Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 9, Number 2, 2025
DOI: 10.30829/zero.v9i2.26257
Page: 662-670

Distribution Models of Claim Frequency and Claim Severity in Determining the Pure Premium of Car Insurance with the Application of a Deductible

[1] **Tohap Manurung**
Universitas Sam Ratulangi, North Sulawesi, 95115, Indonesia

[2] **Rillya Arundaa**
Universitas Sam Ratulangi, North Sulawesi, 95115, Indonesia

[3] **Eliasta Ketaren**
Universitas Sam Ratulangi, North Sulawesi, 95115, Indonesia

Article Info

Article history:

Accepted, 15 November 2025

Keywords:

Car Insurance;
Claim Frequency;
Claim Severity;
Deductible;
Pure Premium.

ABSTRACT

The objective of this study is to determine the pure premium value based on a car damage claim data model from car insurance company X, using data that applies a deductible value. The data used comprises car damage claims with deductibles applied during a year period. Determining the distribution model for insurance claims is one of the relevant techniques for measuring operational risk in insurance companies. In this context, historical claim data is tested against existing distribution models, enabling the calculation of pure premium values for the insurance company. The results show that the claim frequency data follows a Negative Binomial distribution with an expected value of E(N) = 0.0107, and the claim severity data follows a Log-logistic distribution with E(X) = 12,037,950. Therefore, the calculated pure premium value is E(S) = Rp129,205.19. The pure premium obtained serves as the basis for determining the actual premium charged to policyholders, with the addition of loadings.

Corresponding Author:

Tohap Manurung,
Mathematics Department
Universitas Sam Ratulangi, North Sulawesi, Indonesia
Email: tohapm@unsrat.ac.id

1. INTRODUCTION

The use of motor vehicles in Indonesia continues to increase due to the growing need for safe and comfortable transportation. This rise in vehicle numbers also increases the associated risks. Risk refers to uncertainty about future events that may result in desirable or undesirable outcomes [1]. It is essential to anticipate such risks to minimize losses, and one form of risk management is insurance [2]. According to the Commercial Code (Kitab Undang-Undang Hukum Dagang, KUHD) Chapter 9, Article 246 on Insurance and Life-long Coverage, insurance is an agreement between the insured (policyholder) and an insurance company, in which the insurer provides guarantees and compensation for potential loss, damage, or missed expected gains due to uncertain events involving insured assets [3]. Insurance premiums represent the price set by insurers to transfer the risk from the policyholder to the insurance company [4]. Previous research on the determination of pure insurance premiums was conducted by [5] using a distribution model. A similar study was carried out by [6] employing the Fast Fourier Transform method. However, previous studies in determining the distribution and

pure premium did not apply a deductible. To manage risk, insurers often apply deductible [7] provisions. For example, if a policy includes a deductible amount "d" per loss, then if the claim amount "x" is less than "d", the insurer pays nothing; if it exceeds "d", the insurer pays x-d [8]. Determining an appropriate distribution model for insurance claims is crucial in assessing operational risk within an insurance company [9]. This research tests claim history data against statistical distribution models to determine the pure premium value. The pure premium obtained serves the insurance company as the basis for determining the actual premium charged to policyholders, with the addition of loadings.Therefore, this study applies a claim distribution model to determine pure premium values, incorporating deductible levels based on the historical claim data of an insurance portfolio.

2. RESEARCH METHOD

The data used in this study are secondary data obtained from the records of Company X in Manado for the 2019–2020, specifically the total number and amount of motor vehicle insurance claims (inclusive of deductibles). The data provided by the company are confidential; therefore, the author is not permitted to disclose the company's name in this study.

2.1 Determining the Claim Frequency Distribution Model

Calculate the mean and variance of the sample and then determine the appropriate distribution model based on the mean and variance of the sample. The mean, also referred to as the average, is the value obtained from the sum of all data values divided by the number of data points [10]. If each data point has a certain frequency of occurrence, or in other words, if the data is grouped, then the formula to calculate the mean is as follows [11]

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i \cdot x_i}{\sum_{i=1}^{n} f_i} \tag{1}$$

Where:
x_i = the i-th data value
f_i = the i-th frequency

Variance is the average squared deviation of the observed data from its mean, and is denoted by s^2. Then the model for the aggregate loss can be expressed as follows [12]:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 \cdot f_i - \frac{(\sum x_i \cdot f_i)^2}{\sum f_i}}{\sum f_i - 1} \tag{2}$$

Perform a Chi-square goodness-of-fit test

The Chi-Square test, denoted as χ^2, is a test used to examine data derived from a population with a discrete distribution, such as Poisson and Binomial distributions. The hypotheses for the Chi-Square test are defined as follows [13]:

H_0 : The data follows a specified distribution
H_1 : The data follows another distribution
In the calculation, the data is divided into (k) intervals, and the test statistic is given as follows:

$$\chi^2 = \sum_{j=1}^{k} \frac{\left(E_j - O_j\right)^2}{E_j} \tag{3}$$

Compare the test statistic value with the test table value.

The critical value for this test has degrees of freedom equal to (k - 1 - p), where (k) is the number of classes (intervals) and (p) is the number of estimated parameters. $E_j = np_k$ represents the expected number of observations in the (j)-th interval, while Oj is the observed number of observations in the j-th interval. The decision criteria are as follows [13]:

If $\chi^2_{hitung} \geq \chi^2_{tabel}$, then H_0 is rejected.
If $\chi^2_{hitung} < \chi^2_{tabel}$, then H_0 is accepted.

Selecting the appropriate Claim Frequency distribution model.

The selection of a distribution is fundamental in claim frequency modeling, where overdispersed data must be modeled using a distribution that accommodates overdispersion. In general, the mixed Poisson distribution that exhibits overdispersion is the Negative Binomial distribution. Therefore, a test is conducted for the Negative Binomial distribution model.

2.2 Determining the Claim Severity Distribution Model

In contrast to claim frequency modeling, which involves non-negative integer-valued random variables, claim severity modeling is typically represented by non-negative continuous random variables [14]. According to [15], there are many continuous distributions that can be considered for model selection; however, not all of them need to be tested for suitability. The model selection principle proposed by [16] suggests using simpler models whenever possible. Accordingly, four distributions are selected for further analysis in modeling claim severity: the Gamma distribution, the Weibull distribution, the Lognormal distribution, and the Log-logistic distribution. The following are the steps in determining the distribution of claim severity:

Determine the initial parameter values.

The initial parameter values of the claim severity distribution are calculated based on the formulas of each selected distribution, and then followed by parameter estimation.

Estimate the parameter values using Newton–Raphson iteration

According to [17], this method is often used to find the roots of an equation. Given an initial value $x = p_0$, the root of the equation $f(x) = 0$, can be obtained using the following iteration:

$$p_k = g(p_{k-1}) = p_{k-1} - \frac{f(p_{k-1})}{f'(p_{k-1})}, \qquad k = 1,2, \dots \tag{4}$$

where $f'(p_{k-1})$ is defined as:

$$f'(p_{k-1}) = \frac{df(x)}{dx}|_{x=p_{k-1}} \tag{5}$$

The iteration stops once the value of p converges. The convergence criterion is given by:

$$|p_k - p_{k-1}| < \tag{6}$$

The iterative model in equation (4) can be generalized to solve systems of equations as follows:

$$\begin{aligned} f_1(x_1, x_2, \dots, x_k) &= 0 \\ f_2(x_1, x_2, \dots, x_k) &= 0 \\ &\vdots \\ f_k(x_1, x_2, \dots, x_k) &= 0 \end{aligned} \tag{7}$$

With ($\underline{x} = \underline{p}_0$)as the initial value, the system of equations above can be solved using the following iteration:

$$\underline{p}_k = \underline{p}_{k-1} - \left[M'\left(\underline{p}_{k-1}\right)\right]^{-1} \underline{f}\left(\underline{p}_{k-1}\right), \quad k = 1, 2, \dots \tag{8}$$

The iteration will stop when the vector $\underline{p}$ converges. The convergence criterion for the vector $\underline{p}$ is:

$$\|\underline{p}_k - \underline{p}_{k-1}\| < \quad \delta \tag{9}$$

Formulate the hypotheses to be tested.

After obtaining the parameter estimates, the process continues by formulating the hypothesis test for each claim severity distribution.

Calculate the test statistic using the Anderson–Darling test.

This test is similar to the Kolmogorov–Smirnov test but employs two different measures derived from distribution functions [18]. The hypotheses of the Anderson–Darling test are:

H_0 : The data follow a specified distribution

H_1 : The data follow an alternative distribution

The test statistic is defined as follows [16]:

$$A^2 = n \int_t^u \frac{[F_n(x) - F^*(x)]^2}{F^*(x)[1 - F^*(x)]} f^*(x)dx \tag{10}$$

For individual data, the integral in Equation (10) can be simplified as:

$$\begin{aligned} A^2 = & -nF^*(u) + n\sum_{j=0}^{k}\left[1 - F_n(y_j)\right]^2\{ln\ ln\left[1 - F^*(y_j)\right] - ln\ ln\left[1 - F^*(y_{j+1})\right]\} \\ & + n\sum_{j=1}^{k} F_n(y_j)^2[lnF^*(y_{j+1}) - ln\ ln\ F^*(y_j)], \end{aligned} \tag{11}$$

where $t = y_0 < y_1 < \cdots < y_k < y_{k+1} = u$, are unique uncensored data points. When $u = \infty$, the last term of the first summation becomes zero.

The critical values of the Anderson–Darling goodness-of-fit test are 1.933 for significance level α = 1%, 2.492 for α = 5%, and 3.857 for α =10% [16].

Compare the test statistic with the corresponding critical value.

The next step is to compare the test statistic value with the critical table value for each claim severity distribution to determine whether the data follows the specified distribution.

Compare the candidate models.

If, based on the distribution model testing for the claim severity data, two suitable models are obtained, then the most appropriate distribution model is selected by considering the tail weight of those distributions.

Select the most appropriate model to be used.

From the results of comparing each claim severity distribution, the most appropriate distribution for the claim severity data will be determined.

2.3 Calculating Pure Premium Value using Compound Model [20]

If S denotes the sum of random variables where N is a nonnegative integer-valued random variable that is distributed independently of $X_1, \ldots, X_N$, then the model for the aggregate loss can be calculated using the following equation [12]:

$$S = X_1 + \cdots + X_N \tag{12}$$

In this model, the distribution of N is referred to as the primary distribution, while the distribution of X is referred to as the secondary distribution. When these two distributions are combined, they form a new compound distribution [21].

The approaches used for analyzing this model is:

Constructing the distribution model for N based on data.

The distribution model of N represents the distribution of claim frequency based on the obtained claim data.

Constructing the distribution model for X based on data.

The distribution model of X represents the distribution of claim severity based on the obtained claim data.

Performing calculations using the two models above to obtain the distribution of S.

The aggregate loss distribution model (S) is obtained from the combination of the claim frequency distribution and the claim severity distribution.

The mean and variance of S are derived from the moments of N and X [22], as follows:

$$(S) = E(N)E(X) \tag{13}$$

$$Var(S) = E(N)Var(X) + E(X)^2Var(N) \tag{14}$$

Thus, pure premium value is the expected value of the aggregate loss distribution formed from the claim frequency distribution and the claim severity distribution.

3. RESULT AND ANALYSIS

3.1 Claim Frequency Distribution Model

Sample Mean and Variance

To calculate the sample, mean for data with a certain frequency of occurrence, the sample mean can be calculated using Equation (1) and the variance of a sample from grouped data can be determined using Equation (2). Thus, the results of the sample mean and sample variance calculations can be seen in Table 1.

Table 1. Sample Mean and Sample Variance Values

Number of Claims (k)	Number of Policyholders (Nk)	Sample Mean	Sample Variance
0	492		
1	37		
2	5	0,0934	0,1148
3	1		
Total	535		

Based on the claim data of car insurance company X in Manado City for the period 2019–2020 were obtained, as shown in Table 1, there were 492 policyholders who did not file any claims, 37 policyholders who filed one claim, 5 policyholders who filed two claims, and 1 policyholder who filed three claims. The total number of policyholders from 2019 to 2020 was 535. The total number of claims filed within one year was 43.

Based on the obtained values of the sample mean and sample variance, it can be concluded that the data exhibits overdispersion. Overdispersion is a condition where the sample mean of a dataset is smaller than its sample variance. According to [23], the selection of a distribution is fundamental in claim frequency modeling, where overdispersed data should be modeled using a distribution that also exhibits overdispersion. Typically, a mixture Poisson distribution that accounts for overdispersion is the Negative Binomial distribution. Therefore, a test was conducted to evaluate the suitability of the Negative Binomial distribution model.

Chi-Square Test for Negative Binomial Distribution Model

The estimated parameters of the Negative Binomial distribution are:

H_0 : The claim frequency data follows a Negative Binomial distribution ($\hat{\alpha} = 0{,}4084$ dan $\hat{\beta} = 0{,}2288$).

H_1 : The claim frequency data follows an alternative distribution.

Table 2. Chi-Square Test Calculation for Negative Binomial Distribution

Number of Claims (k)	Number of Policyholders (Nk)	O_j	E_j	$\frac{(O_j - E_j)^2}{E_j}$
0	492	492	491,8175151	6.77095×10^{-5}
1	37	529	529,2221349	9.32386×10^{-5}
2	5	534	534,1273313	3.03547×10^{-5}
3	1	535	534,8606603	3.63002×10^{-5}
+4	0	535	534,9770258	9.86608×10^{-5}
	Total			0.00022859

Based on the results in Table 2, the calculated Chi-Square statistic is 0.00022859. The Chi-Square critical value with 2 degrees of freedom (5-1-2) at a 5% significance level is 5.99148. Since the test statistic is less than the critical value (0.00022859 < 5.99148), we fail to reject H_0. Therefore, it can be concluded that automobile insurance claims follow a Negative Binomial distribution.

3.2 Claim Size Distribution Model

Table 3. Statistical Summary Claim Size Data

Total Data	43
Mean	4.267.124,674
Median	2.150.000
Minimum	393.800
Maximum	61.584.655
Range	61.190.855
Standard Deviation	9.320.443,992
Skewness	5,8016
Kurtosis	36,0785

Table 3 presents the summary statistics of the claim size data obtained from the company. Based on this information, the appropriate distribution will be determined using the steps for identifying the claim severity distribution.

Lognormal Distribution Model Test

The initial parameters of the Lognormal distribution (μ, σ) are calculated using Equations:

$$\mu = \sqrt{\ln(t) - 2\ln(m)} \text{ dan } \sigma = \ln(m) - \frac{1}{2}\sigma^2 \quad (15)$$

Where:

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ dan } t = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \quad (16)$$

From the calculations, we obtain:

$$\mu = 1{,}3165 \text{ and } \sigma = 14{,}3997$$

Further, parameter estimates were refined using statistical software, resulting in:

$$\hat{\mu} = 14{,}6032 \text{ and } \hat{\sigma} = 0{,}9789$$

The hypotheses for the distribution test are as follows:
H_0: The claim size data follows a Lognormal distribution ($\hat{\mu} = 14{,}6032$ and $\hat{\sigma} = 0{,}9789$)
H_1: The claim size data follows a different distribution.

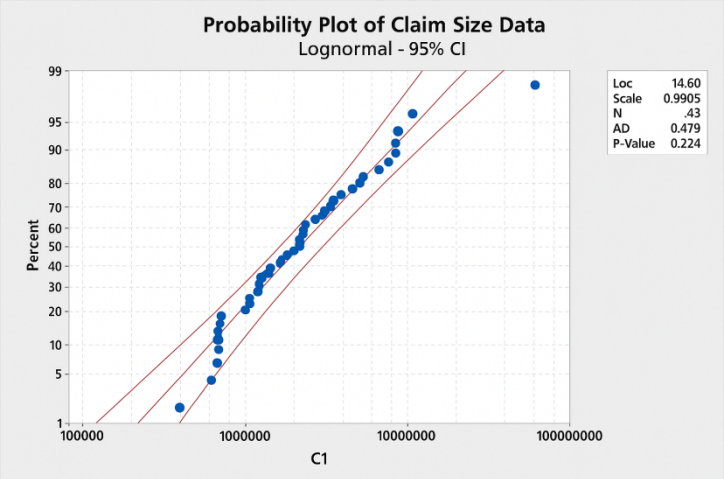

Figure 1. Probability Plot of Claim Size Data – Lognormal Distribution

Based on the results obtained, the test statistic value is 0.479 with a P-Value of 0.224. The critical value for the Anderson–Darling test at the 5% confidence level is 2.492. Since the test statistic is smaller than the critical value and the P-Value > 0.05, the null hypothesis H_0 is accepted. Thus, the claim size data follows a Lognormal distribution.

Log-Logistic Distribution Model Test

The initial parameters of the Log-Logistic distribution (γ, θ) are calculated using Equation [24].

$$\gamma = \frac{2\ln(3)}{\ln(q) - \ln(p)} \; dan \; \theta = \exp\left(\frac{g\, ln(q) - ln(p)}{2}\right) \quad (17)$$

Where:
p is the 25th percentile and q is the 75th percentile
From the calculation, the values obtained are:

$$\gamma = 1{,}2700 \; dan \; \theta = \; 3.014.344{,}582$$

Then, using statistical software, the estimated parameters are:

$$\hat{\gamma} = 14{,}5376 \text{ dan } \hat{\theta} = 0{,}5445.$$

The hypotheses for the test are:
H₀: The claim size data follows a Log-Logistic distribution ($\hat{\gamma} = 14{,}5376$ dan $\hat{\theta} = 0{,}5445$)
H₁: The claim size data follows a different distribution

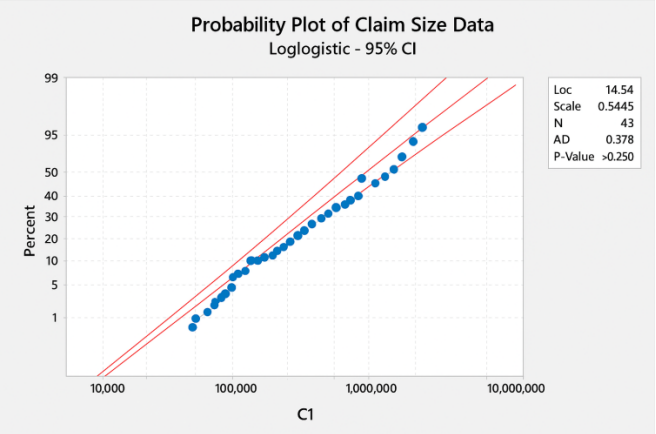

Figure 2. Probability Plot of Claim Size Data – Log-Logistic Distribution

From the results obtained, the test statistic value is 0.378 with a P-Value > 0.250. The critical value for the Anderson–Darling test at the 5% confidence level is 2.492. Since the test statistic is less than the critical value and P-Value > 0.05, the null hypothesis H₀ is accepted. Therefore, the claim size data follows a Log-Logistic distribution.

Selection of Claim Size Distribution Model

Based on the distribution model tests for claim size data, among the four tested distributions, two models were found to be appropriate: the Lognormal and Log-Logistic distributions. Therefore, one of these two models will be selected as the most suitable distribution by considering the tail weight of each distribution. According to [25], tail weight can help narrow down model choices or confirm a model selection during the model selection process. The right tail of a distribution refers to the part of the distribution that corresponds to large values of the random variable. This is intended to understand the likelihood of large loss values, which have the greatest effect on total loss. If the probability of a random variable tends to be higher at larger values, the distribution is said to have a heavier tail. Model selection can apply a relative concept, such as model A having a heavier tail than model B, or vice versa.

The tail weight graphs of the Log-normal and Log-logistic distributions based on their probability density functions are as follows:

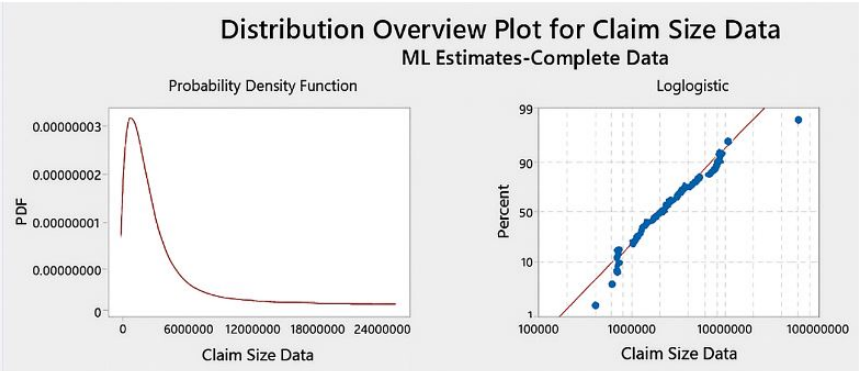

Figure 3. Probability Density Function Graph of the Log-logistic Distribution for Large Claim Data

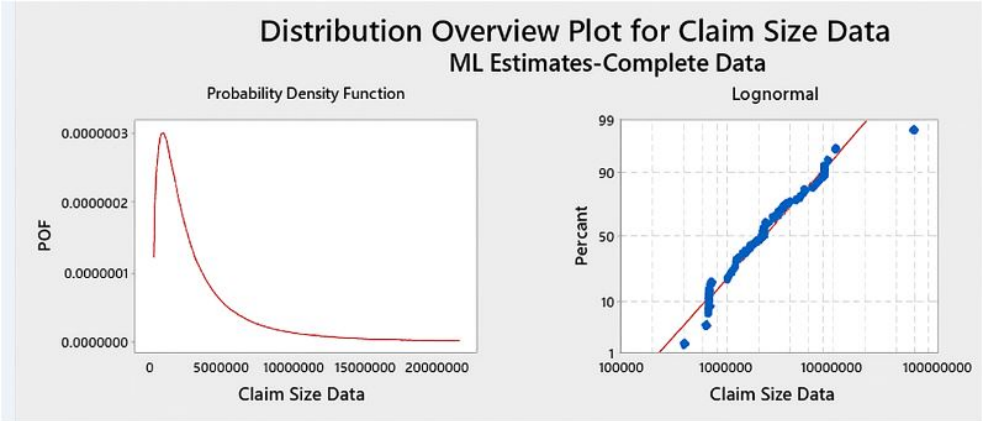

Figure 4. Probability Density Function Graph of the Log-normal Distribution for Large Claim Data

From the probability density function graphs of the two distributions, it can be seen that the comparison is not significantly different in value. However, the log-logistic distribution has a heavier right tail compared to the log-normal distribution. Therefore, the appropriate model for large claim data is the log-logistic distribution.

Since the large claim data follow a log-logistic distribution, the expected value is calculated using Equation (10)

$$E(X) = \theta^k \Gamma\left(1+\frac{k}{\gamma}\right)\Gamma\left(1-\frac{k}{\gamma}\right)$$
$$= (3.014.344{,}582)\Gamma\left(1+\frac{1}{1{,}2700421662}\right)\Gamma\left(1-\frac{1}{1{,}2700421662}\right)$$
$$= 12.037.950$$

Calculation of Pure Premium

Based on the previous discussion, models have been obtained for both claim frequency and claim severity. The resulting compound model is the Negative Binomial – Log-logistic model. Therefore, the pure premium value is derived from the expected value of the Negative Binomial claim frequency and the expected value of the Log-logistic claim size. The pure premium is calculated using Equation (9):

$$E(S) = E(N)E(X) = (\alpha\beta)E(X) = (0{,}010733155)(\,12.037.950)$$
$$= 129.205{,}1873$$

Thus, based on the calculation results, the pure premium *E(S)* for motor vehicle insurance company X in the city of Manado, based on property damage claim data, is Rp. 129,205.19.

4. CONCLUSION

Based on claims data obtained, the claim frequency data for a car insurance company can be modeled using a Negative Binomial Distribution with estimated parameters $\hat{\alpha} = 0{,}4084$ dan $\hat{\beta} = 0{,}22882$.The claim size data can be modeled using a Log-logistic Distribution with estimated parameters $\hat{\gamma} = 14{,}5376$ dan $\hat{\theta} = 0{,}5445$

The pure premium for car insurance based on historical damage claim data is Rp. 129,205.19. This is the basis for the insurance company X to determine the gross premium charged to the policyholder.

5. REFERENCES

[1] R. Maralis and A. Triyono, “Manajemen Resiko,” 2019.
[2] D. Sunyoto and W. Harisa Putri, *Manajemen Risiko dan Asuransi: Tinjauan Teoretis dan Implementasinya*. 2017.
[3] Biro Hukum dan Humas Badan Urusan Administrasi Mahkamah Agung-RI, “Kitab Undang - Undang Hukum Dagang,” 2025.
[4] Y.-K. Tse, *Nonlife Actuarial Models*. Cambridge University Press, 2023. doi: 10.1017/9781009315067.
[5] C. Kireina Waha *et al.*, “Model Distribusi Data Klaim Asuransi Mobil untuk Menentukan Premi Murni,” 2019. [Online]. Available: https://ejournal.unsrat.ac.id/index.php/decartesian
[6] T. Manurung, “Taksiran Distribusi Aggregate Loss Asuransi Mobil Menggunakan Fast Fourier Transform (FFT) dalam Menentukan Premi Murni.”
[7] N. Lewaherilla and G. Haumahu, “Premium Calculation with the Application of Deductible in Actuarial Model for One Year Sickness Insurance,” vol. 1, 2019, [Online]. Available: https://ojs3.unpatti.ac.id/index.php/variance/
[8] J. Cao, D. Li, V. R. Young, and B. Zou, “Strategic underreporting and optimal deductible insurance,” *ASTIN Bulletin*, vol. 54, no. 3, pp. 767–790, Sep. 2024, doi: 10.1017/asb.2024.14.
[9] S. C. K. Lee and X. S. Lin, “Modeling And Evaluating Insurance Losses Via Mixtures Of Erlang Distributions.”
[10] A. fauzi *et al.*, *Metodologi Penelitian*. Pena Persada, 2022.
[11] T. Hidayati, I. Handayani, and I. H. Ikasari, *Statistika Dasar*. Pena Persada, 2019.
[12] A. Natawiria Suryana and Riduwan, *Statistika Bisnis*. Alfabeta, 2010.
[13] “A Review Of Some Goodness-Of-Fit Test For Logistic Regression Model.”
[14] C. O. Omari, S. G. Nyambura, and J. M. W. Mwangi, “Modeling the Frequency and Severity of Auto Insurance Claims Using Statistical Distributions,” *Journal of Mathematical Finance*, vol. 08, no. 01, pp. 137–160, 2018, doi: 10.4236/jmf.2018.81012.
[15] “Loss Data Analytics, Second Edition An open text authored by the Actuarial Community,”
[16] “Stuart A. Klugman, Harry H. Panjer, Gordon E. Willmot-Loss Models_ From Data to Decisions-Wiley (2012) 4th ed”.
[17] R. V. . Hogg and S. A. . Klugman, *Loss distributions*. Wiley, 1984.
[18] “Alternative modelling and inference methods for claim size distributions.” [Online]. Available: www.mathiasraschke.de
[19] G. Marsaglia and J. C. W. Marsaglia, “Evaluating the Anderson-Darling Distribution.”
[20] T. Rahmawati and D. Susanti, “Determining Pure Premium of Motor Vehicle Insurance with Generalized Linear Models (GLM),” *International Journal of Quantitative Research and Modeling*, vol. 4, no. 4, pp. 207–214, 2023.
[21] *Actuarial_Mathematics*.
[22] C. Laudagé, S. Desmettre, and J. Wenzel, “Severity modeling of extreme insurance claims for tariffication,” *Insur Math Econ*, vol. 88, pp. 77–92, Sep. 2019, doi: 10.1016/j.insmatheco.2019.06.002.
[23] J. Tamansari No, K. Tamansari, K. Bandung Wetan, K. Bandung, M. Karim, and A. Komarudin Mutaqin, “Modeling Claim Frequency in Indonesia Auto Insurance Using Generalized Poisson-Lindley Linear Model Pemodelan Frekuensi Klaim Asuransi Kendaraan Bermotor Indonesia Menggunakan Generalized Poisson-Lindley Linear Model,” vol. 16, no. 3, pp. 428–439, 2020, doi: 10.20956/jmsk.v%vi%i.9315.
[24] M. Ahsanullah and A. Alzaatreh, “Parameter Estimation For The Log-Logistic Distribution Based On Order Statistics.”
[25] A. J. Mcneil, R. Frey, and P. Embrechts, “Quantitative Risk Management: Concepts, Techniques and Tools,” 2005.