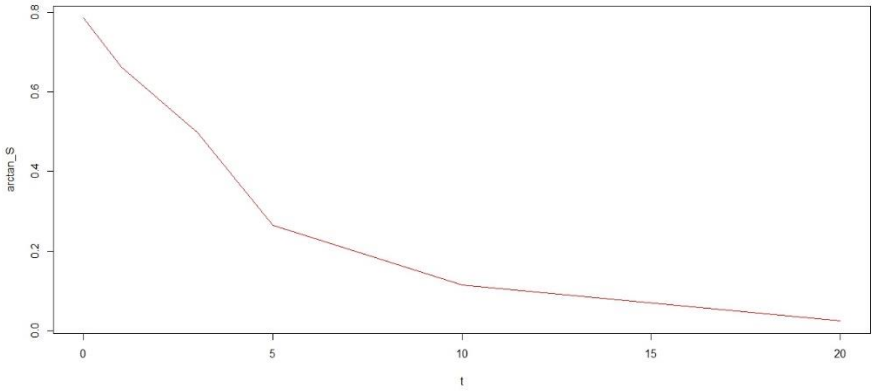
$$\widehat{H(t)} = \sum_{t_i \le t} \frac{d_j}{n_j}$$

- **CALCULATIONS:** The following table shows the estimated values of S(t) for the given time points in the data;

T	S(t)
0	1
1	0.781117514
3	0.544402392
5	0.271114399
10	0.115069453
20	0.023716314

- On transforming the S(t) using f(.)= arctan(.), we obtain the values of the transformed survival function as follows;

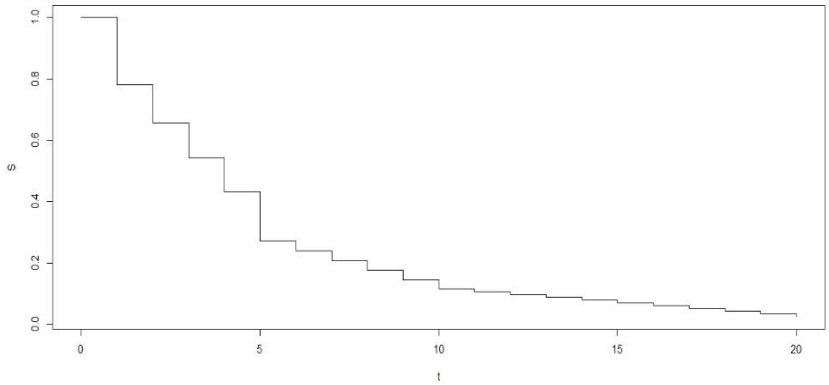
T	S(t)	arctan S(t)
0	1	0.785398163
1	0.781117514	0.663120716
3	0.544402392	0.498535468
5	0.271114399	0.264750222
10	0.115069453	0.114565573
20	0.023716314	0.023711869



The values of the survival function obtained by the method of linear interpolation are given below:

T	arctan S(t)	S(t)=tan{arctan S(t)}
0	0.785398163	1
1	0.663120716	0.781117514
2	0.580828092	0.656352638
3	0.498535468	0.544402392
4	0.407758396	0.431968741
5	0.264750222	0.271114399
6	0.234713292	0.239120582
7	0.204676363	0.207583211
8	0.174639433	0.176436802
9	0.144602503	0.145618879
10	0.114565573	0.115069453
11	0.105480203	0.105873145
12	0.096394832	0.096694512
13	0.087309462	0.087531992
14	0.078224091	0.078384034
15	0.069138721	0.069249097
16	0.060053351	0.060125647
17	0.05096798	0.05101216
18	0.04188261	0.041907117
19	0.03279724	0.032809004
20	0.023711869	0.023716314

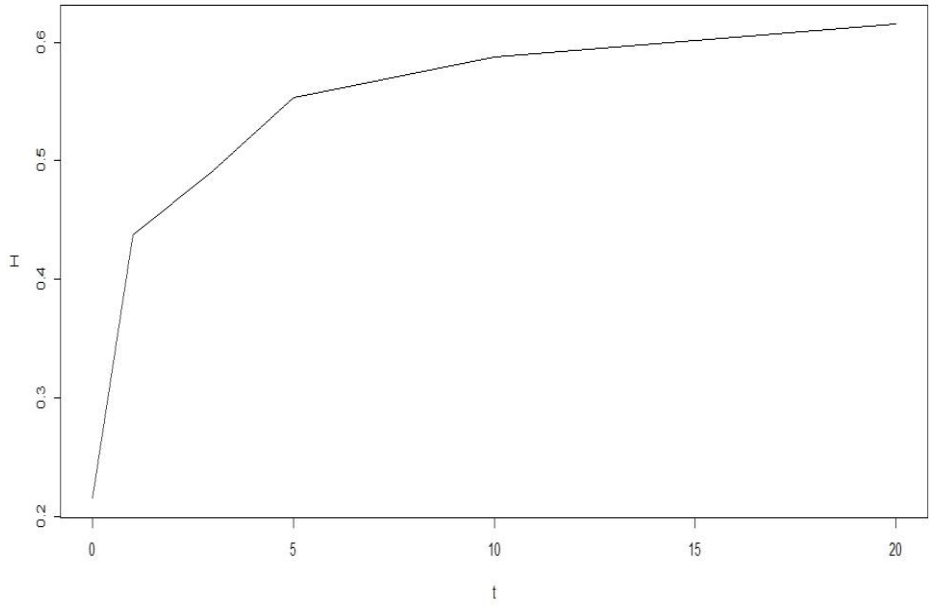
The plot of the survival function now looks like the given graph.



On interpolating, we find that the median of the distribution is found to be 3.46 years, or approximately 3 years and 5 months.
The calculated values of $\widehat{H(t)}$ are tabulated below.

T	H(t)
0	0.215306
1	0.437157
3	0.491932
5	0.553063
10	0.587965
20	0.614996

The plot of the estimated cumulative hazard rate is shown below.



Parametric Approach :

In interval-censored data in which the responses are independent pairs (u_i,v_i), i = 1(1) n, it being known that $u_i < T_i \leq v_i$. An exactly known lifetime is given by taking v_i-. When the individual lifetimes T_i are identically distributed with c.d.f. F(t), the likelihood function becomes

$$L = \prod_{i=1}^{n} [F(v_i) - F(u_i)].^{\delta_i} . (1 - F(v_i))^{1-\delta_i}$$

Where δ_i=number of events in the interval (u_i,v_i).

The interval censoring scheme can be described as follows. Suppose n identical items are put on a life test and let $T_1, \ldots, T_n$ be the lifetime of these items. For the i-th item, there is a random censoring interval (L_i , R_i), which follows some unknown bivariate distribution. Here Li and Ri denote the left and right random end point, respectively, of the censoring interval. The life time of the i-th item, T_i , is observable only if $T_i \notin [L_i , R_i]$, otherwise it is not observable. Define δ_i = I($T_i \notin [L_i, R_i]$), then δ_i = 1 implies the observation is not censored. In that case the actual value of T_i is observed. When δ_i= 0, only, the censoring interval $[L_i, R_i]$ is observed. For all the n items, the observe data is of the form (y_i , δ_i), i = 1, . . . , n, where

$$(y_i , \delta_i) = \begin{cases} (T_i, 1) & if\ T_i \notin [L_i, R_i] \\ ([L_i, R_i], 0) & otherwise \end{cases}$$

In the parametric setup, we make a guess about the underline distribution of the time to judgment random variable from the cumulative hazard rates obtained from the non parametric computation. The cumulative hazard function obtained from the non parametric procedure, looks similar to that of the Weibull and Log Logistic distributions.

Case 1: The Weibull Distribution

In this section, it is assumed that $T_i, \ldots, T_n$ are independent identically distributed (i.i.d.) Weibull random variables with the probability density function

$$f(t; \beta, \lambda) = \begin{cases} \beta\lambda^{\beta-1} e^{-\lambda t^{\beta}} & if \quad t > 0 \\ 0 & if \quad t \leq 0 \end{cases}$$

here $\beta > 0$, $\lambda > 0$ are the shape and scale parameters respectively. From now on, the Weibull distribution with the PDF defined as WE(β, λ). Also it is assumed that the random censoring times L_i and R_i are independent of T_i , and it does not have any information regarding the population parameters β and λ.First we consider the maximum likelihood estimation (MLE) of $\beta$and λ. It is observed that the maximum likelihood estimators (MLEs) do not exist in closed form, and they have to be obtained by solving two non-linear equations. Moreover, the standard Newton-Raphson algorithm may not even converge sometimes.

In this section we provide the MLEs of β and λ. It is assumed that the observed data is as follows. $(T_i, 1), \ldots (T_{n_1} , 1), ([L_{n_1+1}, R_{n_1+1}], 0), \ldots ,([L_{n_1+n_2} , R_{n_1+n_2}], 0)$.

Here n_1 and n_2, denote the number of uncensored and censored observations, respectively, and $n_1 + n_2 =$ n. Based on the assumptions as described in the previous section, the likelihood function can then be written as

$$L(\beta,\lambda|data) = c\beta^{n_1}\lambda^{n_1}\prod_{i=1}^{n_1} t_i{}^{\beta-1}e^{-\lambda\sum_{i=1}^{n_1} t_i{}^{\beta}}\prod_{i=n_1+1}^{n_1+n_2}(e^{-\lambda l_i{}^{\beta}} - e^{-\lambda r_i{}^{\beta}}) .$$

Here c is the normalizing constant independent of β and λ. The log-likelihood function becomes

$$l(\beta,\lambda|data) = \mathrm{l}(\beta,\lambda) = \ln c + n_1 \ln\beta + n_1 \ln\lambda + (\beta - 1)\sum_{i=1}^{n_1}\ln t_i - \lambda\sum_{i=1}^{n_1} t_i{}^{\beta} + \prod_{i=n_1+1}^{n_1+n_2}\ln\left(e^{-\lambda l_i{}^{\beta}} - e^{-\lambda r_i{}^{\beta}}\right).$$

The corresponding normal equations can be written as;

$$\frac{\partial l(\beta,\lambda)}{\partial\lambda} = \frac{n_1}{\beta} + \sum_{i=1}^{n_1} lnt_i - \lambda\sum_{i=1}^{n_1} t_i{}^{\beta} lnt_i + \lambda\sum_{i=n_1+1}^{n_1+n_2}\frac{r_i{}^{\beta}e^{-\lambda r_i{}^{\beta}}lnr_i - l_i{}^{\beta}e^{-\lambda l_i{}^{\beta}}lnl_i}{e^{-\lambda l_i{}^{\beta}} - e^{-\lambda r_i{}^{\beta}}} = 0$$

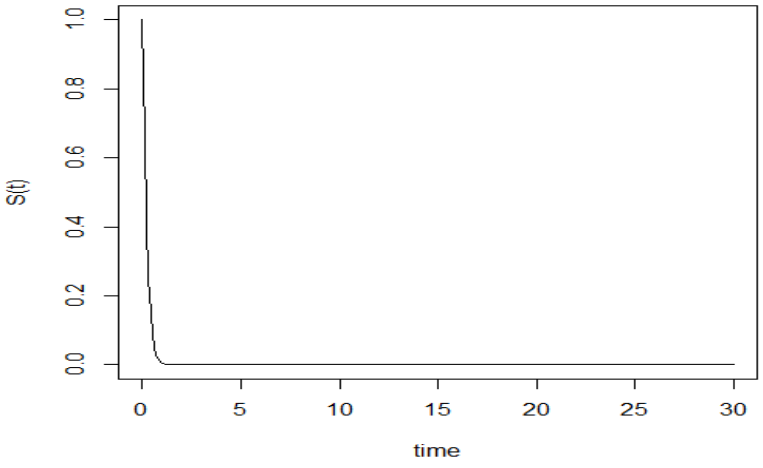
$$\frac{\partial l(\beta,\lambda)}{\partial\lambda} = \frac{n_1}{\beta} - \sum_{i=1}^{n_1} t_i{}^{\beta} + \sum_{i=n_1+1}^{n_1+n_2}\frac{r_i{}^{\beta}e^{-\lambda r_i{}^{\beta}}lnr_i - l_i{}^{\beta}e^{-\lambda l_i{}^{\beta}}}{e^{-\lambda l_i{}^{\beta}} - e^{-\lambda r_i{}^{\beta}}} = 0$$

The MLEs of β and λ are obtained by solving equations simultaneously, using numeric methods like Newton-Raphson. Here the estimates of the parameters are obtained by using the icenReg package in R. The results are as follows:

>Call: ic_par(formula = Surv(L, R, type = "interval2") ~ grp, data = mt,

model = "aft", dist = "weibull")

	Estimate	Exp(Est)	Std.Error
log_shape	0.1006	1.1060	0.0002234
log_scale	1.5070	4.5120	0.0003096
grp	-1.5330	0.2159	0.0003645



So, $\hat{\lambda} = 4.4512$, $\hat{\beta} = 1.106$

Thus the estimated survival function $\widehat{S(t)} = e^{-4.4512t^{1.106}}$ and the estimated cumulative hazard function $\widehat{H(t)} = 4.4512t^{1.106}$.

Case 2: The Log Logistic Distribution

In this section, it is assumed that $T_i, \ldots, T_n$ are independent identically distributed (i.i.d.) Log Logistic random variables with the probability density function

$$f(t) = \begin{cases} \frac{(\frac{\alpha}{\beta})(\frac{t}{\beta})^{\alpha-1}}{(1+(\frac{t}{\beta})^{\alpha})^2} & t > 0, \end{cases}$$

here $\alpha > 0$, $\beta > 0$ are the shape and scale parameters respectively. From now on, the Log Logistic distribution with the PDF defined as LL(β, α). Also it is assumed that the random censoring times L_i and R_i are independent of T_i , and it does not have any information regarding the population parameters α and β.

In this section we provide the MLEs of α and β. It is assumed that the observed data is as follows. (T_i, 1), . . . (T_{n_1} , 1), ([L_{n_1+1}, R_{n_1+1}], 0), . . . ,([$L_{n_1+n_2}$, $R_{n_1+n_2}$], 0).

Here n_1 and n_2, denote the number of uncensored and censored observations, respectively, and $n_1 + n_2$ =n. Based on the assumptions as described in the previous section, the likelihood function can then be written as

The log-likelihood function becomes

$$lnL(\alpha, \beta | data) = \sum_{i=1}^{n} \delta_i \ln \left[\frac{(\frac{u_i}{\alpha})^{\beta}}{1+(\frac{u_i}{\alpha})^{\beta}} - \frac{(\frac{v_i}{\alpha})^{\beta}}{1+(\frac{v_i}{\alpha})^{\beta}}\right] - (1-\delta_i)\ln\left[1 + (\frac{v_i}{\alpha})^{\beta}\right]$$

The corresponding normal equations can be written as;

$$\frac{\partial lnL}{\partial \alpha} = \sum_{i=1}^{n} \delta_i \left[\frac{\beta}{\alpha}\left\{\frac{A - B + C - D}{A + C}\right\}\right] \frac{(1-\delta_i)\frac{\beta}{\alpha}[C - D]}{1 + (\frac{v_i}{\alpha})^{\beta}} = 0$$

$$\frac{\partial lnL}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left[\frac{U - V - X + Y}{U + X}\right] - \frac{(1-\delta_i)(\frac{v_i}{\alpha})^{\beta} \ln(\frac{v_i}{\alpha})}{1 + (\frac{v_i}{\alpha})^{\beta}} = 0$$

Where

A=$\frac{(\frac{u_i}{\alpha})^{\beta}}{1+(\frac{u_i}{\alpha})^{\beta}}$, B=$\frac{(\frac{u_i}{\alpha})^{\beta}}{(1+(\frac{u_i}{\alpha})^{\beta})^2}$, C=$\frac{(\frac{v_i}{\alpha})^{\beta}}{1+(\frac{v_i}{\alpha})^{\beta}}$, D=$\frac{(\frac{v_i}{\alpha})^{\beta}}{(1+(\frac{v_i}{\alpha})^{\beta})^2}$, U=$\frac{\ln(\frac{u_i}{\alpha})(\frac{u_i}{\alpha})^{\beta}}{1+(\frac{u_i}{\alpha})^{\beta}}$,

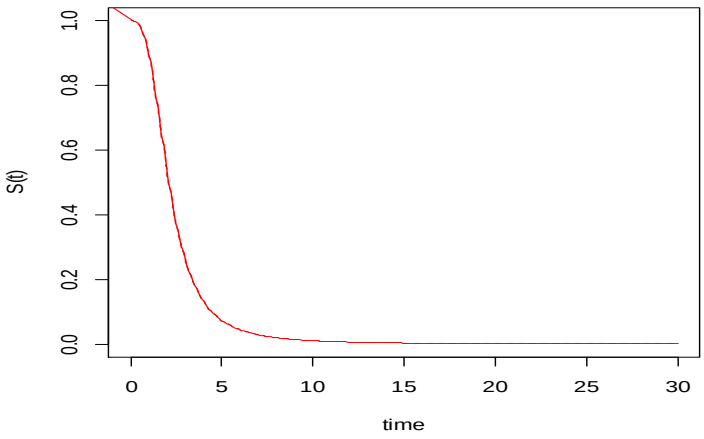
V=$\frac{\ln(\frac{u_i}{\alpha})(\frac{u_i}{\alpha})^{2\beta}}{(1+(\frac{u_i}{\alpha})^{\beta})^2}$, X=$\frac{\ln(\frac{v_i}{\alpha})(\frac{v}{\alpha})^{\beta}}{1+(\frac{v_i}{\alpha})^{\beta}}$, Y=$\frac{\ln(\frac{v_i}{\alpha})(\frac{v_i}{\alpha})^{2\beta}}{(1+(\frac{v_i}{\alpha})^{\beta})^2}$, and δ_i=number of events in the interval (u_i,v_i).

The MLEs of α and β are obtained by solving equations simultaneously, using numeric methods like Newton-Raphson. Here the estimates of the parameters are obtained by using the icenReg package in R. The results are as follows:

>Call: ic_par(formula = Surv(L, R, type = "interval2") ~ grp, data = mt,

model = "aft", dist = "loglogistic")

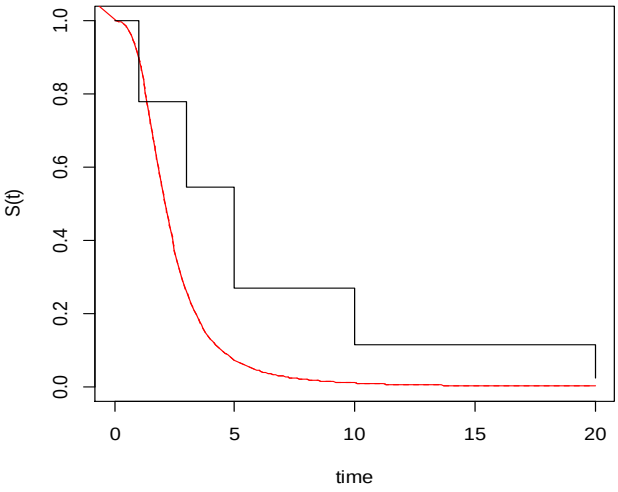
	Estimate	Exp(Est)
log_alpha	1.0770	2.9350
log_beta	0.7418	2.1000
grp	-2.1630	0.1149

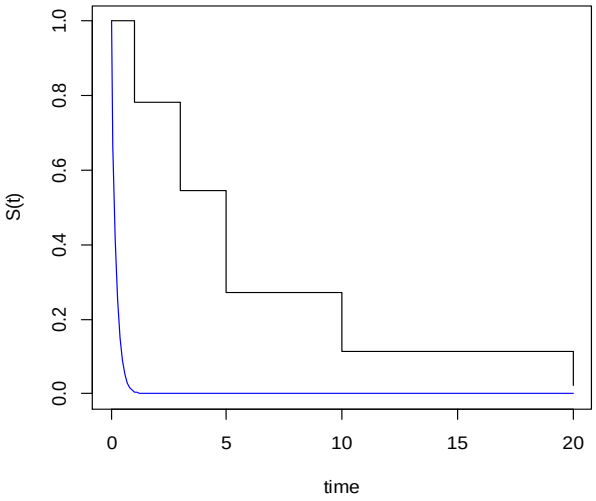


So, $\hat{\alpha} = 2.93$, $\hat{\beta} = 2.10$

- Thus the estimated survival function $\widehat{S(t)} = \frac{1}{1+(\frac{t}{2.10})^{2.93}}$ and the estimated cumulative hazard function $\widehat{H(t)} = \ln\left(1 + \left(\frac{t}{2.10}\right)^{2.93}\right)$

  The plots of the survival functions are given below.





The blue curve is the survival function of the fitted Weibull distribution and the red curve is the survival function of the fitted LogLogistic distribution.

It is evident from the two plots that the LogLogistic distribution is a better fit to the given data.

Thus, it is graphically visible that the LogLogistic distribution is a better fit to the given data.
Now, the expectation of the LogLogistic distribution is given by the integral
$\int_0^\infty t \frac{(\frac{\alpha}{\beta})(\frac{t}{\beta})^{\alpha-1}}{(1+\left(\frac{t}{\beta}\right)^{\alpha})^2} dt$ and the median is α.
Since, α = 2.93, the median time to judgement is about 3 years, and the mean time to judgement is found to be 4.39 years.

R code : The R code used for fitting the distribution is given below:

```
L<-
c(rep(0,360937),rep(1,1288062),rep(1,261895),rep(3,897719),rep(3,64835),rep(5,1088527),rep(5,93282),rep(10,1389719),rep(10,41762),rep(20,291420),rep(20,8346))

R<-
c(rep(1,360937),rep(1,128802),rep(3,261895),rep(3,897719),rep(5,64835),rep(5,1088527),rep(10,93282),rep(10,1389719),rep(20,41762),rep(20,291420),rep(Inf,8346))

grp<c(rep(1,360937),rep(0,128802),rep(1,261895),rep(0,897719),rep(1,64835),rep(0,1088527),rep(1,93282),rep(0,1389719),rep(1,41762),rep(0,291420),rep(1,8346))

mt<-data.frame(L,R,grp)

fit_weibull <- ic_par(Surv(L,R, type="interval2")~grp,data=mt, model = "aft", dist = "weibull")

summary(fit_weibull)

fit_log <- ic_par(Surv(L,R, type="interval2")~grp,data=mt, model = "aft", dist = "loglogistic")

summary(fit_log)
```

Results and Discussions:

The time to judgement of court cases concerning the data , is therefore found to follow the LogLogistic distribution.

The mean time to survival for the waiting time to judgement has been found to be 4.39 years.

The median time to judgement was found to be about 3 years in both the parametric and nonparametric setup.
The standard deviation of the LogLogistic distribution was found to be 3.77 years.

Thus, we find a 95% confidence interval for the time to judgement random variable as (2.88,15.99), using large sample normal approximation and thus the percentage points of the standard normal distribution.

Hence, it could be inferred that the time to judgement of court cases, both civil and criminal fall in the interval (2.88,15.99) 95% of the time.

References:

1. Statistical Models and Methods for Lifetime Data; J.F. Lawless
2. Mara Tableman (p.65)
3. Analysis of interval-censored data with Weibull lifetime distribution; Biswabrata Pradhan & Debasis Kundu