Introduction to Worldwide Earthquake Probability Distributions

Nur Anisah Mohamed[a] and Mohamad Norikmal Fazli Hisam[b]

[a,b]*Institute of Mathematical Sciences,*
Faculty of Science, Universiti Malaya,
50603, Kuala Lumpur, Malaysia.

[a]*nuranisah_mohamed@um.edu.my*
[b]*ikmalfazli23@gmail.com*

Abstract. Modelling the seismicity data is extremely difficult; hence, the assumptions on the distribution of earthquake occurrences play a crucial part in determining seismic hazard. Due to its simplicity and ease of use, the Poisson distribution has been the most common distribution for modelling earthquake data over the past year. Nevertheless, the Poisson distribution appears inefficient due to the diversity of earthquake data and the temporal correlations that are common in many real earthquake sequences. The statistical goodness-of-fit tests using worldwide seismicity data from 1921 to 2021 indicate that earthquake temporal occurrences do not always match the commonly used Poisson distribution in earthquake research. On the other hand, the Negative Binomial distribution was discovered to be a better distribution for observed earthquake magnitude distributions, and it may be applied in seismic analysis.

INTRODUCTION

One of the most significant challenges in seismicity analysis is the earthquakes cannot be predicted. It is impossible to monitor the stress–strain characteristics of the rock that are far away beneath the ground surface, where earthquake focal points are found [1]. Given the situation, a variety of empirical connections for applied seismology and earthquake engineering have been developed.

There have been a variety of empirical correlations proposed in earthquake engineering, such as those between the soil's shear modulus and plasticity index [2], damping ratio and shear strain [3], and shear-wave velocity and SPT-N value (SPT: standard penetration test). There have been more than twenty empirical associations developed using different datasets from different places for shear-wave velocity and SPT-N data alone. A series of empirical models for estimating earthquake magnitudes based on rupture length, rupture width, and other parameters is one of the most extensively cited [4]. This clearly demonstrates the need of empirical models in earthquake engineering research.

Over the year, Poisson distribution has been the most common distribution used in modelling earthquake data because of its simplicity and easy to handle. However, it appears to be ineffectual due to the variety of earthquake data and the temporal correlations that are frequent in many real earthquake sequences [5]. The distribution of earthquakes does not always match the Poisson distribution, according to several research. For example, [6] investigated a statistically significant series of earthquakes in southern California for a Poisson distribution and discovered that the data did not fit a Poisson distribution for several aftershock removal methods. Shlien & Toksoz, [7] used Poisson processes to model earthquake occurrences around the world. Except for deep focal earthquakes, goodness of fit testing rejected the simple Poisson process.

While the Poisson distribution is inefficient, multiple studies have shown that the Negative Binomial model is an appropriate model for describing earthquakes in various locations. [8] evaluated shallow focus earthquakes $(h \leq 60\ km)$ with magnitudes ranging from 4.0 to 6.0 that occurred in various high seismicity zones of the Alpide-Himalayan belt between 1954 and 1975 using Poisson and negative binomial laws. When the conventional Poisson model is no longer applicable in most of the high seismicity zones of the Alpide-Himalayan belt, Negative Binomial entries have proven to be an excellent model for describing earthquake occurrences [8]. However, for the location of Killini in Western Greece, the data yielded three states, each representing a different level of seismicity (low, medium, high). The state with the lowest seismicity has a Poisson distribution, while the other two states (medium and high) have a Negative Binomial distribution [5]. This outcome is consistent with the nature of the data. In medium and high seismicity states, the Negative Binomial distribution adds more variety within each state to the model.

In this study, we aim to investigate the distribution of earthquake occurrence in this study using real earthquake data from the USGS seismicity database. This project presents statistical goodness-of-fit tests on worldwide seismicity data from 1921 to 2021. The tests show that earthquake temporal occurrences do not always correspond to the Poisson distribution, which is commonly used in earthquake research. In contrast, the Negative Binomial distribution was discovered to be a good model for capturing observed earthquake magnitude distributions with statistical significance, and it may be used in seismic analyses that require earthquake magnitude distributions as input.

METHODOLOGY

Earthquake Data

The United States Geological Survey's (USGS) [9] seismicity database was used to extract worldwide earthquake data with magnitudes greater than 6.5. A total of 3991 occurrences occurred in 101 years between 1921 and 2021. We concentrate on earthquakes of greater magnitude because they cause more damage on the earth's surface. Earthquakes with magnitudes greater than 6.0, according to the Michigan Technological University can cause significant damage to buildings and other structures.

From 1921 to 2021, earthquake occurrence statistics are summarised in Table 1. The table shows that as the magnitude increases, the mean and variance decrease. Before examining the distribution of earthquake occurrences, all data was sorted by magnitude. Figure 1 depicts the observed frequencies in years versus the number of earthquakes with magnitudes greater than 6.5 from 1921 to 2021. According to the data, there were 24 earthquakes with magnitudes greater than 6.5 between 1958 and 1924.

.

Table 1 Summary statistics of Earthquake Occurrence from 1921 to 2021.

Magnitude	Min	1st Qu.	Median	Mean	3rd Qu.	Max	StdDev.	Variance
6.5	18	32	39	39.51	46	66	10.58	111.91
6.6	15	24	31	31.16	38	53	8.37	70.03
6.7	12	19	25	25.18	30	42	6.77	45.87
6.8	7	16	19	20.19	24	19	5.65	31.95
6.9	6	13	15	16.14	19	30	4.85	23.48
7	5	10	12	12.97	16	24	4.06	16.49
7.1	4	8	9	10.05	13	19	3.60	12.93
7.2	2	6	8	7.93	10	16	3.05	9.29
7.3	1	4	6	6.20	8	14	2.76	7.60
7.4	0	3	5	4.97	7	12	2.50	6.25
7.5	0	2	4	4.10	6	11	2.17	4.69
7.6	0	2	3	3.23	5	10	2.02	4.10
7.7	0	1	2	2.43	4	9	1.80	3.25
7.8	0	1	1	1.74	3	6	1.38	1.91
7.9	0	0	1	1.15	2	5	1.04	1.09
8	0	0	1	0.83	1	4	0.88	0.78
8.1	0	0	0	0.60	1	3	0.79	0.62
8.2	0	0	0	0.37	1	2	0.62	0.38

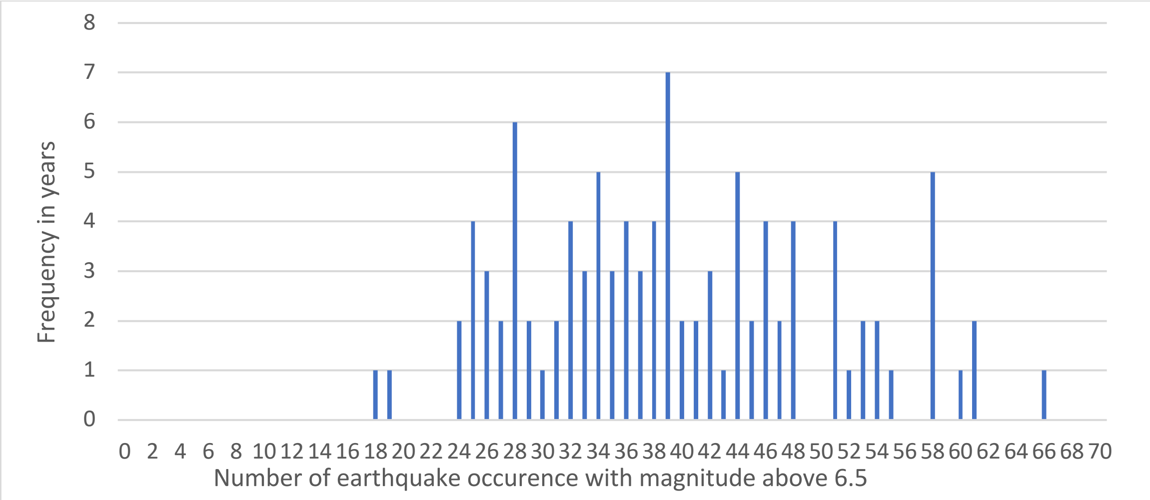

Figure 1 Number of earthquakes occurrence in 101 years

Poisson and Negative Binomial distributions

According to the textbook [10], the probability distribution function of the Poisson distribution is as follows:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0,1,2 \dots \qquad [1]$$

where λ is a Poisson parameter which is the mean annual rate. The common use of this distribution in estimating earthquake occurrence in each period is because it's usually recommended as its suitability in stimulating the temporal occurrence of rare events.

The Poisson Distribution was developed from the Binomial Distribution, and for the Binomial Distribution to hold, the probability, p, of an event E must be constant for all occurrences of the events that make up its context. However, another distribution known as the Negative Binomial Distribution (NB) may offer an even closer "match" if it is known that p is not constant in its context-events [11].

The NB distribution is a sequence of independent Bernoulli trials until exactly r success occurs, assuming that r is a fixed integer [10]. Let a random variable, X be the number of trials needed to observe r^{th} success. Thus, the probability mass function (pmf) of X is the probability of obtaining exactly $r-1$ success in the first $x-1$ trials and the probability of success, p on the r^{th} trial is given by

$$g(x) = \binom{x-1}{r-1} p^{r-1} q^{x-r}, x = r, r+1, \dots. \qquad [2]$$

The main justification for NB distribution is by using Maclaurin's series expansion. Consider $h(w) = (1-w)^{-r}$, where the binomial $(1-w)$ with the negative exponent $-r$. Suppose

$$(1-w)^{-r} = \sum_{k=0}^{\infty} \frac{h^{(k)}(0)}{k!} w^k = \sum_{k=0}^{\infty} \binom{r+k-1}{r-1} w^k, \quad -1 < w < 1 \qquad [3]$$

Let $x = k + r$, then $k = x - r$, and

$$(1-w)^{-r} = \sum_{k=0}^{\infty} \binom{r+x-r-1}{r-1} w^{x-r} = \sum_{k=0}^{\infty} \binom{x-1}{r-1} w^{x-r}, \qquad [4]$$

If we assume $w = q$, then the sum of the probabilities in equation [2] is 1 because,

$$\sum_{x=r}^{\infty} g(x) = \sum_{k=0}^{\infty} \binom{x-1}{r-1} p^r q^{x-r} = p^r \sum_{k=0}^{\infty} \binom{x-1}{r-1} q^{x-r} = p^r (1-q)^{-r} \quad [5]$$

We determine the mean and variance of the NB distribution using the moment generating functions (mgf). It is

$$M(t) = \sum_{x=r}^{\infty} e^{tx} \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

$$= (pe^t)^r \sum_{x=r}^{\infty} \binom{x-1}{r-1} [(1-p)e^t]^{x-r}$$

$$= \frac{(pe^t)^r}{[1-(1-p)e^t]^r}, \text{where } (1-p)e^t < 1 \quad [6]$$

Consider when $t < -\ln(1-p)$. Therefore,

$$M'(t) = (pe^t)^r(-r)[1-(1-p)e^t]^{-r-1}[-(1-p)e^t] + r(pe^t)^{r-1}(pe^t)[1-(1-p)e^t]^{-r}$$

$$= r(pe^t)^r[1-(1-p)e^t]^{-r-1} \quad [7]$$

and

$$M''(t) = r(pe^t)^r(-r-1)[1-(1-p)e^t]^{-r-2}[-(1-p)e^t] + r^2(pe^t)^{r-1}(pe^t)[1-(1-p)e^t]^{-r-1} \quad [8]$$

Hence,

$$M'(0) = rp^r p^{-r-1} = rp^{-1}$$

and

$$M''(0) = rp^{-2}[(1-p)(r+1) + rp] = rp^{-2}(r+1-p)$$

Thus, the mean

$$E(X) = \mu = M'(0) = \frac{r}{p}; \quad [9]$$

and the variance

$$Var\,(X) = M''(0) - [M'(0)]^2 = rp^{-2}(r+1-p) - \frac{r^2}{p^2} = \frac{r[1-p]}{p^2}. \quad [10]$$

Chi-square goodness-of-fit test

The goodness-of-fit test is a statistical test that determines how well sample data fits a population's normal distribution. According to the Central Limit Theorem, for a large sample size, n with a particular distribution approximates a normal distribution with zero mean and one variance [10]. The disparity between actual values and those expected of the model in a normal distribution instance is determined by goodness-of-fit. The Chi-square value is calculated as follows:

$$\chi^2_{(n-1)} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad [11]$$

Where χ^2 denoted the Chi-square value, $k = (n-1)$ is the degree of freedom, O_i is the observed frequency and E_i is the fitted/expected frequencies. This Chi-square value (χ^2) compares the total difference between observed (O_i) and "fitted" (E_i) frequencies. The critical Chi-square value corresponding to a specified level of significance (commonly at 5%); given the total difference smaller than critical value, the statistical inference that the selected model is acceptable. The null hypothesis is not rejected if the expected values are close to the actual values, whereas rejection indicates the opposite.

RESULTS AND DISCUSSIONS

The Poisson Distribution

Based on the worldwide seismicity in a period of 101 years, Figure 2 shows the observed and expected frequencies in terms of the numbers of years in which x earthquakes, $0 < x < 70$, above magnitude 6.5 had occurred. As for example from the data, 24 earthquakes with magnitude above 6.5 occurred only two years which is year 1958 and 1924. The expected frequencies in Fig. 4 were calculated based on 3991 samples from USGS database with sample mean annual rate = 39.5 per year.

Table 2 summarised the Chi-square goodness-of-fit tests on the temporal occurrence of different magnitude of earthquakes. From the table, the Poisson distribution is rejected at the magnitude 6.5, 6.6, 6.7, 6.8 and 7.7. Hence, we can say that Poisson Distribution is unsuitable for magnitude below than the magnitude 7. In the next section, we will review and compare model from Negative Binomial distribution and focus on those magnitude.

Table 2 Summary of the 18 Chi-square tests on using the Poisson distribution to model the earthquake's temporal occurrence

Magnitude of Exceedance	Annual rate	DF	Chi square	critical value	Model accepted/rejected
6.5	39.5	70	1003.1	90.5	Rejected
6.6	31.2	60	239.7	79.1	Rejected
6.7	25.2	50	121.7	67.5	Rejected
6.8	20.2	40	99.6	55.8	Rejected
6.9	16.1	40	42.7	55.8	Accepted
7	13.0	30	17.1	43.8	Accepted
7.1	10.0	19	24.1	30.1	Accepted
7.2	7.9	16	20.2	26.3	Accepted
7.3	6.2	14	16.1	23.7	Accepted
7.4	5.0	13	16.4	22.4	Accepted
7.5	4.1	11	13.4	19.7	Accepted
7.6	3.2	10	12.3	18.3	Accepted
7.7	2.4	9	19.5	16.9	Rejected
7.8	1.7	6	3.1	12.6	Accepted
7.9	1.1	5	2.6	11.1	Accepted
8	0.8	4	0.6	9.5	Accepted
8.1	0.6	3	1.5	7.8	Accepted
8.2	0.4	2	0.8	6.0	Accepted

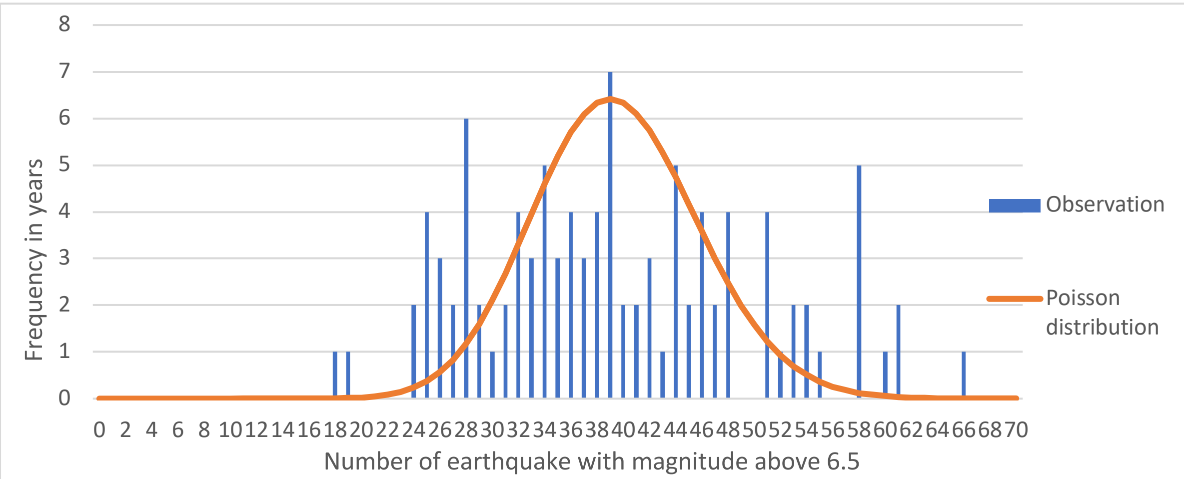

Figure 2 Observed and Expected earthquake frequencies for magnitude above 6.5

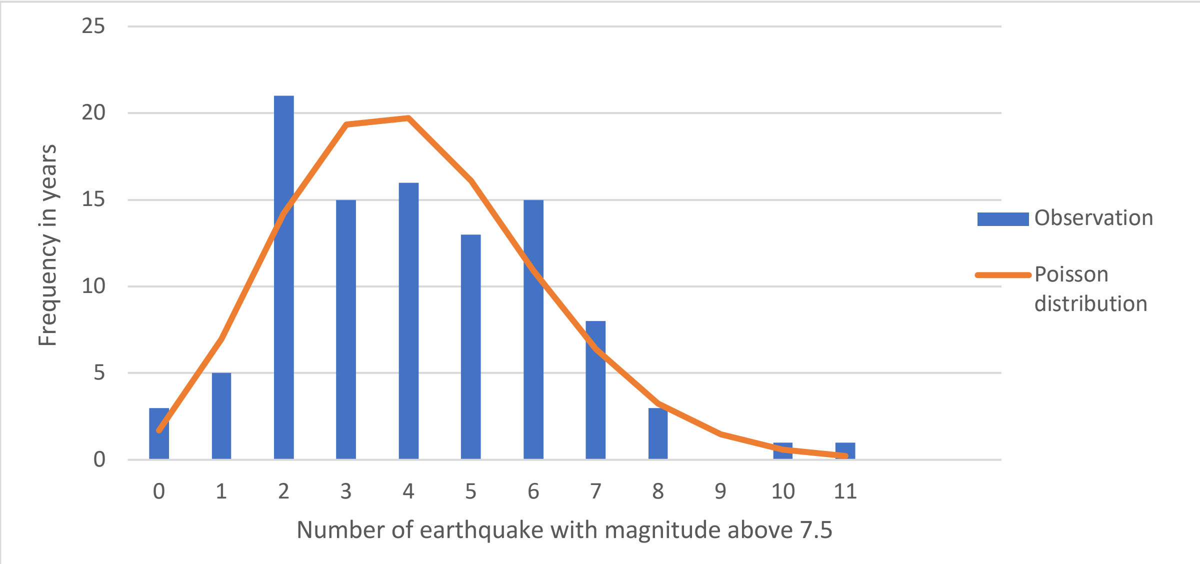

Figure 3 Observed and Expected earthquake frequencies for magnitude above 7.5

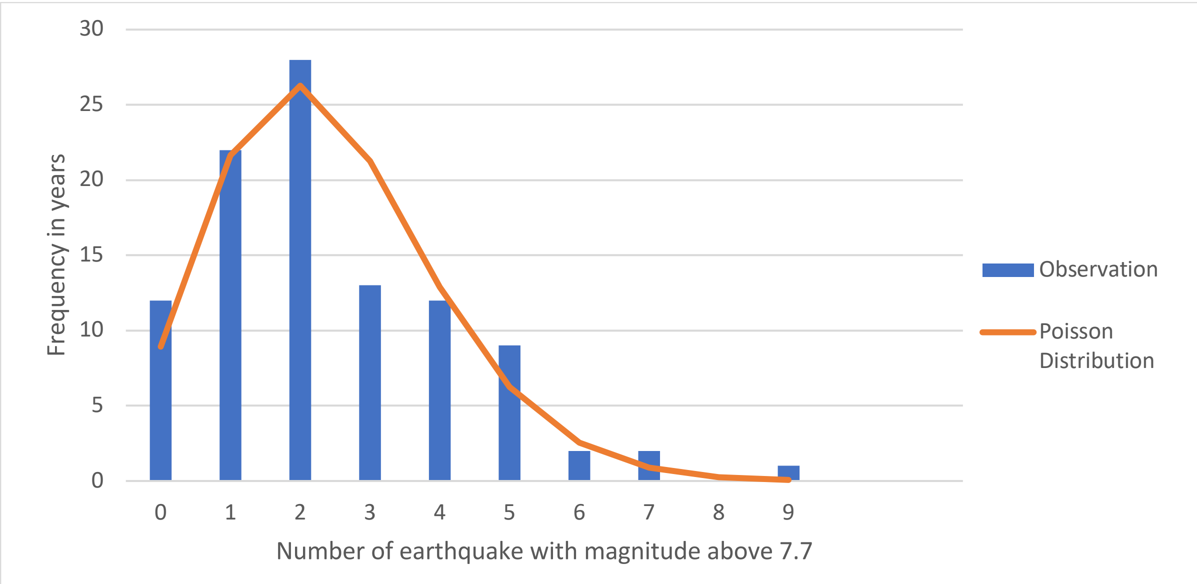

Figure 4 Observed and Expected earthquake frequencies for magnitude above 7.7

In Figure 2, it is certain that Poisson model will be rejected as there is a huge disparity between model and observation. As we calculated the Chi-square value and compared to the critical value, we observed that the critical value is 1002.1 which is much larger than the critical value of 90.5 at level of significant 5%,. Hence, the statistical inference is the model was rejected for modelling the temporal occurrence of earthquake with magnitude above 6.5.

In comparison, Figure 3 shows the comparison between model and observation for earthquakes with magnitude above 7.5. With a total difference = 13.4 which was smaller than the critical value = 19.7, the statistical inference is that using the Poisson distribution to model the temporal occurrence of earthquakes with magnitude above 7.5 (with mean annual rate = 4.7) is accepted. However, for magnitude above 7.7 (Figure 4), the total difference between model and observation is equal to 19.5 which seems to be larger than critical value equal to 13.4 (level of significant 5%) inferring that the model is to be rejected.

The Negative Binomial Distribution

Since the Poisson distribution is statistically rejected for modelling earthquakes occurrence for a few magnitudes, our new objectives now are to improve the distribution of earthquake by referring to the goodness-of-fit tests. Specifically, we propose the use of Negative Binomial distribution due to its ability to handle over dispersed data more effective. We parameterize this distribution in terms of the arithmetic mean (m) and the measure of the heterogeneity of the distribution, k.

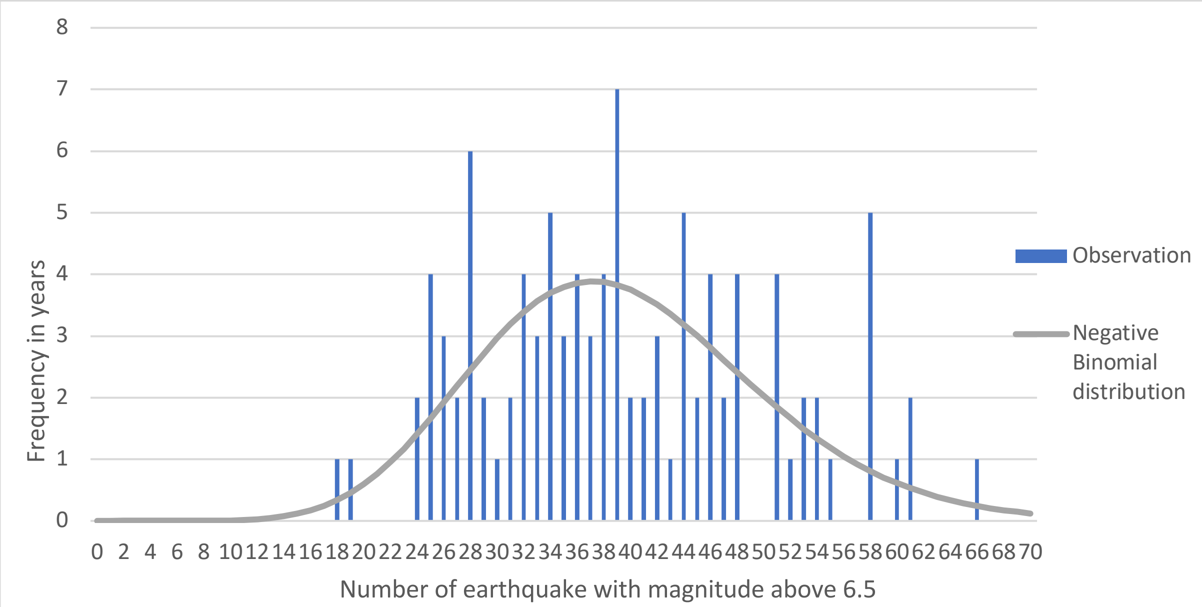

Figure 5 Observed and Expected earthquake frequencies for magnitude above 6.5

Figure 5 shows the result of using the Negative Binomial distribution to model the earthquake data for magnitude above 6.5. From Figure 5, clearly that the goodness-of-fit test between model and observation was improved over the Poisson distribution (Figure 2). Based on the Chi-square test, the model is accepted as the total difference is equal to 68.2 which is lower than the critical value = 90.5 at 5% level of significance.

Next, let us try applying this distribution to a bigger magnitude = 7.7. As we can see from Figure 6, the model seems to be accepted as the total difference is equal to 6.2 which is lower than the critical value = 16.9 at 5% level of significance.

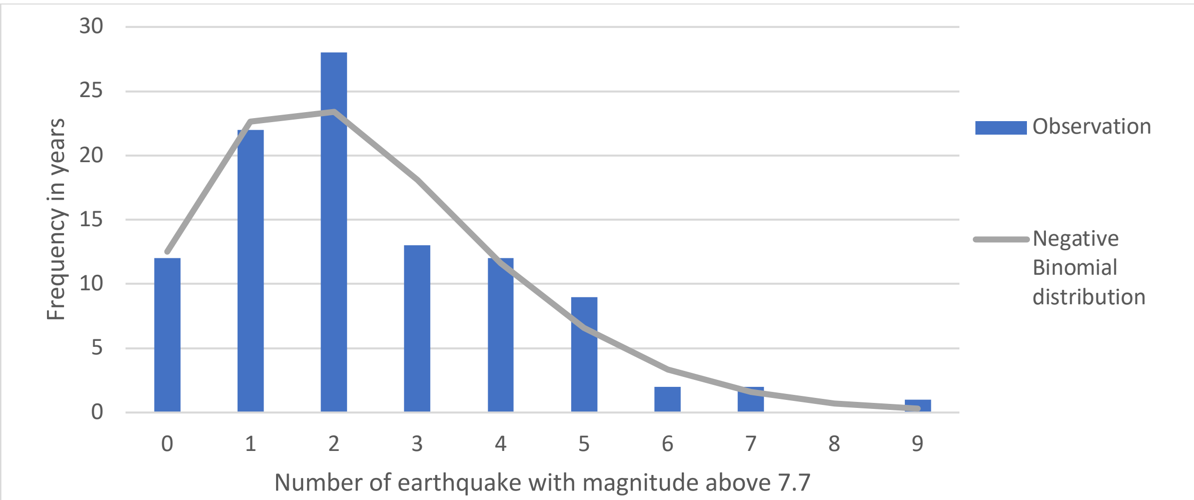

Figure 6 Observed and Expected earthquake frequencies for magnitude above 7.7

Table 3 Summary of the 18 Chi-square tests on using the NB distribution to model the earthquake's temporal occurrence in a given period of time

Magnitude of Exceedance	Annual rate	DF	K-Parameter	Chi square	Critical value	Model accepted/rejected
6.5	39.5	70	21.6	68.2	90.5	Accepted
6.6	31.2	60	25.0	40.5	79.1	Accepted
6.7	25.2	50	30.6	36.0	67.5	Accepted
6.8	20.2	40	34.6	32.0	55.8	Accepted
6.9	16.1	40	35.5	20.5	55.8	Accepted
7.0	13.0	30	47.8	10.7	43.8	Accepted
7.1	10.0	19	35.1	16.6	30.1	Accepted
7.2	7.9	16	46.4	15.6	26.3	Accepted
7.3	6.2	14	27.4	9.7	23.7	Accepted
7.4	5.0	13	22.7	11.6	22.4	Accepted
7.5	4.1	11	27.1	10.6	19.7	Accepted
7.6	3.2	10	12.0	5.7	18.3	Accepted
7.7	2.4	9	7.2	6.2	16.9	Accepted
7.8	1.7	6	17.8	2.3	12.6	Accepted
7.9	1.1	5	-21.7	2.7	11.1	Accepted
8.0	0.8	4	-13.8	0.5	9.5	Accepted
8.1	0.6	3	20.7	1.3	7.8	Accepted
8.2	0.4	2	-22.1	0.9	6.0	Accepted

Table 3 summarises the Chi-square goodness-of-fit tests on the temporal occurrence of earthquakes of various magnitudes. According to the table, the Negative Binomial distribution was acceptable at all magnitudes, indicating a better fit to this data.

Comparison between the Negative Binomial distribution with Poisson distribution

In this section, we will compare the Negative Binomial (NB) and Poisson distributions for earthquake occurrence over 101 years to the actual observed value. Figure 7 depicts a comparison of two different models and observed frequency for magnitudes greater than 6.5. Clearly, the NB distribution has a wider spread than the Poisson distribution, and the graph has a lower disparity when compared to the Poisson. As in previous tests, the NB distribution outperforms the Poisson approach in terms of goodness-of-fit.

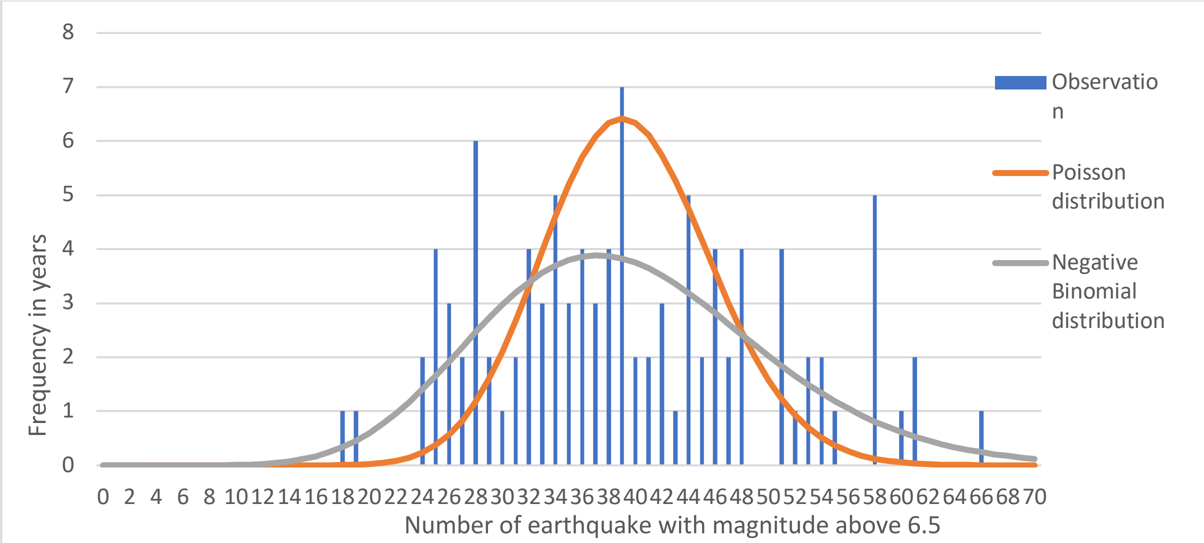

Figure 7 Observed and Expected earthquake frequencies for magnitude above 6.5

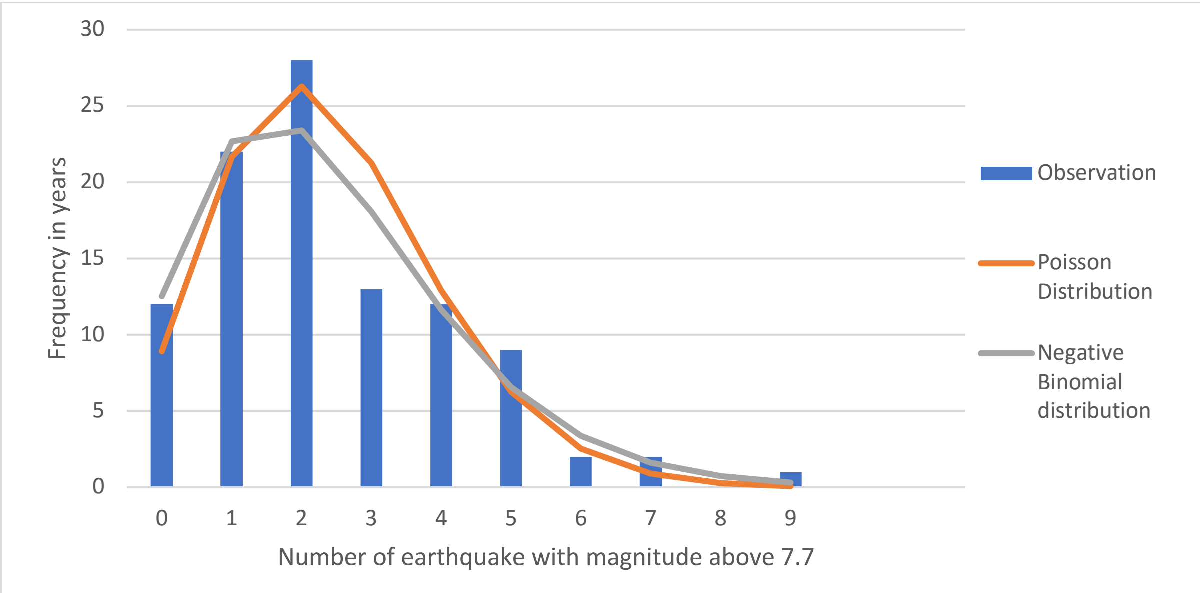

Figure 8 Observed and Expected earthquake frequencies for magnitude above 7.7

Similarly, Figure 8 show another analysis using magnitude above 7.7, with the same conclusion – NB accepted while Poisson rejected – reached. Therefore, the inference from this statistical study is summarized as follows: The Negative Binomial distribution can well describe the earthquake occurrences, with sufficient statistical significance for the hypothesis – earthquake occurrences distribution follows the Negative Binomial distribution – not being rejected by Chi-square tests. By contrast, using Poisson distribution few magnitudes seem to reject this distribution by Chi-square test. More details of the Chi-square test are given in Table 4.

Table 4 Summary of the 18 Chi-square tests on using the Poisson and NB distribution to model the earthquake's temporal occurrence in a given period of time

Magnitude of Exceedance	Annual rate	DF	Poisson distribution			Negative Binomial distribution			
			Chi square	critical value	Model accepted/ rejected	K-Parameter	Chi square	critical value	Model accepted/ rejected
6.5	39.5	70	1003.1	90.5	Rejected	21.6	68.2	90.5	Accepted
6.6	31.2	60	239.7	79.1	Rejected	25.0	40.5	79.1	Accepted
6.7	25.2	50	121.7	67.5	Rejected	30.6	36.0	67.5	Accepted
6.8	20.2	40	99.6	55.8	Rejected	34.6	32.0	55.8	Accepted
7.7	2.4	9	19.5	16.9	Rejected	7.2	6.2	16.9	Accepted

CONCLUSIONS

Although it has been used in numerous studies, the Poisson distribution is not always adequate for simulating an earthquake's temporal occurrences over a specific time period. The Poisson distribution, on the other hand, is only appropriate for modelling powerful, infrequent earthquakes. [12] on his studies demonstrate the negative binomial distribution's parity with other models, however due to the sample size limitations, no general inferences can be made. In this paper, the observed distribution of earthquake occurrences is well described by the Negative Binomial distribution. As a result, for modelling earthquake occurrence distributions, the Negative Binomial distribution is a more statistically significant alternative to the Poisson distribution. As a result, it is strongly advised that the Negative Binomial assumption be used to model the high earthquake distributions.

REFERENCES

[1] R. J. Geller, D. D. Jackson, Y. Y. Kagan and F. Mulargia, "Earthquakes cannot be predicted," *Science 275,* p. 1616, 1997.

[2] B. O. Hardin and V. P. Drnevich, "Shear modulus and damping in soil: design equations and curves," *Soil Mechanics and Foundations Division,* pp. 667-692, July 1972.

[3] H. B. Seed and I. M. Idriss, "Soil Moduli and Damping Factors for Dynamic Response Analyses," Berkeley, 1970.

[4] Coppersmith, D. L. Wells and J. Kevin, "New Empirical Relationships among Magnitude, Rupture Length,," *Bulletin of the Seismological Society of America,* vol. 84, no. 4, pp. 974-1002, August 1994.

[5] K. Orfanogiannaki and D. Karlis, "Modeling earthquake numbers by Negative Binomial Hidden Markov models," in *EGU General Assembly*, 2020.

[6] L. Knopoff, "The statistics of earthquakes in Southern California," *Bulletin of the Seismological Society of America,* pp. 1871-1873, 1964.

[7] S. Shlien and M. N. Toksöz, "A clustering model for earthquake occurrences," *Bulletin of the Seismological Society of America,* p. 1765–1787, 1970.

[8] N. M. Rao and K. L. Kaila, Application of the negative binomial to earthquake, Uppal Road, Hyderabad 500 007, India: National Geophysical Research Institue, 1985.

[9] 1 1 2022. [Online]. Available: https://earthquake.usgs.gov/earthquakes/search/ .

[10] R. V. Hogg, E. A. Tanis and D. L. Zimmerman, Probability and Statistical Inference, United States of America: Pearson Education, 2015.

[11] G. Bazigos, "Appendix I: The Negative Binomial Distribution," in *A Manual on Acoustic Surveys: Sampling Methods for Acoustic Surveys*, Rome, 1981.

[12] A. Tejedor, J. B. Gomez and A. F. Pacheco, "The Negative Binomial Distribution as a Renewal Model for the Recurrence," *Pure and Applied Geophysics,* 2014.

[13] Baker and W. Jack, Introduction to Probabilistic Seismic Hazard Analysis., 2015.

[14] M. H. Wu, J. P. Wang and K. W. Ku, "Earthquake, Poisson and Weibull distributions," Elsevier B.V., Taiwan, 2019.