Kolmogorov-Smirnov Goodness-of-Fit test for identifying distribution of the number of earthquakes and the losses due to earthquakes in Indonesia

Pratya P Oktaviana¹ and Irhamah²
¹ Department of Actuarial Science, Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
² Department of Statistics, Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia

E-mail: paramitha.oktaviana@gmail.com

Abstract. Earthquakes of high magnitude often result in large casualties and huge economic losses in Indonesia. An earthquake risk assessment must be carried out. The risk assessment process requires a structured approach. This paper is containing of the first step of risk assessment using actuarial-statistical method. The data used as research variables in this paper are the number of earthquakes and the losses due to earthquakes. The earthquakes data used are earthquakes with magnitude ≥ 5 Richter Scale. Data is collected within the last 40 years, from 1980 to 2019. Data is collected from Indonesian Disaster Data and Information, and NOAA National Centers for Environmental Information (NCEI). Descriptive statistics analysis obtains that the most frequently earthquakes with magnitude ≥ 5 Richter Scale occurred in Indonesia in 2004, where as it is known that at the end of 2004 an earthquake with a magnitude of 9.1 Richter Scale occurred in Aceh which caused the highest loss, which was around 147485.500 billion rupiah. Normality test of the variables results that the number of earthquakes and the losses due to earthquakes are not normally distributed. Kolmogorov-Smirnov Goodness-of-Fit test results that the number of earthquakes follows Gamma Distribution with shape parameter $\alpha = 1.800$ and scale parameter $\beta = 2.236$; while the losses due to earthquakes follows Power Function Distribution with shape parameter $\alpha = 0.072$, and boundary parameter $a = 0$, $b = 2.810 \times 10^5$.

1. Introduction
In Indonesia, there is a database called Indonesian Disaster Data and Information. This database was created by the National Disaster Management Agency in Indonesia. This database is an example of the application of technology that has been adapted for local data and indicators, and has been used in risk mapping and risk indexing of disasters that have occurred in Indonesia [1]. There is other site that provides data about natural disasters in the world, namely NOAA National Centers for Environmental Information (NCEI). NOAA NCEI is responsible for monitoring, assessing and providing public access to geophysical data and information.

This paper is containing of the first step of risk assessment using actuarial-statistical method. The data used as research variables in this paper are the number of earthquakes and the losses due to earthquakes. The earthquakes data used is earthquakes with magnitude ≥ 5 Richter Scale. Data is
collected within the last 40 years, from 1980 to 2019. Data is collected from Indonesian Disaster Data and Information, and NOAA National Centers for Environmental Information (NCEI).

The analysis that have been carried out include: the first is to conduct descriptive statistical analysis and test data normality from the data that has been obtained to describe how the Indonesian earthquake disaster. The data that has been pre-processed is then used for the second stage of analysis, namely the selection of distribution for each variable \( N \) (number of earthquakes) and \( X \) (losses caused by earthquakes-billion Rupiah). The selection of this distribution is done through testing the distribution of data using the One Sample Kolmogorov-Smirnov (KS) test. The results of the second stage are then used for the third stage, namely making a histogram from a Probability Distribution Function of variables \( N \) (number of earthquakes) and \( X \) (losses). This histogram is used to show more clearly how many earthquakes and losses due to earthquakes are in diagram form according to the data distribution.

2. Literature review

2.1 Kolmogorov-Smirnov Goodness-of-Fit test

The Kolmogorov-Smirnov one-sample distribution test or it may be called the Kolmogorov-Smirnov goodness-of-fit test usually involves examining a random sample of several unknown distributions to test the null hypothesis that the known distribution is actually a known and determined function. The Kolmogorov-Smirnov test [2] is statistics test used to decide if a sample comes from a population with a specific distribution. Given \( N \) ordered data \( X_1, X_2, \ldots, X_N \), the empirical distribution function is defined as:

\[
E_N = n(i)/N
\]  

(1)

\( n(i) \) is the number of points less than \( X_i \) and the \( X_i \) are sorted to the largest value.

The Kolmogorov-Smirnov test is defined by:

\( H_0 \): The data follow a specified distribution

\( H_1 \): The data do not follow the specified distribution

Test Statistic: The Kolmogorov-Smirnov test statistic is defined as

\[
D = \max_{i} \left( F(X_i) - \frac{i-1}{N}, \frac{i}{N} - F(X_i) \right)
\]  

(2)

\( F \) is the theoretical cumulative distribution of the distribution. \( H_0 \) is rejected if \( D > d \) (Kolmogorov-Smirnov Table)

2.2 Gamma distribution

Gamma distribution in modeling is used to predict the waiting time until a future event occurs. That is, by finding that the variable number of earthquakes that occur is a gamma distribution, it can be described as follows:

\[
f(n) = \frac{n^{(\alpha - 1)}}{\beta^\alpha \Gamma(\alpha)} \exp(-n/\beta)
\]  

(3)

where the parameters \( \alpha \) and \( \beta \) satisfy \( \alpha > 0, \beta > 0 \). The standard gamma distribution has \( \beta = 1 \) [3].

2.3 Power Function distribution

The Power Function distribution is a special form of the Uniform distribution. Also Power Function is special form of beta distribution and inverse Gaussian distributions [4]. This distribution is very suitable to describe the lifetime distribution of the failure dataset. By finding that the distribution of the loss variable follows this distribution, it can be written as follows:

\[
f(x) = \frac{\alpha(x-a)^{\alpha-1}}{(b-a)^\alpha}
\]  

(4)
The Power Function distribution is flexible as it is able to model the various types of data. It is usually used for the reliability analysis, lifetime and income distribution data. Likewise, many probability models are also used to assess patterns of income distribution but these models are mathematically more complex to handle [5].

3. Data and research variables
The data used as research variables in this paper are the number of earthquakes and the losses due to earthquakes. The earthquakes data used are earthquakes with magnitude ≥ 5 Richter Scale. The research variables are shown in Table 1.

| Variable | Explanation | Scale |
|----------|-------------|-------|
| N        | The number of earthquakes ≥ 5 Richter Scale | Ratio |
| X        | The losses due to earthquakes ≥ 5 Richter Scale (Billion Rupiah) | Ratio |

Data structure of this research is shown in Table 2.

| Observation | N   | X   |
|-------------|-----|-----|
| 1           | N₁  | X₁  |
| 2           | N₂  | X₂  |
| 3           | N₃  | X₃  |
| 4           | N₄  | X₄  |
| 5           | N₅  | X₅  |
| 6           | N₆  | X₆  |
| ...         | ... | ... |
| 40          | N₄₀ | X₄₀ |

4. Result and discussion
4.1 Descriptive statistics
Indonesia should face the risk of earthquakes of high magnitude often resulting in heavy casualties and huge economic losses. This descriptive statistics analysis shows the visualization of the variables, which are the number of earthquake magnitude ≥ 5 Richter scale and losses due to earthquakes magnitude ≥ 5 Richter scale. The data visualization is presented in Figure 1 and Figure 2.

Figure 1 shows that the most frequently earthquakes with magnitude ≥ 5 Richter scale occurred in Indonesia in 2004, where as it is known that at the end of 2004 an earthquake with a magnitude of 9.1 Richter scale occurred in Aceh. In 1988 and 1993 the graph shows 0 events, which does not mean that there were no earthquakes in that year, but it means that that year there were no earthquakes with a magnitude ≥ 5 Richter scale. Based on Figure 2, it can be seen that the range of losses due to earthquakes in Indonesia is increasing from year to year from 1980 to 2004, then after that the range of losses decreases but remains high and the last is 2018 which causes the highest losses. In 2018 it was known that earthquakes occurred in Lombok (6.9 Richter scale) and Sulawesi (7.5 Richter scale) which caused high losses.

From Figure 1 and Figure 2, it can conclude that the data of the number of earthquake magnitude ≥ 5 Richter scale and losses due to earthquakes magnitude ≥ 5 Richter scale are varieties. Some of the data are too high but also some other is not.
4.2 Normality test

Identifying that variable $X$ and $N$ are normally distributed or not was performed by using Kolmogorov-Smirnov Normality Test. The results are shown in Figure 3, and then clarified in the next Table 3:

Figure 1. Graph of the number of earthquakes magnitude $\geq 5$ richter scale in Indonesia.

Figure 2. Graph of the losses due to earthquakes $\geq 5$ richter scale in Indonesia.

Figure 3. Normal probability plot of: (a) the number of earthquakes $\geq 5$ richter scale, (b) the losses due to earthquakes $\geq 5$ richter scale in Indonesia.
Table 3. Normality test using kolmogorov-smirnov of research variables.

| Variable                                      | P-Value |
|-----------------------------------------------|---------|
| The Number of Earthquakes ≥ 5 Richter Scale   | 0.014   |
| The Losses Due To Earthquakes ≥ 5 Richter Scale | <0.010 |

The results in Figure 3 and Table 3 show that the variable number of earthquakes and losses due to earthquakes ≥ 5 Richter Scale does not follow the normal distribution because the p-value of both variables is smaller than α (0.05). Because the distribution of the two variables does not follow the Normal distribution, the next step is to test the distribution of each of these variables in order to find a suitable distribution.

4.3 Kolmogorov-Smirnov Goodness-of-Fit test

Distribution testing by using the Kolmogorov-Smirnov Goodness-of-fit test for each variable is presented in Table 4 and the Probability Density Function of each distribution variable is made in Figure 4 as follows:

Table 4. Goodness-of-Fit test of variables.

| Variable                                      | Distribution | Parameter | Kolmogorov-Smirnov Statistics | P-Value |
|-----------------------------------------------|--------------|-----------|-------------------------------|---------|
| The Number of Earthquakes ≥ 5 Richter Scale   | Gamma        | α = 1.800 | 0.118                         | 0.587   |
|                                               |              | β = 2.236 |                               |         |
| The Losses Due To Earthquakes ≥ 5 Richter Scale | Power       | α = 0.072 | 0.179                         | 0.137   |
|                                               |              | α = 0.000 |                               |         |
|                                               |              | b = 2.810 × 10^3 |                   |         |

Figure 4. Probability density function plot: (a) the number of earthquakes ≥ 5 richter scale, (b) the losses due to earthquakes ≥ 5 richter scale in Indonesia.
In testing this distribution, the significance level $\alpha = 0.05$ was used. The $p$-value of each test distribution variable, compared with the $\alpha$ value. In Table 4, it is shown that the $p$-value is 0.587 and 0.137, respectively, both of which are greater than the $\alpha$ value. So it can be decided Failed to Reject $H_0$, that is, the data follows the selected distribution. By referring to these results, it can be concluded that the variable number of earthquakes $\geq 5$ Richter Scale follows the Gamma distribution, with the shape parameters $\alpha = 1.800$ and scale $\beta = 2.236$; whereas for the variable Loss due to the Earthquake $\geq 5$ Richter Scale, it follows the Power Function distribution, with the continuous shape parameter $\alpha = 0.072$, and the continuous boundary parameter $a = 0$, $b = 2.810 \times 10^5$. Figure 4 shows each -Each histogram from the PDF of the distribution of selected variables for each.

Gamma distribution in modeling is used to predict the waiting time until a future event occurs. That is, by finding that the variable number of earthquakes that occur is a gamma distribution, it can be described as follows:

\[ f(n) = \frac{n^{(\alpha-1)} \exp(-n/\beta)}{\beta^n \Gamma(\alpha)} \]

\[ f(n) = \frac{n^{0.8}}{2.236^{1.8} \Gamma(1.8)} \exp(-n/2.236) \]

\[ f(n) = \frac{n^{0.8}}{3.964} \exp(-n/2.236), \text{ where } 0 < n < \infty \]

Based on the $f(n)$ graph, it can be concluded that the highest probability of an earthquake occurring in Indonesia is when $n$ is between 2 and 3. Looking at the graph that peaks and then slopes, it is concluded that the waiting time for an earthquake $\geq 5$ Richter Scale in the future does not take time. This is in sync with the condition of Indonesia, where earthquakes occur almost every year. This shows that the gamma distribution is right to choose.

The Power Function distribution is a special form of the Uniform distribution. This distribution is very suitable to describe the lifetime distribution of the failure dataset. By finding that the distribution of the loss variable follows this distribution, it can be written as follows:

\[ f(x) = \frac{\alpha(x-a)^{\alpha-1}}{(b-a)^\alpha} \]
\[ f(x) = \frac{0.072(x)^{0.9281}}{(2.810 \times 10^5)^{0.072}}, \text{ where } a \leq x \leq b \text{ or } 0 \leq x \leq 2.810 \times 10^5 \]

**Figure 6.** Plot of \( f(x) \).

Based on the \( f(x) \) graph, it is found that a very large loss (> 200000 billion rupiah) will have less chance, while the highest chance is in the loss between around 0-5000 billion rupiah.

**5. Conclusion**

Descriptive statistics analysis obtains that the most frequently earthquakes with magnitude \( \geq 5 \) Richter scale occurred in Indonesia in 2004, where as it is known that at the end of 2004 an earthquake with a magnitude of 9.1 Richter scale occurred in Aceh which caused the highest loss, which was around 147485.500 billion rupiah. Normality test of the variables results that the number of earthquakes and the losses due to earthquakes are not normally distributed. Kolmogorov-Smirnov Goodness-of-Fit test results that the number of earthquakes follows Gamma Distribution with shape parameter \( \alpha = 1.800 \) and scale parameter \( \beta = 2.236 \); while the losses due to earthquakes follows Power Function Distribution with shape parameter \( \alpha = 0.072 \), and boundary parameter \( a = 0 \), \( b = 2.810 \times 10^5 \).

**Acknowledgments**

We are grateful to Directorate of Research and Community Service, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia, for supporting this research.

**References**

[1] Wibowo A, Surbakti I and Yunus R 2013 Presentation file: Indonesia Disaster Database, Data Informasi Bencana Indonesia (DIBI) http://dibi.bnpb.go.id/

[2] Chakravarti, Laha and Roy 1967 Handbook of Methods of Applied Statistics Vol.1 John Wiley and Sons 392-394 (in Engineering Statistics Handbook)

[3] Devore J L 2012 Probability & Statistics for Engineering and the Sciences – Eighth Edition Brooks/Cole Cengage Learning USA

[4] Haq M A u, Usman R M, Bursa N and Ozel G 2018 McDonald Power Function Distribution with Theory and Applications International Journal of Statistics and Economics Vol.19 Issue Number:2

[5] Shahzad M N, Asghar Z, Shehzad F and Shahzadi M 2015 Parameter Estimation of Power Function Distribution with TL-moments Revista Colombiana de Estadistica Vol. 38 Issue 2 321-334