REPORT ON

Problem 6: Earthquake Interevent Time Distribution

MATH F432 - APPLIED STATISTICAL METHODS

FIRST SEMESTER 2020-21

Submitted by: BAYES GROUP (Group-8)

Supervised by: Dr. Sumanta Pasari

Submitted on: 27th November 2020

Group Members

Aditya Tulsyan: 2017B4A70740P

Ishan Rai: 2017B4A20596P

Manan Soni: 2017B4A70495P

Sarthak Bansal: 2017A4PS0311P

Aishwarya Kondaveeti: 2017B4A70697P

Sonam Verma: 2017B4TS1224P

Bhavika Sajnani: 2017B4A80789P

1. Introduction

Earthquake:

Earthquakes are phenomena of the movement of the earth’s surface, it happens when two blocks of the earth instantaneously pass over one another. The surface where they slip is called the fault plane. Earthquakes are measured by an instrument named Seismograph. The size of an earthquake is called its magnitude. Scientists use the triangulation method to determine the exact location of an earthquake since it takes three seismographs to locate an earthquake.

Why is Indonesia prone to earthquakes?

Indonesia is located on the “**PACIFIC RING OF FIRE**” (a path along the Pacific Ocean, an area with a high degree of tectonic activities and nearly ninety percent of Earth’s earthquakes occur along its path). Over the past 20 years, earthquakes are the biggest threat in terms of natural disasters in Indonesia.

Earthquakes with a magnitude of 5.0 or lower occur almost daily in Indonesia. Sumatra is the worst affected island of Indonesia since it lies at the boundary of two of Earth's tectonic plates, hence making it highly prone to earthquakes. In 2004, the 9.1-magnitude earthquake of Sumatra is considered to be the largest earthquake ever recorded. The earthquake and subsequent tsunami had killed 227,898 and left 1.7 million people displaced across the Pacific region. In recent times, on September 28, 2018, a magnitude of 7.5 earthquakes struck Indonesia, more than 2000 people are known to have died and about 4400 seriously injured [2].

Difference between prediction and forecasting

Earthquake forecasting is concerned with the probabilistic assessment of general earthquake seismic hazard, including the frequency and magnitude of damaging earthquakes in a given area over years or decades. Trend methods are generally thought to be useful for forecasting.

Earthquake prediction is concerned with finding the date and time, location, and magnitude of an earthquake. There’s not enough scientific evidence to predict an earthquake. However, studies on prediction focus on its empirical analysis, i.e identifying distinctive precursors to earthquakes. Precursor methods are pursued largely because of their potential utility for short-term earthquake prediction.

2. Data Filtering

We collected the data from the ISC bulletin for the years from **1900.01.01 to 2020.10.20** with a magnitude ranging between 6-8 containing total 347 entries. A sample of the data is shown in table 2.1.

	EVENTID	AUTHOR	DATE	TIME	LAT	LON	DEPTH	DEPFIX	AUTHOR.1	TYPE	MAG
0	16957943	ISC	1907-01-04	05:19:11.80	1.8725	94.2091	15.0	TRUE	ISC	MS	7.8
1	913990	ISC	1914-06-25	19:07:25.91	-3.9242	101.8203	35.0	TRUE	ISC	MS	7.6
2	913286	ISC	1918-09-22	09:55:00.13	-1.4570	100.0902	35.0	TRUE	ISC	MS	6.5
3	912712	ISC	1919-01-18	05:52:24.87	-4.5766	101.4670	35.0	TRUE	ISC	MS	6.3
4	912756	ISC	1919-04-02	00:34:59.58	-5.4963	104.4870	20.0	TRUE	ISC	MS	6.5

Table 2.1 Raw data from the ISC bulletin

Earthquakes are followed by a large number of aftershocks (identifiable by their occurrence in limited intervals of time, space, and small magnitude) and preceded by foreshocks. Hence to obtain completely reliable data for further analysis the aftershocks, foreshocks and need to be filtered out and thus obtained data is independent and will not affect the result [4].

We wrote code for a **dynamic window-based Spatio-temporal filtering algorithm** to remove **foreshocks, aftershocks and seismic clusters** from our data. The filtered table contained only 164 entries which are i.i.d and is shown in table 2.2.

	EVENTID	AUTHOR	DATE	TIME	LAT	LON	DEPTH	DEPFIX	AUTHOR.1	TYPE	MAG
0	16957943	ISC	1907-01-04	05:19:11.80	1.8725	94.2091	15.0	TRUE	ISC	MS	7.8
1	913990	ISC	1914-06-25	19:07:25.91	-3.9242	101.8203	35.0	TRUE	ISC	MS	7.6
2	913286	ISC	1918-09-22	09:55:00.13	-1.4570	100.0902	35.0	TRUE	ISC	MS	6.5
3	912712	ISC	1919-01-18	05:52:24.87	-4.5766	101.4670	35.0	TRUE	ISC	MS	6.3
4	912756	ISC	1919-04-02	00:34:59.58	-5.4963	104.4870	20.0	TRUE	ISC	MS	6.5

Table 2.2 Filtered Data containing 164 entries as compared to 347 entries in the original dataset.

The window that we have taken for selection is –

- Search Radius, r = exp(-1.024+0.804M) ±15
- Time Window, t = exp(-2.870 + 1.235M) ± 60

M: the magnitude of the event, r is radius taken in kilometers, t is the time taken in days(M>6).

A separate column was added with title INTEREVENT TIME which contained the time difference between the current occurrence of an earthquake and the last occurrence.

The following table (Table: 2.3.1) shows the intervening time for the filtered data.

	EVENTID	AUTHOR	DATE	TIME	LAT	LON	DEPTH	DEPFIX	AUTHOR.1	TYPE	MAG	INTEREVENT TIME
1	913990	ISC	1914-06-25	19:07:25.91	-3.9242	101.8203	35.0	TRUE	ISC	MS	7.6	7.580556
2	913286	ISC	1918-09-22	09:55:00.13	-1.4570	100.0902	35.0	TRUE	ISC	MS	6.5	4.305556
3	912712	ISC	1919-01-18	05:52:24.87	-4.5766	101.4670	35.0	TRUE	ISC	MS	6.3	0.327778
4	912756	ISC	1919-04-02	00:34:59.58	-5.4963	104.4870	20.0	TRUE	ISC	MS	6.5	0.205556
5	912979	ISC	1919-10-12	21:48:31.74	-4.1786	101.7521	25.0	TRUE	ISC	MS	6.5	0.536111

Table 2.3: Containing the newly added column of Interevent Time

3. Methodology

Our approach to solving the problem at hand includes the following three major steps-

- Model description
- Parameter Estimation
- Model Validation

3.1 Model Description

Here, we've used four major different probability distributions for the inter-event times and they are:

(i) Exponential distribution

(ii) Gamma distribution

(iii) Lognormal distribution

(iv) Weibull distribution

To calculate the conditional probability of occurrence of an earthquake in the time interval: $(\alpha, \alpha + v)$, assuming that the elapsed time after the last major event is α and there is no major event until the time α.

$$P(V \leq \alpha + v \mid V \geq \alpha) = [F(\alpha + v) - F(\alpha)] / [1 - F(\alpha)]$$

Where V is the random variable corresponding to waiting time v.

The following table consists of the probability density function and the parameters of the four distributions(time-dependent Poisson distributions) we have taken for analysis.

Exponential	$\frac{1}{\alpha}e^{-\frac{t}{\alpha}}$	$t > 0$	α −scale	$\alpha > 0$
Gamma[a]	$\frac{1}{\Gamma(\beta)}\frac{t^{\beta-1}}{\alpha^{\beta}}e^{-\frac{t}{\alpha}}$	$t > 0$	α −scale β −shape	$\alpha > 0$ $\beta > 0$
Lognormal	$\frac{1}{t\beta\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\ln t-\alpha}{\beta}\right)^2\right]$	$t > 0$	α −log-scale β −shape	$-\infty < \alpha < \infty$ $\beta > 0$
Weibull[a]	$\frac{\beta}{\alpha^{\beta}}t^{\beta-1}e^{-\left(\frac{t}{\alpha}\right)^{\beta}}$	$t > 0$	α −scale β −shape	$\alpha > 0$ $\beta > 0$

Table 3.1.1 Probability density function and parameter corresponding to them.

3.2 Parameter Estimation

Parameter estimation for our problem was done using two methods namely Maximum Likelihood Estimation (MLE) and Method of Moments (MOM).

3.2.1 Method of Moments (MoM):

Method of Moments estimation starts by expressing the population moments as functions of the parameters to be estimated. Those expressions are then set equal to the sample moments. The number of moments taken is the same as the number of parameters in the distributions. In cases where we estimate the parameters of an unknown probability distribution, this estimation is chosen over the Maximum Likelihood estimate[7].

Assumptions for Method of Moments estimation: Independent and identically distributed sample.

For fitting the distributions using the Method of Moments we have used **fitdistrplus** library in R. The values are shown in table 3.2.1.

MOM Parameter Estimation

Model	Parameter Values	
Exponential	rate	1.594322
Gamma	shape	0.4913494
	rate	0.7833695
Lognormal	meanlog	-1.021589
	sdlog	1.053699
Weibull	shape	0.7153732
	scale	0.505648

Table 3.2.1: Parametric value corresponding to distributions as obtained from MoM method

Figure 3.2.1 shows the plot of the probability density function against the inter-event time. We can observe that most of the values are concentrated in the lower-medial range.

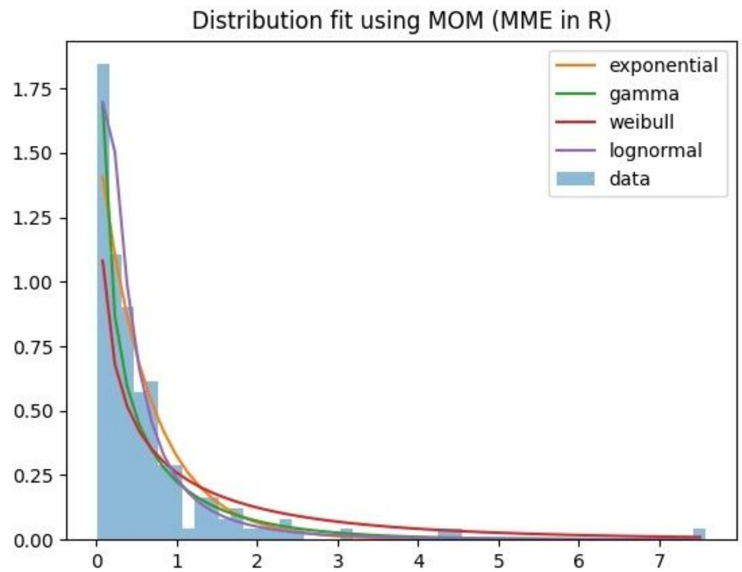


Figure 3.2.1 Result Visualizations

3.2.2 Maximum Likelihood Estimation (MLE)

The goal of maximum likelihood estimation is to make inferences about the population that is most likely to have generated the sample. The likelihood function is defined as the joint density function of the observed data. Finally, we want to maximize this likelihood function and obtain the optimum values of our parameters. Computations are made easier by taking the log of the likelihood function to obtain a log-likelihood function. This function is then maximized to get the optimized values of the parameters.

Some of the key highlights of Maximum Likelihood Estimation:

- Maximum Likelihood estimators are asymptotically unbiased, and the bias tends to zero as the sample size increased.
- These are asymptotically efficient i.e. they achieve the Cramer-Rao lower bound as the sample size approaches infinity.

Assumptions for Maximum Likelihood Estimation:

- Boundedness of the expected likelihood
- Differentiability of the log-likelihood

- Uniform integrability

We have used the **fitdistrplus** library in R for calculating the fitting values and the data is shown in table 3.2.2.

MLE Parameter Estimation		
Model	**Parameter Values**	
Exponential	rate	1.594323
Gamma	shape	0.8055547
	rate	1.2842414
Lognormal	meanlog	-0.4754889
	sdlog	1.3596
Weibull	shape	0.8456672
	scale	0.5691232

Table 3.2.2: Parametric value corresponding to distributions obtained from the MLE method

The following plot shows the inter-event time against the probability density function of the chosen distributions in the figure 3.2.2

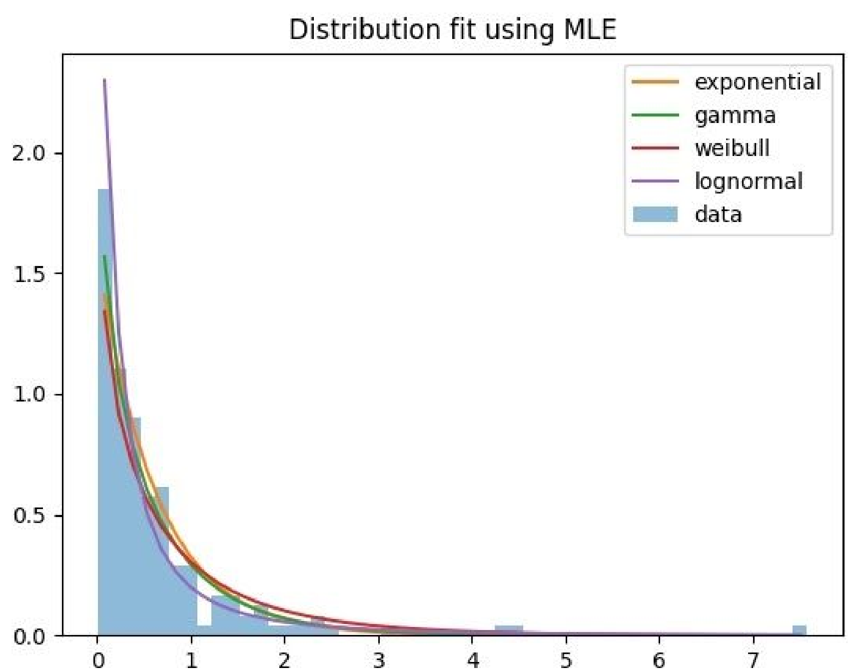


Figure 3.2.2 Result Visualizations

3.3 Model Validation

We use two popular goodness of fit tests to verify the results obtained from using the Parameter Estimation approaches are Akaike Information Criteria (AIC), and Kolmogorov-Smirnov (K-S) minimum distance criterion.

3.3.1 Akaike Information Criterion (AIC)

This criterion validates a model's efficiency for a given set of data. The work is based on information theory as whenever a model represents data it will never be exact and some information will always be lost, AIC estimates the information lost and compares the model with other suitable options. The risk of both over fittings and underfitting is taken into account by minimizing the loss instead of maximizing the likelihood function. If k be the number of estimated parameters and L be the maximum value of the likelihood function, then –

$$\mathbf{AIC = 2k - 2ln(L)}$$

The following assumptions need to be fulfilled for the AIC:

- The models use the same data between them
- The same outcome variable is measured between models
- The sample size is infinite

In practice, the last assumption is satisfied if the ratio of the number of samples to the number of parameters is more than 40. Otherwise, we need to use an adjusted version of the test known as AIC_c. In our case, we have about 160 samples and the number of parameters varies from 1 to 3. So the ratio is always greater than 40 and thus usage of this test is justified.

Now the preferred model will be with a minimum value of AIC. The values of AIC for MOM and MLE are calculated using **fitdist** function in the **fitdistrplus** library in R.

The values of MOM and MLE estimation for four distributions using the AIC test are shown in table 3.3.1 and table 3.3.2.

3.3.2 Kolmogorov-Smirnov(K-S) Test

This test is also known as the non-parametric test used to validate a continuous fully defined distribution for a sample of data given. This test draws a comparison between the empirical and cumulative distribution function. Our approach is to assume two hypotheses:

H_0: The sample belongs to the hypothesized distribution

H_a : The sample doesn't belong to the hypothesized distribution

The statistic that we use for hypothesis testing also known as the Kolmogorov – Smirnov for a cumulative distribution function *F(x)* is-

$$D_n = \sup |F_n(x) - F(x)|$$

Where sup is the supremum of the set of distances.

The test statistic represents the maximum error between the measured and the expected values of the cumulative function.

The K-S test is a **non-parametric** test and it also makes some implicit assumptions. Some of them are:

- The test only applies for univariate distributions
- The distribution must be continuous
- The samples should be i.i.d (independent and identically distributed)

The initial data did not satisfy the last assumption as the data included foreshocks and aftershocks. Thus we performed data preprocessing to make the resulting samples independent from each other.

The minimum KS distance is calculated by using the **K-S test function** in R.

The following tables show the result of the KS test on MOM and MLE estimation.

	Maximum Likelihood		K-S min distance	
Distribution	ln L	AIC	K-S distance	p-value
Exponential	-85.90173	173.8035	0.10542	0.05584
Gamma	-83.21791	170.4358	0.36928	2.20E-16
Lognormal	-84.28948	172.579	0.76515	2.20E-16
Weibull	-81.36024	166.7205	0.22504	1.66E-07

Table 3.3.1 AIC and KS test for MLE estimation

	Maximum Likelihood		K-S min distance	
Distribution	ln L	AIC	K-S distance	p-value
Exponential	-85.90173	173.8035	0.10542	0.05584
Gamma	-94.85328	193.7066	0.18734	2.47E-05
Lognormal	-99.15332	202.3066	0.70951	2.20E-16
Weibull	-85.11699	174.234	0.22398	1.93E-07

Table 3.3.2 AIC and KS test for MOM estimation

4. Forecasting

The conditional probability of an earthquake occurring in the time interval $(\alpha, \alpha+\nu)$ is given by the formula discussed in section 3.1. From the last major event, the elapsed time is relative because the plots for the 4 distributions (Exponential, Logarithmic, Gamma, Weibull) for MLE and MOM are shown below for different values of α.

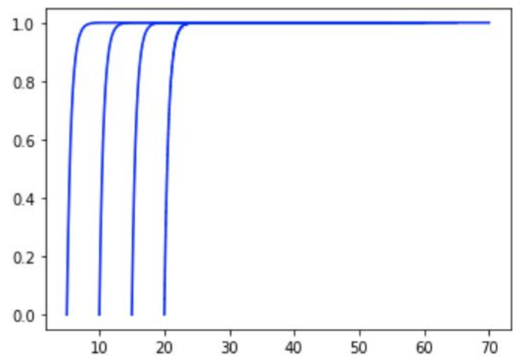


Fig 4.1.1 Exponential (MOM)

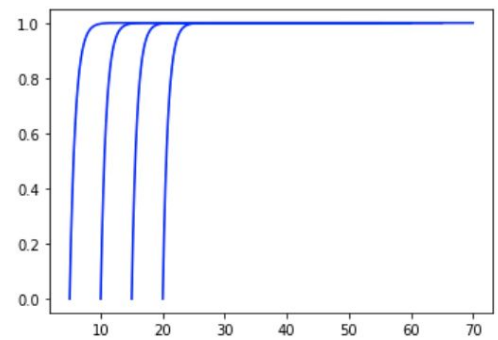


Fig 4.1.2 Gamma(MOM)

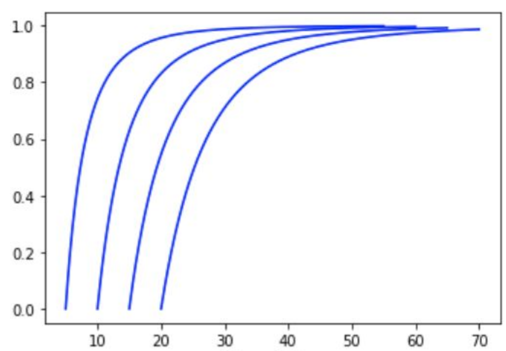


Fig 4.1.3 Lognormal(MOM)

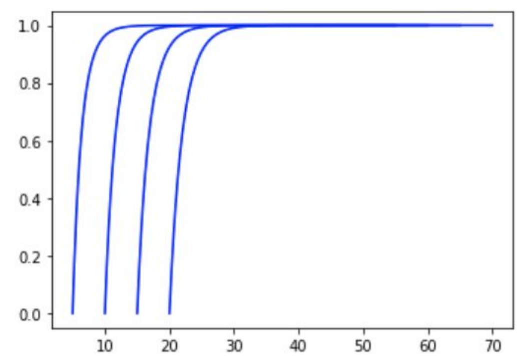


Fig 4.1.4 Weibull(MOM)

Fig 4.1 Conditional Probability curves for MOM parameters

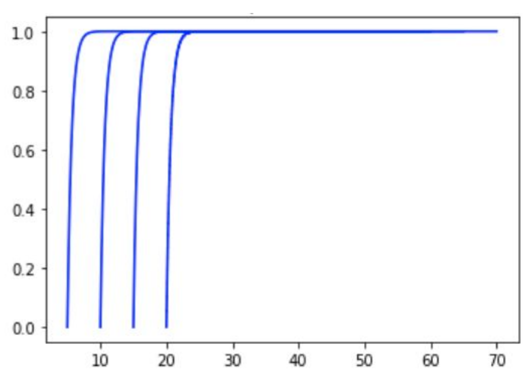


Fig 4.2.1 Exponential(MLE)

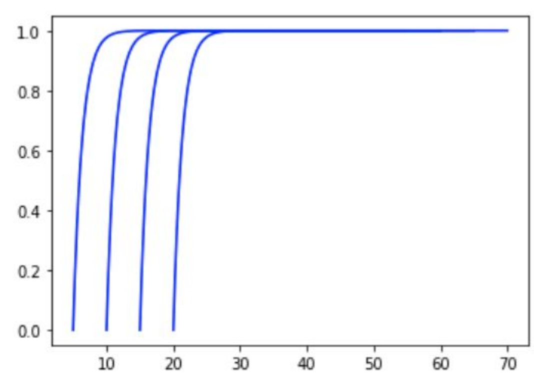
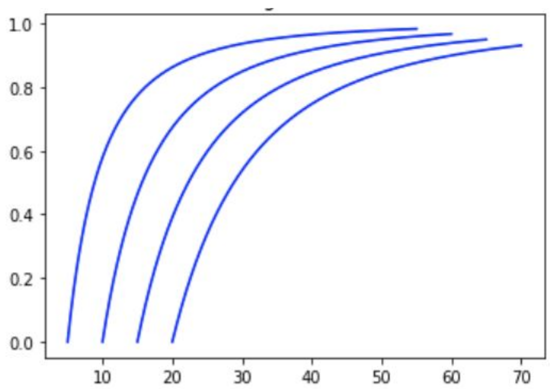


Fig 4.2.2 Gamma(MLE)



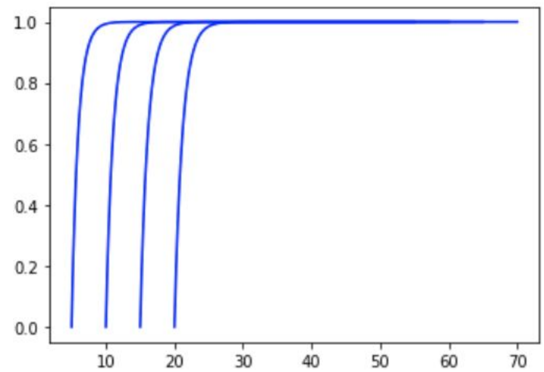


Fig 4.2.3 Lognormal(MLE) **Fig 4.2.4 Weibull(MLE)**

Fig 4.2 Conditional Probability curves for MLE parameters

5. Conclusion and Observations

As per parameters obtained from Maximum Likelihood Estimation and Method of Moments, and the two tests conducted to check the best fit model, Table 5.1 contains the conclusion drawn.

	Parameter Estimation Technique	
	MLE	**MOM**
AIC Test	Weibull	Exponential
KS Test	Exponential	Exponential

Table 5.1 Best fit probability distribution model to predict earthquakes

From table 5.1, we conclude that we can take two distributions, namely Exponential and Weibull distribution, to forecast the earthquakes. However, given that Exponential distribution is time-independent, Poisson distribution has memoryless property; hence, the probability will be constant, which demonstrates a **constant hazard function**.

The following table shows earthquake forecasting using the exponential distribution function. The columns represent waiting time, **v**, and the rows represent α, the time after the last major earthquake.

	v = 1	v = 3	v = 5	v = 7	v = 9	v = 11	v = 13	v = 15	v = 17	v = 19
alpha = 1	0.7969534468	0.9916288164	0.9996548737	0.9999857712	0.9999994134	0.9999999758	0.999999999	1	1	1
alpha = 3	0.7969534468	0.9916288164	0.9996548737	0.9999857712	0.9999994134	0.9999999758	0.999999999	1	1	1
alpha = 5	0.7969534468	0.9916288164	0.9996548737	0.9999857712	0.9999994134	0.9999999758	0.999999999	1	1	1
alpha = 7	0.7969534468	0.9916288164	0.9996548737	0.9999857712	0.9999994134	0.9999999758	0.999999999	1	1	1
alpha = 9	0.7969534468	0.9916288165	0.9996548737	0.9999857712	0.9999994133	0.9999999758	0.9999999991	1	1	1
alpha = 11	0.7969534488	0.991628818	0.9996548737	0.9999857695	0.9999994124	0.999999977	1	1	1	1
alpha = 13	0.7969534924	0.991628818	0.9996548334	0.999985748	0.9999994433	1	1	1	1	1
alpha = 15	0.7969536156	0.9916278442	0.999654311	0.9999864965	1	1	1	1	1	1
alpha = 17	0.7969343639	0.9916153544	0.9996724748	1	1	1	1	1	1	1
alpha = 19	0.7965023847	0.9920508744	1	1	1	1	1	1	1	1

Table 5.1: Probabilistic forecast of interevent time between earthquakes[1]

On the contrary, the Weibull distribution parameter, **shape, value is less than 1 (0.8457762 for MLE and 0.7153 for MOM)**, signifying that Weibull distribution demonstrates a **decreasing hazard function.** Hence Weibull distribution is preferred for earthquake forecasting over exponential distribution.

The following table shows earthquake forecasting using the Weibull distribution function.

	v = 1	v = 3	v = 5	v = 7	v = 9	v = 11	v = 13	v = 15	v = 17	v = 19
alpha = 1	0.7230394092	0.9724329847	0.9967161734	0.9995638366	0.9999374142	0.9999904695	0.9999984773	0.9999997468	0.9999999564	0.9999999923
alpha = 3	0.6748078733	0.9612625978	0.9948548333	0.999261711	0.9998875742	0.9999820376	0.9999970126	0.9999994858	0.9999999088	0.9999999834
alpha = 5	0.6491051468	0.9533935576	0.99331236	0.9989816139	0.999837291	0.9999729391	0.999995342	0.9999991738	0.9999998495	0.9999999719
alpha = 7	0.6314600426	0.9471175568	0.9919471229	0.9987133803	0.9997860159	0.9999631668	0.9999934669	0.9999988101	0.9999997781	0.9999999577
alpha = 9	0.6180239516	0.9418331303	0.990706596	0.9984543679	0.9997339492	0.9999528107	0.999991405	0.9999983969	0.9999996945	0.9999999406
alpha = 11	0.6071823931	0.9372389688	0.9895619014	0.9982032825	0.9996813171	0.9999419554	0.9999891736	0.9999979368	0.999999599	0.9999999207
alpha = 13	0.5981018401	0.9331583221	0.9884944936	0.9979592741	0.9996283043	0.9999306719	0.9999867881	0.9999974325	0.999999492	0.9999998978
alpha = 15	0.5902950815	0.929477196	0.9874914057	0.9977216978	0.9995750545	0.9999190179	0.9999842623	0.9999968861	0.9999993733	0.9999998723
alpha = 17	0.583452474	0.9261171746	0.9865430598	0.9974900318	0.9995216738	0.9999070445	0.9999816075	0.9999962981	0.9999992458	0.999999846
alpha = 19	0.5773649538	0.9230216965	0.9856421231	0.9972638107	0.9994682628	0.9998947888	0.9999788241	0.9999956858	0.9999991189	0.9999998177

Table 5.2

We got two different distributions because the data is not explicit and contains entries of magnitudes varying from 6 to 8 hence the irregularity. If taken for only one M value, Weibull would have shown significant results, and for the other might not.

To conclude the distribution used for forecasting the earthquakes in Indonesia's Sumatra region, we will use the Weibull distribution.

[1] The values shown as 1 in the figure do not signify the probability is one but is due to the limitation up to which decimal place software can calculate, and after that, the value is rounded off.

6. References

1. Pasari, S., Dikshit, O. Earthquake interevent time distribution in Kachchh, Northwestern India. *Earth Planet Sp* **67,** 129 (2015).
2. Ram Bichar Singh Yadav, Jayant Nath Tripathi, Bal Krishna Rastogi, Mridul Chandra Das, Sumer Chopra (2010), "Probabilistic Assessment of Earthquake Recurrence in Northeast India and Adjoining Regions".
3. https://earthquake.usgs.gov/education
4. https://www.tandfonline.com/doi/full/10.1080/19475705.2018.1466730
5. Sieh, K., & Natawidjaja, D. (2000). Neotectonics of the Sumatran fault, Indonesia. *Journal of Geophysical Research: Solid Earth*, *105*(B12), 28295-28326.
6. Subarya, C., Chlieh, M., Prawirodirdjo, L. *et al.* Plate-boundary deformation associated with the great Sumatra–Andaman earthquake. *Nature* 440, 46–51 (2006)
7. Utsu, T. (1970). Aftershocks and earthquake statistics (1): Some parameters which characterize an aftershock sequence and their interrelations. *Journal of the Faculty of Science, Hokkaido University. Series 7, Geophysics*, *3*(3), 129-195.

Appendix

All R and Python codes, along with results can be found here