Applications of a Stationary Weibull Process method with an R-package

Debarghya Jana

MTH599A Presentation

Supervisor:

Dr. Debasis Kundu

Department Of Mathematics and Statistics

Indian Institute Of Technology, Kanpur

Date: November 11, 2023

Presentation Contents

Introduction

Methodology

Data Application

Creating R Package

Appendix

References

Introduction

- We build a discrete time and continuous state space Markov stationary process denoted as $\{X_n; n = 1, 2, \ldots\}$, where X_n follows a two-parameter Weibull distribution, and most importantly, X_n's are dependent and they hold a positive probability for X_n being equal to X_{n+1} for $n = 1, 2, \ldots$
- The motivation came from several instances where a time-series data came from a lag-1 stationary process in which $X_n = X_{n+1}$, which can not be ignored. There are several positive valued stationary processes available in the existing literature but none of them can be applied for the case $X_n = X_{n+1}$, because in all these cases $Pr(X_n = X_{n+1}) = 0$.

Introduction(Continued)

- Based on the Profile likelihood method, we find the MLEs of the unknown parameters of the model and the Parametric bootstrap method has been used to obtain the confidence interval of unknown parameters.
- Parametric bootstrap is also used to find the test statistic and corresponding p-values for the models which we build up using the bootstrap samples, for the Goodness of fit test by which we can detect which model provides a good fit to the particular dataset.
- An R-package**"stnweib"** is made where a function named **"stnwep"** is created which will take the time-series data from the user and will return the test statistic and the associated p-value for the established models after finding the MLEs and parametric bootstrap and will detect which model provides a good fit to the data best. It will also return the histograms of the generated test statistics for both of the established models which are obtained from each bootstrap sample.
- A GitHub repository https://github.com/debarghya2000/stnwep is created for this R-package.

Weibull Process

- A Weibull random variable, characterized by the shape parameter $\alpha > 0$ and the scale parameter $\lambda > 0$, possesses the following probability density function (PDF):

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

- Now,if $U_0, U_1, \ldots,$ are i.i.d. and each following a uniform distribution $U(0,1)$. Now, for given parameters $\lambda_0 > 0$, $\lambda_1 > 0$, and $\alpha > 0$, let's introduce a new sequence of random variables $\{X_n; n = 1, 2, \ldots\}$ defined as:

$$X_n = \min\left\{ \left[-\frac{1}{\lambda_0} \ln U_n \right]^{\frac{1}{\alpha}}, \left[-\frac{1}{\lambda_1} \ln U_{n-1} \right]^{\frac{1}{\alpha}} \right\}. \tag{1}$$

- This sequence of random variables, denoted as a Weibull process, is referred to as $\mathrm{WEP}\,(\alpha, \lambda_0, \lambda_1)$. [**kundu2022stationary**]

Maximum Likelihood Estimation

- The maximum likelihood estimators of the unknown parameters of a Weibull process based on a sample of size n i.e $x_1, x_1, ..., x_n$ is,

  CASE - I: $\lambda_0 = \lambda_1$

  In this case it is assumed that $\lambda_0 = \lambda_1 = \lambda$. Our problem is to estimate α and λ based on $\mathcal{D} = \{x_1, \ldots, x_n\}$.$I = \{1, \ldots, n-1\}$
  $I_1 = \{i : i \in I, x_i < x_{i+1}\}, I_2 = \{i : i \in I, x_i > x_{i+1}\}, I_0 =$
  $\{i : i \in I, x_i = x_{i+1}\}$ The number of elements in I_0, I_1 and I_2 are denoted by n_0, n_1 and n_2, respectively.

- For a given α, the MLE of λ, say $\widehat{\lambda}(\alpha)$ can be obtained as,

$$\widehat{\lambda}(\alpha) = \frac{n_1 + n_2 + 1}{g_1(\alpha \mid \mathcal{D})} \tag{2}$$

MLE(Continued)

- MLE of α, say $\widehat{\alpha}$ can be obtained by maximizing

$$h(\alpha) = (n_1 + n_2 + 1)\ln\alpha + (n_1 + n_2 + 1)\left(\ln(n_1 + n_2 + 1) - \ln g(\alpha \mid \mathcal{D})\right) + \alpha\left(\ln x_1 + \sum_{i \in I_1 \cup I_2} \ln x_{i+1}\right). \tag{3}$$

where,

$$g_1(\alpha \mid \mathcal{D}) = \sum_{i \in I_1}\left(x_i^{\alpha} + 2x_{i+1}^{\alpha}\right) + \sum_{i \in I_2}\left(2x_i^{\alpha} + x_{i+1}^{\alpha}\right) + 3\sum_{i \in I_0} x_i^{\alpha} - 2\sum_{i=2}^{n-1} x_i^{\alpha} \tag{4}$$

- Once, $\widehat{\alpha}$ is obtained, then the MLE of λ, say $\widehat{\lambda}$ can be obtained as $\widehat{\lambda}(\widehat{\alpha})$.

MLE(Continued)

- **CASE 2:** $\lambda_0 \neq \lambda_1$
  In this case when λ_0 and λ_1 are arbitrary. We use the following notations

$$\begin{aligned} I_1(\beta) &= \{i : i \in I, \beta x_i < x_{i+1}\} \\ I_2(\beta) &= \{i : i \in I, \beta x_i > x_{i+1}\} \\ I_0(\beta) &= \{i : i \in I, \beta x_i = x_{i+1}\} \end{aligned} \tag{5}$$

  and $n_0(\beta) = |I_0(\beta)|, n_1(\beta) = |I_1(\beta)|$ and $n_2(\beta) = |I_2(\beta)|$. Here, $\beta = (\lambda_0/\lambda_1)^{1/\alpha}$and define, $\gamma = \frac{\lambda_0}{\lambda_1}$.

- We use profile likelihood method to find the MLE's of α ,λ_0 ,λ_1.

- For fixed γ and α (β is also fixed in that case), first, we maximize (22) with respect to λ_1, say $\widehat{\lambda}_1(\gamma, \alpha)$, and it can be obtained in explicit form as,

$$\widehat{\lambda}_1(\gamma, \alpha) = \frac{n_1(\beta) + n_2(\beta) + 1}{g_2(\alpha, \gamma \mid \mathcal{D})} \tag{6}$$

where,

$$g_2(\alpha, \gamma \mid \mathcal{D}) = \sum_{i \in I_1(\beta)} x_i^{\alpha} + (1+\gamma) \left(\sum_{i \in I_1(\beta)} x_{i+1}^{\alpha} + \sum_{i \in I_2(\beta)} x_i^{\alpha} \right) + \gamma \sum_{i \in I_2(\beta)} x_{i+1}^{\alpha}$$
$$+ \left(\gamma^2 + \gamma + 1\right) \sum_{i \in I_0(\beta)} x_i^{\alpha} - (1+\gamma) \sum_{i=2}^{n-1} x_i^{\alpha}$$

- The MLEs of γ and α, say $\widehat{\gamma}$ and $\widehat{\alpha}$, respectively, can be obtained by maximizing $l\left(\gamma, \widehat{\lambda}_1(\gamma, \alpha), \alpha\right)$.
- Finally, the MLE of λ_1 can be obtained as $\widehat{\lambda}_1(\widehat{\gamma}, \widehat{\alpha})$.

Goodness of fit

- We want to test the following null hypothesis

$$H_0 : \{X_1, \ldots, X_n\} \sim \text{WEP}\,(\alpha, \lambda_0, \lambda_1) \tag{7}$$

- We use the following statistic for goodness of fit test.

$$W_n = \max_{1 \le i \le n} |X_{i:n} - a_i| \,. \tag{8}$$

where, $X_{1:n} < \ldots < X_{n:n}$ as the ordered $\{X_1, \ldots, X_n\}$ and $a_i = E_{H_0}\,(X_{i:n})$.

- **Test criterion:**Reject H_0 if $W_n > c_n(\beta)$,where $c(\beta)$ is such that

$$P_{H_0}\,(W_n > c_n(\beta) = \beta. \tag{9}$$

- It is difficult to obtain $c_n(\beta)$ theoretically even for large n. Hence, we will use the parametric bootstrap technique to approximate $c_n(\beta)$ from a given observed sample $\{x_1, \ldots, x_n\}$.

Goodness of fit(Continued)

Algorithm 1 Algorithm for Parametric Bootstrap

1: Obtain $\widehat{\alpha}, \widehat{\lambda}_0, \widehat{\lambda}_1$, the MLEs of $\alpha, \lambda_0, \lambda_1$, respectively, based on $\{x_1, \ldots, x_n\}$

2: Generate a sample of size n from a $\mathrm{WEP}\left(\widehat{\alpha}, \widehat{\lambda}_0, \widehat{\lambda}_1\right)$, order them. Let us denote them as $\left(x^1_{1:n}, \ldots, x^1_{n:n}\right)$. Repeat this procedure B times, and obtain $\left\{\left(x^b_{1:n}, \ldots, x^b_{n:n}\right); b = 1, \ldots, B\right\}$.

3: Obtain estimates of $a_1, \ldots, a_n$ as

$$\widehat{a}_i = \frac{1}{B}\sum_{b=1}^{B} x^b_{i:n}; \quad i = 1, \ldots, n$$

4: Compute

$$w^b = \max_{1 \le i \le n} \left\{\left|x^b_{i:n} - \widehat{a}_i\right|\right\}; \quad b = 1, \ldots, B$$

5: Order $\left\{w^1, \ldots, w^B\right\}$ as $w^{(1)} < \ldots < w^{(B)}$, then $\widehat{c}_n(\beta) = w^{[(100(1-\beta))]}$ is an estimate of $c_n(\beta)$.

Data Application

- We made two synthetic data sets to see the performance of the models.
- **Synthetic Data 1:**It has been generated using the following model specification: $\alpha = 2.0, \lambda_0 = \lambda_1 = 1.0, n = 75$. The data set is presented in Figure 1. Based on the profile maximization we obtain the MLEs of α and λ as $\widehat{\alpha} = 1.889$ and $\widehat{\lambda} = 1.025$. The associate 95% confidence intervals are $(1.743, 2.056)$ and $(0.863, 1.264)$, respectively.

Data Application(Continued)

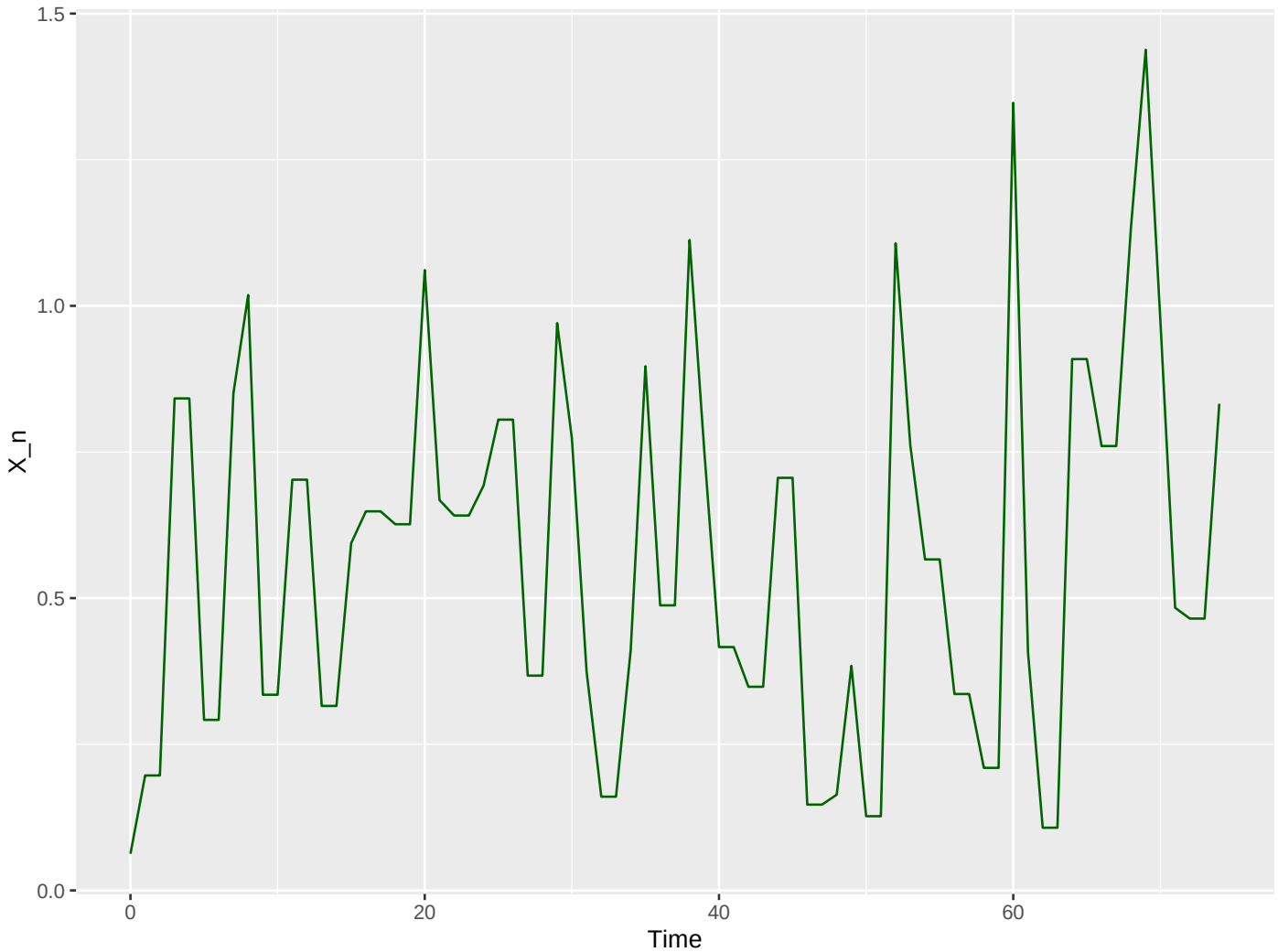

Figure 1: Synthetic data set with $\alpha = 2.0, \lambda_0 = \lambda_1 = 1$

Simulation study(Data Set 1

- **Simulation Study:**Now we have done a simulation study and find the MLEs of α and λ for that model specification, where $\lambda_0 = \lambda_1 = \lambda$ for different sample size i.e for $n = 75, 100, 125, 150, 175, 200$ in such a manner that we first draw a sample from the model specification where $\alpha = \widehat{\alpha}$ and $\lambda = \widehat{\lambda}$.Then we find MLE's of α and λ from here and again use those MLE's of α and λ in the model specification to generate dataset.Then, We repeat this procedure 1000 times where we find MLE's of α and λ at each time and then use that MLE's in the model specification to generate again an another dataset and use that new dataset to find the MLE's of α and λ again.

Simulation study(Dataset 2)

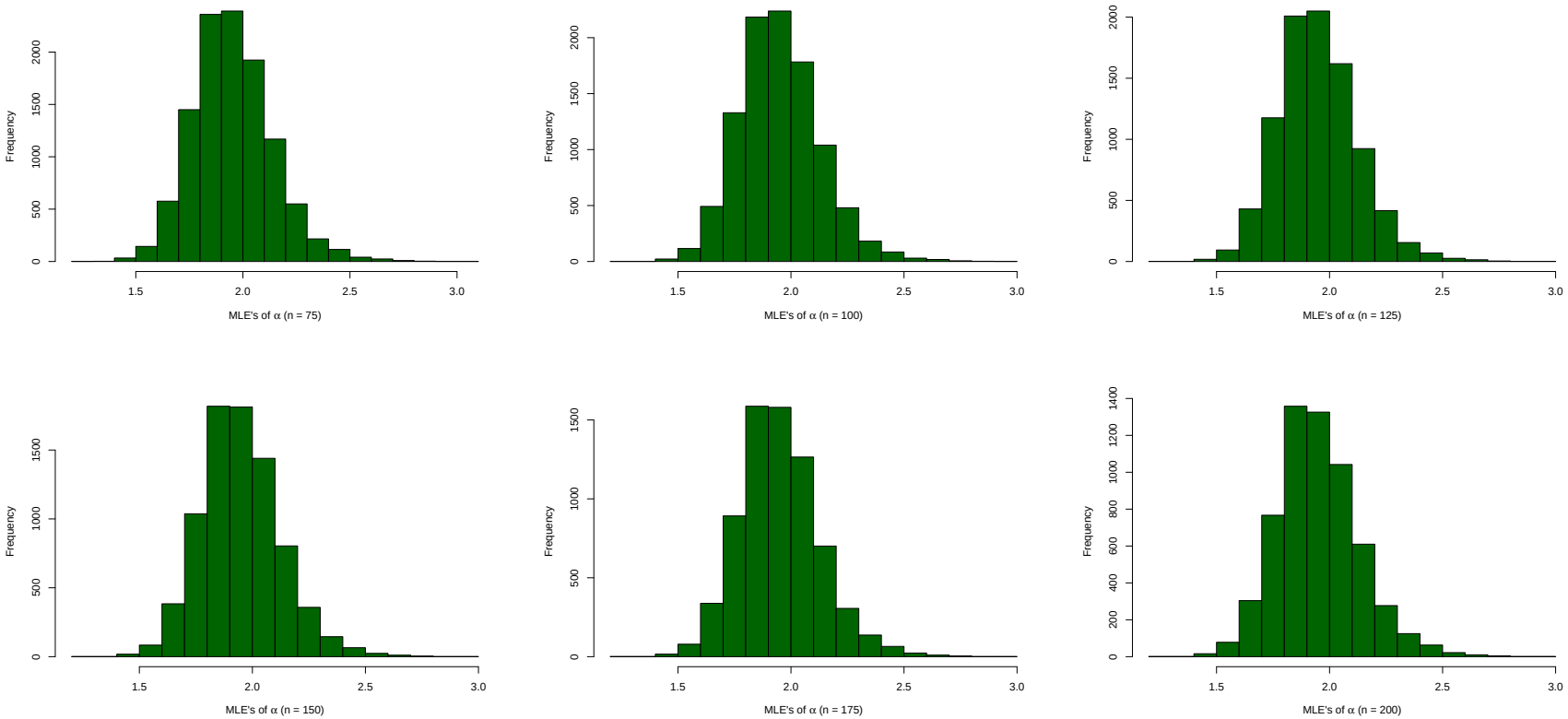

Figure 2: Histograms of MLE of α

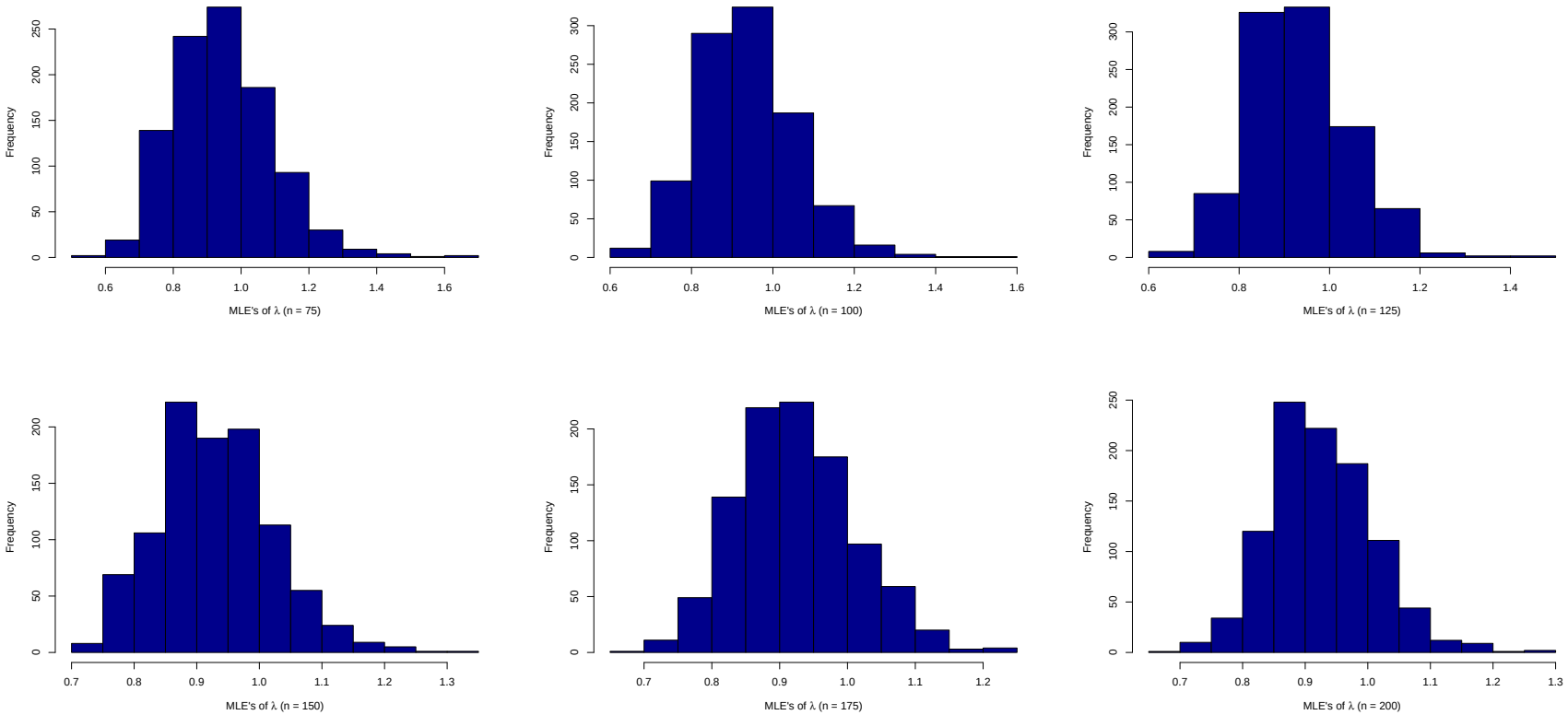

Figure 3: Histograms of MLE of λ

Data Application(Continued)

- **Synthetic Data 2:**It has been generated using the following model specification: $\alpha = 3.0, \lambda_0 = 0.15, \lambda_1 = 0.04$ and $n = 75$. The data set is presented in Figure 4. Based on the profile maximization we obtain the MLEs of α, λ_0 and λ_1 become $\widehat{\alpha} = 3.295, \widehat{\lambda}_0 = 0.1429$ and $\widehat{\lambda}_1 = 0.024$. The associated 95% bootstrap confidence intervals become $(2.876, 3.954), (0.124, 0.152)$ and $(0.014, 0.035)$, respectively.

Data Application(Continued)

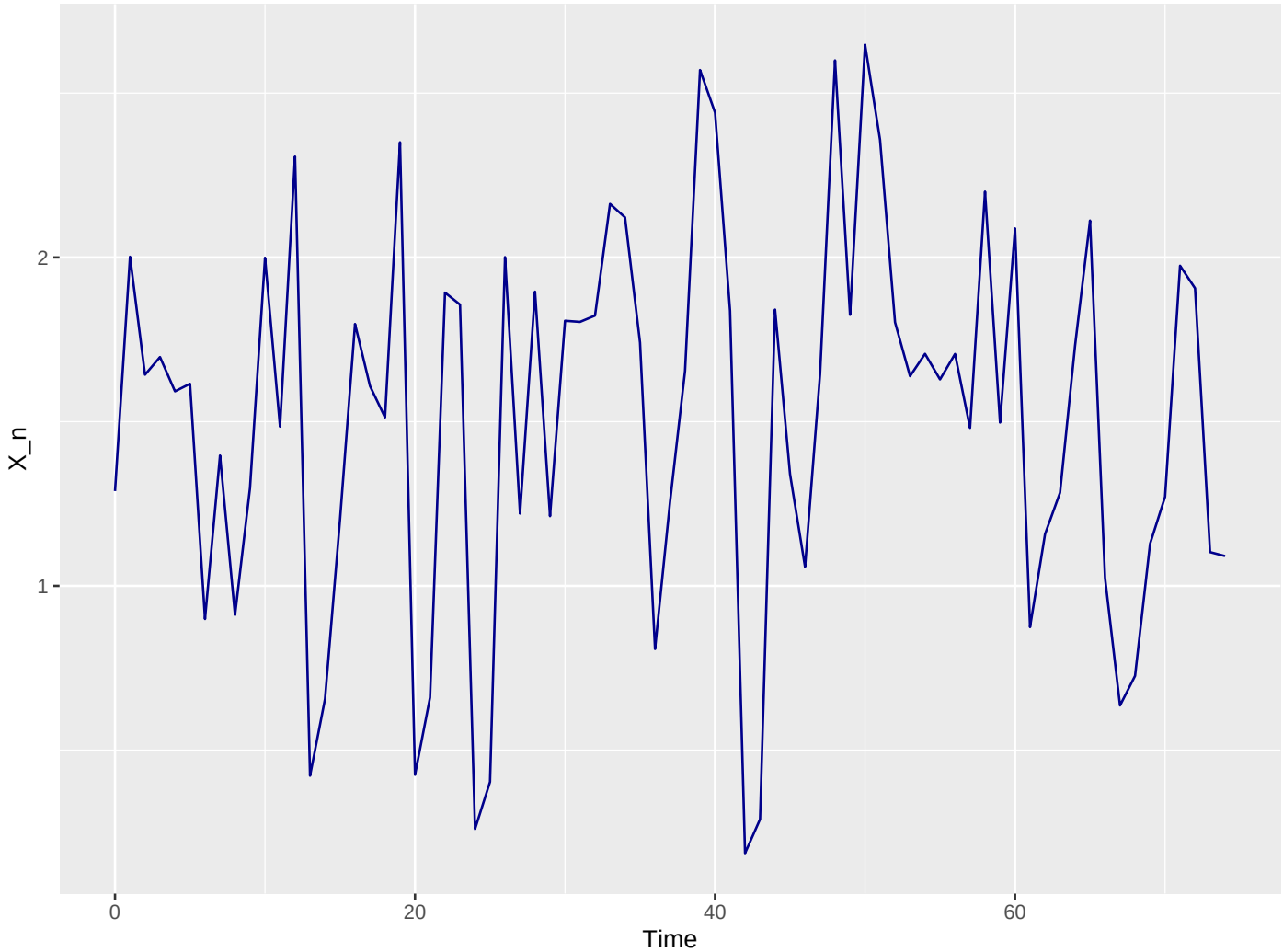

Figure 4: Synthetic data set with $\alpha = 3.0, \lambda_0 = 0.15 and \lambda_1 = 0.04$.

Simulation Study(Data Set 2)

- **Simulation Study:**Now we have done a simulation study and find the MLEs of α, λ_0 λ_1 for that model specification where λ_0 and λ_1 are different for different sample sizes, $n = 50, 75, 100, 125, 150, 175, 200$ in such a manner that we first draw a sample from the model specification where $\alpha = \widehat{\alpha}$,$\lambda_0 = \widehat{\lambda_0}$ and $\lambda_1 = \widehat{\lambda_1}$.Then we find MLEs of α,λ_0,λ_1 from here and again use those MLEs of α,λ_0 and λ_1 in the model specification to generate the dataset. Then, We repeat this procedure 1000 times where we find MLEs of α and λ at each time and then use those MLEs in the model specification to generate another dataset and use that new dataset to find the MLEs of α,$\lambda_0$andλ_1 again.

Simulation study(Dataset 2)

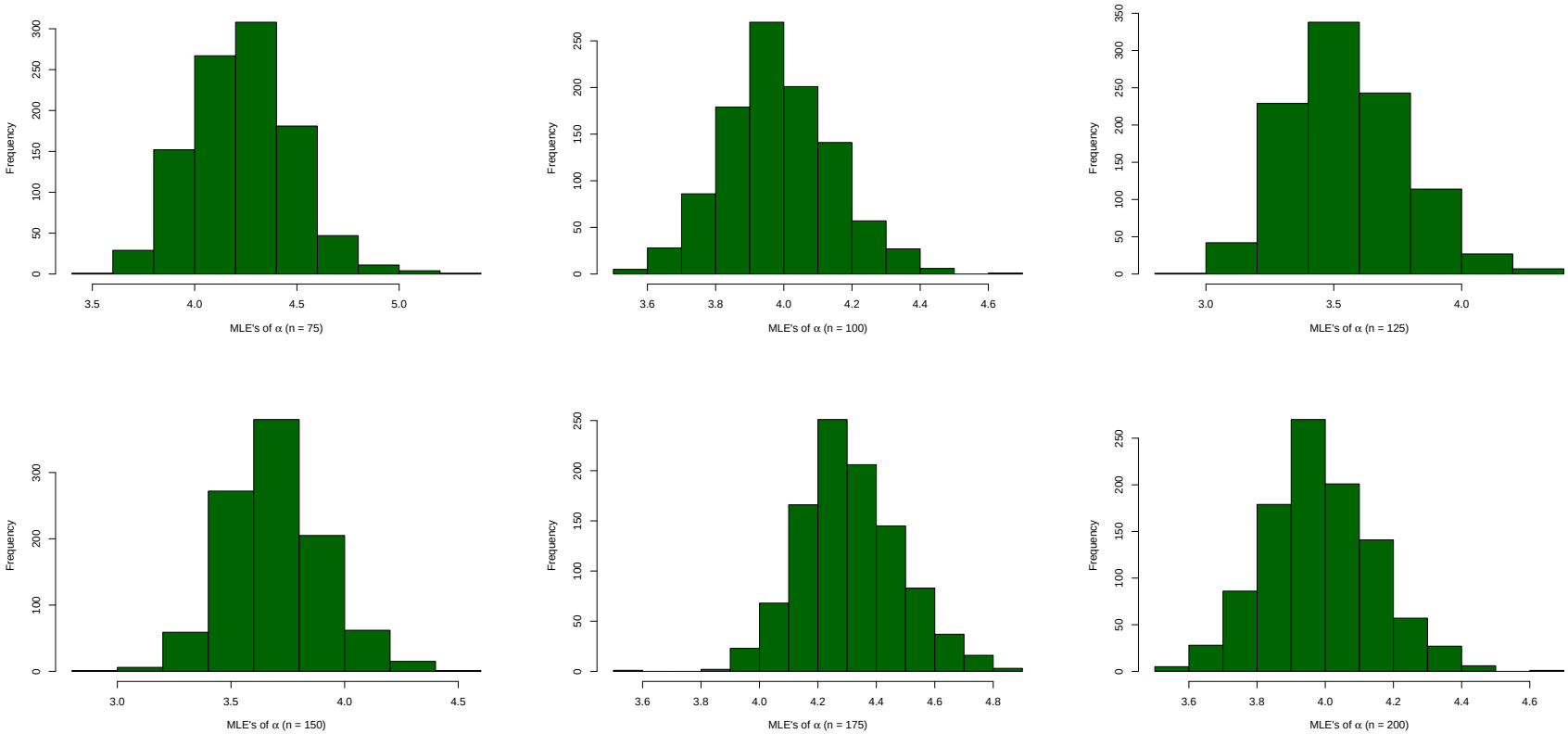

Figure 5: Histograms of MLE of α

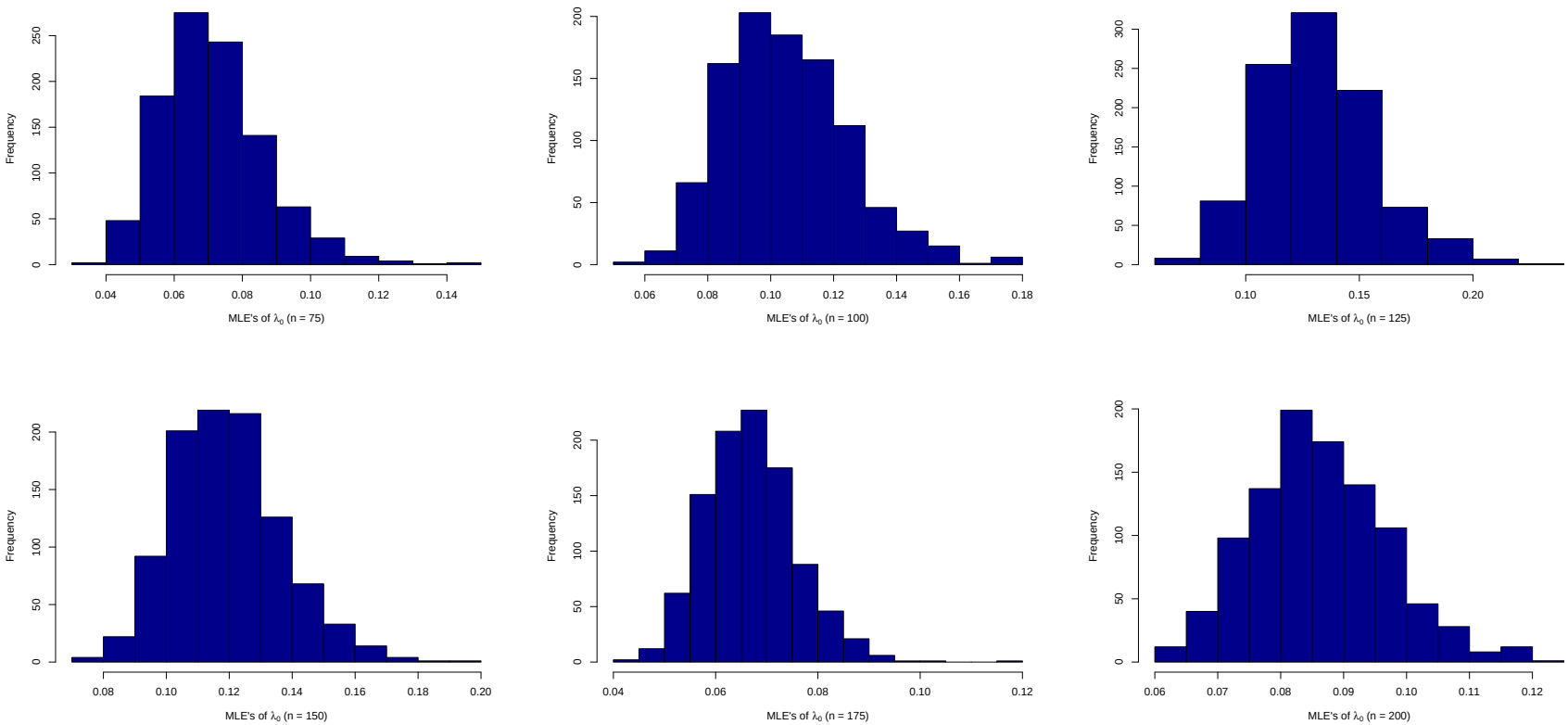

Figure 6:Histograms of MLE of λ_0

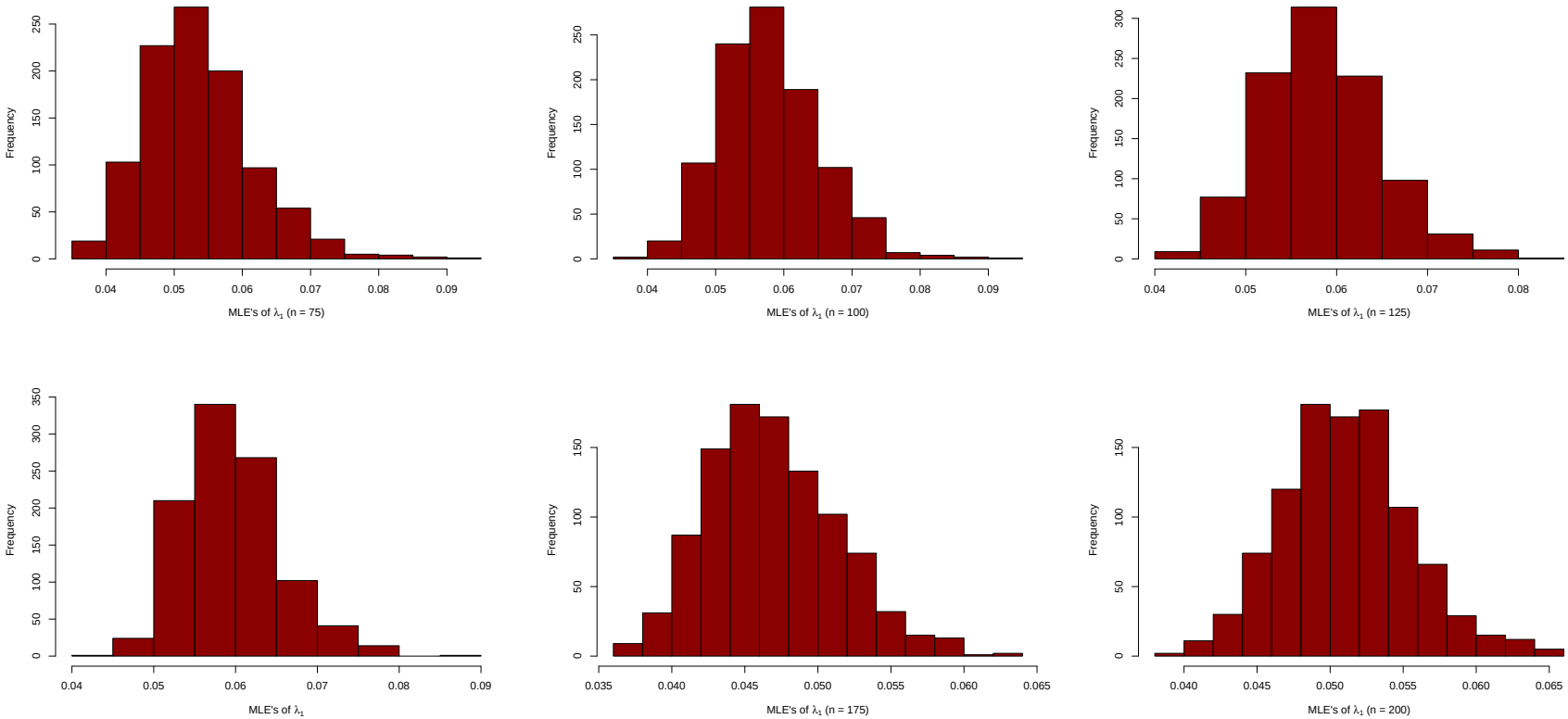

Figure 7:Histograms of MLE of λ_1

R Package

- Our main goal was to make an R-Package where the user will provide time-series data and our function **stnwep** in the package will implement the dataset in the two above-mentioned models and will find MLEs of α and λ for the the model with specific setting that $\lambda_0 = \lambda_1 = \lambda$ and also will find the MLE's of α, λ_0 and λ_1 for the model with specific setting where λ_0 is not equal to λ_1.
- Then it will generate test statistic using parametric bootstrap method and will give histograms for both model.Ultimately,it will provide the test statistics which is used to test whether the data is coming from the model 1 or it is coming from model 2.Then it will provide the corresponing p-values to measure how good the fit is.We made a GitHub repository of our R-package.The name of our package is **stnweib**.
- Link: https://github.com/debarghya2000/stnweib

R Package test run results

When $\alpha = 3$, $\lambda_0 = 0.15$, $\lambda_1 = 0.04$(figure 8) then,

R-package test for WEP(n,α, λ_0, λ_1)						
	$n = 75$	$n = 100$	$n = 125$	$n = 150$	$n = 175$	$n = 200$
Test Statistic	1.27073e-01	1.50618e-01	1.430506e-01	9.76388e-02	9.91640e-02	8.52095e-02
P-value	0.953	0.863	0.832	0.968	0.957	0.982

R-package test for WEP(n,α, λ, λ)						
	$n = 75$	$n = 100$	$n = 125$	$n = 150$	$n = 175$	$n = 200$
Test Statistic	1.28159e-01	1.58759e-01	1.02728e-01	1.8353e-01	1.4329e-01	1.8492e-01
P-value	0.024	0.037	0.026	0.021	0.031	0.042

R Package test run results

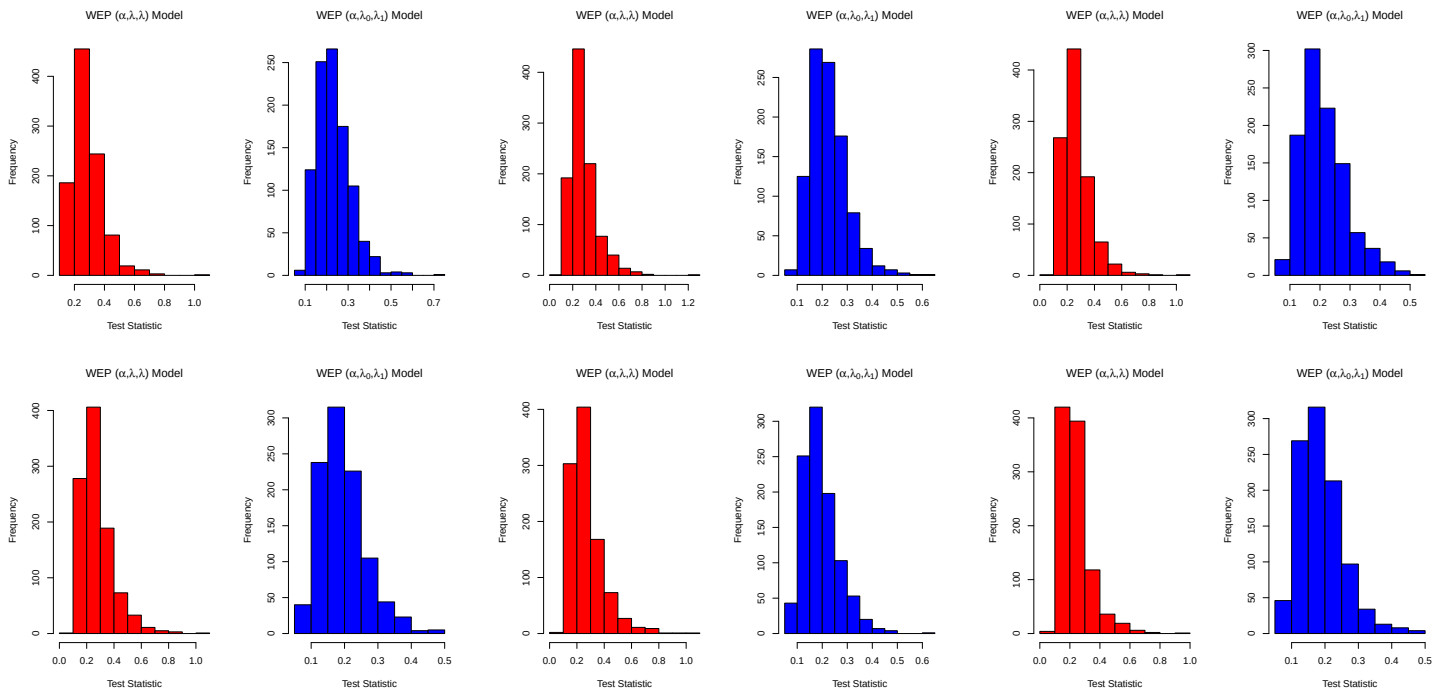

Figure 8:Histograms of Test Statistic

R Package test run results

When $\alpha = 5$, $\lambda_0 = 0.25$, $\lambda_1 = 0.07$(figure 9) then,

R-package test for WEP(n,α, λ_0, λ_1)						
	$n = 75$	$n = 100$	$n = 125$	$n = 150$	$n = 175$	$n = 200$
Test Statistic	9.48934e-02	6.34619e-02	7.11130e-02	3.70319e-02	5.06981e-02	5.62673e-02
P-value	0.861	0.965	0.9	0.997	0.975	0.93

R-package test for WEP(n,α, λ, λ)						
	$n = 75$	$n = 100$	$n = 125$	$n = 150$	$n = 175$	$n = 200$
Test Statistic	1.3957e-01	1.2138e-01	1.4287e-01	1.2985e-01	1.3628e-01	1.1832e-01
P-value	0.34	0.034	0.026	0.032	0.021	0.25

R Package test run results

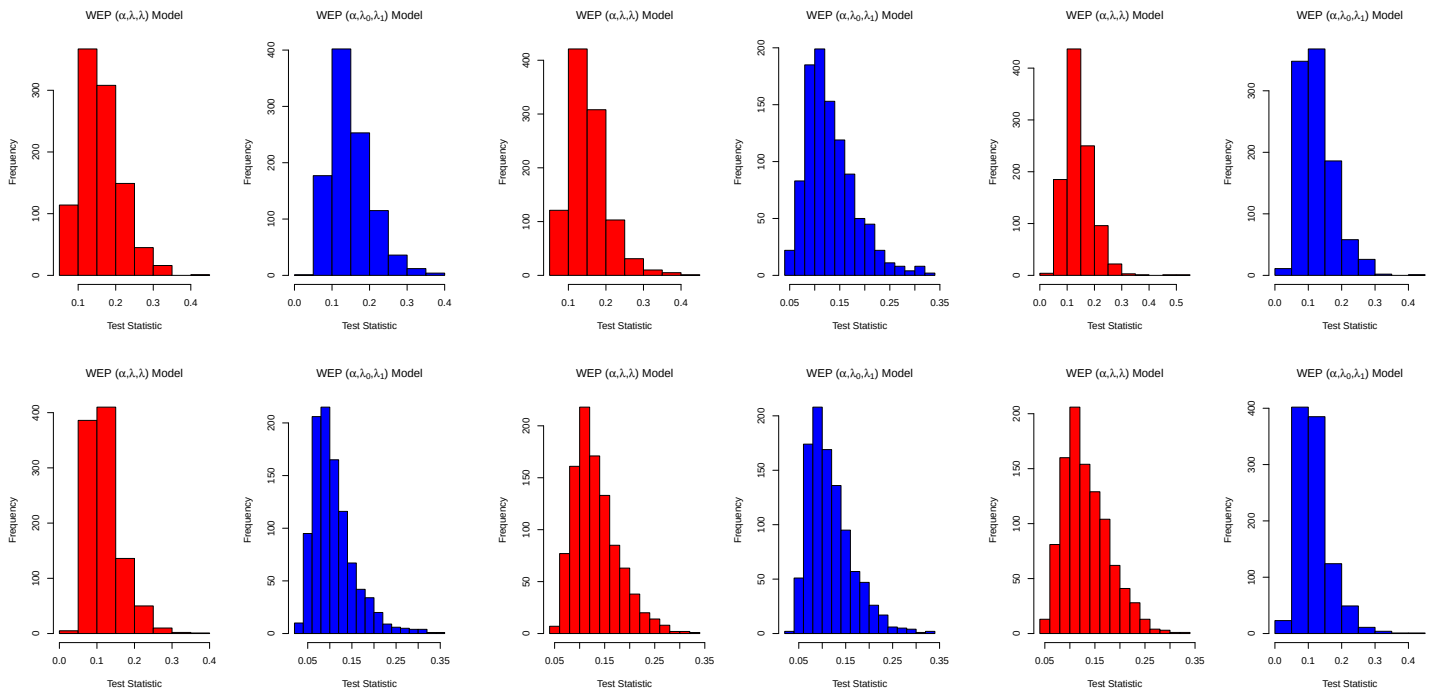

Figure 9: Histograms of Test Statistics

R Package test run results

When $\alpha = 4$, $\lambda_0 = 0.2$, $\lambda_1 = 0.2$(figure 10) then,

R-package test for WEP(n,α, λ, λ)						
	$n = 75$	$n = 100$	$n = 125$	$n = 150$	$n = 175$	$n = 200$
Test Statistic	1.01350e-01	9.69882e-02	8.13595e-02	7.82217e-02	7.5234e-02	8.24579e-02
P-value	0.954	0.903	0.97	0.981	0.979	0.935

R-package test for WEP(n,α, λ_0, λ_1)						
	$n = 75$	$n = 100$	$n = 125$	$n = 150$	$n = 175$	$n = 200$
Test Statistic	1.3983e-01	1.6529e-01	1.2139e-01	1.2658e-01	1.3429e-01	1.7592e-01
P-value	0.262	0.118	0.038	0.329	0.023	0.262

R Package test run results

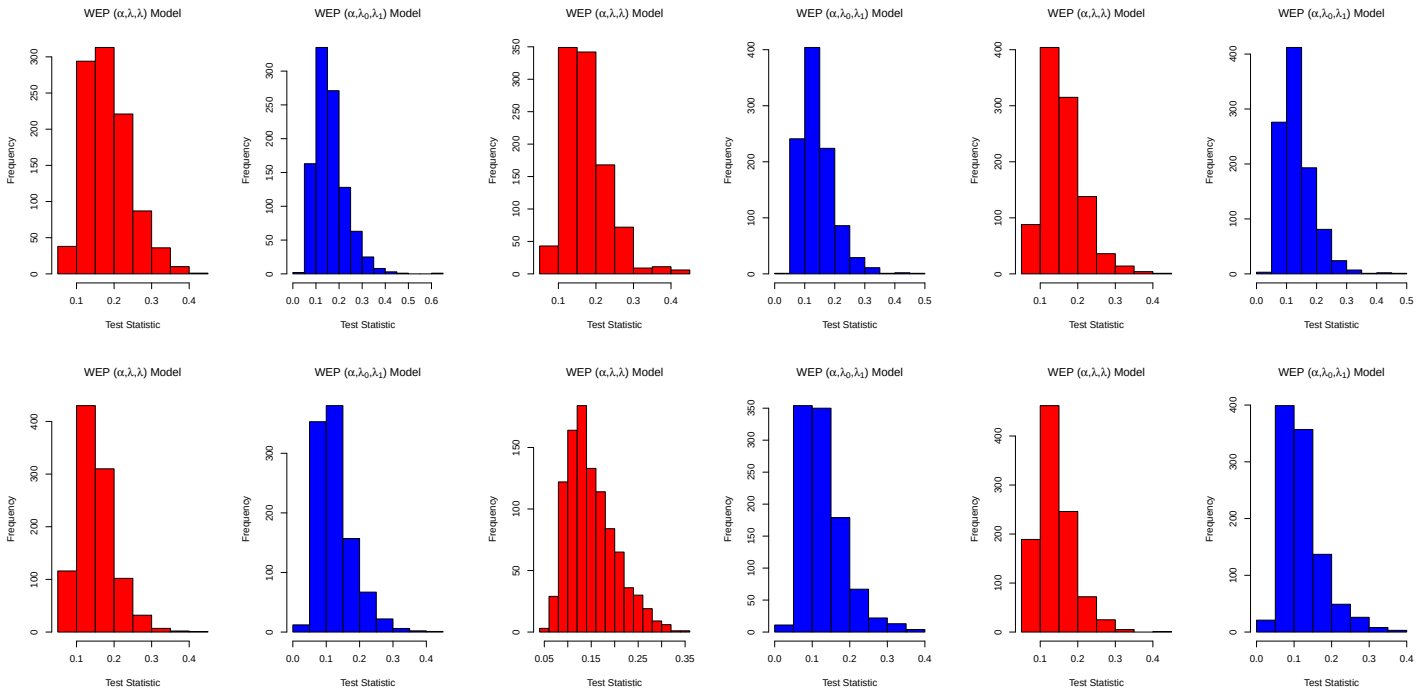

Figure 10:Histograms of Test Statistics

Frame Title

When $\alpha = 8$, $\lambda_0 = 0.7$, $\lambda_1 = 0.7$(figure 11) then,

R-package test for WEP(n,α, λ, λ)						
	$n = 75$	$n = 100$	$n = 125$	$n = 150$	$n = 175$	$n = 200$
Test Statistic	1.01350e-01	9.69882e-02	8.1628e-02	7.3298e-02	7.8734e-02	8.9238e-02
P-value	0.977	0.987	0.932	0.9	0.956	0.949

R-package test for WEP(n,α, λ_0, λ_1)						
	$n = 75$	$n = 100$	$n = 125$	$n = 150$	$n = 175$	$n = 200$
Test Statistic	1.28159e-01	1.58759e-01	1.58759e-01	1.03536e-01	1.43619e-01	1.545802e-01
P-value	0.262	0.118	0.038	0.329	0.883	0.262

R Package test run results

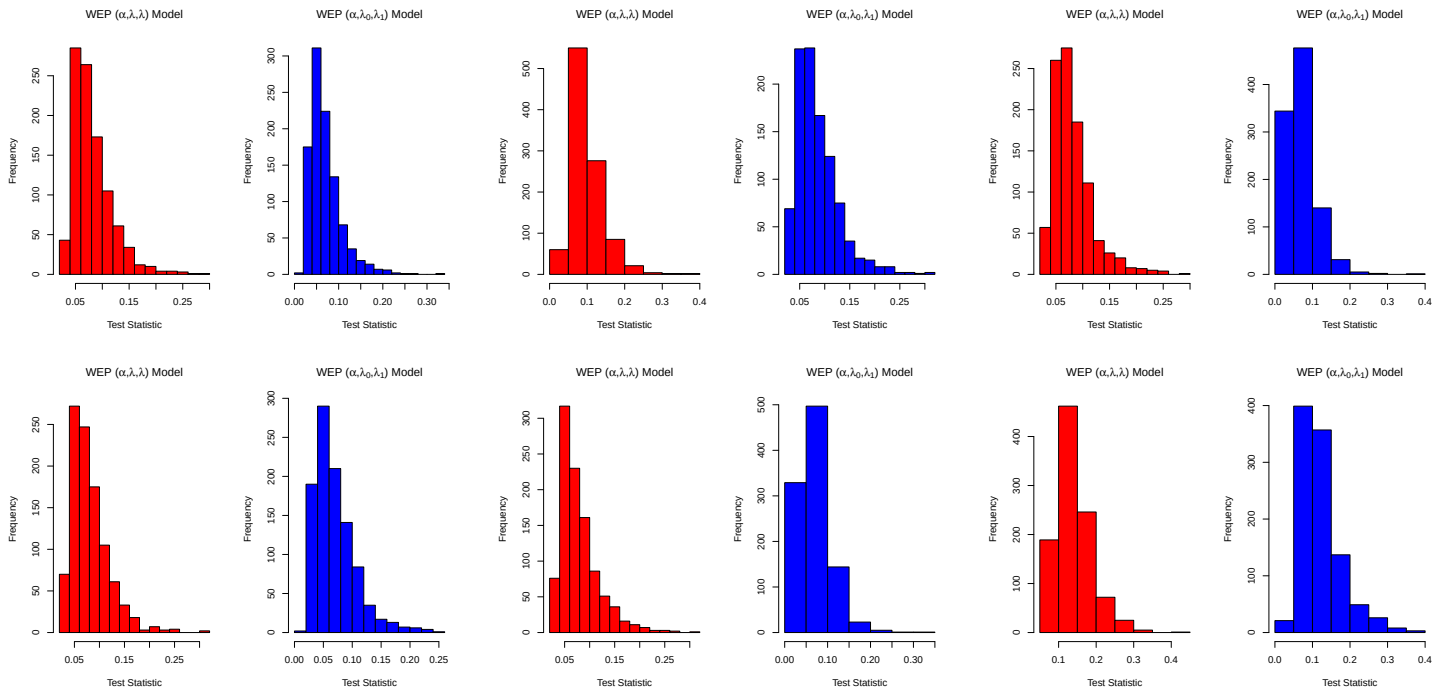

Figure 11:Histograms of Test Statistics

Appendix

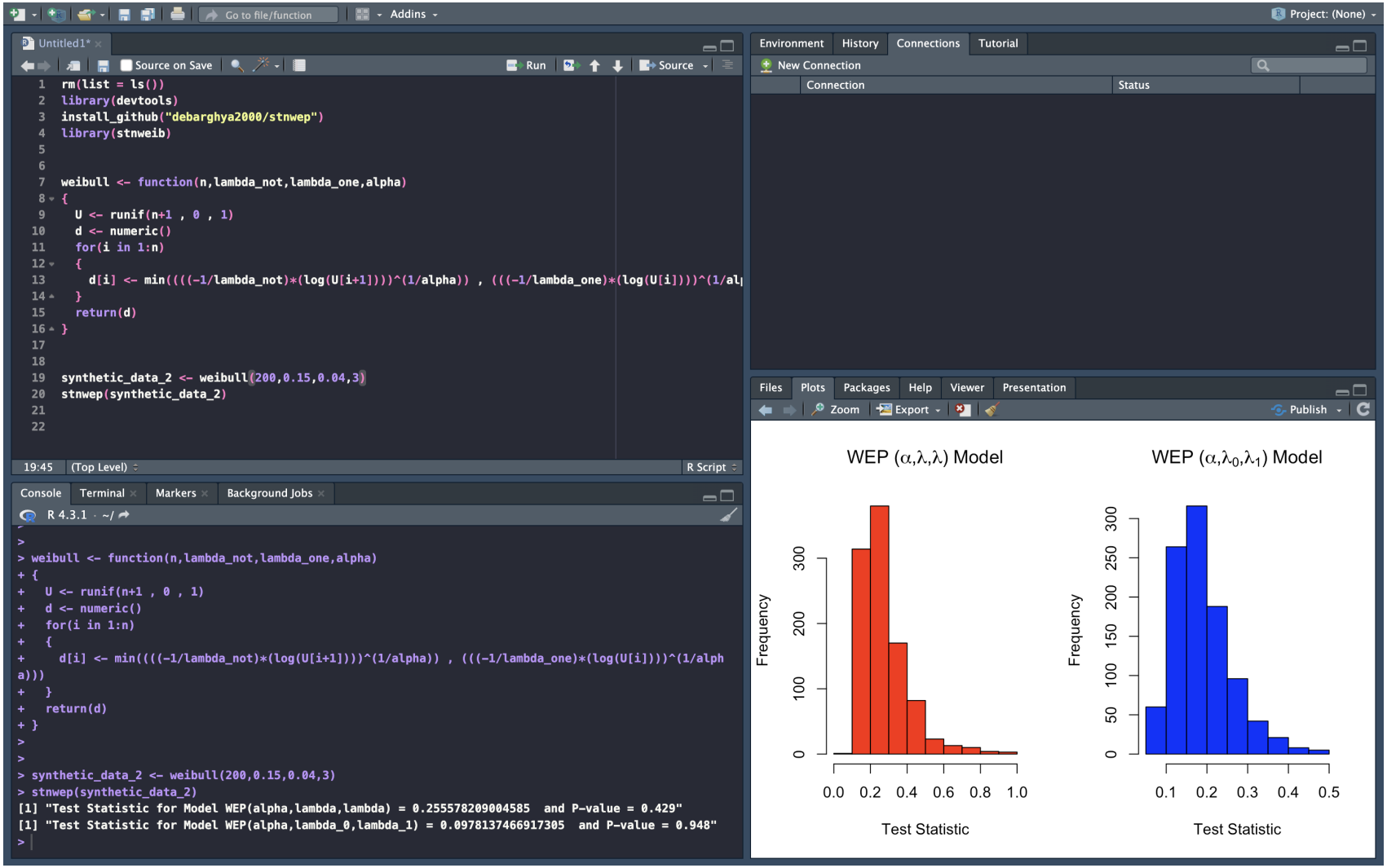

Figure 10:R package run in R Studio

References

[1] Kundu, D. (2022). A stationary Weibull process and its applications. Journal of Applied Statistics 1–20.

[2] Kundu, D. and Dey, A. K. (2009). Estimating the parameters of the Marshall–Olkin bivariate Weibull distribution by EM algorithm. Computational Statistics Data Analysis 53 956–965.

[3] Novikov, A. and Shiryaev, A. (2007). On a solution of the optimal stopping problem for processes with independent increments. Stochastics An International Journal of Probability and Stochastic Processes 79 393–406.

[4] Bemis, B. M., Bain, L. J. and Higgins, J. J. (1972). Estimation and hypothesis testing for the parameters of a bivariate exponential distribution. Journal of the American Statistical Association 67 927–929.

Thank you!