# Analysis of interval-censored data with Weibull lifetime distribution

**Shatrughna Chaurasia**
Under the Supervision of
**Dr. Prof. Debasis Kundu**

- **Introduction**
- **Data extraction**
- **MLE of Parameters**
- **Conclusion**
- **Reference**

# Introduction

- Lifetime data analysis is a valuable technique applied across various domains to examine data related to the time elapsed between two events. This approach goes by several names, including event history analysis, survival data analysis, reliability analysis, and time-to-event analysis. In lifetime data analysis, it's common for the data to be censored. Censoring occurs in different forms.
- Lifetime data analysis often the data are censored. The event time is right censored, when follow-up is curtailed without observing the event. Left censoring arises when the event occurs at some unknown time prior to a known specified time point. The event time is considered to be interval censored when an event occurs within some interval of time

# Continued...

- **Right Censoring**: This happens when the follow-up period terminates before the event of interest is observed. In other words, we know that the event will occur at some point in the future, but we don't observe it within the study period.
- **Left Censoring**: Left censoring occurs when the event of interest actually occurred before a specific, known time point, but the exact timing is unknown.
- **Interval Censoring**: In interval censoring, we know that the event took place within a certain time interval, but we don't have precise information about when it occurred.

# Weibull family of distribution

**Weibull family of distributions:** The Weibull distribution is also widely used in reliability as a model for time to failure. It generalizes the exponential model to include nonconstant failure rate functions. In particular, it encompasses both increasing and decreasing failure rate functions and has been used successfully to describe both initial burning failures as well as failures attributed to wearout. The two-parameter Weibull distribution is given by

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

Here, $\alpha > 0$ and $\lambda > 0$ are the shape and scale parameters, respectively. From now on, the Weibull distribution with the PDF defined as above will be denoted as WE$(\alpha, \lambda)$.

# Data extraction

- **Package Installation and Loading:** First of all installing and loading the flexsurv package, which is used for survival analysis, including interval-censored data.
- **Data Generation:** We generate random samples from a Weibull distribution with shape parameter (shape) 1.5 and scale parameter (scale) 1. These samples represent the survival times T.
- **Generating Censoring Times L, Z, and R:** We take to parameters $\theta 1$ and $\theta 2$. and generated random sample L from an exponential distribution with rate $\theta 1 = 0.5$
- Then generated Z from exponential distribution with rate $\theta 2 = 0.75$, It calculates the censoring times R as the sum of L and Z.

# Cont...

- **Censoring Data:** In this section we separate the generated data into censored and uncensored data.
- For each value of T, it checks whether it falls within the interval [L, R]. If it does, it's considered censored, and the interval [L, R] is recorded.
- If T doesn't belong to the interval, it's considered uncensored, and the exact time T is recorded.
- Now from this simulation process we get the censored and uncensored data in which some of the observations are uncensored and some of them are in interval data.

# MLE of Parameters

- The log-likelihood function is given by:

$$l(\alpha,\lambda) = \ln c + n_1 \ln \alpha + n_1 \ln \lambda + (\alpha-1)\sum_{i=1}^{n_1} \ln t_i - \lambda \sum_{i=1}^{n_1} t_i^{\alpha} + \sum_{i=n_1+1}^{n} \ln\left(e^{-\lambda l_i^{\alpha}} - e^{-\lambda r_i^{\alpha}}\right)$$

- The corresponding normal equations are:

$$\frac{\partial l(\alpha,\lambda)}{\partial \alpha} = \frac{n_1}{\alpha} + \frac{\sum_{i=1}^{n_1} \ln t_i - \lambda \sum_{i=1}^{n_1} t_i^{\alpha}}{\alpha \sum_{i=1}^{n_1} t_i^{\alpha} \ln t_i + \lambda \sum_{i=n_1+1}^{n} r_i^{\alpha} e^{-\lambda r_i^{\alpha}} \ln r_i - l_i^{\alpha} e^{-\lambda l_i^{\alpha}} \ln l_i e^{-\lambda l_i^{\alpha}} - e^{-\lambda r_i^{\alpha}}}$$

$$\frac{\partial l(\alpha,\lambda)}{\partial \lambda} = \frac{n_1}{\lambda} - \frac{\sum_{i=1}^{n_1} t_i^{\alpha} - \sum_{i=n_1+1}^{n} r_i^{\alpha} e^{-\lambda r_i^{\alpha}} - l_i^{\alpha} e^{-\lambda l_i^{\alpha}}}{\lambda\left(\sum_{i=1}^{n_1} t_i^{\alpha} - \sum_{i=n_1+1}^{n} r_i^{\alpha} e^{-\lambda r_i^{\alpha}} - l_i^{\alpha} e^{-\lambda l_i^{\alpha}}\right)}$$

# Cont...

- The MLEs for $\alpha$ and $\lambda$ require solving these simultaneous nonlinear equations. It's important to note that explicit solutions cannot be obtained from these equations.
- Now Furthermore, they cannot be simplified like Type-I, Type-II, or progressive censoring cases. Therefore, a suitable numerical technique must be applied to solve these equations. Common methods include the Newton-Raphson or Gauss-Newton methods, or their variants.
- Newton-Raphson algorithm converges between 82%-85% of the times, where as the EM algorithm converges all the times, so we will use the Expectation maximization algorithm to find th MLE of $\alpha$ and $\lambda$.

# EM algorithm for censored data

- The likelihood function, denoted as $L_c(\alpha, \lambda)$, is expressed as:

$$L_c(\alpha,\lambda) = \frac{\alpha^n}{n\lambda^n} \prod_{i=1}^{n_1} t_i^{\alpha-1} \prod_{i=n_1+1}^{n_1+n_2} z_i^{\alpha-1} \cdot e^{-\lambda\left(\sum_{i=1}^{n_1} t_i^{\alpha} + \sum_{i=n_1+1}^{n_1+n_2} z_i^{\alpha}\right)}$$

Here, for $i = n_1 + 1, \ldots, n_1 + n_2$, $z_i$ represents the expected value of $T$ given that $T$ lies within the interval $(L_i, R_i)$. Its formula is given by:

$$z_i = \int_{L_i}^{R_i} x^{\alpha} \lambda e^{-\lambda x^{\alpha}} dx / e^{-\lambda L_i^{\alpha}} - e^{-\lambda R_i^{\alpha}}$$

# Cont...

- The pseudo log-likelihood function, denoted as $l_c(\alpha, \lambda)$, can then be written as:

$$l_c(\alpha,\lambda) = n\ln(\alpha)+n\ln(\lambda)+(\alpha-1)\left(\sum_{i=1}^{n_1}\ln(t_i)+\sum_{i=n_1+1}^{n_1+n_2}\ln(z_i)\right)-\lambda\left(\sum_{i=1}^{n_1}t_i^{\alpha}+\sum_{i=n_1+1}^{n_1+n_2}z_i^{\alpha}\right)$$

In the 'M-step' of the EM algorithm, we maximize the pseudo log-likelihood function ($l_c(\alpha, \lambda)$) with respect to $\alpha$ and $\lambda$ to obtain the next iterates. If $(\alpha^{(k)}, \lambda^{(k)})$ is the estimate of $(\alpha, \lambda)$ at the $k$-th stage of the EM algorithm, then $(\alpha^{(k+1)}, \lambda^{(k+1)})$ can be obtained by maximizing:

$$l_c^*(\alpha,\lambda) = n\ln(\alpha)+n\ln(\lambda)+(\alpha-1)\left(\sum_{i=1}^{n_1}\ln(t_i)+\sum_{i=n_1+1}^{n_1+n_2}\ln(z_i^{(k)})\right)-\lambda\left(\sum_{i=1}^{n_1}t_i^{\alpha}+\sum_{i=n_1+1}^{n_1+n_2}z_i^{(k)\alpha}\right)$$

# Cont...

- here we get the pseudo log likelihood function. by maximizing we can get the estimate of $\alpha$ at k+1 th stage iteration by using stopping rule the iteration continues untill converges.

$$\lambda^{(k+1)}(\alpha) = \frac{n\sum_{i=1}^{n_1} t_i^{\alpha} + \sum_{i=n_1+1}^{n_1+n_2} z_i^{\alpha}(\alpha, \lambda^{(k)})}{\sum_{i=1}^{n_1} t_i^{\alpha} + \sum_{i=n_1+1}^{n_1+n_2} z_i^{\alpha}(\alpha, \lambda^{(k)})}$$

- Clearly, for a given $\alpha$, $\lambda^{(k+1)}(\alpha)$ is unique, and it maximizes (11). The next step is to find $\alpha^{(k+1)}$ by maximizing $l_c^*(\alpha, \lambda^{(k+1)}(\alpha))$, which is the 'pseudo-profile log-likelihood function,' with respect to $\alpha$. By utilizing a similar argument as presented.

# Cont...

- If $\alpha^{(k+1)}$ maximizes $l_c^*(\alpha, \lambda^{(k+1)}(\alpha))$, it's immediate that $(\alpha^{(k+1)}, \lambda^{(k+1)}(\alpha^{(k+1)}))$ maximizes $l_c^*(\alpha, \lambda)$ since:

$$l_c^*(\alpha, \lambda) \leq l_c^*(\alpha, \lambda^{(k+1)}(\alpha)) < l_c^*(\alpha^{(k+1)}, \lambda^{(k+1)}(\alpha^{(k+1)}))$$

- The maximization of $l_c^*(\alpha, \lambda^{(k+1)}(\alpha))$ with respect to $\alpha$ can then be performed by solving a fixed-point type equation:

$$g^{(k)}(\alpha) = \alpha$$

Where:

$$g^{(k)}(\alpha) = \frac{n[\sum_{i=1}^{n_1} t_i^{\alpha} \ln t_i + \sum_{i=n_1+1}^{n_1+n_2} z_i^{\alpha}(\alpha^{(k)}, \lambda^{(k)}) \ln z_i(\alpha^{(k)}, \lambda^{(k)})}{\sum_{i=1}^{n_1} t_i^{\alpha} + \sum_{i=n_1+1}^{n_1+n_2} z_i^{\alpha}(\alpha^{(k)}, \lambda^{(k)})} -$$

# Cont...

- The simple iterative process can then be employed to compute $(\alpha^{(k+1)}, \lambda^{(k+1)})$ from $(\alpha^{(k)}, \lambda^{(k)})$. In the $k$-th step, begin by solving above equation using an iteration method of the form $\alpha^{(i+1)} = g^{(k)}(\alpha^{(i)})$. The iteration continues until it converges. Once $\alpha^{(k+1)}$ is determined, $\lambda^{(k+1)}$ is obtained as follows.
- The simulation is carried out for sample sizes $n = 20, 30, 50,$ and $100$, and for different choices of $(\theta_1, \theta_2)$. We choose $(\theta_1, \theta_2) = (0.50, 0.75)$ (Scheme 1), $(1.25, 1.0)$ (Scheme 2), and $(1.50, 0.25)$ (Scheme 3) for the simulation study. These three schemes correspond to different proportions of censored observations.
- For each set of simulated data, we generate observations and calculate different estimates of $\alpha$ and $\lambda$. In each case, MLEs are obtained by the Expectation-Maximization (EM) algorithm. We replicate the process $500$ times.

# Final result

Table 1: Estimate of $\alpha$ and $\lambda$.

| $X$ | $\theta$ | $PC$ | $n$ | $\alpha$ | $\lambda$ |
|---|---|---|---|---|---|
| 1 | 0.50, 0.75 | 0.24 | 20 | 0.2395791 | 0.057499025 |
| 2 | 0.50, 0.75 | 0.24 | 30 | 0.1696775 | 0.035954631 |
| 3 | 0.50, 0.75 | 0.24 | 50 | 0.1238654 | 0.012773317 |
| 4 | 0.50, 0.75 | 0.24 | 100 | 0.0987910 | 0.001662262 |
| 5 | 1.25, 0.75 | 0.36 | 20 | 0.2922956 | 0.064680646 |
| 6 | 1.25, 0.75 | 0.36 | 30 | 0.2258795 | 0.03402968 1 |
| 7 | 1.25, 0.75 | 0.36 | 50 | 0.1611673 | 0.015852528 |
| 8 | 1.25, 0.75 | 0.36 | 100 | 0.1415520 | 0.004253471 |
| 9 | 1.50, 0.25 | 0.55 | 20 | 0.7493564 | 0.406387605 |
| 10 | 1.50, 0.25 | 0.55 | 30 | 0.6008459 | 0.223034431 |
| 11 | 1.50, 0.25 | 0.55 | 50 | 0.5088860 | 0.149195393 |
| 12 | 1.50, 0.25 | 0.55 | 100 | 0.4551829 | 0.10508174 |

# Conclusion

- Here in this work we have used the EM algorithm for finding the MLE of the parameter $\lambda$ and $\alpha$. with the different censoring intervals.
- we have considered three scheme as different censoring interval. So in every sceme I calculated the the parameters and repeated the algorithms 500 times and then calculated the average value of parameter and also observed the MSE. then we saw that as sample size increases the mse was decreasing .

# References


Berger, J.O., Sun, D., 1993. Bayesian analysis for the Poly-Weibull distribution. Journal of the American Statistical Association, 88, 1412 - 1418.

Chen, M.-H., Shao, Q.-M., 1999. Monte Carlo estimation of Bayesian credible and HPD intervals. Journal of Computational and Graphical Statistics, 8, 69-92.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. Biometrics, 42, 845-865.

Gelman, A., Carlin, J., Stern, H., Rubin, D., 1995. Bayesian Data Analysis, Text in Statistical Science, Chapman  Hall, London.

Gomez, G., Calle, M.L., Oller, R., 2004. Frequentist and Bayesian approaches for interval-censored data. Statistical Papers, 45, 139 - 173.