JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 8, No. 4, October 2024, pp. 1233-1248

The Value at Risk Analysis using Heavy-Tailed Distribution on the Insurance Claims Data

Utriweni Mukhaiyar[1], Aprilia Dianpermatasari[2], Azizah Dzakiya[2], Sasqia Bunga Widyani[2], Husnul Khatimah Syam[2]
[1]Statistics Research Division, Institut Teknologi Bandung, Indonesia
[2]Master Program in Actuarial Science, Institut Teknologi Bandung, Indonesia
utriweni.mukhaiyar@itb.ac.id

Article History:
Received : 09-07-2024
Revised : 09-10-2024
Accepted : 13-10-2024
Online : 18-10-2024

Keywords:
Value at risk;
Heavy-tailed distribution;
Claim distribution.

ABSTRACT

The insurance has often been involved to minimize financial losses. As the product providers, the insurance companies must effectively manage risks to prevent errors in risk measurement. The amount of risk or loss experienced by the policyholder refers to the claim amount. The Value at Risk (VaR) is commonly used to measure risk. The VaR is calculated from the probability function, which can be obtained by evaluating the distribution of claims data. Most claim frequencies are small, but occasionally, huge claims appear. Therefore, the appropriate distribution would be characterized by a heavy-tailed. Thus, this research aims to model and evaluate insurance claims data using exponential, Weibull, Pareto, and lognormal distributions to assess financial risk through VaR. The insurance claims data were collected from a single insurance company and include 1,326 claims. This research specifically examines variables such as gender, diabetic status, smoking status, the number of claims, and the level of confidence. The data were analysed using descriptive statistics, Maximum Likelihood Estimation for parameter estimation, and Goodness of Fit tests to determine the best-fitting distribution, along with VaR calculations based on the results. The suitability of the distribution model is assessed through the VaR and is analysed based on the appropriate distribution of insurance claims data. It is obtained that the Weibull and lognormal distributions appropriately model insurance claims data. The highest VaR is observed in the claim data for female non-diabetic smokers, with a level of confidence of 99.5%. The lowest VaR is obtained from the claim data for male diabetic non-smokers, with a level of confidence of 90%. This approach enhances the prediction of large potential losses for specific demographic groups, aiding more informed decision-making in premium pricing and risk management. The integration of heavy-tailed distributions in risk assessment, with a particular focus on demographic specificity, constitutes a substantial and novel contribution to this research.

https://doi.org/10.31764/jtam.v8i4.25053

A. INTRODUCTION

Insurance claims usually have data with small claim sizes but in large frequencies, moreover, data with large claim sizes usually has small frequencies. The most suitable and widely used models for claim sizes are distributions of continuous random variable that assume positive values only and have “heavy tails”, that is distributions which allow for occasional occurrences of very large values (Gray & Pitts, 2012). The heavy-tailed behaviour is usually associated with large values of quantities such as the coefficient of variation, the skewness, and the kurtosis (Klugman et al., 2019). The heavy-tailed distribution is characterized by the characteristics of insurance claim data that form a long tail, the mean values usually greater than the median so that they have a positive skewness, as well as a slower probability decrease

to values far from the mean characterized by larger kurtosis values. Heavy-tailed distributions include exponential, Weibull, Lognormal, Log-gamma, and generalized Pareto distributions (Riad et al., 2023; Xie, 2023). The Pareto distribution and the Weibull distribution with $\tau < 1$ have heavy tails dan thus relatively larger extreme quantiles (Klugman et al., 2019). The lognormal, Pareto, Weibull (of which the exponential is a sub-family), log-gamma families, and the three-parameter Burr family are considered as important distributions used as models in practice to modelling claim sizes (Gray & Pitts, 2012). The other families of heavy-tailed distributions such as normal (Gaussian) distribution, exponential and gamma are useful for reference and comparison purposes, and are included for completeness (Gray & Pitts, 2012). Hence, the heavy-tailed distributions that will be used in this paper are lognormal, Pareto, Weibull, and exponential.

The risks faced by insurance companies tend to be related to the claim size. The number of insurance claims that the insurance company can provide to the insured party can be calculated based on the distribution model of the frequency of claims and the size of claims (Rohiim & Mutaqin, 2023). Hence, the risk measurement needs to be done so that the risk is at a controlled level to reduce the occurrence of losses to insurance companies. To avoid the risk one of the measuring tools that can be used to calculate the risk of large insurance claims is Value-at-risk (VaR) and T-VaR that had been studied by Adiyansyah et al. (2023) to modelling the claim sizes using tailed distribution on sample data from Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan. The VaR is a common and widely used risk measure, and it may be selected to determine the potential magnitude of risks in the future (Syuhada et al., 2023). The other application of Value-at-Risk is to analysing risk using automated threshold selection method in property insurance (Pahrany et al., 2024). Several researchers in the fields of mathematics and statistics have modelled the Value-at-Risk on insurance claims data, including Yousof et al. (2023) who modelled new distributions based on insurance data through key risk indicators, which include VaR, tail variance, conditional VaR, Tail Value at Risk, dan tail-mean variance. Yildirim (2015) uses VaR, historical simulation, and the Monte Carlo simulation method to determine insurance company losses caused by foreign exchange risk. Keçeci & Sarul (2014); Zhao et al. (2022) applied various VaR models to insurance claim data and demonstrated the appropriateness of distributions in measurement accuracy.

Analysis of VaR on insurance claims data through heavy-tailed distribution needs to be reviewed to determine the reliability of the insurance risk model through heavy-tailed distribution according to the nature of each insurance claim data, identify potential significant losses that may occur to insurance companies due to event, and provide a more realistic estimate of the possibility of extreme losses (Afify, et.al, 2020). Propose new heavy-tailed distribution called alpha power exponentiated exponential (APExE), then doing risk measures using VaR for the unemployment insurance data. Research about quantifying risk of South African taxi claims data and the Danish fire loss data using VaR and T-VaR for 19 standard parametric distribution including heavy-tailed distribution had already presented by (Marambakuyana & Shongwe, 2024). Moreover, those researches having the same approach using VaR for heavy-tailed distribution data on one category data.

In this paper, insurance claim data will be modelled through selected heavy-tailed distributions, parameter estimation, the suitability of distribution models, and VaR through

appropriate distributions on insurance claim data. The data consisted of some categories, such as gender, diabetic status, and smoking habits. In addition to comparing the fit of different heavy-tailed distributions, this research adds a more comprehensive analysis of factors that affect insurance claim risks including the influence of real-world factors on claim amounts. This research further seeks to investigate how gender, diabetic status, or smoking habits may affect the VaR of the insurance claims, thereby enabling insurance companies to offer more tailored risk pricing. These categories will help the insurers to differentiate the risk profile of policyholders. Thus, would lead to estimating the risk more accurate compared to previous approaches that has only one category data. These risk estimates are more personalized and nuanced and so help insurers in the ability to set their rates and reserves in such a way that they can better choose the risk they are taking.

B. METHODS

The data was collected through documentation studies to acquire secondary data on insurance claims. The use of dummy data in this study is a practical choice for preliminary testing and modelling. It is justified to use dummy data in this research because dummy data allows for methodological testing, simplification, and controlled experimentation when real-world data is unavailable or inaccessible. It provides a cost-effective way to explore the theoretical aspects of VaR estimation and model-fitting, which can be later be validated and calibrated with real insurance claim data. The key is to use dummy data as a stepping stone for more accurate once actual data is obtained. However, it is crucial to acknowledge the limitations of dummy data in representing real-world insurance claims. While useful for initial testing and theoretical validation, using dummy data is limited by its inability to fully capture the complexity, dependencies, and extremes present in real-world insurance claims. This limits the generalizability of the findings, the accuracy of risk estimates, and the reliability of the models when applied to real insurance settings. To overcome these limitations, researchers must eventually test their models on actual claim data to ensure the results are robust, realistic, and applicable in practical insurance risk management. For this research, the analytical process involves the following steps:

1. Analyse insurance claims data and its distribution patterns using descriptive statistical methods.
2. Categorize the insurance claims data into multiple groups.
   For this research, the data that were used divided into 8 categories: male diabetic smokers (male D-S), male diabetic non-smokers (male D-NS), male non-diabetic smokers (male ND-S), male non-diabetic non-smokers (male ND-NS), female diabetic smokers (female D-S), female diabetic non-smokers (female D-NS), female non-diabetic smokers (female ND-S), and female non-diabetic non-smokers (female ND-NS).
3. Estimate the parameters for each data category using the Maximum Likelihood Estimation (MLE) technique.
4. Select the suitable model to identify the best-fitting distribution for insurance claims data using the Goodness of Fit test (GoF). The GoF test in this research is conducted using the exponential, Weibull, Pareto, and lognormal distributions, all of which are heavy-tailed distributions. These were selected specifically because of their ability to model

extreme values, as they capture the positive skewness, high kurtosis, and long tails that are characteristic of the insurance claims data used in this study. This makes them suitable for accurately reflecting the presence of rare but severe claims, which is a key feature of heavy-tailed data. The GoF test used is Akaike's Information Criterion (AIC). The model or best distribution is selected based on the AIC value, with the distribution with the lowest AIC value being the best model or distribution.

5. Calculate the risk amount in insurance claims data using the VaR distribution that aligns with the GoF test outcomes. For this research, the VaR is calculated using the VaR formula for each selected distribution with the high level of confidence of 90%, 95%, 99%, and 99.5%.
6. Draw the conclusions based on the analysis conducted. The analysis methods could be drawn as a flowchart on Figure 1.

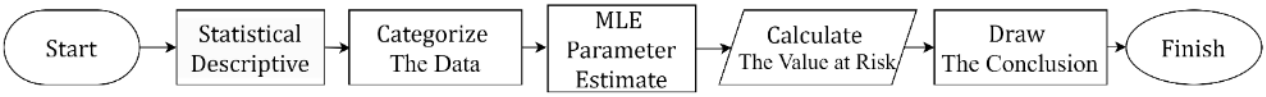

Figure 1. Flowchart of the Analysis Methods

The insurance claim data used in this research is dummy data which contains 1,326 insurance claims obtained from the kaggle.com website (Shukla, 2022). The data consists of information such as the claim amount, age, gender, BMI (Body Mass Index), blood pressure, number of children, diabetic status, smoking habits, and residential area of the insured individuals.

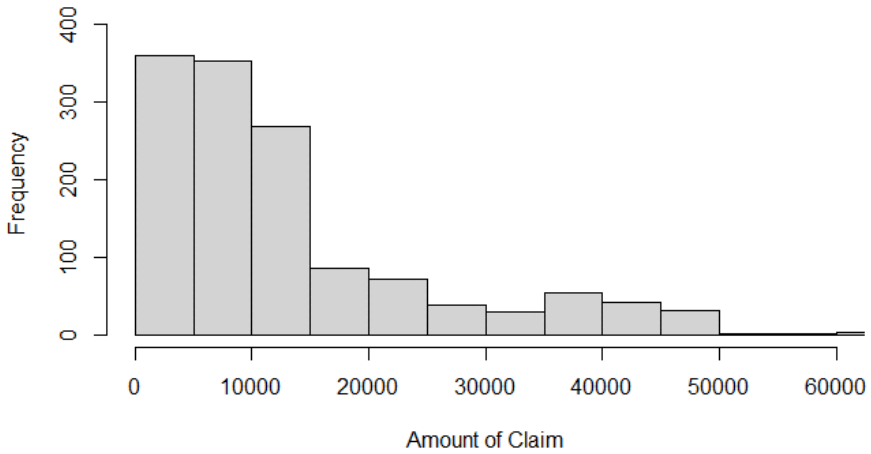

Figure 2. Histogram of insurance claims data

The histogram in Figure 2 illustrates the claim amount towards claims frequency. This histogram clearly illustrates that the insurance claim data is long tailed. Most claims fall within the lower range (small amounts), with the highest frequency occurring around the lower claim amounts (close to zero). As the claim amounts increase, the frequency of occurrence decreases, which is indicative of a long tail. This pattern is typical of heavy-tailed distributions, where a large number of small claims are observed, but there is also a non-negligible probability of

extreme (high) claims. This justifies the use of heavy-tailed distributions conducted in this research. The following is a Box Diagram for each Data Category, as shown in Figure 3.

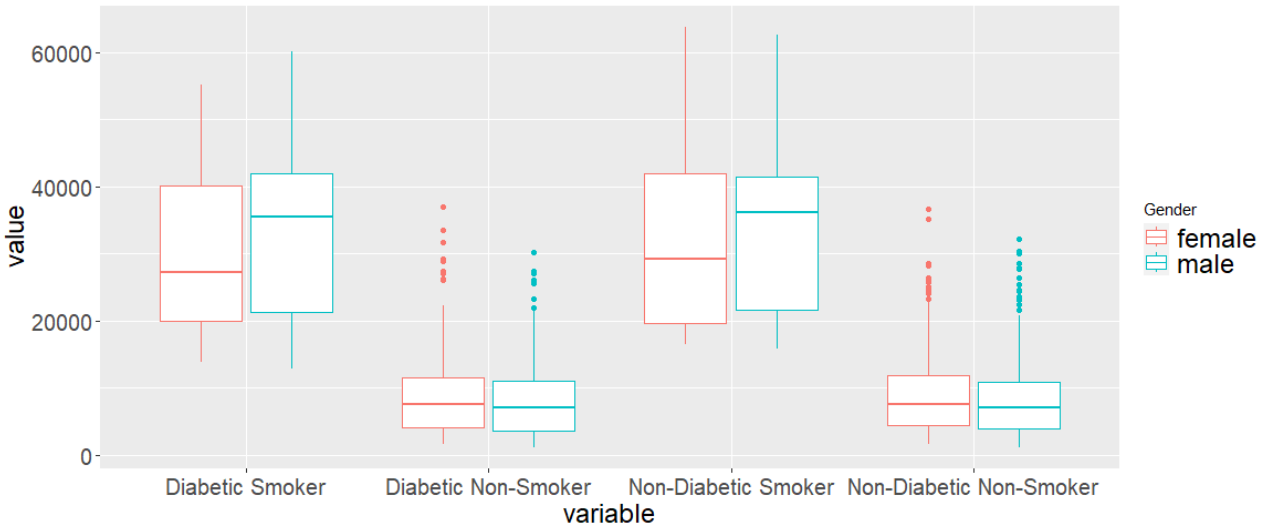

Figure 3. Boxplots for Each Data Category

Figure 3 shows the boxplot of each category data to see whether there are outliers in each category data. It can be seen from Figure 3 that the distribution of insurance claim values across various categories based on gender, diabetic status, and smoking status. Initial conclusions reveal that smokers, particularly diabetic smokers, tend to have higher and more variable claim values compared to non-smokers. Smokers generally exhibit a higher median and wider interquartile range (IQR), indicating more variability and larger claims. In contrast, non-smokers have relatively lower claim amounts for both diabetic and non-diabetic groups, characterized by smaller IQRs and lower medians.

The analysis also highlights subtle gender differences, with males generally having slightly higher claims in certain categories. For example, in categories like non-diabetic smokers, males appear to have a slightly higher median claim amount than females. The diabetic non-smoker and non-diabetic non-smoker categories almost have similar values while the non-diabetic non-smoker category has more outliers than the diabetic non-smoker category. Likewise for the diabetic smokers and non-diabetic smokers category but without the outlier, the female non-diabetic smoker category has slightly greater value than the diabetic smoker category and the male non-diabetic smokers category has a slightly lower value than the diabetic smoker category.

In terms of outliers, there are several points outside the whiskers of the boxes, particularly in male non-smokers for both diabetic and non-diabetic, indicating exceptionally high claim values. These outliers suggest that a small number of individuals in these categories possess unusually high claims compared to others in the same category. Overall, the plot suggests that smoking status and diabetic status are significant factors influencing claim values, with smokers, particularly diabetics, experiencing higher claim amounts. Table 1 and Table 2 show that most of the data categories are right-skewed. However, the categories of D-S and ND-S for males can be said to be skewed to the left because the skewness of the data is negative.

Table 1. Statistic Descriptive of Each Categories Data of Male

Statistic Descriptive	All of the Data	Male			
		D-S	D-NS	ND-S	ND-NS
Variance (USD)	146,642,600	124,520,000	32,024,770	126,986,100	37,350,120
n	1,335	76	240	83	274
Mean (USD)	13,252.75	32,314.22	8,001.01	33,708.41	8,240.91
Std. Deviation (USD)	12,109.61	11,158.85	5,659.04	11,268.81	6,111.47
Median (USD)	9,369.61	35,538.61	7,073.40	36,197.70	7,066.31
Min (USD)	1,121.87	12,829.46	1,121.87	15,817.99	1,131.51
Max (USD)	63,770.43	60,021.40	30,166.62	62,592.87	32,108.66
Range (USD)	62,648.56	47,191.94	29,044.75	46,774.88	30,977.15
Skewness	1.51	-0.08	1.3	-0.04	1.56
Kurtosis	4.61	1.84	5.06	1.98	5.86
1st Quartile	4,720	21,219	3,632	21,601	3,851
3rd Quartile	16,604	41,990	10,968	41,387	10,804
Sum (USD)	17,758,680	2,455,881	1,920,242	2,797,798	2,258,009
Std. Error	330.81	1,280.01	365.29	1,236.91	369.21

According to Table 1, the aggregate insurance claims data obtained a mean of 13,252.75 and a median of 9,369.61, which indicates that the mean surpasses the median. Generally, when the mean exceeds the median, it suggests right skewness. Moreover, all of the insurance claim data have a skewness of 1.51 and a kurtosis of 4.61. With a kurtosis value exceeding 3, the insurance claims data exhibits heavy tails. Table 1 shows that the standard deviation of all the categories data is smaller than the mean, implying that the data are clustered closely around the mean value. Most of the data categories have positive skewness except for male diabetic smokers and male non-diabetic smokers.

Table 2. Statistic Descriptive of Each Categories Data of Female

Statistic Descriptive	Female			
	D-S	D-NS	ND-S	ND-NS
Variance (USD)	132,207,700	35,831,290	153,697,800	37,646,770
n	60	263	55	284
Mean (USD)	29,964.44	8,593.77	31,458.52	8,918.36
Std. Deviation (USD)	11,498.16	5,985.92	12,397.49	6,135.70
Median (USD)	27,285.91	7,626.99	29,141.36	7,645.54
Min (USD)	13,844.51	1,607.51	16,420.49	1,621.88
Max (USD)	55,135.40	36,910.61	63,770.43	36,580.28
Range (USD)	41,290.89	35,303.10	47,349.94	34,958.40
Skewness	0.27	1.66	0.44	1.54
Kurtosis	1.74	6.98	2.14	6.07
1st Quartile	19,923	4,144	19,567	4,355
3rd Quartile	55,135	11,525	41,887	11,841
Sum (USD)	1,797,866	2,260,161	1,730,218	2,532,816
Std. Error	1,484.41	369.11	1,671.68	364.09

From Table 1 and Table 2, most categories have a kurtosis value of more than 3, except for the category's diabetic smokers and non-diabetic smokers. It can be inferred that most Insurance claim data categories have heavy-tailed distributions. The maximum value of the data is from the category of female ND-S, while the minimum value of the data is from male D-

NS. Besides, the highest value of the first quartile is male ND-S, while the third quartile is female diabetics smokers. The category that has the highest value of sum is the "all of the data" category, however, the lowest value is female ND-S. On the other hand, the female ND-S category has the highest value of the standard error and the lowest value of the standard error is "all of the data". It could be caused by the number of "all of the data" consisting of 1,340 datums and the number of female ND-S consisting of 55 datums. Thus, this implies that the larger number of the data would lead to a small standard error value.

C. RESULT AND DISCUSSION

1. The Goodness of Fit

Let X represent the claim amount for each type of data. As illustrated in Table 1 and Table 2, most insurance claims data categories are skewed to the right and heavy-tailed. The distribution is called a heavy-tailed if and only if its tail function fails to be bounded by any exponentially decreasing function (Dinh et al., 2016). Apart from that, distribution is said to have a heavy right tail if there are only positive moments up to a specific value or no positive moments. Distribution can also be classified as heavy-tailed if there is a decreasing hazard rate function (Klugman et al., 2019).

Table 3. The Probability Density Function and Cumulative Distribution Function of Weibull, Pareto, Lognormal, and Exponential Distributions

X Distribution	Probability density function	Cumulative distribution function
Weibull (θ,τ)	$f(x)=\begin{cases}\frac{\tau\left(\frac{x}{\theta}\right)^{\tau}e^{-\left(\frac{x}{\theta}\right)^{\tau}}}{x}, x\geq 0\\ 0, \quad x<0\end{cases}$	$F(x)=\begin{cases}1-e^{-\left(\frac{x}{\theta}\right)^{\tau}}, x\geq 0\\ 0, \quad x<0\end{cases}$
Pareto (α,θ)	$f(x)=\begin{cases}\frac{\alpha\theta^{\alpha}}{(x+\theta)^{\alpha+1}}, x\geq 0\\ 0, \quad x<0\end{cases}$	$F(x)=\begin{cases}1-\left(\frac{\theta}{x+\theta}\right)^{\alpha}, x\geq 0\\ 0, \quad x<0\end{cases}$
Lognormal (μ,σ^2)	$f(x)=\begin{cases}\frac{1}{x\sigma\sqrt{2\pi}}exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right), x\geq 0\\ 0, \quad x<0\end{cases}$	$F(x)=\begin{cases}\Phi\left(\frac{\ln(x)-\mu}{\sigma}\right), x\geq 0\\ 0, \quad x<0\end{cases}$
Exponential $\left(\frac{1}{\theta}\right)$	$f(x)=\begin{cases}\frac{1}{\theta}e^{-\frac{x}{\theta}}, x\geq 0\\ 0, \quad x<0\end{cases}$	$F(x)=\begin{cases}1-e^{-\frac{x}{\theta}}, x\geq 0\\ 0, \quad x<0\end{cases}$

Insurance data sets are usually positive and have a unimodal shape and are skewed to the right also with heavy-tailed (Afify et al., 2020). These two properties are the properties of heavy-tailed distributions, so to model the extensive insurance claim data, lognormal distributions, exponential distributions, Pareto distributions, and Weibull distributions will be used (Riad et al., 2023; Xie, 2023). The probability density function and cumulative distribution function of those distributions will be shown in Table 3. Maximum likelihood estimation is one of the most critical approaches to estimation in all statistical inference (Walpole et al., 2012). This research uses maximum likelihood estimation to estimate the distribution parameter values based on data. Let $X_1, X_2, \ldots, X_n$ as n sized random sample of a distribution with a parameter θ with unknown value with probability density function $f(x;\theta)$, the likelihood function for the random sample is as follows.

$$L(\theta;x)=\prod_{i=1}^{n} f(x_i;\theta) \tag{1}$$

with $X = (x_1, x_2, \ldots, x_n)$ and $x_1, x_2, \ldots, x_n$ assumed to be independent.

$$l(\theta) = log\ L(\theta; x) = \sum_{i=1}^{n} log\ f(x_i; \theta) \quad (2)$$

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{\partial \sum_{i=1}^{n} log\ f(x_i; \theta)}{\partial \theta} = 0 \quad (3)$$

by deriving the log-likelihood function θ, it will obtain $\hat{\theta}_{MLE}$ (Hogg et al., 2019). Parameter estimates were obtained using the maximum likelihood method from several heavy-tailed distributions for each data category.

The Goodness of Fit test tests the model's suitability or distribution to the data. The Goodness of Fit index shows the difference between the observed value and the expected value from the model (Snipes & Taylor, 2014). This research uses the Goodness of Fit test to show which distribution best fits the data. Some Goodness of Fit tests, such as Kolmogorov-Smirnov and Anderson-Darling, can be used to determine the best data distribution. Kolmogorov-Smirnov is a statistical test that measures the most significant discrepancy between the observed and hypothesized distribution (Zeng et al., 2015). Other than the Kolmogorov-Smirnov and Anderson-Darling test, there are also Goodness of Fit Test criteria for models such as Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). AIC is used to find the best approximation model for the unknown data-generating process.

Using R studio, Akaike's Information Criterion values were obtained for each data category using Goodness of Fit tests. The most suitable distribution from the four heavy-tailed distributions is determined by reviewing the smallest AIC value of each distribution. AIC has the following equation:

$$AIC = -2\ln(L) + 2p \quad (4)$$

with L being the likelihood under a model that is used, and p is the number of parameters of the model. The model or best distribution is the one that has the lowest AIC value. So, the following results are obtained in Table 4 (Snipes & Taylor, 2014)(Zeng et al., 2015).

Table 4. The result of parameter estimation and the AIC's value of Weibull and lognormal distribution

k	Data Categories	Weibull					Lognormal		
		τ_k	θ_k	μ_k	σ_k	AIC_k	μ-log$_k$	σ-log$_k$	AIC_k
1	Male D-S	3.31	36,145.70	32,428.18	10,790.39	932.30	5.71	0.38	941.13
2	Male D-NS	1.49	8,891.43	8,033.29	5,488.10	2,534.31	4.12	0.77	2,537.69
3	Male ND-S	3.41	37,632.47	33,813.48	10,954.31	1,019.86	5.76	0.37	1,028.47
4	Male ND-NS	1.44	9,132.62	8,288.75	5,844.14	2,915.56	4.14	0.77	2,909.72
5	Female D-S	2.90	33,736.57	30,082.55	11,270.68	739.31	5.63	0.40	738.90
6	Female D-NS	1.55	9,615.51	8,648.13	5,697.36	2,799.92	4.22	0.70	2,783.02
7	Female ND-S	2.79	35,465.15	31,575.71	12,247.04	685.84	5.67	0.40	682.53
8	Female ND-NS	1.56	9,984.71	8,974.16	5,877.36	3,041.96	4.26	0.70	3,027.84

Table 4 shows that based on the smallest AIC value, the Weibull distribution is best described the data categories of male insureds who are smokers and suffer from diabetes with scale parameters $\theta_1 = 36{,}145.70$ and shape parameters $\tau_1 = 3.31$, male insureds who are non-smokers but suffer from diabetes with scale parameter $\theta_2 = 8{,}891.43$ and shape parameter $\tau_2 = 1.49$, and male insureds who smoke but do not suffer from diabetes with scale parameter $\theta_3 = 37{,}632.47$ and shape parameter $\tau_3 = 3.41$.

On the other hand, also based on the lowest AIC value, the lognormal distribution is best described the data categories of male insureds who are non-smokers and do not suffer from diabetes with parameters μ_4-log= 4.14 and σ_4-log = 0.77, female D-S insureds with parameters μ_5-log = 5.63 and σ_5-log = 0.40, female insureds who do not smoke but suffer from diabetes with parameters μ_6-log = 4.22 and σ_6-log = 0.70, female insured smokers without diabetes with parameters μ_7-log = 5.67 and σ_7-log = 0.40, and female insureds who are non-smokers without diabetes with parameters μ_8-log = 4.26 and σ_8-log = 0.70.

Beside of the AIC value, if we compared some statistics values such as mean, median, and skewness of the categories based on the estimated parameter obtained from the fitted distribution with statistic descriptive derived in Table 1 and 2, it gave almost similar value. For example of the Weibull distribution that best described the data categories of male diabetic smokers with estimated parameters of $\theta = 36{,}145.70$ and $\tau = 3.31$ have estimated mean of 32,428.18 while the actual mean value is 32,314.22, estimated median of 32,356.96 with the actual value of 35,538.61, and estimated skewness of 0.075 with actual value of 0.08 and skewed to the left.

As presented in Table 4, the highest estimated parameter from the categories that fit with Weibull distribution is male non-diabetic smokers. In contrast, the lowest estimated parameter is male diabetic non-smokers. Those results have aligned with the highest and the lowest mean and standard deviation of that category. Moreover, the male ND-S has 2,797,798.37 USD total amount of claims, which is the highest total amount of claims among other categories. As can be seen in Table 4, for the categories that fit lognormal distribution, the female non-diabetic smoker has the highest mean log but has the lowest standard deviation log. It could happen because female non-diabetic smokers have 1,730,218.38 USD total amount of claims from 55 claims, the lowest value of the total amount of claims among other categories that fit the lognormal distribution. Whereas male ND-NS had the lowest mean log and the highest standard deviation log with the second highest total amount of claims and number of claims among the lognormal fitted categories.

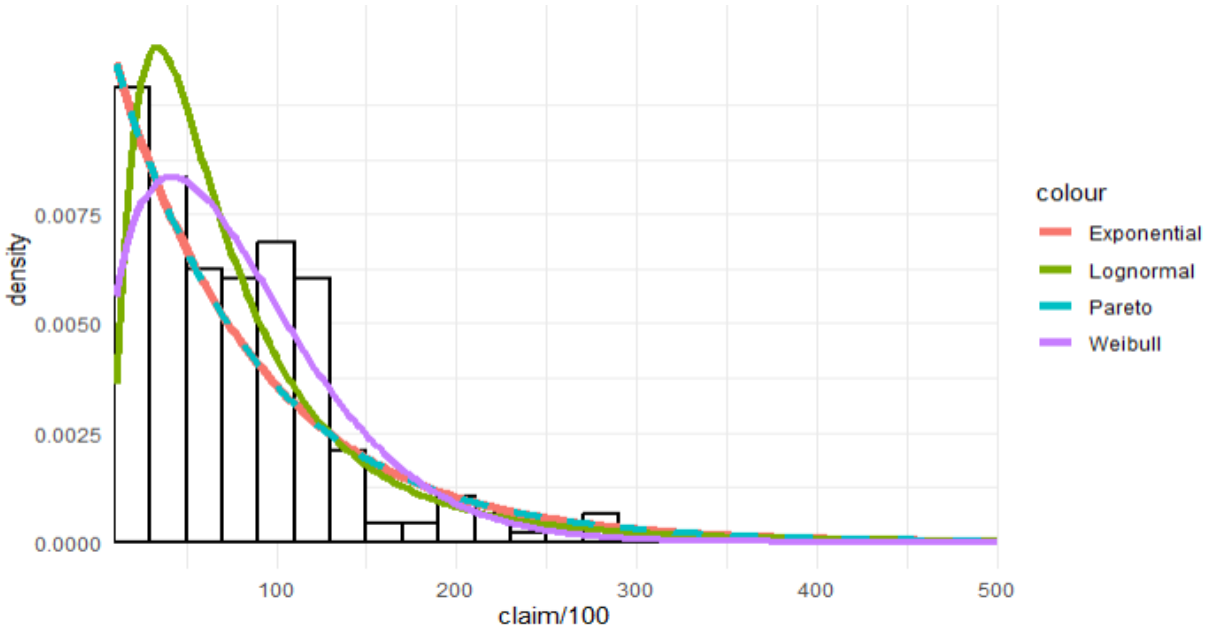

Figure 4. The result of the Goodness of Fit test on the data of male diabetic non-smokers

One of the categories suitable for modelling with the Weibull distribution is data on male insureds who suffer from diabetes and are non-smokers. Figure 4 shows that among the four distributions, the Weibull distribution is the distribution that best suits the shape of the data histogram diagram. The Weibull distribution has two parameters, shape (τ) and scale (θ), which make it possible to describe various shapes of distribution curves, such as highly skewed or symmetrical, depending on the parameters (Wu et al., 2019). The curves of the Weibull distributions become bell-shaped if $\tau > 1$ (Walpole et al., 2012). From the estimated model parameters for male diabetic non-smoker data, the shape parameter τ = 8,891.43 is obtained, which is more than 1, so the data distribution curve is bell-shaped. Due to those characteristics, Weibull distribution is commonly used in many research and applications to establish proper application in insurance, finance, engineering, economics, and biostatistics (Jurić, 2017). It could also be used to model the particle size distribution of a dust sample (Abousrafa et al., 2024) and predict wind conditions and the inverse of wind direction (Aljeddani & Mohammed, 2023). The Weibull distribution is the most widely used as it provides the characteristics of both the exponential and Rayleigh distributions (Ahmad & Hussain, 2017). The Weibull distribution can change exponentially when the shape parameter τ = 1 (Walpole et al., 2012).

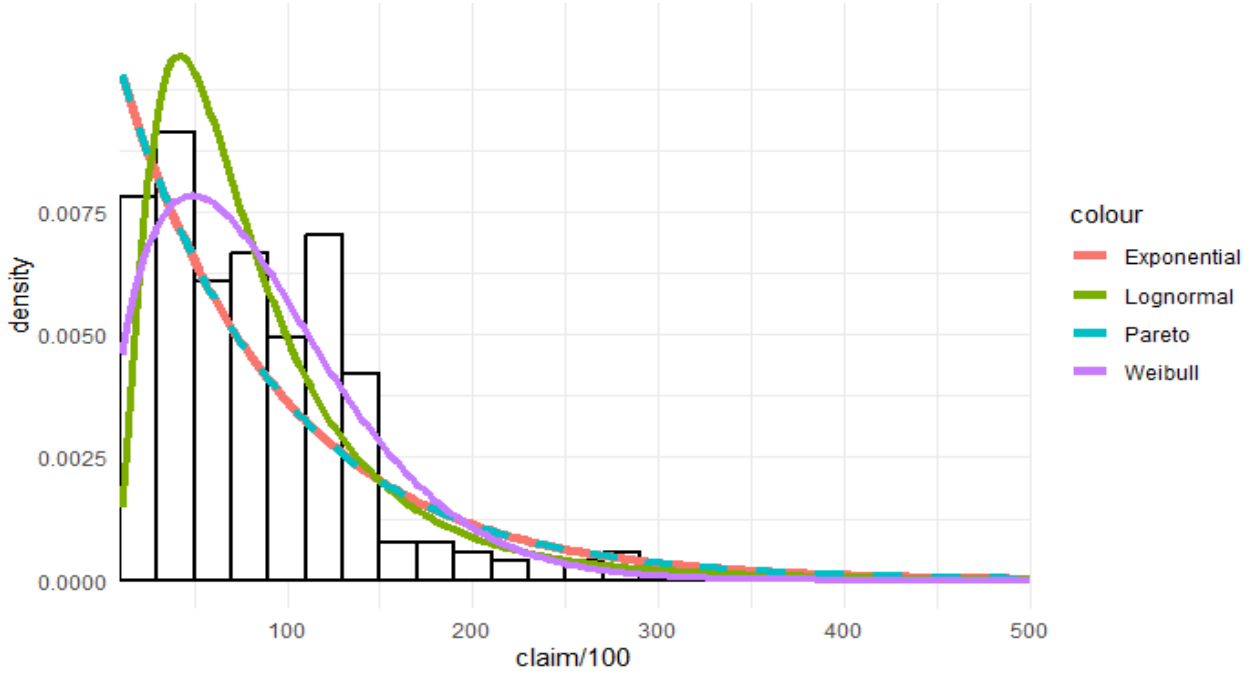

Figure 5. The result of the Goodness of Fit test of female diabetic non-smoker insurer data

On the other hand, data suitable to be modelled using the lognormal distribution is data on female non-diabetic smoker insureds. The histogram in Figure 5 shows that all the data for female diabetic non-smokers insured have long tails, but the graph for exponential and Pareto distribution gave the biggest claim frequency in the smallest claim amount while the Weibull and lognormal distribution gave the quite similar graph to the data's histogram. In contrast, the boxplot of the claims data for female diabetic non-smokers insured in Figure 3 shows that the data has many outliers. The lognormal distribution is more suitable because it can accommodate long tails, which can handle the possibility of large claims that rarely occur but have a significant impact and cover the many outliers in the data. The lognormal distribution is a common approach for modelling long-tailed type data, such as automobile claim data (Nadarajah & Kwofie, 2022). In neuroscience and computer science, the long-tailed distribution of the strength or amplitude of a connection between two nodes can be modelled using the lognormal distribution (Teramae & Fukai, 2014).

2. The Value at Risk

After obtaining a suitable distribution for each group of claims data, the VaR value will be determined for each data category. The VaR is a risk measurement intensively used in business, finance, and insurance. The VaR is defined as a value or amount of capital required or determined to ensure, with a high level of confidence, that a company does not experience a technical bankruptcy (Klugman et al., 2019). The VaR may be defined as the quantile of asset returns distribution conditional on the last observation (Syuhada, 2020). From the result of the goodness of fit test, it is known that the suitable distributions are the Weibull and lognormal. The VaR for the Weibull distribution can be determined by:

$$VaR_p(X) = \theta[-ln(1-p)]^{1/\tau} \tag{5}$$

Then, the VaR value for the lognormal distributions could be obtained with:

$$VaR_p(X) = e^{\mu+Z_p\sigma} \tag{6}$$

Based on those equations, with the high level of confidence (p) of 90%, 95%, 99%, and 99.5% (Obadović et al., 2016) obtained VaR values for each data category.

Table 5. The Value-at-Risk for each data category with respecify distribution clarity in Table 4

Data Categories		Dist.	Amount of Claim	Level of Confidence			
				90% (USD)	95% (USD)	99% (USD)	99.5% (USD)
Male	D-S	Weibull	2,455,880.62	46,488.39	50,330.04	57,302.20	59,778.33
	D-NS		1,920,242.01	15,568.48	18,578.25	24,798.57	27,247.31
	ND-S		2,797,798.37	48,058.47	51,913.55	58,889.32	61,360.81
	ND-NS	Lognormal	2,258,008.66	16,928.49	22,510.37	37,981.56	46,065.45
Female	D-S		1,797,866.19	46,235.45	53,542.61	70,093.25	77,414.95
	D-NS		2,260,161.15	16,707.81	21,606.30	34,637,96	41,224.87
	ND-S		1,730,218.38	48,408.22	56,031.48	73,285.93	80,914.45
	ND-NS		2,532,815.51	17,392.84	22,501.61	36,101.10	42,978.48

It can be seen from Table 5, that based on the level of confidence 90%, 95%, 99%, and 99.5%, the highest VaR in the male categories is VaR of the non-diabetic smokers with Weibull distribution. On the contrary, the lowest VaR is from male diabetic non-smokers with Weibull distribution. On the female categories, it is obvious that all the categories have the same distribution, which is a lognormal distribution. On the level of confidence 90%, 95%, 99%, and 99.5%, the highest VaR in female categories is VaR of non-diabetic smokers while the lowest VaR is diabetic non-smokers. Based on those results, it is interesting that the gender (male and female) of insureds gave the same categories of VaR regarding how the non-diabetic smoker status is more affected than the diabetic non-smoker status. That condition is affected by the awareness of their health condition, because insureds who suffer from diabetes, whether the insureds are smokers or not, usually have more awareness of their health condition and tend to have regular medical check-ups so the amount of claim for insureds who is a diabetic non-smoker has the lowest VaR. Meanwhile, the insureds who are non-diabetic smokers are less aware of their health condition because they feel healthy and ignore the initial symptoms, thus making the insureds who are non-diabetic smokers, whether male or female have the highest VaR. Regarding medical check-ups, the study found that smokers with diabetes are less likely to have visited a hospital clinic or seen a diabetes nurse in the last year compared to non-smokers (Gucciardi et al., 2011).

Generally, the female non-diabetic smoker with a 99.5% level of confidence has the highest VaR at 80,914.45 USD, and the male diabetic non-smoker with a 90% level of confidence has the lowest VaR at 15,568.48 USD. It shows that women with non-diabetic smoker status experience have a high average claim size. This situation is influenced by the metabolism of women's bodies, which are more susceptible to disease than men. In addition, women with non-diabetic smoker status tend to be more ignorant of the early symptoms of disease caused by

smoking, so the claims offered tend to be larger. In contrast, male diabetic non-smokers tend to be more aware of the importance of health care to reduce the risk of complications, so they have a low number of claims but frequently do medical check-ups to diminish the risk of sudden large claims.

3. Claim Insurance Data

After obtaining the VaR for each data category, the graph on Figure 6 compares the VaR for male and female data.

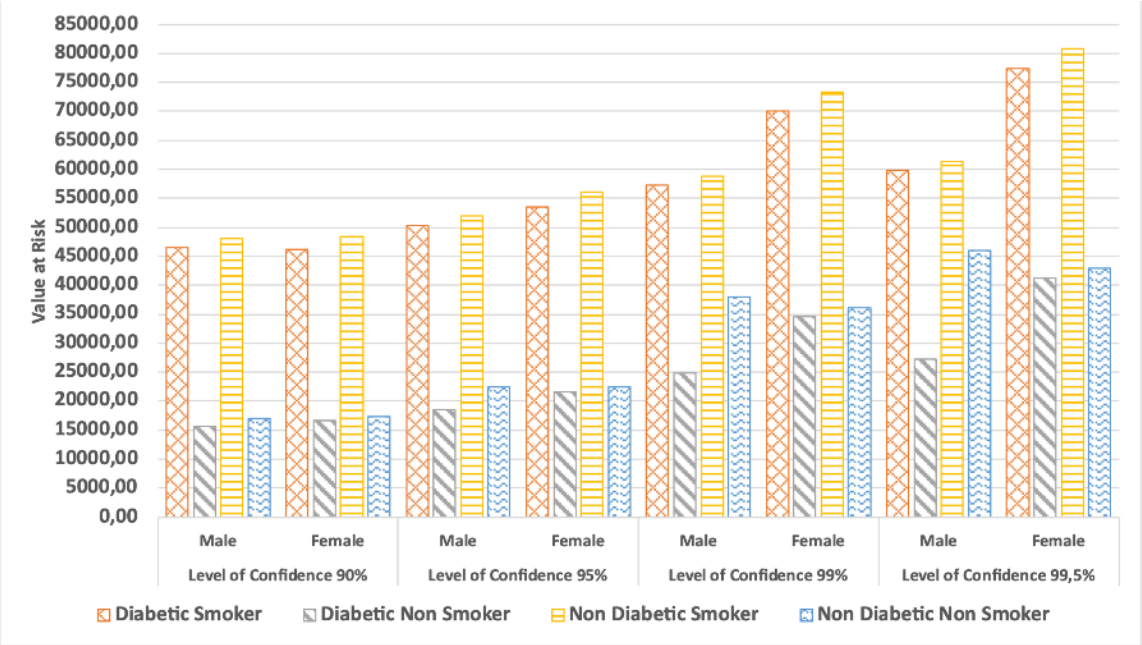

Figure 6. Comparison of the VaR values for each data category

From Figure 6, the factors of gender, diabetic status, smoking status, and the level of confidence used influence the VaR. The greater VaR could be obtain if given a high level of confidence. For male insureds, the higher the level of confidence, the higher the VaR, but not as much as female insureds. It can be seen from Figure 6 that female tend to have higher VaR than male. As for female, the biggest VaR value is 80,914.45 derived from the female with non-diabetic smoker with level of confidence of 99.5%, while for the male the biggest VaR value is 61,360.81 derived from male non-diabetic smoker with level of confidence of 99.5%. There is big increase in VaR with the higher level of confidence for female insureds. It can also be seen that the diabetic status and smoking status factors are divided into four sub-categories. For male insureds, it appears that insureds who do not have diabetes and are smokers have the highest VaR for every level of confidences whether its 90%, 95%, 99%, or 99.5%. It is apparently seen that insured persons who have diabetics and are non-smokers have the lowest VaR. However, for female insureds, insureds who do not have diabetes and are smokers have the highest VaR. Also, the insurers who have diabetics and are non-smokers have the lowest VaR. From these results, it can be concluded that the observed factors influence the VaR, namely gender, diabetic status, and smoking status.

D. CONCLUSION AND SUGGESTIONS

This research used heavy-tailed distributions to fit the data which describes the characteristics of insurance claim data that form a long tail. The heavy-tailed distribution used in this research are Weibull, Pareto, lognormal, and exponential distribution. Based on the results of the conducted research, the following conclusions can be drawn that the most appropriate distribution for modelling insurance claims data across various categories Weibull distribution that best describes the following categories, male diabetic smokers, male diabetic nonsmokers, and male nondiabetic smoker, and lognormal distribution is best describes the following categories, male non-diabetic non-smoker, female diabetic smoker, female diabetic non-smoker, female non-diabetic smoker, and female non-diabetic non-smoker. The Weibull and lognormal distribution are the best fitted distribution for the data based on the lowest AIC's value obtained for each data categories that include gender and disease amongst the distributions used. Besides, comparing the statistics values such as mean, median, and skewness of the data that derived in Table 1 and 2 with value obtained using estimated parameter for the distribution gave quite similar results.

The Value at Risk (VaR) obtained is influenced by the insured's gender, diabetic status, smoking status, and the level of confidence utilized. Specifically, when considering gender alone, the highest VaR value is observed in female claims data with a level of confidence of 99.5%. However, when accounting for gender, smoker status, and diabetic status, the highest VaR value is found in female non-diabetic smoker claims data with a level of confidence of 99.5%. On the contrary, the lowest VaR value was obtained in claims data for male diabetic non-smokers with a level of confidence of 90%. The Value at Risk values obtained can be utilized to identify demographic categories that present higher risk and are more likely to file claims. The higher VaR value indicates greater potential risk and larger potential losses for the company. Consequently, insurance companies must calculate appropriate premiums and reserves based on the VaR results, while implementing effective risk management strategies to mitigate these risks. This research is constrained using dummy data, which limits its applicability. The incorporation of real data in future studies is expected to offer a more accurate assessment of financial risk conditions. Furthermore, the current simplistic approach to VaR analysis could be enhanced by employing a Generalized Linear Model (GLM) to achieve more robust financial risk estimations, providing a more precise quantification of the associated risks.

ACKNOWLEDGEMENT

The authors acknowledge Statistics Research Division, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung for supporting this study. This research is also supported by LPPM ITB through Riset Unggulan ITB 2024, Contract No. 959/IT1.B07.1/TA.00/2024. The authors also thank the reviewers for their insightful comments and recommendations, which have enhanced the article.

REFERENCES

Abousrafa, A., Olewski, T., & Véchot, L. (2024). Use of a two-parameter Weibull distribution for the description of the particle size effect on dust minimum explosible concentration. *Journal of Loss Prevention in the Process Industries*, *88*(2024). https://doi.org/10.1016/j.jlp.2024.105269

Adiyansyah, F., Widodo, V. R., Rais Anwar, Y., & Sari, K. N. (2023). Pemodelan Besar Klaim menggunakan Distribusi Berekor dan Tail-Value-at-Risk (TVaR) pada Data Sampel Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan. *Jurnal Statistika Dan Aplikasinya*, *7*(2). https://doi.org/10.21009/JSA.07207

Afify, A. Z., Gemeay, A. M., & Ibrahim, N. A. (2020). The Heavy-Tailed Exponential Distribution: Risk Measures, Estimation, and Application to Actuarial Data. *Mathematics*, *8*(8), 1276. https://doi.org/10.3390/math8081276

Ahmad, Z., & Hussain, Z. (2017). Flexible Weibull Extended Distribution. *MAYFEB Journal of Materials Science* , *2*(2017), 5–18. https://mayfeb.com/index.php/MAT/article/view/188/188

Aljeddani, S. M. A., & Mohammed, M. A. (2023). A novel approach to Weibull distribution for the assessment of wind energy speed. *Alexandria Engineering Journal*, *78*(18), 56–64. https://doi.org/10.1016/j.aej.2023.07.027

Dinh, V., Si, L., Ho, T., Nguyen, D., & Nguyen, B. T. (2016). Fast learning rates with heavy-tailed losses. In D. Lee, M. Sugiyama, U. Luxburg I. Guyon and, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29 (NIPS 2016)* (pp. 1–9). https://proceedings.neurips.cc/paper_files/paper/2016/file/63923f49e5241343aa7acb6a06a751e7-Paper.pdf

Gray, R. J., & Pitts, S. M. (2012). *Risk Modelling in General Insurance : From Principles to Practice*. Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9781139033756

Gucciardi, E., Mathew, R., Demelo, M., & Bondy, S. J. (2011). Profiles of smokers and non-smokers with type 2 diabetes: Initial visit at a diabetes education centers. *Primary Care Diabetes*, *5*(3), 185–194. https://doi.org/10.1016/j.pcd.2011.03.001

Hogg, R. V., McKean, J. W. , & Craig, A. T. (2019). *Introduction to Mathematical Statistics* (8th edition). Pearson Education. https://minerva.it.manchester.ac.uk/~saralees/statbook2.pdf

Jurić, V. (2017). Univariate Weibull Distributions and Their Applications. *ENTRENOVA - Enterprise Research Innovation Conference*, 451–458. https://ssrn.com/abstract=3282606

Keçeci, N. F., & Sarul, L. S. (2014). Application of Value at Risk (VaR) Models on Insurance Claim Data. *9th International Statistic Days Symposium*, 87–94. http://josunas.selcuk.edu.tr/login/index.php/josunas/article/view/509

Klugman, S. A. , Panjer, H. H. , & Willmot, G. E. (2019). *Loss models: from data to decisions.* (Fourth edition). A John Wiley & Sons, Inc., Publication. https://www.wiley.com/en-nl/Loss+Models%3A+From+Data+to+Decisions%2C+5th+Edition-p-9781119523789

Marambakuyana, W. A., & Shongwe, S. C. (2024). Quantifying Risk of Insurance Claims Data Using Various Loss Distributions. *Journal of Statistics Applications and Probability*, *13*(3), 1031–1044. https://doi.org/10.18576/jsap/130315

Nadarajah, S., & Kwofie, C. (2022). Heavy tailed modeling of automobile claim data from Ghana. *Journal of Computational and Applied Mathematics*, *405*(2022). https://doi.org/10.1016/j.cam.2021.113947

Obadović, M., Petrović, E., Vunjak, N., & Ilić, M. (2016). Assessing the accuracy of delta-normal VaR evaluation for Serbian government bond portfolio. *Economic Research-Ekonomska Istrazivanja* , *29*(1), 475–484. https://doi.org/10.1080/1331677X.2016.1174391

Pahrany, A. D., Aeli, L. W., & Mukhaiyar, U. (2024). Value at Risk Analysis using Automated Threshold Selection Method in Property Insurance. *The 3rd International Conference on Mathematics and Its Applications (ICoMathApp) 2022*, 020022. https://doi.org/10.1063/5.0193668

Riad, F. H., Radwan, A., Almetwally, E. M., & Elgarhy, M. (2023). A new heavy tailed distribution with actuarial measures. *Journal of Radiation Research and Applied Sciences*, *16*(2), 100562. https://doi.org/10.1016/j.jrras.2023.100562

Rohiim, M. D. N., & Mutaqin, A. K. (2023). Pemodelan Data Frekuensi Klaim Asuransi Kendaraan Bermotor untuk Cakupan Third Party Liability Menggunakan Distribusi Poisson-Aradhana.

Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika, *4*(2), 1376–1385. https://doi.org/10.46306/lb.v4i2.414

Shukla, S. K. (2022). *Insurance Claim Analysis: Demographic and Health*. https://www.kaggle.com/datasets/thedevastator/insurance-claim-analysis-demographic-and-health/.

Snipes, M., & Taylor, D. C. (2014). Model selection and Akaike Information Criteria: An example from wine ratings and prices. *Wine Economics and Policy*, *3*(1), 3–9. https://doi.org/10.1016/j.wep.2014.03.001

Syuhada, K., Hakim, A., & Nur'aini, R. (2023). The expected-based value-at-risk and expected shortfall using quantile and expectile with application to electricity market data. *Communications in Statistics: Simulation and Computation*, *52*(7), 3104–3121. https://doi.org/10.1080/03610918.2021.1928191

Teramae, J. N., & Fukai, T. (2014). Computational implications of lognormally distributed synaptic weights. *Proceedings of the IEEE*, *102*(4), 500–512. https://doi.org/10.1109/JPROC.2014.2306254

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). *Probability & Statistics for Engineers & Scientists* (9th Edition). Pearson Education Limited. https://spada.uns.ac.id/pluginfile.php/221008/mod_resource/content/1/ProbabilityStatistics_for_EngineersScientists(9th_Edition)_Walpole.pdf

Wu, M. H., Wang, J. P., & Ku, K. W. (2019). Earthquake, Poisson and Weibull distributions. *Physica A: Statistical Mechanics and Its Applications*, *526*(C). https://doi.org/10.1016/j.physa.2019.04.237

Xie, S. (2023). Modelling auto insurance Size-of-Loss distributions using Exponentiated Weibull distribution and de-grouping methods. *Expert Systems with Applications*, *231*(1), 120763. https://doi.org/10.1016/j.eswa.2023.120763

Yildirim, I. (2015). Financial Risk Measurement for Turkish Insurance Companies Using VaR Models. *Journal of Financial Risk Management*, *04*(03), 158–167. https://doi.org/10.4236/jfrm.2015.43013

Yousof, H. M., Tashkandy, Y., Emam, W., Ali, M. M., & Ibrahim, M. (2023). A New Reciprocal Weibull Extension for Modeling Extreme Values with Risk Analysis under Insurance Data. *Mathematics*, *11*(4), 966. https://doi.org/10.3390/math11040966

Zeng, X., Wang, D., & Wu, J. (2015). Evaluating the Three Methods of Goodness of Fit Test for Frequency Analysis. *Journal of Risk Analysis and Crisis Response*, *5*(3), 178–187. https://www.jracr.com/index.php/jracr/article/view/151

Zhao, W., Gao, Y., & Wang, M. (2022). Measuring liquidity with return volatility: An analytical approach based on heavy-tailed Censored-GARCH model. *North American Journal of Economics and Finance*, *62*(C). https://doi.org/10.1016/j.najef.2022.101774