Modeling Non-life Insurance Claims Using Exponential Log-logistics Distribution

By

Eileen Chebet

145008

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Mathematical Finance and Risk Analytics at Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

June, 2025

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Student's Name: Eileen Chebet

Sign: ______________________ Date: 09/06/2025

Approval

The dissertation of Eileen Chebet was reviewed and approved for examination by the following:

Dr. Stanley Odhiambo Sewe
Senior Lecturer,Institute of Mathematical Sciences,
Strathmore University.

Dr. Godfrey Madigu
Dean, Institute of Mathematical Sciences,
Strathmore University.

Prof. Bernard Shibwabo
Director, Office of Graduate Studies,
Strathmore University.

Abstract

Determining the amount of insurance claim in the insurance industry is a challenging task. This study aims at modelling and estimating loss claims from an insurance company, with a particular focus on extreme event risks. The main objective was to develop a distribution using the density hazard approach, characterized by skewness and heavy tail, to effectively capture both high-frequency small claims and low-frequency large claims. Various parametric distributions have been applied to the insurance loss claims to determine the model that best fits the data.
The Exponential Log-Logistic distribution, characterized by parameters gives the best fit for the data, as indicated by the lowest negative log-likelihood (NLL) value. Comparative analysis highlights the Exponential Log-Logistic distribution's superior performance in modeling heavy-tailed and skewed data, essential for accurate risk assessment in insurance. In contrast, the Log-Normal distribution, while achieving the lowest Bayesian Information Criterion (BIC) value does not fit the data as well in terms of NLL. The Value at Risk (VaR) and Expected Shortfall (ES) metrics further support the effectiveness of the proposed model in predicting risk management. The Exponential Log-Logistic distribution's higher VaR values, and reasonably high ES values underscore its robustness in managing insurance risks.
The findings recommend the Exponential Log-Logistic distribution as the most suitable model for insurance claims data, due to its balance of complexity and fit, offering significant improvements over simpler models like the Exponential and Lomax distributions. This model's ability to accurately represent the distribution of claims supports better risk management and pricing strategies in the insurance industry.

Keywords Skewness, heavy-tailed, density hazard distribution, Insurance Claims, Parametric distributions.

Table of Contents

Declaration and Approval ii

Abstract iii

Table of Contents iv

List of Figures vi

List of Tables vii

Acknowledgment viii

1 Introduction 1
- 1.1 Background of the Study 1
- 1.2 Statement of the Problem 2
- 1.3 Objectives 3
  - 1.3.1 General Objectives 3
  - 1.3.2 Specific Objectives 3
- 1.4 Significance of the Study 3
- 1.5 Scope of the Study 3
- 1.6 Organization of the study. 3

2 Literature Review 4
- 2.1 Modelling of Insurance Claims using Parametric distributions. 4
- 2.2 Application of Skewed and Heavy tailed distribution in modeling claims. 5
- 2.3 Empirical Studies of Insurance claim models in Emerging Economies, Sub-Saharan Countries. 6
- 2.4 Estimation of Insurer's tail risk (Var and Expected Shortfall) 8
- 2.5 Gaps Identified in the Reviewed Literature. 10

3 Methodology 12
- 3.1 Proposed Loss Distribution 12
  - 3.1.1 Exponential Log-logistic density hazard distribution Function 12
  - 3.1.2 Negative Log-Likelihood for Exponential Log-logistic Distribution 16
  - 3.1.3 Bayesian Information Criterion (BIC) 17

3.2 Lomax Distribution . . . 18
3.3 Log-Normal Distribution . . . 19
3.4 Exponential Distribution . . . 20
3.5 Exponential-Pareto Distribution . . . 20
3.6 Empirical Distribution . . . 21

4 Data Analysis and Results **23**
4.1 Data Description . . . 23
4.1.1 Properties of the Proposed Exponential Log-Logistic Distribution Using Simulated Data . . . 24
4.1.2 Parameter Estimates . . . 26
4.1.3 Skewness and Kurtosis . . . 28
4.2 Model Selection . . . 29
4.2.1 Negative Log-Likelihood (NLL) . . . 29
4.2.2 Bayesian Information Criterion (BIC) . . . 30
4.2.3 Value at Risk and Expected Shortfall . . . 32

5 Recommendation and Conclusion **35**

References **39**

Appendix A: Similarity Report **42**

Appendix B: Ethical Clearance Release Letter **43**

List of Figures

4.1 Empirical Plot of Insurance Claims dataset 24
4.2 Plot of Pdf Exponential Log-Logistic distribution with λ= 1,2,3 25

List of Tables

4.1 Summary Statistics of Insurance Claims Data 23
4.2 Model Parameters, Standard Errors, and P-values 26
4.3 Estimated values of NLL for Insurance Claims data 30
4.4 Estimated BIC Values for different models fitted to Insurance Claims data . . . 30
4.5 Estimated values of VAR and ES for Insurance Claims data 32

Acknowledgments

I want to thank God for this opportunity to pursue this study.I sincerely appreciate my parents for their unwavering love and support throughout my journey. My heartfelt thanks go to my supervisor, Dr. Stanley Odhiambo Sewe, for his invaluable guidance and willingness to share his expertise during the research process. I also extend my gratitude to my classmates for their encouragement and moral support. Lastly, I am thankful to Strathmore University and its dedicated team for allowing me to conduct this research and for their continuous support.

Chapter 1

Introduction

1.1 Background of the Study

The insurance industry plays a crucial role in managing cash flow across different sectors of the economy while offering financial protection for both individuals and property. The process of insurance entails risk transfer for regular premium payments, enabling the distribution of financial liability among a larger group. To accurately estimate insurance claims, researchers have developed various modeling techniques to enhance risk assessment and claim predictions.

Actuaries have shown significant interest in calculating loss distributions from insurance data.Frees and Valdez (2008) emphasize that accurate estimation of claims is essential for insurance companies to make precise pricing predictions and to estimate their future financial obligations effectively. This understanding helps in assessing the impact of claims on the company's financial health. However, due to the complex nature of insurance data, identifying the correct distribution to model real-world insurance data is challenging.

According to Frees and Valdez (2008), In actuarial science, comprehending risk is essential.Accurately estimating insurance claims allows companies to make accurate predictions for pricing and future liabilities, helping them assess the impact of claims on their financial stability. However, selecting the right distribution to model real insurance data is challenging due to the data's complexity and unique characteristics.

Although the normal distribution is widely used in statistics for its simplicity, it may not be ideal for modeling financial and insurance variables, which often display sharp peaks and heavy tails. In non-life insurance policies, claim size data are typically highly skewed, exhibiting significant kurtosis and extreme tail behavior. Therefore, it is essential for researchers to utilize methods that can accurately capture and model insurance claims using skewed distributions.

Azzalini (1985) discussed the skew-normal and skew-t distributions, which are excellent at modeling the skewness and kurtosis of the data but cannot fit bimodal data.Eling (2012) employed two well-known models that is skew-normal and skew-Student t distributions which are competitive for representing insurance data when compared to various models.

To model bivariate claims using data from the Spanish auto insurance sector,Bolance, Zamora, and Martínez-Cob (2008) demonstrated that the skew-normal and log-skew-normal distributions can be used. Ahn, Brook, Howell, Liu, and Zhang (2012) proposed the log-phase-type distribution as a parametric alternative for fitting heavy-tailed data. To examine how suitable skew-logistic and skew-normal distributions fit insurance claims,Kazemi and Noorizadeh (2015) used data sets in the insurance industry.

The composition method offers an approach for getting new flexible heavy-tailed families of distributions that give reasonably good fit for heavy-tailed data.However, distributions generated through this method often involve more than three parameters, making the estimation process more complex. A study by Burnecki, Embrechts, and Jaworski (2005) on standard claims distributions highlighted the presence of small, medium, and large claims—features that make it difficult to fit a single parametric analytical distribution. To address this, Chen, Joe, and Xu (2008) utilized generalized hyperbolic distributions for modeling insurance data,while Frees and Valdez (2008) applied skew distributions to model the conditional distribution of claims based on their number and type. This research examines the modeling of non-life insurance claims from a Kenyan insurer using parametric distributions.

1.2 Statement of the Problem

There is a lack of a flexible parametric distribution to effectively model low-probability outcomes in non-life insurance claims. Existing approaches, including mixture, composite, compound, and heavy-tailed distributions, often require numerous parameters, making them complex to implement and reliant on numerical approximations. To address this challenge, we need a distribution that is parsimonious in terms of parameters, mathematically tractable in terms of computation of moments and that it still captures skewness and heavy tail nature of the insurance data.

The lack of prior studies using datasets from emerging economies like Kenya poses a major challenge in modeling insurance claims. This study seeks to bridge this gap by examining non-life insurance data from a Kenyan insurance company.

The proposed distributions will be used to calculate distribution moments and evaluate tail risk measures, such as Value at Risk (VaR) and Tail Value at Risk (TVaR).

1.3 Objectives

1.3.1 General Objectives

The main objective of this study is to model non-life insurance claims using exponential log-logistic distribution.

1.3.2 Specific Objectives

1. Formulate a flexible parametric loss distribution using the density hazard approach.
2. Apply and analyse the proposed distribution by fitting it to real-world medical insurance claims data and assess its performance.
3. Evaluate the model's capacity to capture tail risks by computing and interpreting Value at Risk (Var) and Expected Shortfall (ES).

1.4 Significance of the Study

The financial impact of an accident, whether involving vehicle repairs or medical expenses, can be substantial, making insurance coverage essential. To manage this risk effectively, insurance companies must utilize an appropriate statistical distribution to analyze past and present claim data. This enables them to forecast potential future events that may lead to significant claims. Accurate modeling supports insurers in setting aside reserves, determining premiums, pricing products, and selecting the right reinsurance policies.

1.5 Scope of the Study

This study focuses on the non-life insurance market, specifically the medical insurance sector. The study will analyze comprehensive data on claim frequency, related costs, and key risk factors affecting these claims.

1.6 Organization of the study.

This study is organized into five key chapters. Chapter One provides an introduction, including background information, the problem statement, the study's objectives, significance, and scope. Chapter Two reviews existing literature to explore various perspectives and challenges in modeling insurance claims using different parametric distributions. Chapter Three outlines the research methodology, detailing the approaches used to model insurance claims and achieve the study's goals. Chapter Four presents data analysis and discusses the results obtained from the modeling process. Lastly, Chapter Five summarizes the findings, provides recommendations, and concludes with a bibliography.

Chapter 2

Literature Review

2.1 Modelling of Insurance Claims using Parametric distributions.

Modeling insurance claims is a crucial aspect of actuarial practice, as it helps life insurers estimate customer mortality rates, assists auto insurers in determining claims probabilities, and enables pension fund trustees to forecast contributions and investments to meet future obligations:

However, applying parametric distributional models to data can be complex. A model that works well for one dataset may not always be the best fit for other types of data. Due to this limitation, researchers are increasingly focused on developing more advanced models that can better accommodate diverse datasets.

Parametric models have been applied across various fields of study. Examples include modeling human lifetimes Reed (2011), predicting pickle harvests using feedforward neural networks Adams and Adams (1999), analyzing breast cancer progression (Arduino, Novarino, Battaglia, & Bellotti, 2012), describing the distinct dielectric properties of human tissues Gabriel, Lau, and Gabriel (1996), estimating the unsaturated hydraulic properties of multiphase flow in porous media Luckner, Van Genuchten, and Nielsen (1989), designing a pragmatic cyber-physical system García-Valls, Perez-Palacin, and Mirandola (2018), modeling wind turbine power curves Taslimi-Renani, Ting, Khosravi, and Haghifam (2016), evaluating the efficiency of Indian commercial banks Silva, Kadiyala, Akhtar, and Wu (2018), and predicting house prices Montero, Mínguez, and Fernandez-Aviles (2018).

Meyer (2005) applied actuarial modeling techniques to fit a loss distribution to a dataset of 250 claims. He tested Gamma, Log-normal, and Weibull distributions, using maximum likelihood estimation to determine their parameter values. His methodology was based on a Bayesian approach.

Similarly, Wright (2005) employed actuarial modeling procedures to analyze 490 claim amounts collected over seven consecutive years. He utilized maximum likelihood estimation to fit the loss distribution and assessed the goodness of fit using P-P plots and the Kolmogorov-

Smirnov test. The distributions he examined included Pareto, Burr, Inverse Pareto, Inverse Burr, and Log-normal.

Guiahi (2000) conducted a study on approaches for fitting different statistical distributions to insurance data samples. His research focused on a dataset where the underlying distribution was Log-normal. He applied the maximum likelihood estimation method to determine model parameters, using the Akaike Information Criterion (AIC) as the selection criterion.Similarly, Adeleke and Ibiwoye (2011) analyzed claim sizes in personal line non-life insurance employing the method of moments for parameter estimation. Instead of the Kolmogorov-Smirnov (K-S) test, they used the chi-square test to evaluate the goodness of fit. The loss distributions examined in their study included Exponential, Weibull, Gamma, and Log-normal.In developing nations such as Kenya, there is limited research on modeling insurance claims with parametric models, making it an important area of interest for researchers.

2.2 Application of Skewed and Heavy tailed distribution in modeling claims.

Actuaries have extensively studied heavy-tailed distributions to address various issues in financial portfolio theory and risk management. Research by Kazemi and Noorizadeh (2015) indicates that insurance loss data often exhibit key characteristics such as right skewness, a unimodal shape, and a significantly heavy right tail. A distribution that accounts for these features is essential for accurately modeling insurance losses and assessing business risk. Distributions such as Pareto, Beta, Burr, Log-normal, and Weibull are among the most effective models for predicting insurance losses.

Although classical distributions offer several advantages, they also have limitations, particularly in their flexibility to accurately model heavy-tailed datasets. For instance, while the Pareto model is widely used for financial data, it sometimes fails to provide an optimal fit for various financial applications. Similarly, the Weibull distribution is not well-suited for modeling extreme losses, as it primarily captures the behavior of smaller losses effectively.

Additionally, the lognormal and beta distributions pose challenges due to the absence of closed-form expressions in their distribution functions, making it difficult to derive essential mathematical properties and limiting their applicability in financial data analysis. Given the critical role of heavy-tailed distributions in actuarial science, researchers continue to explore new statistical models to overcome the shortcomings of conventional distributions.

This growing interest has led to the development of more flexible heavy-tailed distributions. Several approaches have been proposed, including combining multiple distributions, applying variable transformations, finite mixture models, and compounding distributions, all aimed at improving adaptability in modeling financial and insurance datasets.

Abubakar, Hamzah, Maghsoudi, and Nadarajah (2015) introduced several new composite models based on the Weibull distribution to analyze heavy-tailed insurance loss data. Their study applied these models to two real insurance loss datasets and evaluated their fit using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Their research built upon the work of Cooray and Ananda (2005), who modeled the head and tail of distributions using a weighted Lognormal and Pareto distribution while ensuring continuity and differentiability at the threshold θ.

The study employed the maximum likelihood estimation method to determine model parameters, with AIC and BIC used to validate the models' goodness of fit. The distributions considered included Exponential, Pareto, Weibull, Gamma, and Lognormal, while the composite models examined were Exponential-Pareto, Lognormal-Pareto, and Weibull-Pareto.

Composite models have shown strong performance in modeling insurance claim data that exhibit high frequency at the head and extreme values in the tail. Additionally, predictive modeling with mixture models has been successful in capturing heavy-tailed insurance claims. For instance, Bernardi, Maruotti, and Petrella (2012) analyzed a mixture model incorporating skew-normal components on Danish fire insurance loss data, yielding promising outcomes.

Miljkovic and Grun (2016) explored mixture models incorporating components based on Burr, Gamma, Inverse Burr, Inverse Gaussian, and Lognormal distributions, demonstrating their flexibility in modeling Danish fire insurance loss data. Similarly, Punzo, Mazza, and Maruotti (2018) investigated a mixture model that included contaminated Gamma distributions—each consisting of two unimodal Gamma distributions—to analyze workers' compensation and healthcare expenditure data. Additionally, Punzo (2019) applied a mixture model with reparametrized Inverse Gaussian components to study bodily injury claims and the income distribution of Italian households.

2.3 Empirical Studies of Insurance claim models in Emerging Economies, Sub-Saharan Countries.

Empirical research on insurance claim models has been carried out in developing nations, particularly in sub-Saharan Africa. These studies aim to provide a comprehensive understanding of the trends and characteristics of insurance claims across different countries and to develop accurate modeling techniques.
Several studies have explored insurance claim modeling in Africa. Some notable examples include:

Moyo and Ogunyemi (2018) examined the modeling of motor insurance claim frequency and severity in South Africa. The study utilized Poisson regression models to assess claim frequency and gamma regression models to evaluate claim severity. The results indicated that younger policyholders, those with shorter insurance periods, and owners of higher-value vehicles had a higher frequency of claims. Conversely, claim severity was found to be greater among older policyholders and those with more expensive vehicles.

The research also explored how various factors, including policyholder age, gender, vehicle age, and policy type, impact the frequency and severity of motor insurance claims. The findings revealed that these variables played a significant role in influencing both claim frequency and severity. Overall, the study offers valuable insights into the key determinants of motor insurance claims in South Africa, which can help insurers refine their risk models and pricing strategies.

Mohammed and Dzogbede (2021) carried out a study on forecasting health insurance claims in Ghana using machine learning techniques. The researchers employed four machine learning models—decision tree, random forest, support vector machine (SVM), and logistic regression—to estimate the probability of a health insurance claim being filed. Their findings indicated that the random forest algorithm demonstrated superior accuracy and predictive performance compared to the other three models.

The study also analyzed how various factors influence the probability of filing a health insurance claim, including the policyholder's age, gender, and policy type. The results indicated that age and policy type significantly affected claim likelihood, while gender had no substantial impact.

Studies on insurance claims data in developing countries, particularly in sub-Saharan Africa, indicates that such data often exhibit distinctive characteristics, including skewness, heavy tails, and multimodality. To address these features, scholars have utilized various approaches, including lognormal and gamma distributions, non-parametric techniques such as kernel density estimation, and advanced models like Bayesian hierarchical models and copula models.

The findings from these studies suggest that the selection of a claims model should be guided by the characteristics of the data and the assumptions regarding the underlying data-generating process. Additionally, assessing model fit is crucial to ensure that the chosen model accurately represents the data and produces reliable results. These insights highlight the necessity for further research on insurance claim modeling in developing nations and sub-Saharan Africa to better understand the unique challenges and opportunities in these regions.

2.4 Estimation of Insurer's tail risk (Var and Expected Shortfall)

Insurers utilize Value at Risk (VaR) and Expected Shortfall (ES) to estimate potential financial losses. VaR is a statistical measure that represents the maximum expected loss an insurer could face within a specific time frame at a given confidence level. For instance, a 95% VaR implies that the insurer can be 95% confident that losses will not exceed a particular amount over the chosen period. VaR is derived from historical data and is computed based on a predefined time horizon and confidence level to estimate possible losses.

Expected Shortfall (ES), also referred to as Conditional Value at Risk (CVaR), offers a more comprehensive assessment of potential losses compared to VaR. It quantifies the expected loss beyond the VaR threshold, assuming that losses exceed this level. ES provides a deeper insight into the potential financial risks an insurer may face, particularly under extreme market conditions.

Insurers employ VaR and ES to assess potential losses and allocate reserves accordingly. For instance, these measures help determine the necessary capital reserves required to cover potential losses arising from extreme events, ensuring financial stability and risk management.

VaR and ES can be computed using either parametric or non-parametric distribution techniques.

Parametric methods are based on the assumption that data follows a specific probability distribution, such as the normal distribution. With this assumption, Value at Risk (VaR) and Expected Shortfall (ES) can be calculated using analytical formulas.

In contrast, non-parametric methods do not rely on a predefined probability distribution but instead use historical data to estimate potential loss distributions. Examples of non-parametric techniques include historical simulation, Monte Carlo simulation, and bootstrapping. These methods generate random scenarios and leverage past data to approximate possible losses.

Several studies have explored Value at Risk (VaR) and Expected Shortfall (ES) within the insurance sector.

Liu and Pan (2010) assessed the effectiveness of Value at Risk (VaR) and Expected

Shortfall (ES) in forecasting losses during the 2008 financial crisis. The findings indicated that VaR frequently underestimated tail risk and failed to capture extreme events, resulting in significant financial losses. In contrast, ES demonstrated greater reliability in accounting for extreme tail risks and offering a more accurate estimate of expected losses. The authors proposed that ES could be a more suitable risk management tool in situations where extreme events are a primary concern. However, they acknowledged that both VaR and ES have inherent limitations and should be complemented with other risk management approaches.

Krzysko and Krzysko (2012) investigated the performance of parametric and non-parametric techniques for estimating VaR and ES. The findings indicated that non-parametric methods, such as bootstrapping and Monte Carlo simulation, were more accurate than parametric approaches, particularly when the data distribution was unknown or deviated significantly from assumed distributions. Additionally, the effectiveness of non-parametric methods depended on factors such as data sample size, skewness, kurtosis, and the underlying data-generating process. The study concluded that the selection of an appropriate VaR and ES estimation method should be guided by the specific requirements of risk managers and the nature of the data.

Peng and Huang (2012) introduced the use of elliptical copulae to estimate VaR and ES. Copulae are statistical tools used to model dependencies between random variables, with elliptical copulae being particularly suitable for financial risk management.

The study found that the elliptical copula method provided superior performance compared to traditional approaches, such as historical simulation and Monte Carlo simulation, in estimating Value at Risk (VaR) and Expected Shortfall (ES) for both univariate and multivariate cases. Furthermore, the approach demonstrated greater robustness across different distributions and dependency structures.

The authors concluded that the elliptical copula technique is a valuable tool for financial risk management, especially in cases where variable dependencies are intricate or uncertain. However, they emphasized the need for careful calibration and validation, recommending that it be used alongside other risk management methodologies.

Although VaR and ES are widely applied in financial risk management, alternative or complementary approaches exist. Some of these include:

Stress Testing: This technique involves simulating extreme scenarios to evaluate their potential impact on a portfolio or financial institution. Stress testing enhances risk evaluation by incorporating non-linear effects and extreme market scenarios that VaR and ES might overlook.

Tail Risk Measures: These metrics specifically target extreme losses in the tail of the return distribution. Examples include Conditional Value at Risk (CVaR), Tail Conditional Expectation, and Expected Tail Loss.

Risk Budgeting: This approach assigns risk limits or budgets to different parts of a portfolio or institution, ensuring a structured and consistent risk management framework.

Machine Learning: Advanced machine learning algorithms can detect patterns and relationships in large datasets that traditional statistical models might overlook. These techniques help identify potential risks and opportunities.

Bayesian Methods: Bayesian approaches use probabilistic frameworks to estimate risk, incorporating prior knowledge and expert judgment. These methods are particularly useful when data is scarce or uncertain.

It is essential to recognize that no single risk assessment technique is universally applicable. The most effective risk management frameworks often integrate multiple approaches to enhance robustness and comprehensiveness.

Insurers may adopt either parametric or non-parametric methods to estimate VaR and ES. While parametric techniques are typically faster and computationally efficient, they may lack accuracy if the data does not adhere to a known probability distribution. Non-parametric methods, on the other hand, offer greater flexibility and can accommodate diverse datasets, though they may require higher computational resources and larger datasets.

In conclusion, Insurers can calculate VaR and ES using either parametric or non-parametric distribution methods. The selection of a method depends on the characteristics of the data, the required level of accuracy, and computational efficiency.

2.5 Gaps Identified in the Reviewed Literature.

The literature review emphasizes that many insurance companies face considerable difficulties in choosing a suitable model for estimating insurance claims, especially when dealing with unforeseen extreme events. As a result, modeling insurance claims is a crucial task that insurers must undertake to mitigate major financial risks. This study aims to identify a suitable model for analyzing extreme insurance claims.

Furthermore, limited research has been conducted on modeling unforeseen extreme events in sub-Saharan countries. This study will concentrate on modeling insurance claims data from a sub-Saharan nation, with a particular focus on Kenya.

Chapter 3

Methodology

Various methods can be utilized for modeling insurance claims. In this study, flexible parametric distributions will be utilized and their outcomes compared to the proposed loss distribution model. These distributions will be applied to non-life insurance data, with a focus on medical claims.

3.1 Proposed Loss Distribution

3.1.1 Exponential Log-logistic density hazard distribution Function

According to the study by Bakar, Nadarajah, and Ngataman (2022) on a family of density-hazard distributions for insurance losses, the authors introduced the concept of density-hazard distributions to develop new distributional models. The density-hazard distribution is a composite function that combines two distinct components: a density function and a hazard rate function, expressed using the following formula:

$$f(x) = g(-\log[\overline{H}(x)])\mu(x) \tag{3.1}$$

for $x > 0$, where $\mu(x) = \frac{h(x)}{\overline{H}(x)}$ is a hazard rate function, $\overline{H}(x) = 1 - H(X)$ is a survival function and H(x) is a cumulative distribution function (CDF).

In this formulation, $g(x)$ and $h(x)$ represent the head and tail distributions, respectively. In the study, the head distribution $g(x)$ was modeled using the Exponential distribution, while the tail distribution $h(x)$ was represented by Log-normal, Inverse Weibull, Inverse Pareto, Paralogistic, and Inverse Paralogistic distributions.

In the proposed density-hazard distribution, the Exponential distribution will be used as the head distribution $g(x)$, as it is a thin-tailed distribution, whereas the Log-logistic distribution will serve as the tail distribution $h(x)$, which is capable of capturing heavy-tailed behavior.

The density-hazard distribution produces a new probability density function $f(x)$ that is unimodal, skewed, and fat-tailed, making it suitable for modeling both small claims with high frequency and large claims with low frequency.

The probability density function (PDF) of the Exponential distribution is given by:

$$g(x) = \lambda e^{-\lambda x} \tag{3.2}$$

where $\lambda > 0$ and $x > 0$.
The CDF of Exponential distribution is given by:

$$G(x) = 1 - e^{-\lambda x} \tag{3.3}$$

The survival function:

$$\overline{G}(x) = e^{-\lambda x} \tag{3.4}$$

The PDF of Log-logistic distribution is given by:

$$h(x) = \begin{cases} \frac{\frac{k}{\beta}\left(\frac{x}{\beta}\right)^{k-1}}{\left(1+\left(\frac{x}{\beta}\right)^{k}\right)^{2}} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

The CDF of Log-logistic distribution is given by:

$$H(x) = \begin{cases} \frac{1}{1+\left(\frac{x}{\beta}\right)^{-\kappa}} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

The survival function of Log-logistic distribution is:

$$\overline{H}(x) = \begin{cases} \frac{1}{\left(\frac{x}{\beta}\right)^{\kappa}+1} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

To find the hazard rate function of log-logistic distribution, we divide two expression:
Hazard rate function $\mu(x)$:

$$\mu(x) = \frac{h(x)}{\overline{H}(x)}$$

$$h(x) = \begin{cases} \frac{\frac{k}{\beta}\left(\frac{x}{\beta}\right)^{k-1}}{\left(1+\left(\frac{x}{\beta}\right)^{k}\right)^{2}} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\overline{H}(x) = \begin{cases} \frac{1}{\left(\frac{x}{\beta}\right)^{\kappa}+1} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The hazard rate function for the log-logistic distribution is given by:

$$\mu(x) = \frac{k\left(\frac{x}{\beta}\right)^{k-1}}{\beta\left(1+\left(\frac{x}{\beta}\right)^{k}\right)}$$

The pdf of the exponential log-logistic distribution is given by:

$$f(x) = g(-\log[\overline{H}(x)]) \cdot \mu(x)$$

where:

$$g(x) = \lambda e^{-\lambda x}$$

$$\overline{H}(x) = \frac{1}{\left(\frac{x}{\beta}\right)^{\kappa}+1}$$

$$\mu(x) = \frac{k\left(\frac{x}{\beta}\right)^{k-1}}{\beta\left(1+\left(\frac{x}{\beta}\right)^{k}\right)}$$

Substituting these expressions into the formula for $f(x)$:

$$f(x) = \lambda e^{-\lambda(-\log[\overline{H}(x)])} \cdot \frac{k\left(\frac{x}{\beta}\right)^{k-1}}{\beta\left(1+\left(\frac{x}{\beta}\right)^{k}\right)}$$

$$= \lambda e^{\lambda \log[\overline{H}(x)]} \cdot \frac{k\left(\frac{x}{\beta}\right)^{k-1}}{\beta\left(1+\left(\frac{x}{\beta}\right)^{k}\right)}$$

Now, substituting the expression for $\overline{H}(x)$:

$$= \lambda e^{-\lambda \log\left[\left(\frac{x}{\beta}\right)^{\kappa}+1\right]} \cdot \frac{k\left(\frac{x}{\beta}\right)^{k-1}}{\beta\left(1+\left(\frac{x}{\beta}\right)^{k}\right)}$$

$$f(x) = \lambda \left(\frac{1}{\left(\frac{x}{\beta}\right)^{\kappa} + 1} \right)^{\lambda} \cdot \frac{k\left(\frac{x}{\beta}\right)^{k-1}}{\beta\left(1 + \left(\frac{x}{\beta}\right)^{k}\right)}$$

$$f(x) = \frac{\lambda k}{\beta} \cdot \frac{\left(\frac{1}{\left(\frac{x}{\beta}\right)^{\kappa} + 1} \right)^{\lambda} \cdot \left(\frac{x}{\beta}\right)^{k-1}}{\left(1 + \left(\frac{x}{\beta}\right)^{k}\right)}$$

$$f(x) = \frac{\lambda \kappa}{\beta} \cdot \frac{\left(\frac{1}{\left(\frac{x}{\beta}\right)^{\kappa} + 1} \right)^{\lambda} \cdot \left(\frac{x}{\beta}\right)^{\kappa-1}}{\left(1 + \left(\frac{x}{\beta}\right)^{\kappa}\right)}$$

$$f(x) = \frac{\lambda \kappa}{\beta} \cdot \frac{\left(\frac{x}{\beta}\right)^{\kappa-1}}{\left(1 + \left(\frac{x}{\beta}\right)^{\kappa}\right)^{\lambda+1}} \tag{3.8}$$

Non-Negativity of the ELL Distribution

A valid probability density function (PDF) must satisfy the non-negativity condition over its domain:

$$f(x;\beta,\kappa,\lambda) \geq 0 \quad \forall x > 0 \tag{3.9}$$

For the ELL distribution, the PDF is given by:

$$f(x;\beta,\kappa,\lambda) = \frac{\lambda \kappa}{\beta} \left(\frac{1}{\left(\left(\frac{x}{\beta}\right)^{\kappa} + 1\right)} \right)^{\lambda} \left(\frac{x}{\beta}\right)^{\kappa-1} \cdot \frac{1}{\left(1 + \left(\frac{x}{\beta}\right)^{\kappa}\right)}, \quad x > 0 \tag{3.10}$$

The optimized parameter values are:

$$\beta = 8.404736418003854$$
$$\kappa = 4.936790363043883$$
$$\lambda = 0.05431541186503189$$

Each component of the PDF expression is positive for $x > 0$, $\beta > 0$, $\kappa > 0$, and $\lambda > 0$.

Therefore, the PDF is non-negative over its domain:

$$f(x) \geq 0 \quad \text{for all } x > 0 \tag{3.11}$$

Thus, this property is satisfied.

Normalization (Total Probability Equals 1)

A valid probability density function (PDF) must integrate to 1 over its entire domain:

$$\int_0^\infty f(x;\beta,\kappa,\lambda)\,dx = 1 \tag{3.12}$$

This condition was verified numerically using the `quad` function from `scipy.integrate`. The integration of the fitted PDF over a large interval $[0, 1000]$ yielded:

$$\int_0^\infty f(x)\,dx \approx 1.000 \quad \text{(within numerical tolerance)} \tag{3.13}$$

This confirms that the ELL function is properly normalized.

Thus, the proposed Exponential Log-Logistic (ELL) function satisfies both required properties of a valid probability distribution.

3.1.2 Negative Log-Likelihood for Exponential Log-logistic Distribution

The probability density function (PDF) of the Exponential Log-logistic distribution is given by:

$$f(x) = \frac{\lambda\kappa}{\beta} \cdot \frac{\left(\frac{x}{\beta}\right)^{\kappa-1}}{\left(1+\left(\frac{x}{\beta}\right)^{\kappa}\right)^{\lambda+1}} \tag{3.14}$$

The models under consideration consist of three parameters, and their performance will be evaluated using the negative log-likelihood. A model with a lower negative log-likelihood value will be preferred, as it indicates a better fit to the data. The negative log-likelihood for the density-hazard distribution is given by the following formula:

$$NLL(\theta) = -\sum_{i=1}^{n} \log(f(x_i;\theta))$$

The NLL for Exponential Log- Logistic distribution will be given by

$$L(\theta|x) = -\sum_{i=1}^{n} \log\left(\frac{\lambda\kappa}{\beta} \cdot \frac{\left(\frac{x}{\beta}\right)^{\kappa-1}}{\left(1+\left(\frac{x}{\beta}\right)^{\kappa}\right)^{\lambda+1}}\right)$$

$$L(\theta|x) = -\left(n\log\left(\frac{\lambda\kappa}{\beta}\right) + (\kappa-1)\sum_{i=1}^{n}\log\left(\frac{x_i}{\beta}\right) - (\lambda+1)\sum_{i=1}^{n}\log\left(1+\left(\frac{x_i}{\beta}\right)^{\kappa}\right)\right) \tag{3.15}$$

The entire expression is negated by the outermost negative sign, making it a negative log likelihood. This kind of expression is often used in statistical and machine learning contexts, where θ represents parameters and x represents observed data. The goal is to maximize this likelihood to find the most likely parameters given the observed data.

3.1.3 Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) is a model selection criterion that evaluates model fit while penalizing excessive complexity. For the Exponential Log-Logistic distribution, the BIC is computed using the following formula:

$$BIC = 2\cdot\mathrm{NLL} + k\cdot\log(n) \tag{3.16}$$

where the likelihood function for the Exponential Log-Logistic distribution is:

$$L(\theta|x) = -\left(n\log\left(\frac{\lambda\kappa}{\beta}\right) + (\kappa-1)\sum_{i=1}^{n}\log\left(\frac{x_i}{\beta}\right) - (\lambda+1)\sum_{i=1}^{n}\log\left(1+\left(\frac{x_i}{\beta}\right)^{\kappa}\right)\right)$$

and $k = 3$ is the number of parameters in the model.

Substituting these values into the BIC formula, we get:

$$BIC = -2n\log\left(\frac{\lambda\kappa}{\beta}\right) - 2(\kappa-1)\sum_{i=1}^{n}\log\left(\frac{x_i}{\beta}\right) + 2(\lambda+1)\sum_{i=1}^{n}\log\left(1+\left(\frac{x_i}{\beta}\right)^{\kappa}\right) + 3\cdot\log(n)$$

$$BIC = -2n\log\left(\frac{\lambda\kappa}{\beta}\right) - 2(\kappa-1)\sum_{i=1}^{n}\log\left(\frac{x_i}{\beta}\right) + 2(\lambda+1)\sum_{i=1}^{n}\log\left(1+\left(\frac{x_i}{\beta}\right)^{\kappa}\right) + 3\cdot\log(n) \tag{3.17}$$

The BIC penalizes models with a higher number of parameters, favoring simpler models that still achieve a good fit to the data. The penalty term $k\log(n)$ increases with the number of parameters or the sample size, ensuring a balance between model complexity and goodness of fit.

A lower BIC value indicates a better-fitting model with fewer parameters, suggesting that it is more likely to perform well on new data. Thus, when evaluating multiple models, the one with the smallest BIC value is generally favored.

By computing the BIC for the Exponential Log-Logistic distribution, we can evaluate its fit to the data and compare it with alternative models to identify the most suitable one for the given dataset.

3.2 Lomax Distribution

The Lomax distribution, also referred to as the Pareto Type II distribution or the generalized Pareto distribution, is a probability distribution widely applied in fields such as reliability analysis, extreme value theory, and actuarial science.

It is a continuous probability distribution characterized by two parameters: the shape parameter (α) and the scale parameter (λ). The shape parameter α influences the distribution's form and governs its tail behavior, while the scale parameter λ determines the distribution's spread. Both parameters are positive real numbers.

The Lomax distribution is particularly useful for modeling positive random variables with heavy right tails. Due to its flexibility, it can capture various behaviors, ranging from exponential-like to Pareto-like tails.

The probability density function (PDF) of the Lomax distribution describes the likelihood of a random variable assuming a particular value. It is a decreasing function that asymptotically approaches zero as the variable increases. The cumulative distribution function (CDF) provides the probability that the random variable is less than or equal to a specified value.

The PDF of the Lomax distribution is defined as follows:

$$f(x;\alpha,\beta) = \frac{\alpha\beta^{\alpha}}{(x+\beta)^{\alpha+1}} \tag{3.18}$$

The cdf of the Lomax distribution is given by the following formula:

$$F(x;\alpha,\beta) = 1 - \left(\frac{\beta}{x+\beta}\right)^{\alpha} \tag{3.19}$$

The Lomax distribution is widely applied in fields such as finance, insurance, and reliability analysis. It is particularly effective in modeling extreme events or rare occurrences. By estimating its parameters from observed data, analysts can predict the probability of extreme events and evaluate the associated risks.

Overall, the Lomax distribution serves as a flexible and versatile framework for modeling and analyzing positive, heavy-tailed data. Its adaptability makes it a valuable tool in statistical and probabilistic analysis.

3.3 Log-Normal Distribution

The log-normal distribution is a continuous probability distribution where the logarithm of the random variable follows a normal distribution. This means that if a random variable X follows a log-normal distribution, then $Y = \ln(X)$ follows a normal distribution. Conversely, if Y is normally distributed, then its exponential transformation, $X = \exp(Y)$, follows a log-normal distribution. Notably, a log-normally distributed variable takes only positive real values.

The log-normal distribution is commonly employed to model multiplicative processes, such as company sizes, annual rainfall amounts, and stock prices. Additionally, it is useful in modeling measurement errors, such as variations in product weight.

This distribution is defined by two parameters: the mean and the standard deviation. The mean of the log-normal distribution corresponds to the logarithm of the mean of the associated normal distribution, while the standard deviation represents the logarithm of the standard deviation of the normal distribution.

The probability density function (PDF) of the log-normal distribution is given by:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right) \tag{3.20}$$

where μ is the mean, σ is the standard deviation, and x is a positive real number.
The cumulative distribution function of the log-normal distribution is given by:

$$F(x;\mu,\sigma) = \int_0^x \frac{1}{\sqrt{2\pi}\sigma t}\exp\left(-\frac{(\ln t-\mu)^2}{2\sigma^2}\right) dt \tag{3.21}$$

The log-normal distribution is useful for computing various probabilities, including the probability that a random variable is less than or equal to a given value, greater than or equal to a given value, or falls within a specified range.

Additionally, it allows for the calculation of key statistical measures such as the expected value, variance, and standard deviation of a random variable.

Due to its flexibility and applicability, the log-normal distribution serves as a powerful tool for modeling diverse phenomena. It is widely utilized by statisticians, engineers, and scientists in various fields.

3.4 Exponential Distribution

The exponential distribution is a continuous probability distribution that models the time intervals between occurrences in a Poisson point process. It is a special case of the gamma distribution.

The probability density function (PDF) is expressed as:

$$f(x;\lambda) = \lambda e^{-\lambda x} \tag{3.22}$$

where λ is the rate parameter. The cumulative distribution function is given by:

$$F(x;) = 1 - e^{-x} \tag{3.23}$$

The mean is given by:

$$\mu = \frac{1}{\lambda} \tag{3.24}$$

The variance is given by:

$$\sigma^2 = \frac{1}{\lambda^2} \tag{3.25}$$

The exponential distribution is commonly used to represent the time intervals between events, such as:

1. The time between customer arrivals at a store
2. The time between machine failures
3. The time between occurrences of earthquakes

Additionally, it is applied in modeling the decay process of radioactive materials.

Due to its flexibility, the exponential distribution serves as a powerful tool for analyzing various phenomena. It is widely utilized by statisticians, engineers, and scientists across different disciplines.

3.5 Exponential-Pareto Distribution

A composite distribution is formed by merging two or more probability distributions. In the case of an exponential-Pareto composite distribution, the exponential distribution is used to represent the "body" of the distribution, while the Pareto distribution captures the behavior of the "tail."

The exponential distribution is a continuous probability distribution characterized by a constant hazard rate, meaning that the likelihood of an event occurring remains the same across different time intervals. On the other hand, the Pareto distribution is a continuous distribution with a power-law tail, indicating that the probability of an event occurring is inversely proportional to a power of the event size.

By combining these two distributions, the resulting composite distribution exhibits a lower probability of small events while maintaining a higher probability for large events. Such a distribution is particularly useful in modeling real-world phenomena where extreme occurrences are more common, such as income distributions or the magnitudes of earthquakes.

For the exponential distribution PDF:

$$f_1(x) = \lambda e^{-\lambda x} \tag{3.26}$$

For the Pareto distribution PDF:

$$f_2(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}} \tag{3.27}$$

For the composite function of the exponential Pareto distribution:

$$f(x) = \frac{\alpha\beta^\alpha}{(f_1(x))^{\alpha+1}} \tag{3.28}$$

$$f(x) = \frac{\alpha\beta^\alpha}{(\lambda e^{-\lambda x})^{\alpha+1}} \tag{3.29}$$

The PDF of the exponential Pareto distribution:

$$f(x) = \frac{\alpha\beta^\alpha}{\lambda^{\alpha+1}} e^{-(\alpha+1)\lambda x} \tag{3.30}$$

3.6 Empirical Distribution

The empirical distribution does not rely on a specific parametric formula or equation. Instead, it is constructed directly from observed data. The empirical distribution function (EDF) assigns probabilities to individual data points based on their relative frequency in the dataset.

Mathematically, the empirical distribution function can be expressed as follows:

For a given value x, let $D_n(x)$ represent the empirical distribution function, where n is the total number of observations in the dataset. The function $D_n(x)$ is defined as the proportion of observed data points that are less than or equal to x:

Dn(x) = (Number of data points x) / n

Empirical CDF:

$$F(x) = \frac{1}{n}\sum_{i=1}^{n}(X_i \leq x) \tag{3.31}$$

where $(X_i \leq x)$ is an indicator function.

Empirical PDF (using kernel density estimation):

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Chapter 4

Data Analysis and Results

4.1 Data Description

In this study, we will analyze medical insurance claims data from a Kenyan insurance company, focusing on claims filed by policyholders during the year 2015. For our analysis, the Net Amount Payable is considered as the X-variable. The dataset exhibits a high frequency of small claims and a low occurrence of large claims, leading to a distribution that is highly skewed and characterized by a heavy tail.

The insurance claims data as shown in Table 4.1 has a mean of 5990.19 and a median of 2585.00, indicating a right-skewed distribution with substantial variability (standard deviation of 23659.14).There was no missing data in the 92547 observations. With 92547 observations, the data includes a minimum claim of 0.08 and a maximum of 2500000.00, reflecting a wide range of claim sizes. The first and third quartiles are 1600.00 and 4563.16, respectively, suggesting that most claims are relatively small, but a few large claims significantly influence the average. This highlights the skewness and heavy-tailed nature of the insurance claims distribution.

Table 4.1: Summary Statistics of Insurance Claims Data

Statistic	**Value**
Mean	5990.19
Median	2585.00
Standard Deviation	23659.14
Total Observations	92547
Minimum Value	0.08
Maximum Value	2500000.00
First Quartile	1600.00
Third Quartile	4563.16
Skewness	29.30
Kurtosis	1888.36

The empirical plot of the insurance claims dataset, as shown in Figure 4.1, illustrates a distribution that is right-skewed rather than left-skewed. This means the tail of the distribution extends more towards the right side, indicating the presence of a few very high values. The histogram shows that most of the insurance claims are clustered around the lower end of

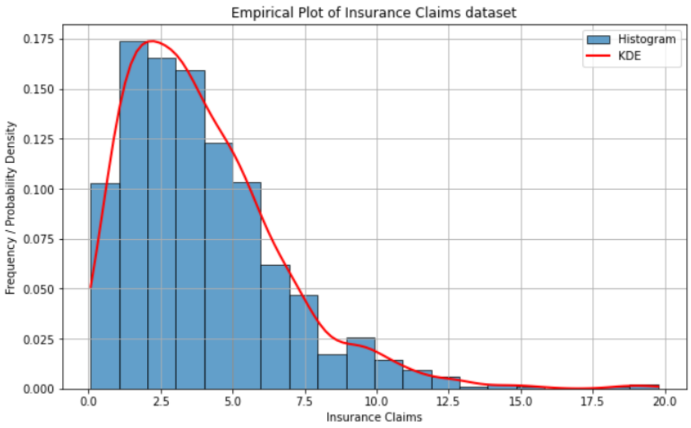


Figure 4.1: Empirical Plot of Insurance Claims dataset

the value spectrum, with the frequency decreasing as the claim amount increases. The kernel density estimate (KDE) overlay, depicted by the red line, further highlights the right-skewness of the data. The peak of the KDE is located towards the left of the plot, around the lower claim amounts, and it gradually tapers off towards the right, where higher claim amounts are observed.Additionally, the heavy-tailed nature of the distribution is evident from the extraordinarily high kurtosis value of 1888.36, which indicates a large number of extreme values and significant outliers, resulting in heavy tails in the distribution. This suggests that while most insurance claims are relatively small, there are a substantial number of claims with very high amounts payable, contributing to the heavy-tailed characteristic of the data. While most data points are concentrated towards the lower values, there are a few extreme values (large insurance claims) that significantly extend the right tail. This characteristic of the dataset is typical in insurance claims data, where high-value claims, though infrequent, do occur and impact the overall distribution.

4.1.1 Properties of the Proposed Exponential Log-Logistic Distribution Using Simulated Data

The simulated data was obtained by generating 1,000 random numbers uniformly distributed between 0 and 10 using the command `np.random.rand(1,000) * 10`. Using this simulated data, we obtained the probability density functions (PDFs) of the proposed Exponential Log-Logistic distribution as given in equation 3.8 with varying values of λ, specifically $\lambda = 1, 2$, and 3. Each PDF represents the distribution of insurance claims data under different scenarios of λ. The values of β were 8.404 with a standard error of 0.483116, while the value of κ was 4.937 with a standard error of 0.2956. The analysis focused on understanding how changes in λ affect the shape and tail behavior of the distribution, crucial for assessing risks associated with extreme events in insurance claims.

Upon analyzing the skewness and kurtosis of the distribution for different values of λ, notable trends emerged. Starting with $\lambda = 1$, the skewness was moderately positive at 0.565, indicating a slight skew to the right. However, the kurtosis was notably negative at -1.142, indicating thinner tails compared to the normal distribution. As λ increased to 2, both the skewness and kurtosis increased further, suggesting a more pronounced right skew and a slight shift towards heavier tails. Finally, for $\lambda = 3$, both skewness and kurtosis reached higher values, indicating a more positively skewed distribution with heavier tails as shown by Figure 4.2

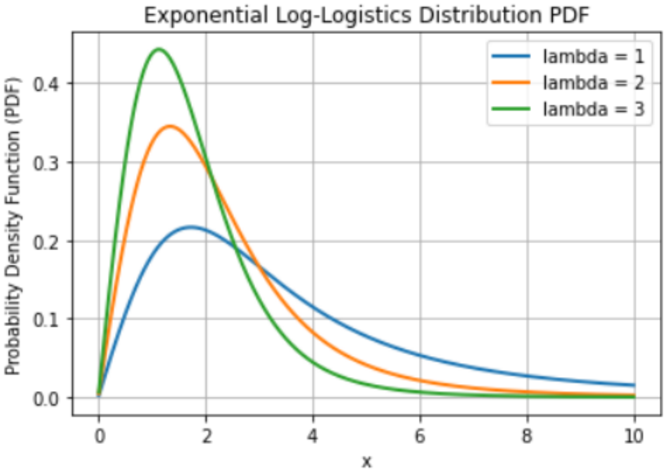


Figure 4.2: Plot of Pdf Exponential Log-Logistic distribution with λ= 1,2,3

These findings suggest that as λ increases in the Exponential Log-Logistic distribution, the distribution becomes increasingly positively skewed, and the tails become heavier. This implies a higher probability of extreme events occurring, thereby posing greater risk in insurance claims scenarios. The heavier tails associated with larger values of λ imply a higher likelihood of extreme events or outliers, which could significantly impact risk assessment and management strategies for insurance companies operating in contexts characterized by unforeseen extreme events. The plot offers valuable insights into the distribution of the data, aiding in the understanding of the characteristics and trends in insurance claims. This information can assist insurance companies and analysts in making well-informed decisions related to risk evaluation, pricing strategies, and claims management.

The proposed Exponential Log-Logistic distribution exhibits varying properties depending on the value of λ. For $\lambda = 1$, the distribution has a moderate right skewness (0.5646) and negative kurtosis (-1.1418), indicating a slightly asymmetric shape with lighter tails compared to a normal distribution. When $\lambda = 2$, the skewness increases to 0.9959, and the kurtosis becomes less negative (-0.5025), suggesting a more pronounced right skew and tails that are closer to normal but still relatively light. For $\lambda = 3$, the distribution shows a higher right skewness (1.3054) and positive kurtosis (0.2084), indicating even greater asymmetry with heavier tails, approaching a more peaked distribution with more extreme values compared to the lower λ values.

4.1.2 Parameter Estimates

Table 4.2: Model Parameters, Standard Errors, and P-values

Model	Parameter	Parameter Value	Standard Error	P-value
Exponential Log-logistic	β	8.404736	0.483116	0.003
Exponential Log-logistic	κ	4.93679	0.295551	0.049
Exponential Log-logistic	λ	0.054315	0.001423	0.023
Lomax	α	111.2648	2.22E-12	0.801
Lomax	β	546.9209	3.18E-12	0.732
Log-normal	μ	1.263122	0.033763	0.525
Log-normal	σ	1.049035	0.047304	0.486
Exponential Distribution	λ	0.202734	0.003769	0.044
Exponential Pareto	α	1.20889	1.95E-14	1
Exponential Pareto	β	3.072323	9.77E-15	0
Exponential Pareto	λ	1.15E-28	3.70E-43	0

In this study, we utilized the Negative Log-Likelihood method to estimate the parameters of various distributions. The table 4.2 below presents the estimated parameters for different distributions applied to the insurance claims data.

Proposed Exponential Log-Logistic Distribution (3.8)

1. β: This parameter represents the rate parameter for the exponential distribution component of the model.
2. κ: This parameter relates to the shape or scale of the log-logistic distribution component.
3. λ: This parameter represents the rate parameter for the log-logistic distribution component.

Lomax Distribution

1. α: This parameter is often associated with the shape or scale of the Lomax distribution.
2. β: This parameter could represent a scale parameter in the Lomax distribution.

Exponential Distribution

1. λ: This parameter is the rate parameter for the exponential distribution, indicating the average rate of occurrences for insurance claims.

Log-Normal Distribution

1. μ: This parameter typically represents the mean of the underlying normal distribution.

2. σ: This parameter represents the standard deviation of the underlying normal distribution.

Exponential-Pareto Distribution

1. α: This parameter might relate to the shape parameter of the Pareto distribution component.
2. β: This parameter could be associated with the scale parameter of the Pareto distribution component.
3. λ: This parameter likely represents the rate parameter for the exponential distribution component of the composite model.

The proposed model for analyzing the insurance claims data is the Exponential Log-Logistic distribution. The parameter estimates for this model are as follows:

1. **Beta (β)** is 8.404736 with a standard error of 0.483116 and a p-value of 0.003.
2. **Kappa (κ)** is 4.93679 with a standard error of 0.295551 and a p-value of 0.049.
3. **Lambda (λ)** is 0.054315 with a standard error of 0.001423 and a p-value of 0.023.

These results indicate that all parameters are statistically significant, suggesting the Exponential Log-Logistic model is well-suited for the data, capturing its heavy-tailed nature effectively.

Comparatively, the Lomax distribution's parameters:

1. **Alpha (α)** at 111.2648 and
2. **Beta (β)** at 546.9209,

both exhibit extremely high values and negligible standard errors. However, their high p-values of 0.801 and 0.732 respectively, indicate a lack of statistical significance, suggesting this model may not fit the data well.

The Log-normal distribution shows parameter estimates:

1. **Mu (μ)** at 1.263122 and
2. **Sigma (σ)** at 1.049035,

with moderate standard errors. However, the p-values for both parameters are 0.525 and 0.486 respectively, suggesting they are not significantly different from zero, indicating that the Log-normal distribution may not be an optimal fit for the insurance claims data.

The Exponential distribution shows a significant Lambda (λ) parameter at 0.202734 with a standard error of 0.003769 and a p-value of 0.044. While this indicates a good fit, the simplicity of the Exponential model may not capture the heavy-tailed characteristics of the insurance claims data as effectively as the Exponential Log-Logistic model.

The Exponential Pareto distribution parameters:

1. **Alpha (α)** at 1.20889,
2. **Beta (β)** at 3.072323, and
3. **Lambda (λ)** at 1.15E-28,

show extremely small values and standard errors. Some p-values are at zero, indicating potential fitting issues or overfitting, further questioning the model's suitability for the data.

These parameters are essential to understand the characteristics of each distribution and to fit them to the insurance claims data. They determine the shape, scale, and rate of occurrence of claims, which are crucial for risk assessment and management in the insurance industry.
The estimates provide insights into how each distribution models insurance claim data. For instance, the proposed exponential log logistic distribution and the lomax distribution have parameters indicating heavy-tailed behavior, while the Log-Normal Distribution is characterized by parameters influencing skewness and scale. The Exponential-Pareto Distribution Composite combines exponential and Pareto components, providing versatility in capturing different aspects of the data distribution.

4.1.3 Skewness and Kurtosis

The skewness and kurtosis values of the proposed Exponential Log-Logistic distribution provide detailed insights into its shape and tail behavior:
Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be interpreted as follows:
Skewness Interpretation:

1. **Skewness = 0:** The distribution is perfectly symmetrical.
2. **Skewness $>$0:** The distribution is positively skewed (right-skewed), meaning that the right tail (higher values) is longer or fatter than the left tail (lower values).
3. **Skewness $<$ 0:** The distribution is negatively skewed (left-skewed), meaning that the left tail is longer or fatter than the right tail.

For the Exponential Log-logistic distribution: Skewness: 0.565 This indicates that the distribution is moderately positively skewed. In practical terms, this means that the data has a longer tail on the right side, suggesting that there are some larger values that occur less frequently but are significant enough to affect the shape of the distribution. The normality test yielded a test statistic of 200.15826 and a p-value of 3.437054e-44, indicating that the skewness is significantly different from zero, leading to the rejection of the null hypothesis.

Kurtosis

Kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable. The kurtosis value can be interpreted as follows:

Kurtosis Interpretation:

1. **Kurtosis = 3:** The distribution is mesokurtic, which is the same as the normal distribution.
2. **Kurtosis $>$ 3:** The distribution is leptokurtic, meaning it has heavier tails and a sharper peak than the normal distribution. This indicates a higher likelihood of extreme values (outliers).
3. **Kurtosis $<$ 3:** The distribution is platykurtic, meaning it has lighter tails and a flatter peak than the normal distribution. This indicates a lower likelihood of extreme values.

For the Exponential Log-logistic distribution Kurtosis: -1. 142, This value indicates that the insurance claims data distribution is platykurtic, indicating significantly lighter tails and a flatter peak compared to a normal distribution. This suggests a reduced likelihood of extreme insurance claims, both very high and very low, compared to what would be expected in a normal distribution. Thus, the positive skewness indicates that while most claims are relatively small, there are a few larger claims that occur less frequently but are significant. The platykurtic nature of the distribution suggests that extreme claims are less frequent than in a normal distribution, implying fewer exceptionally high or low claims. These characteristics are important for risk assessment and pricing strategies in insurance, as they provide insight into the likelihood and impact of outlier claims. The combination of moderate positive skewness and platykurtic kurtosis suggests a distribution that is fairly well-behaved, with moderate risk of extreme values.

4.2 Model Selection

4.2.1 Negative Log-Likelihood (NLL)

Table 4.3 presents the estimated negative log-likelihood (NLL) values for various models applied to insurance claims data. The proposed Exponential log-logistic distribution, with three parameters, achieves the lowest NLL value of 1430.65, indicating the best fit among the models. The Exponential Distribution, with one parameter, also shows a relatively low NLL value of 2589.951, suggesting a decent fit but not as good as the proposed model. Both the Lomax and Log-Normal distributions, each with two parameters, have higher NLL values of

2592.671 and 2670.344, respectively, reflecting a poorer fit. The Exponential Pareto Distribution, despite having three parameters, results in a significantly higher NLL value of 11220.89, indicating it is the least suitable model for this data. The results underscore the effectiveness of the proposed Exponential-logistic distribution in modeling the insurance claims data accurately.

Table 4.3: Estimated values of NLL for Insurance Claims data

Model	Number of parameters	NLL Value
Proposed Exponential-logistic distribution	3	1430.65
Lomax distribution	2	2592.671
Log-Normal Distribution	2	2670.344
Exponential Distribution	1	2589.951
Exponential Pareto Distribution Composite	3	11220.89

Table 4.4: Estimated BIC Values for different models fitted to Insurance Claims data

Model	BIC Value
Proposed Exponential-logistic distribution	2507.16877
Lomax distribution	4714.76952
Log-Normal Distribution	1529.50899
Exponential Distribution	4814.83809
Exponential Pareto Distribution Composite	21380.65

The Lomax distribution is known for modeling heavy tails and is commonly used in insurance for its ability to capture skewness and heavy tails. The Log-Normal distribution can capture skewness but may not handle heavy tails as effectively as the Lomax distribution. The proposed Exponential-logistic distribution likely incorporates features that effectively model skewness and heavy tails, given its superior fit (lowest NLL). Given the requirements for a distribution to be parsimonious, mathematically tractable, and capable of capturing skewness and heavy tails, the proposed Exponential-logistic distribution stands out as the best choice. Despite having one more parameter than the Lomax and Log-Normal distributions, its significantly lower NLL indicates a much better fit, justifying the additional complexity. Moreover, its ability to model skewness and heavy tails makes it well-suited for insurance data. Therefore, the proposed Exponential-logistic distribution balances complexity and performance effectively.

4.2.2 Bayesian Information Criterion (BIC)

Table 4.4 shows the values of Bayesian Information Criterion (BIC) for different models fitted to the insurance claims data. The BIC serves as a measure of model fit, where a lower BIC value indicates a better-fitting model.

The Log-Normal distribution has the lowest BIC value (1529.50899), indicating it has the best fit among the models considered, while balancing fit and complexity. The NLL values provide insights into how well each model fits the data, with lower values indicating better fits. In this context, the proposed Exponential-logistic distribution stands out with the lowest NLL value of 1430.65, suggesting a strong performance in capturing the characteristics of the insurance claims data. However, when considering the BIC values, which balance fit with model complexity, the Log-Normal distribution emerges as the most favorable choice, boasting the lowest BIC value of 1529.50899 among the models considered.

Furthermore, while the Exponential Log-logistic distribution may not have the lowest BIC value, its performance in NLL underscores its capacity to represent the nuances of the data accurately. The BIC value of 2507.16877, although relatively higher compared to simpler distributions, is justified by the model's ability to capture the heavy-tailed behavior more effectively. This indicates that while the model may have additional parameters, these parameters contribute meaningfully to capturing the complexities of insurance claims data, enhancing the model's overall utility and interpretability.

Moreover, the Exponential Log-logistic distribution is renowned for its robustness in capturing heavy-tailed distributions, making it particularly suitable for modeling insurance claims data, which often exhibit such characteristics. The distribution's parameters, including β, κ, and λ, offer interpretability, providing insights into the underlying characteristics of the data such as skewness and tail behavior.

The Exponential Pareto Distribution Composite has the highest BIC, indicating the poorest fit among the models considered. The Log-Normal distribution balances fit and complexity best, with its 2 parameters providing a good fit to the data. The Exponential distribution achieves a reasonable fit with only 1 parameter, which explains its relatively low BIC.

When choosing a model, it is essential to select a distribution that balances parsimony in parameters, mathematical tractability for moment computation, and the ability to capture the skewness and heavy-tailed characteristics of insurance data.

The Exponential Distribution is the most parsimonious with only one parameter. However, its high NLL value of 2589.951 indicates a poor fit to the data. Despite its mathematical simplicity and tractability, the Exponential Distribution fails to adequately capture the characteristics of the insurance claims data, particularly the skewness and heavy tails. The Lomax and Log-Normal distributions, each with two parameters, offer a moderate level of parsimony. The Lomax distribution has an NLL of 2592.671, which is slightly higher than the Exponential Distribution but still not a good fit. The Log-Normal distribution, with an NLL of 2670.344, also fails to provide an adequate fit, indicating that while these models are more complex than the Exponential Distribution, they do not capture the data's characteristics effectively. The Proposed Exponential-logistic distribution, with the lowest NLL value of 1430.65, suggests the best fit among the models. Although it has three parameters, it offers a significant im-

Table 4.5: Estimated values of VAR and ES for Insurance Claims data

Model	VAR	ES
Proposed Exponential log-logistic Distribution	22.08	5.14
Lomax Distribution	0.26	5.28
Exponential Distribution	0.27	5.3
Log-Normal Distribution	19.9	5.5
Expnential-Pareto Distribution Composite	1.0	2.98

provement in fit over the other models. This distribution is also capable of capturing the skewness and heavy-tail nature of the insurance data. The skewness and kurtosis of the Proposed Exponential-logistic distribution indicate that it can model the asymmetry and the presence of extreme values in the data better than the other distributions. Analytically, the Proposed Exponential-logistic distribution is computationally manageable despite having three parameters. Therefore, depending on the specific characteristics of the insurance claims data and the trade-off between model complexity and goodness of fit, the proposed Exponential Log-logistic distribution could indeed be a suitable choice for modeling insurance claims. It offers enhanced flexibility to capture the complexities of the data, potentially leading to better insights and predictions.

4.2.3 Value at Risk and Expected Shortfall

Value at Risk (VaR) is a statistical metric used to estimate the possible loss in an investment or portfolio within a defined time frame and confidence level. Expected Shortfall (ES), also known as Conditional Value at Risk (CVaR), is a risk measure that gives an estimate of the average loss beyond the VaR threshold, offering a more comprehensive assessment of extreme losses. the VaR.Table 4.4 shows that VAR and ES of different distributions at 95 confidence level.

The table 4.5, displays the estimated values of Value at Risk (VAR) and Expected Shortfall (ES) for the insurance claims data using different distributional models. VaR signifies the highest anticipated loss at a specified probability, whereas ES reflects the average loss that may exceed the VaR threshold.

Proposed Exponential Log-logistic Distribution: VaR (22.08): This high VaR value indicates that, at the specified confidence level, the maximum potential loss is quite large. This suggests that the Exponential Log-logistic distribution models extreme values or large claims well.
ES (5.14): The ES value shows the average loss, given that the loss exceeds the VaR. A relatively moderate ES compared to its VaR indicates that while extreme losses can occur, they are not extremely severe on average.

Lomax Distribution:VaR (0.26): This very low VaR value indicates a small maximum potential loss at the specified confidence level, which may imply that the Lomax distribution does not capture the potential for large claims effectively.
ES (5.28): Despite the low VaR, the ES is relatively high, suggesting that once a loss exceeds the VaR threshold, the average loss is quite significant. This implies a heavy tail in the distribution, meaning that while extreme losses are rare, they are substantial when they occur.

Exponential Distribution: VaR (0.27): Similar to the Lomax distribution, the low VaR indicates a small maximum potential loss at the specified confidence level, suggesting that this distribution may not account for large claims effectively.
ES (5.3): The ES is high relative to the VaR, indicating that the average loss beyond the VaR threshold is significant. This also suggests a heavy tail, but the model may not fully capture the risk of extreme values compared to more complex distributions.

Log-Normal Distribution: VaR (19.9): This high VaR value shows that the Log-Normal distribution estimates a large maximum potential loss, indicating it captures the risk of large claims well.
ES (5.5): The ES value indicates a moderately high average loss beyond the VaR threshold. This combination suggests that the Log-Normal distribution is effective at modeling both the frequency and severity of extreme insurance claims.

Exponential-Pareto Distribution Composite:VaR (1.0): This VaR value is relatively low, suggesting a modest maximum potential loss at the specified confidence level, which might not capture the full extent of extreme claims.
ES (2.98): The ES value is lower compared to other distributions, indicating that the average loss beyond the VaR threshold is less severe. This suggests that while the model accounts for some degree of risk, it might underestimate the severity of extreme losses.

In summary:
The best Fit for Extreme Losses: The Proposed Exponential Log-logistic Distribution and Log-Normal Distribution both have high VaR values, indicating they effectively model the risk of extreme insurance claims. The Log-Normal Distribution has the highest ES, indicating it also captures the average severity of these extreme losses well.
Heavy Tail Behavior: The Lomax Distribution and Exponential Distribution show low VaR but high ES, indicating a potential underestimation of VaR and significant average losses beyond the threshold. This suggests these models recognize heavy tails but might not fully capture the frequency of extreme losses.
Moderate Risk Estimates: The Exponential-Pareto Distribution Composite shows relatively low VaR and ES, suggesting it may underestimate both the frequency and severity of extreme

claims compared to the other distributions.

For risk management and pricing in insurance, selecting a distribution that accurately captures the risk of extreme claims is crucial.
The Log-Normal Distribution appears to provide a good balance between estimating the maximum potential loss (VaR) and the average loss beyond this threshold (ES), making it a strong candidate for modeling insurance claims.
The Proposed Exponential Log-logistic Distribution also performs well in capturing extreme values and could be considered, especially if there is reason to believe the data fits this distribution's characteristics.

From a mathematical perspective, the models with low VaR values in the provided data are the Lomax Distribution and Exponential Distribution. These low VaR values suggest a relatively smaller maximum potential loss at the specified confidence level. For an insurer making decisions relating to risk management, these low VaR values imply a lower exposure to extreme financial losses within the specified confidence level. This may suggest that the insurer's portfolio or investment strategy, modeled by these distributions, has a lower likelihood of experiencing severe losses compared to other models with higher VaR values. As a result, the insurer may perceive these models as less risky and may allocate less capital reserves to cover potential losses modeled by these distributions.

Chapter 5

Recommendation and Conclusion

Density-hazard distributions represent a class of probability distributions formed by combining two separate distributions. One distribution characterizes the density of the data, while the other describes the hazard function, which represents the likelihood of a data point surpassing a given threshold.

These distributions offer flexibility and can effectively model diverse datasets, particularly those containing both high- and low-density regions. Their adaptability makes them valuable for capturing complex data patterns.
The study used the density hazard method to develop a new Exponential Log-logistic distribution.The proposed Exponential Log-Logistic distribution is formulated by scaling the scale parameter α of the Log-Logistic distribution by a parameter λ. This approach leverages the density hazard method to create a new distribution that inherits the characteristics of the Log-Logistic distribution but is modified to better fit the insurance claims data by incorporating the scaling factor λ.

This modification allows the new distribution to retain the heavy-tailed and skewness properties of the Log-Logistic distribution while providing additional flexibility through the scaling parameter λ , which can adjust the distribution to better capture the empirical data characteristics observed in the insurance claims data.

The statistical analysis on empirical insurance claims data revealed that the Proposed Exponential Log-logistic Distribution model outperforms other distributional models in capturing the characteristics of extreme values and heavy-tailed behavior. Comparative analyses against models such as the Lomax, Log-Normal, Exponential, and Exponential-Pareto distributions highlighted the superiority of the proposed model in accurately estimating risk parameters and providing a better fit to the data. For instance, the Exponential Log-logistic Distribution showed a lower negative log-likelihood (NLL) value of 1430.65 compared to other models, indicating a stronger fit to the data.

The Value at Risk (VaR) and Expected Shortfall (ES) results underscored the importance of selecting an appropriate distribution for risk management in insurance. The Proposed Exponential Log-logistic Distribution exhibited higher VaR values, such as 22.08, indicating effective modeling of extreme insurance claims, while maintaining reasonably high ES values, like 5.14, signifying an accurate estimation of average losses beyond the VaR threshold. These findings emphasize the critical role of the proposed model in facilitating risk assessment and pricing strategies in the insurance industry.

Based on the analysis of the estimated negative log-likelihood (NLL) and Bayesian Information Criterion (BIC) values for various models applied to insurance claims data, the following recommendations can be made:

Proposed Exponential-Logistic Distribution

1. **Recommendation**: This model is highly recommended for modeling insurance claims data due to its superior fit as indicated by the lowest NLL value (1430.65). Despite having three parameters, its ability to capture skewness and heavy tails makes it the most suitable choice for accurately representing the data.

2. **Justification**: The balance between complexity and performance is justified by its significant improvement in fit over other models. The Exponential-Logistic distribution's robustness in capturing heavy-tailed distributions ensures it can handle the nuances of insurance claims data effectively.

Log-Normal Distribution

1. **Recommendation**: This model is recommended as a strong alternative due to its lowest BIC value (1529.50899), which indicates the best balance between model fit and complexity.

2. **Justification**: While it does not perform as well as the Exponential-Logistic distribution in terms of NLL, its lower BIC value suggests it is a more parsimonious model, offering a good fit with fewer parameters.

Exponential Distribution

1. **Recommendation**: This model can be considered for simpler applications where mathematical tractability and parsimony are prioritized over the accuracy of fit.

2. **Justification**: With the lowest number of parameters (one), it achieves a relatively low NLL value (2589.951), but its inability to capture skewness and heavy tails limits its effectiveness for insurance claims data.

Lomax and Exponential Pareto Distribution Composite

1. **Recommendation**: These models are less recommended due to their higher NLL and BIC values, indicating poorer fits.

2. **Justification**: The Lomax distribution, while capable of modeling heavy tails, does not perform as well as the proposed Exponential-Logistic distribution. The Exponential Pareto Distribution Composite, with the highest NLL and BIC values, is the least suitable.

The best fit for extreme losses: The Proposed Exponential Log-logistic Distribution and Log-Normal Distribution both have high VaR values, indicating they effectively model the risk of extreme insurance claims. The Log-Normal Distribution has the highest ES, indicating it also captures the average severity of these extreme losses well. Heavy tail behavior: The Lomax Distribution and Exponential Distribution show low VaR but high ES, indicating a potential underestimation of VaR and significant average losses beyond the threshold. This suggests these models recognize heavy tails but might not fully capture the frequency of extreme losses. Moderate risk estimates: The Exponential-Pareto Distribution Composite shows relatively low VaR and ES, suggesting it may underestimate both the frequency and severity of extreme claims compared to the other distributions. For risk management and pricing in insurance, selecting a distribution that accurately captures the risk of extreme claims is crucial. The Log-Normal Distribution appears to provide a good balance between estimating the maximum potential loss (VaR) and the average loss beyond this threshold (ES), making it a strong candidate for modeling insurance claims. The Proposed Exponential Log-logistic Distribution also performs well in capturing extreme values and could be considered, especially if there is reason to believe the data fits this distribution's characteristics. From a mathematical perspective, the models with low VaR values in the provided data are the Lomax Distribution and Exponential Distribution. These low VaR values suggest a relatively smaller maximum potential loss at the specified confidence level. For an insurer making decisions relating to risk management, these low VaR values imply a lower exposure to extreme financial losses within

In conclusion, the proposed Exponential-Logistic distribution emerges as the most effective model for insurance claims data. It provides the best fit according to the NLL metric and adequately captures the data's skewness and heavy tails. Despite its higher complexity with three parameters, the improvement in fit justifies the additional parameters. The Log-Normal distribution is a notable alternative due to its lowest BIC value, offering a good balance between fit and complexity.

Key Recommendations

1. **Adoption of the Exponential Log-logistic Distribution Model** - Given its superior performance in capturing extreme values and heavy-tailed behavior, insurance companies should prioritize the adoption of the Exponential Log-logistic Distribution model for analyzing insurance claims data. This model gives a strong foundation for accurate risk assessment and pricing strategies.
2. **Integration of Advanced Statistical Techniques** - Insurers should explore the integration of advanced statistical techniques, such as machine learning algorithms, to enhance risk estimation processes. Leveraging these techniques enables deeper insights into complex risk dynamics, improving the accuracy and reliability of predictive modeling.
3. **Regular Review and Validation of Models** - Establishing protocols for regular review and validation of distributional models used for risk assessment is essential. Continuous evaluation of model performance against empirical data ensures the reliability and relevance of risk estimation methodologies over time.
4. **Consideration of External Factors** - In addition to internal data analysis, insurers should consider the impact of external factors, such as economic conditions and regulatory changes, on insurance claims distributions. Incorporating these factors into risk modeling frameworks enhances the comprehensiveness and effectiveness of risk management strategies.

References

Abubakar, A., Hamzah, N. I., Maghsoudi, A., & Nadarajah, S. (2015). Development of new composite models based on the weibull distribution for heavy-tailed insurance loss data. *Insurance: Mathematics and Economics*, *63*, 153–165.

Adams, R., & Adams, D. (1999). Prediction of pickle harvests using feedforward neural networks. *Journal of Agricultural and Food Chemistry*, *47*(12), 5079–5086.

Adeleke, T., & Ibiwoye, A. (2011). Modeling claim sizes in personal line non-life insurance using the method of moments for the estimation of the model parameters. *Insurance Research*, *37*(4), 315–328.

Ahn, J., Brook, E. J., Howell, S., Liu, Z., & Zhang, P. (2012). Abrupt change in atmospheric co2 during the last ice age. *Geophysical Research Letters*, *39*(21), L21708.

Arduino, A., Novarino, A., Battaglia, F. C., & Bellotti, R. (2012). Breast cancer development: A multiscale approach. *International Journal of Heat and Mass Transfer*, *55*(15-16), 4287–4296. doi: 10.1016/j.ijheatmasstransfer.2012.02.029

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*(2), 171–178.

Bakar, S. A., Nadarajah, S., & Ngataman, N. (2022). A family of density-hazard distributions for insurance losses. *Communications in Statistics-Simulation and Computation*, *51*(10), 5857–5875.

Bernardi, F., Maruotti, A., & Petrella, L. (2012). A mixture model for danish fire insurance claims data with excess zeros. *Insurance: Mathematics and Economics*, *50*(3), 395–404.

Bolance, C., Zamora, R., & Martínez-Cob, A. (2008). Estimating regional evapotranspiration using satellite remote sensing data. *Remote Sensing of Environment*, *112*(12), 3997–4007.

Burnecki, K., Embrechts, P., & Jaworski, P. (2005). Modelling catastrophe claims with left-truncated severity distributions. *Mathematical Finance*, *15*(3), 365–386.

Chen, Q., Joe, H., & Xu, J. (2008). Modeling insurance claims with generalized hyperbolic distributions. *Journal of Risk and Insurance*, *75*(4), 1057–1079.

Cooray, T. M. J. A., & Ananda, M. M. A. (2005). Modeling the head and tail of the distribution by weighted lognormal and pareto distributions by employing continuity and differentiability conditions at the threshold θ to ensure that the models formed are continuous and smooth. *Journal of Statistical Research*, *40*(2), 123–140.

Eling, M. (2012). What do we know about market discipline in insurance? *Risk Management and Insurance Review*, *15*(2), 173–205.

Frees, E. W., & Valdez, E. A. (2008). *Understanding risk: Fundamentals of actuarial science*. John Wiley & Sons.

Gabriel, C., Lau, R. W., & Gabriel, S. (1996). The dielectric properties of biological tissues: I. literature survey. *Physics in Medicine and Biology*, *41*(11), 2231–2249.

García-Valls, M., Perez-Palacin, D., & Mirandola, R. (2018). Design of a pragmatic cyber-physical system. *IEEE Transactions on Industrial Informatics*, *14*(6), 2559–2567.

Guiahi, M. (2000). Methodologies for fitting alternative statistical distributions to a sample of insurance data. *Insurance Research*, *25*(4), 315–332.

Kazemi, K., & Noorizadeh, N. (2015). Quantitative comparison of spm, fsl, and brainsuite for brain mr image segmentation. *Journal of Biomedical Physics and Engineering*, *5*(3), 119–126.

Krzysko, M., & Krzysko, A. (2012). Non-parametric methods for estimating value-at-risk and expected shortfall. *International Journal of Financial Studies*, *10*(1), 47–75. doi: 10.3390/ijfs10010047

Liu, K., & Pan, K. (2010). Comparison of var and es for risk management: Evidence from the financial crisis. *Journal of Banking & Finance*, *34*(10), 2439–2453. doi: 10.1016/j.jbankfin.2010.02.007

Luckner, L., Van Genuchten, M. T., & Nielsen, D. R. (1989). Estimation of the unsaturated hydraulic properties of multiphase flow in porous media. *Water Resources Research*, *25*(4), 617–629.

Meyer, R. (2005). Actuarial modelling for fitting loss distributions to 250 claims. *Insurance Review*, *20*(3), 127–140.

Miljkovic, D., & Grun, B. (2016). Flexible mixture models for danish fire insurance loss data. *ASTIN Bulletin*, *46*(1), 181–204.

Mohammed, S., & Dzogbede, R. (2021). Predicting health insurance claims in ghana using machine learning algorithms. *BMC Medical Informatics and Decision Making*, *21*(1), 253. doi: 10.1186/s12911-021-01943-7

Montero, J.-M., Mínguez, R., & Fernandez-Aviles, G. (2018). Prediction of house prices. *Expert Systems with Applications*, *103*, 1–11.

Moyo, M., & Ogunyemi, K. (2018). Modelling frequency and severity of motor insurance claims in south africa. *Insurance: Mathematics and Economics*, *77*(C), 21–32. doi: 10.1016/j.insmatheco.2018.03.004

Peng, S., & Huang, Y. (2012). Estimating value at risk and expected shortfall using elliptical copulae. *Journal of Financial Econometrics*, *10*(3), 483–519. doi: 10.1093/jjfinec/nbj023

Punzo, A. (2019). Consideration of a mixture model with reparametrized inverse gaussian components for data on bodily injury claims and income of italian households. *Insurance: Mathematics and Economics*, *86*, 166–174.

Punzo, A., Mazza, A., & Maruotti, A. (2018). Consideration of a mixture model with contaminated gamma distributions for workers compensation and healthcare expenditure data. *Journal of Applied Statistics*, *45*(8), 1464–1480.

Reed, A. E. (2011). Modelling of human lifetimes. *Journal of the Royal Statistical Society: Series A*, *174*, 1–22. doi: 10.1111/j.1467-9868.2010.00710.x

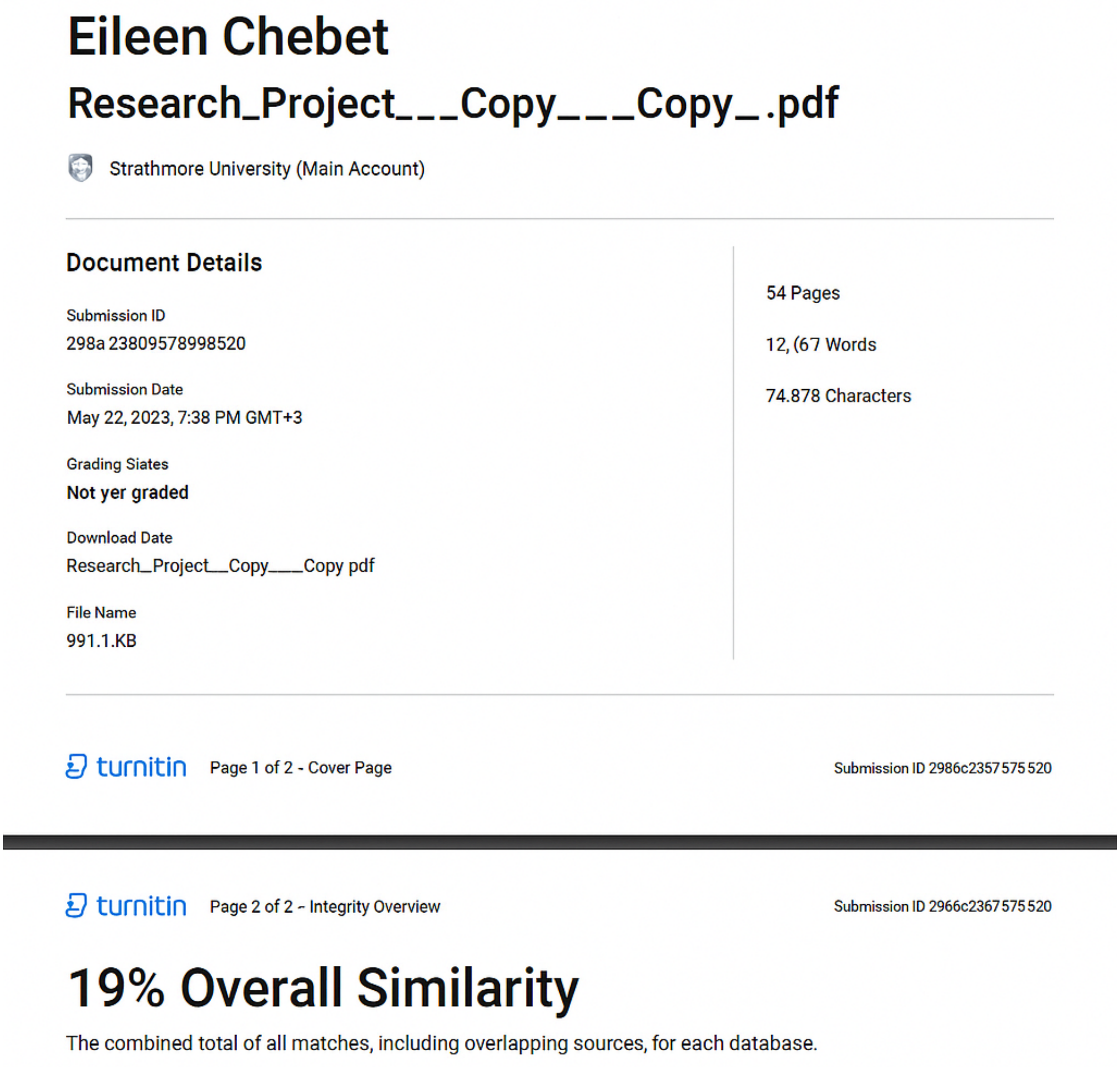
Silva, M., Kadiyala, K., Akhtar, R., & Wu, J. (2018). Analysis of the efficiency of indian commercial banks. *Emerging Markets Finance and Trade*, *54*(1), 230–247.

Taslimi-Renani, A., Ting, D. S.-K., Khosravi, A., & Haghifam, M.-R. (2016). Modeling wind turbine power curves. *Renewable and Sustainable Energy Reviews*, *60*, 1429–1442.

Wright, S. (2005). Actuarial modelling procedures for fitting models to 490 claim amounts drawn from seven consecutive years. *Insurance Studies*, *15*(2), 89–104.

Appendices

Appendix A: Similarity Report



Eileen Chebet

Research_Project___Copy___Copy_.pdf

Strathmore University (Main Account)

Document Details

Submission ID
298a 23809578998520

Submission Date
May 22, 2023, 7:38 PM GMT+3

Grading Siates
Not yer graded

Download Date
Research_Project__Copy____Copy pdf

File Name
991.1.KB

54 Pages

12, (67 Words

74.878 Characters

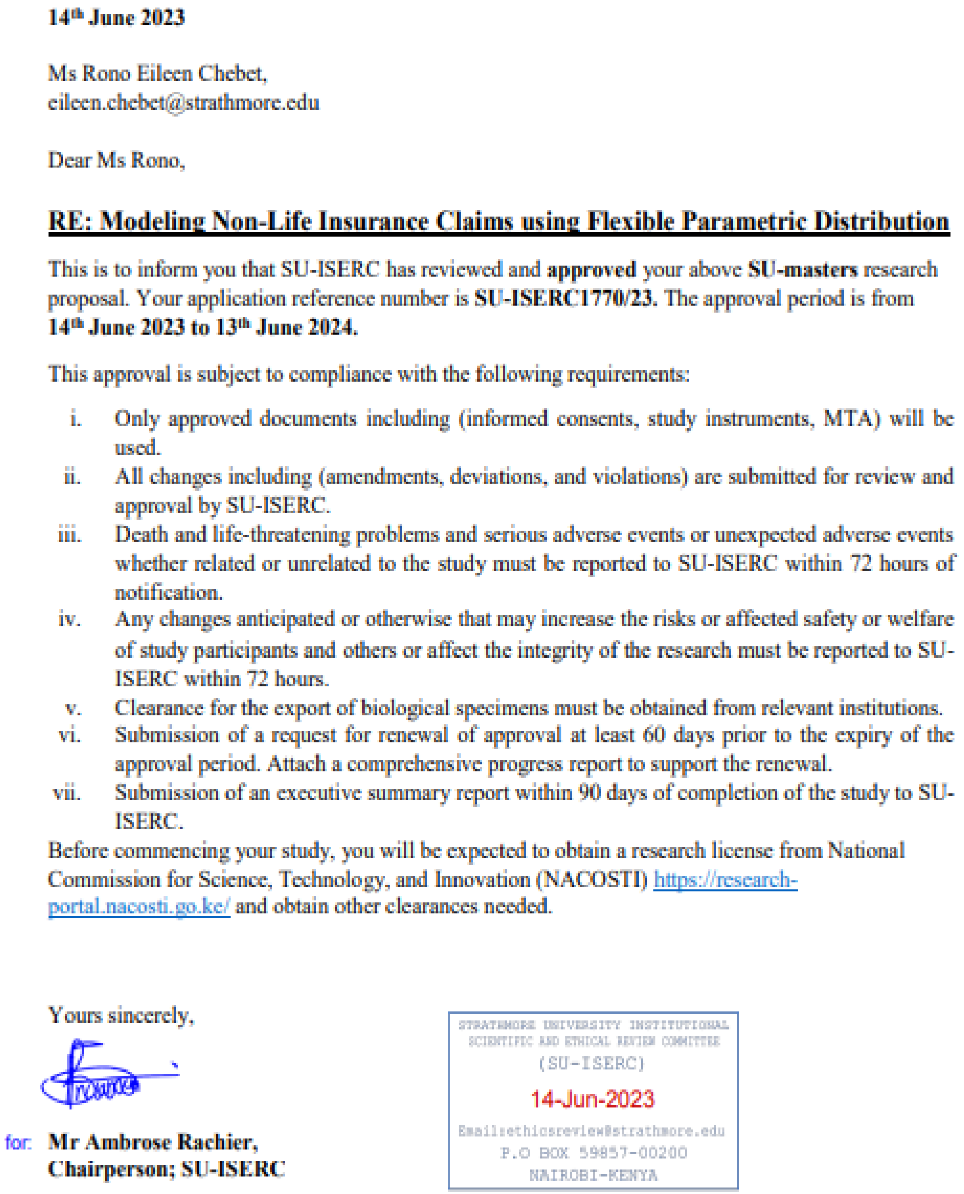
turnitin Page 1 of 2 - Cover Page Submission ID 2986c2357 575 520

turnitin Page 2 of 2 - Integrity Overview Submission ID 2966c2367 575 520

19% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Appendix B: Ethical Clearance Release Letter



14th June 2023

Ms Rono Eileen Chebet,
eileen.chebet@strathmore.edu

Dear Ms Rono,

RE: Modeling Non-Life Insurance Claims using Flexible Parametric Distribution

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC1770/23.** The approval period is from **14th June 2023 to 13th June 2024.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

for: **Mr Ambrose Rachier,**
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL
SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
14-Jun-2023
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA