UNIVERSITY OF NAIROBI

Loss Distributions For Motor Insurance Claim Severity

Case Study: Kenya

Group Members:

	Name	Registration Number
1.	Lillian Ayoo	I07/0817/2018
2.	Nelvine Anyango	I07/0811/2018
3.	Joy Kanyi	I07/132677/2018
4.	Kennedy Mwavu	I07/0807/2018
5.	Rachael Kanini	I07/0878/2018

Contents

List of Figures

List of Tables

1 Acknowledgements

We would love to appreciate our supervisors, Professor Patrick G.O. Weke and Dr. Carolyne Adhiambo Ogutu, for their tremendous support, practical advice and insightful comments that have been a guiding light in our research and writing of this project. Their professional expertise and vast knowledge in their respective fields has enabled us to successfully complete this project.

2 Introduction

2.1 Background Study

Insurance dates back to early human society. There are two known types of economies in human societies: natural or non-monetary and monetary economies. Insurance in the former case entails agreements of mutual aid. Granaries embody an early form of insurance to indemnify against famines. These types of insurance have survived to the present day in countries or areas where a modern money economy with its financial instruments is not widespread. According to Vaughan (1997) the first methods of transferring or distributing risk in a monetary economy were practiced by Chinese and Babylonian traders in the 3rd and 2nd millennia BC, respectively.

Insurance protects policyholders from possible losses. The growth and development of the insurance industry is highly motivated by the general demands of the society for the need of having protection against various types of risks of unpleasant random events with a major economic impact; Mihaela (2015). The underlying concept of insurance is to create a fund to which the insured members contribute known amounts of premium for a given risk level. When the random events that policyholders are protected against occur giving rise to claims then claims are settled from the fund. The insurer is needed to settle the claim, and this is referred to as loss. Insurers are keen with the results of the random outcome of claims instead of the existence of the claims. They are concerned with the loss rather than the circumstances that give rise to the loss; Achieng, O. M. (2010). A desirable feature of such an arrangement is that the insured members are faced with a homogeneous set of risks that are independent of each other. The pooling together of risks enables members to benefit from the law of large numbers.

Motor industry encompasses the management of large numbers of risk events. These arise due to instances of theft, fire and damage to vehicles due to accidents or other causes as well as the extent of damage to the parties involved. In most countries, the auto insurance industry is growing rapidly due to legislations that make motor insurance compulsory for all vehicles. The aggregate amount of claims in a given duration is a measure that is vital to the operations of an insurance company.

Fundamentally, an actuary in charge of general insurance claims needs to

understand various risk models comprising of the aggregate claim amount overdue in a given period. Boland (2006), found out that these models enlighten a company and allow it to decide on things such as premiums charged, anticipated profits, required reserves that will guarantee profitability with a high likelihood and the effect of reinsurance and policy excess. The claims data contains, among other things, the frequency and size of claims that a company has received within a given period. Based on the claims data, mathematical methods can be applied to model individual claims.

Actuarial models assist insurance companies to deal with these large amounts of data. The main challenge when using these kinds of data is the uncertainty that comes when trying to predict the future motor insurance claims. This uncertainty necessitates the use of statistical methods when trying to model the occurrence of claims, the timing of the settlement and the severity of the claims. The mathematical models are known as loss distributions.

Until recent years, auto insurance premiums have been determined by classical characteristic variables like age, years of driving experience, value of vehicle, and many other factors as shown in Table 1. Premiums can also be affected by a poor driving record, known as a Bonus-Malus System (BMS), or by requesting more coverage. However, you can reduce your premiums by agreeing to take on more risk, which means increasing your deductible.

Most general insurance companies base their estimations of claim frequency and severity on their own historical claims data. This is sometimes complemented with data from external sources and it is used as a base for managerial decisions. In analyzing data, the focus is usually on two concerns. First, it is a consensus of the importance of identifying critical explanatory variables for rating purposes. Insurance companies frequently take up a "risk-factor rating system" in determining premiums for motor insurance, so that identifying these important risk factors forms a critical process in developing insurance rates, Frees W., E., & Valdez, E. A. (2012). The second concern is to be able to predict claims as accurately as possible. Actuaries require accurate predictions for pricing, estimating future company liabilities, and for understanding the implications of these claims to the solvency of the company.

Table 1: Factors Affecting Car Insurance Premiums

Factor	Description
Age	• Data shows that young drivers are more likely to be involved in accidents. • Insurance costs should noticeably drop when a driver reaches around 21 years old, as long as they haven't been involved in an accident.
Driving Experience	• The more experience you have, the cheaper your car insurance premium. • Points on your license for speeding will result in a higher premium in the next year.
Vehicle Driven	• The make, model, age, security, value and size of your car all affect the price of your insurance i.e. sports cars are more likely to be involved in accidents – higher risk. • Repairing powerful cars is going to be a long and expensive process, adding to the cost of a premium.
Previous Claim History	• Insurers use data on previous claims to calculate your premium. • Insurers have developed systems which reward/penalize policyholders depending on the number of claims they make in a year. This is known as a No-Claim Discount (NCD) or, as stated earlier, a Bonus Malus System (BMS).
Location	• Rural and urban areas will have different premium costs. • Where the car is parked (road or garage) will also affect the price.
Miles Driven Annually	• Some insurance companies will ask for the amount of miles driven in the previous year as an indicator to gauge the level of risk they may be exposed to.

A loss distribution is the associated probability distribution of a claim-size variable. The claim-size is a non-negative continuous random variable since the claim arising from a covered incident can be measured in the lowest unit of currency e.g. cents. Loss distributions are usually positively skewed and long-tailed. They are vital as they are used for many purposes which include: premium rating, reserving, reviewing reinsurance arrangements and testing for solvency.

The number of claims in a discrete portfolio makes discrete standard distributions appropriate since their probabilities are explained on non-negative numbers. Many actuarial models for claims amounts are established on continuous distributions. The Lognormal and Gamma distributions fall mostly among the commonly used distributions for modeling claim amounts, Bahnemann (2015). Various distributions used to model claim severity are the Exponential, Weibull, and Pareto distributions.

Achieng, O. M. (2010) carried out a research on a model of claim amounts from First Assurance Company Limited, Kenya for motor comprehensive

policy consisted of the lognormal distribution which was chosen as the most suitable model that would provide a good fit to the motor insurance claims size data. Nduwayezu (2016) found out that the exponential distribution is suitable to model insurance data. These are examples of parametric methods which assume that the data set used is quantitative; the population in the data set has a normal distribution and the sample size is large. On the contrary, non-parametric methods make no assumption on the population distribution and sample size. Generally, conclusions drawn from non-parametric methods are not as reliable as the parametric ones. However, according to Hesse, J.B, & E.N. (2017) since non-parametric methods make fewer assumptions, they tend to be more flexible and applicable to non-quantitative data

Notably, most statistical distributions suggested for modelling claims severity are general and not exhaustive since they are based on the sample data that was being used by the various researchers. In particular, the aim of this research is to take a closer look into the Kenyan market and recommend specific statistical distributions that could be used for modelling auto insurance claim severity based on the motor insurance claims data in Kenya.

2.2 Problem Statement

When it comes to accurately forecasting future claims experience, most non-life insurers are faced with challenges on how to precisely estimate the likely prospective claims experience and therefore charging suitable premiums and setting aside sufficient reserves Omari et al. (2018).

Although the empirical distribution functions can be useful tools in understanding claims data, there is always a desire to "fit" a probability distribution with reasonably tractable mathematical properties to the claims data. There is need therefore to have a good estimate of loss distributions which entails selecting a suitable statistical distribution that fits the claims data.

Most important, is the question "If the insured event occurs, what will be the cost to the insurer?" When determining motor insurance claims distributions, we often associate the value of claims with two elements: the occurrence of an accident and the claim amount in case of an accident Frees W., E., & Valdez, E. A. (2012).

The following types of claims are recorded in the motor insurance database:

third-party liability claims, and damage claims to the policyholder, comprising property damage, injury, theft, and fire. This, therefore, implies that for every accident, it is probable for multiple types of claims to be incurred; thus, increasing the claim severity of an insurance company for every single accident. This creates a need for having good models of loss distributions that will enable an insurer to plan accordingly to lower the probability of incurring such a loss and reducing the claim severity incurred.

Motor Insurance Portfolio is an important premium source for all insurers. It constitutes nearly 48% of Insurers' total GDPI. It is also the single largest contributor to the underwriting losses of the insurers. Insurance of motor vehicles against third party risks is compulsory in Kenya, therefore for every new vehicle purchase in the country, a motor insurance policy is added to the existing motor insurance policies for the vehicles on the Kenyan roads. Motor insurance plays a huge role in the Kenyan insurance industry and the economy at large. This implies that motor insurance companies need to have good models that will enable them to accurately forecast future claims experience and thus be able to set aside enough reserves and avoid the occurrence of ruin.

Motor insurance while being one of the largest general insurance classes, it is also known for the massive losses it makes. This has forced the country to recently double the motor insurance premiums in an attempt to mitigate the losses. This could be partly attributed to the fact that motor insurance companies in Kenya do not have good models for loss distributions, and thus cannot be able to correctly forecast future claims experience. This leads them to undergo huge losses because of failing to plan accordingly to lower the probability of incurring such losses.

We therefore, seek to address this niche in the Kenya motor insurance industry by providing a good model of loss distribution for claim severity. This will, in turn, help insurers to precisely estimate prospective claims experience and thus plan accordingly to reduce their huge losses and the chances of them making such losses.

2.3 Research Objectives

The objective of this research paper is to present an appropriate statistical distribution that approximately fits claims severity losses of motor insurance

in Kenya and can be used to accurately forecast future claims experience.

2.4 Justification of the Study

This research will help motor insurance companies in Kenya to correctly forecast future claims experience.

In this way, they will be able to:

- Rate premiums to be paid by policyholders correctly.
- Reserve correctly, that is know the right amount of money to be retained in order to offset the cost of claims made by policyholders.
- Correctly test for solvency, which is simply evaluating the insurance financial condition.
- Review reinsurance arrangements.

Consequently, policyholders' premiums will reduce because the risk associated with claim severity would have been reduced.

Generally, the research will provide motor insurance in Kenya with the most appropriate loss distribution for claim severity for a better performance in future.

3 Literature Review

This section has a detailed overview of the actuarial modelling in insurance claim severity. This uses probability distributions as discussed by other research studies, i.e. published books, reports, and prior collected opinions. This chapter seeks to widen the study scope hence revealing the information gap.

3.1 Introduction

Loss distribution is the probability distribution associated with either the loss or the amount paid due to the loss. This paper involves the steps taken in actuarial modelling to find a suitable probability distribution for the claims data observed and testing for the goodness of fit of the supposed distribution.

R. V. Hogg, S. A. Klugman (1984) gives good introduction on fitting distributions to losses. Emphasis is on the distribution of single losses related to claims made against various types of insurance policies. These models are informative to the company and they enable it make decisions on amongst other things: premium loading, expected profits, reserves necessary to ensure (with high probability) profitability and the impact of reinsurance and deductibles.

Dutta K. and Perry J (2006) in the recent past, carried out an empirical study on loss distributions using Exploratory Data Analysis and empirical approaches to estimate the risk. However, due to lack of flexibility and poor results, they rejected the idea of using exponential, gamma and Weibull distributions, pointing out that "one would need to use a model that is flexible enough in its structure." Hence it is imperative to develop models either from the existing distributions or a new family of models to cater insurance loss data, financial returns, etc.

Achieng, O. M. (2010), did actuarial modeling for insurance claim severity in motor comprehensive policy in 2010, she used claim amounts from First Assurance Company Limited, Kenya. The modeling process established one statistical distribution that could efficiently model the claim amounts, she then did a goodness of fit test both mathematically and graphically using Akaike Information Criterion (AIC) and Quantile-Quantile Plots (Q-Q plots) respectively. The study established that the log-normal distribution was a

suitable model for the claims data.

Packova, V., & Brebera, D. (2015), applied Pareto model in reinsurance. They used data from Czech insurance company for compulsory motor third-party liability insurance. Goodness-of-fit test concluded that the Pareto distribution is a good model for large claims.

Selvakumar V et al. (2022) modelled motor insurance extreme claims using various distributions. They fitted gamma, Pareto, lognormal, Weibull, log-logistic distributions to the claims data. A goodness-of-fit test, Q-Q Plots test and P-P Plots test were then performed on them and it was concluded that the lognormal distribution was the best fitting distribution for overall claim amounts.

Omar et al. (2018) used the Maximum Likelihood Estimation method to obtain parameter estimates for the fitted models. With Auto Collision, data Car, and data Ohlsson as variables they modelled a sample of the automobile portfolio datasets obtained from the insurance Data package in R. After carrying out a goodness-of-fit test, Akaike Information Criterion, Bayesian Information Criterion were applied on the claims data. It was concluded that the lognormal distribution provides a good model for claims severity on a short-term basis.

Since we aim to fit a suitable loss distribution to the claim severity data for motor insurance companies using the company-specific data set, it is important to explain the theoretical framework that will be applied throughout this research. This includes parameter estimation, goodness-of-fit test, standard normal distributions and model selection criteria. These then form the basis of the subsequent parts of the study that will eventually yield an appropriate loss distribution model.

3.2 Estimation Of Parameters

Here, we use the Maximum Likelihood Estimation (MLE) technique.

The MLE is a commonly applied method of estimation in a variety of problems. We opt for this method because it often yields better estimates compared to other methods like Least-squares Estimation (LSE), the method of quantile and method of moments especially when the sample size is large.

Boucher et al., (2007) argued that the MLE approach fully utilizes all the

information about the parameters contained in the data and yields highly flexible estimator with better asymptotic properties.

Likelihood function, say $L(\theta)$, is the probability or probability density function of the observed data expressed as a function of the unknown parameter θ.

The principle of maximum likelihood provides a means of choosing an asymptotically efficient estimator for a parameter or a set of parameters as the value for the unknown parameter that makes the observed data "most probable". The maximum likelihood estimate (MLE) of a parameter θ is obtained through maximizing the likelihood function.

3.2.1 Exponential Distribution

The exponential distribution is a continuous distribution that is usually used to model the time until the occurrence of an event of interest in the process. A continuous random variable, say X, is said to have an exponential distribution if its probability density function (pdf) is given by:

$$f(x) = \lambda e^{-\lambda x},\ x > 0 \tag{1}$$

Some people have used this distribution to model claim severity and some of their conclusions about it are discussed below:

Achieng, O. M. (2010) after using exponential distribution, she concluded that the distribution was not a good fit. Its density value of its probability density function graphical plot and log-likelihood were low hence the conclusion.

Omari et al (2018) used this distribution to model an automobile dataset. The distribution had the largest values among distributions used for the Kolmogorov-Smirnov and Anderson Darling Tests hence rejected null hypothesis.

Mazviona, B. W., & Ch.iduza, T (2013), also, rejected null hypothesis after using this distribution to model motor dataset. This distribution failed to fit the data very closely based on the critical value for the chi-square test.

3.2.2 Lognormal Distribution

Suppose X represents the claim size, and $Y = \log(X)$ has a normal distribution, then X is said to have a lognormal distribution. When X has a lognormal distribution with parameters μ and σ , then:

$$X \sim \log\ N(\mu, \sigma^2) \tag{2}$$

The pdf of the lognormal distribution is given by:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2},\ \text{for } 0 < x < \infty \tag{3}$$

This distribution proves to be ideal in most cases because of its heavy-tailed and highly skewed nature. Achieng, O. M. (2010), while modelling claim severity motor comprehensive data of First Assurance Company Limited Kenya (June 2006 – June 2007), concluded that the lognormal distribution had the smallest Akaike Information Criterion (AIC) value. Hence the lognormal distribution was found to be the best statistical distribution to model the claim amounts at 99% level of confidence.

Omar et al. (2018), found out that the lognormal distribution was the most suitable model to fit an automobile dataset, since it had the lowest Akaike Information Criterion and Bayesian Information Criterion values.

Selvakumar V et al. (2022), carried out a goodness-of-fit test, Q-Q Plots test and P-P Plots test while modelling motor insurance extreme claims. They found out that the lognormal distribution was the best fitting distribution for overall claim amounts.

3.2.3 Gamma Distribution

A random variable, say X, has a gamma distribution with parameters $\alpha > 0$ and $\lambda > 0$ written as $X \sim Gamma(\alpha, \lambda)$.

Its probability density function is given by:

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x},\ x > 0 \tag{4}$$

α is referred to as the shape parameter because it changes the shape of the graph while λ is referred to as scale parameter since it affects the x scale of the graph.

Since majority of claim data arising from the insurance industry are positive skewed and heavy tailed, as noted by Merz and Wuthrich, Achieng, O.M (2010) used the gamma distribution to model claim severity in First Assurance Company Limited, Kenya because it has the two properties. In his study, he came to a conclusion that the distribution was not a suitable model for the claims data based on its Q-Q plot.

Mazviona, B. W., & Ch.iduza, T (2013), in their study to model a motor dataset, used the gamma distribution. They rejected null hypothesis because the distribution, based on the critical value for the chi-square test, failed to fit the data very closely.

Packova, V., & Brebera, D. (2015) used this distribution to model data obtained from a Czech insurance company for compulsory motor third-party liability insurance. They used this distribution because it is specifically applicable for modelling claim severity. However, based on Anderson-Darling test value, their study found out that gamma distribution failed to be a suitable model for this.

3.2.4 Pareto Distribution

Suppose the random variable say X is the claim size and has the Pareto distribution with parameters $\alpha > 0$ and $\lambda > 0$; that is $X \sim Pa(\alpha, \lambda$. The distribution function is given by:

$$F(x) = 1 + \left(\frac{\lambda}{\lambda + x}\right)^{\alpha}, \; x > 0 \tag{5}$$

The pdf is given by:

$$f(x) = \frac{\alpha\lambda^{\alpha}}{(\alpha + x)^{\alpha+1}}, \; x > 0 \tag{6}$$

Packova, V., & Brebera, D. (2015), modelled data from Czech insurance company for compulsory motor third-party liability insurance using Pareto distributions. This is because the distribution frequently models insurance

losses required to obtain well-fitted tails. The study established that based on the tests carried out, for large claims, Pareto distribution is a good model.

Omar et al. (2018) also used Pareto distribution to model an automobile dataset. Based on the tests carried out, the Pareto distribution was discarded since its values were extremely out of range.

Selvakumar V et al. (2022) modelled motor insurance extreme claims using Pareto distribution as one of the distributions to fit the data. However, it was found that it was not a good fit based on the tests carried out and that the lognormal distribution best fits the data.

3.2.5 Weibull Distribution

This distribution is particularly useful in modelling left-truncated claim severity distributions. It has been used to model excess of loss treaty over automobile insurance.

The continuous variable X has Weibull distribution with parameters c, γ and probability density function is given by:

$$f(x) = c\gamma x^{\gamma-1} e^{-cx^{\gamma}},\ x > 0 \tag{7}$$

The distribution function of the Weibull distribution is given by:

$$F(x) = 1 - e^{-cx^{\gamma}}$$

Ahmad et al. (2020) used Weibull distribution to model vehicle insurance loss data and provide greater accuracy in data fitting. Numerical results from their study show that the somewhat improved distribution outperforms other existing long-tail distributions under the different measures of model assessment considered in respect to vehicle insurance loss data.

Packová and Brebera (2015) used Weibull distribution to model data for compulsory motor third-party liability insurance. The study found out that the Weibull distribution failed to be a good model for the losses based on the Anderson-Darling test value.

Achieng, O. M. (2010) applied the Weibull distribution due to its heavy-tailed and highly skewed nature to model claim severity motor comprehensive data

of First Assurance Company Limited, Kenya (June 2006 - June 2007). The study established that the Weibull distribution is not a suitable model for the claims data.

3.2.6 Burr Distribution

The probability density function of Burr random variable X is defined by:

$$f(x) = \frac{\alpha\gamma\lambda^{\alpha}x^{\gamma-1}}{(\lambda + x^{\gamma})^{\alpha+1}}, \text{ for } \alpha, \lambda, \gamma, x > 0 \tag{8}$$

The distribution function of Burr random variable X is defined as:

$$F(x) = 1 - \left(\frac{\lambda}{\lambda + x^{\gamma}}\right)^{\alpha}$$

Hakim A. R et al. (2021) applied the Burr distribution due its similarity to Pareto distribution in terms of tail index. Several numerical computations and Kolmogorov–Smirnov test in the study revealed that the Burr distribution fits well to model the claims data.

Burnecki et al. (2010) used the Burr distribution to model fire losses dataset. The study failed to suggest the Burr distribution as a good model since it failed to pass the applied tests therein.

It is worth to mention that while this distribution can be used to model claim size distributions, most studies fail to utilize it among their chosen distributions.

3.3 Goodness Of Fit Tests

A goodness-of-fit test is a statistical procedure that describes how well a distribution fits a set of observations by measuring the quantifiable compatibility between the estimated theoretical distributions against the empirical distributions of the sample data (Omari et al., 2018). This enables one "to determine whether the observed sample was drawn from a population that follows a particular probability distribution" (Dodge, 2008).

The tests are effectively based on either of the two distribution functions: the probability density function (PDF) or cumulative distribution function

(CDF) in order to test the null hypothesis that the unknown distribution function is, in fact, a known specified function. The tests considered for testing the suitability of the fitted distributions to claims data include; the Chi-Square goodness of fit test, Kolmogorov-Smirnov (K-S) test, and the Anderson-Darling (A-D) test.

Here, the Kolmogorov-Smirnov and Anderson-Darling tests are used because they are suitable for performing an exact test on continuous distributions.

For all the Goodness of Fit tests, the hypotheses of interest are:

- H_0: The claims data sample follows a particular distribution.
- H_1: The claims data samples do not follow the particular distribution.

3.3.1 Kolmogorov-Smirnov Test

Named after Russian mathematicians Andrey Kolmogorov and Nikolai Smirnov, the Kolmogorov-Smirnov test is a non-parametric (does not rely on any distribution to be valid) goodness-of-fit test and is used to determine whether an underlying probability distribution differs from a hypothesized distribution.

Consider an independent random sample $(x_1, x_2, \ldots, x_n)$, a sample of size n with unknown distribution function $F(x)$ coming from a population with a specific and known distribution function $F_0(x)$. The hypothesis to test is as follows:

$$H_0 : F(x) = F_0(x)$$
$$H_1 : F(x) \neq F_0(x)$$

If $F(x)$ is the empirical distribution function of the random sample, then the statistical test T_n is defined as the greatest vertical distance between $F_0(x)$ and $F(x)$.

The decision rule is to reject H_0 at the significance level α if T_n is greater than the value of the Kolmogorov table having for the parameters n and $1 - \alpha$, which is denoted by $t_{n,1-\alpha}$ i.e, if:

$$T_n > t_{n,1-\alpha}$$

3.3.2 Anderson-Darling Test

The Anderson–Darling test is a goodness-of-fit test which allows to control the hypothesis that the distribution of a random variable observed in a sample follows a certain theoretical distribution (Dodge, 2008).

Consider a random variable X, which follows a particular distribution, and has a distribution function $F_0(x;\theta)$, where θ is a parameter (or a set of parameters) that determine, F_0. We further assume θ to be known. An observation of a sample of size n issued from the variable X gives a distribution function $F(x)$. The Anderson-Darling statistic, denoted by A^2, is then given by the weighted sum of the squared deviations $F_0(x;\theta) - F(x)$.

Starting from the fact that A^2 is a random variable that follows a certain distribution over the interval $[0;+\infty]$, it is possible to test, for a significance level that is fixed a priori, whether $F(x)$ is the realization of the random variable $F_0(X;\theta)$; that is, whether X follows the probability distribution with the distribution function $F_0(x;\theta)$.

To compute the A^2 statistic, arrange the observations $x_1, x_2, \ldots, x_n$ in the sample obtained from X in ascending order i.e., $x_1 < x_2 <$ ůůů $< x_n$. The A^2 is then computed as:

$$A^2 = -\frac{1}{n}\left(\sum_{i=1}^{n}(2i-1)(ln(z_i)) + ln(1-z_{n+1-i})\right) - n$$

where $z_i = F_0(x_i;\theta), (i = 1, 2, ..., n)$

The null hypothesis is rejected beyond the limiting values of A^2 depending on the significance level based on the Anderson-Darling Test Table.

3.4 Information Criterion

An information criterion is a measure of the quality of a statistical model. It takes into account:

- How well the model fits the data
- The complexity of the model

Information criteria are used to compare alternative models fitted to the same data set. All else being equal, a model with a lower information criterion is

superior to a model with a higher value.

3.4.1 Akaike Information Criterion

Akaike's information criterion (AIC) developed by Akaike, H. (1974) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. The AIC is not a test of the distribution in the sense of hypothesis testing; rather it compares between distributions—a tool for distribution selection. Given a data set, several fitted distributions may be ranked according to their AIC. The fitted distribution with the smallest AIC is selected as the most appropriate distribution for modeling the claims data. The AIC value for a model is calculated as follows:

$$AIC = 2k - 2ln(L)$$

where k is the number of estimated parameters in the model and L is the maximum value of the likelihood function of the model.

3.4.2 Bayesian Information Criterion

The Bayesian Information Criterion (BIC) also known as Schwarz Information Criterion (also SIC, SBC, SBIC) is a criterion for model selection among a finite set of models; models with lower BIC are generally preferred. It is based, in part, on the likelihood function and it is closely related to the AIC. The appropriate model that is preferred is one that has the lowest BIC value since it implies a lower penalty term. In contrast to the AIC the Bayesian information criterion (BIC), comprises of the number of observations in the penalty term.

The BIC value for a model is calculated as follows:

$$BIC = k \times ln(n) - 2ln(L)$$

where k is the number of estimated parameters in the model, n is the number of observations and L is the maximized value of the likelihood function of the model.

4 Methodology

4.1 Introduction

This chapter discusses the methods used in the research. That is, the population and the sample, sources of data or the data collection methods and the design of our research. The data analysis process of our research is also discussed in this chapter.

We will take the claim size as our variable of interest in the motor insurance industry, and focus on data for Kenya motor insurance companies from 2016 to 2020. The companies that will form the population of the study are the licensed insurance companies providing motor insurance in Kenya as of 2020. Since the data is readily available from Insurance Regulatory Authority (IRA) and the population size is relatively small, we focus on the whole population. Various loss distributions such as exponential, gamma, lognormal, Pareto and Weibull are fitted on the data collected. These distributions are then tested for their goodness-of-fit before recommending an appropriate model.

IRA provides annual reports on activities within the insurance industry by various insurance companies. These reports are available in Microsoft Excel Binary File Format, hence it's easier to extract the relevant data from the report using Microsoft Excel and R. The data published in these reports is what we use as secondary data in our study.

Every insurer has a duty to price premiums profitably, and this can be guided if the insurer has a clue on which probability law better approximates the risk posed by the policyholders, Packova, V., & Brebera, D. (2015). This, according to researchers guide the insurer in making proper evaluations and predictions to avoid or minimize the potential losses. Unlike developed countries where auto insurers have robust pricing systems that capture policyholders' claims, most Kenyan motor insurance firms use a pricing system that pays little attention to the importance of claims histories. However, drawing from the above literature, the danger posed by policyholders' claims can never be undermined.

Notwithstanding, the claims ratios from most market players in Kenya according to the Insurance Regulatory Authority (IRA) annual reports are below the internationally accepted standards. The claim ratio is calculated as the net claims incurred divided by the Net Earned Premiums. It is an influ-

ential ratio indicating the strength an insurer exercise in paying claims and to some extent, how well policyholders are treated. Aside from the market claims ratio, another influential indicator of probability is the total expense ratio i.e. management expense and commission expense. This ratio is determined as a percentage of the Net Earned Premium with an internationally accepted ratio usually less than 40%. As this ratio becomes larger, it implies that the company is inefficiently discharging its duties, and this is more likely to impact its prompt payment of claims to policyholders.

Therefore, we argue that for an insurer to be financially solvent and avoid eroding policyholders trust, the claims from policyholders' must not be taken for granted. Hence, it's important to properly evaluate and predict policyholder's claims distribution to help offset the market's inefficiencies. Therefore, to guide the insurer in making proper evaluations and predictions to avoid or minimize potential losses that could end up eroding trust and to attain financial solvency, this study seeks to investigate the type of loss distribution function that best approximate the policyholders' claims in Kenya using real data from major insurance companies.

In compliance with the research objectives, this study seeks to find an appropriate model that fits claims size of motor insurance companies. The following steps will be followed when fitting a suitable model to claims data:

1. Select a family of distributions for the claims model.
2. Estimate the parameters for the model.
3. Specify a selection criterion to determine the appropriate distribution from the family of distributions.
4. Carry out a goodness-of-fit test on the selected appropriate distribution.

The first step that will be carried out on the data is to find its descriptive statistics such as mean, variance and skewness. These values will be essential when comparing them with the results obtained from the various models to select an appropriate loss distribution. The study also explains the framework that will be applied in the process of analyzing data. This will include parameter estimation, standard continuous distributions, goodness-of-fit test, and model selection criteria. These will then form the basis of data analysis that will eventually yield an appropriate loss distribution model.

4.2 Maximum Likelihood Estimation

The parameters of the chosen loss distribution in this study will be estimated using the Maximum Likelihood method. The most important stage in applying the method is that of writing down the likelihood:

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{9}$$

In most cases taking logs greatly simplifies the determination of the maximum likelihood estimator (MLE).

The following steps are used when determining a maximum likelihood estimate (MLE):

1. Specify the likelihood function for the available data.

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{10}$$

2. Simplify the algebra using natural logs.

$$I(\theta) = \log_L(\theta) = \sum_{i=1}^{n} f(x_i; 0) \tag{11}$$

3. Maximise the log-likelihood function by differentiating the log-likelihood function with respect to each of the unknown parameters and equating the resulting expression(s) to zero.

4. The MLEs of the parameters are obtained by solving the resulting equation(s). To ensure that the obtained values maximize the likelihood function, differentiate a second time.

4.3 Standard Continuous Distributions

4.3.1 Exponential Distribution

The distribution function is given by:

$$F(x) = 1 - e^{-\lambda x}, \; x > 0 \tag{12}$$

The pdf is given by:

$$f(x) = \lambda e^{-\lambda x},\ x > 0 \tag{13}$$

The mean and the variance are given by:

$$E(x) = \frac{1}{\lambda} \text{ and } var(x) = \frac{1}{\lambda^2} \tag{14}$$

The likelihood function is given by:

$$L = \prod_{i=1}^{n} \lambda e^{-\lambda x} = \lambda^n e^{-\lambda \sum x_i} = \lambda^n e^{-\lambda n \bar{x}} \tag{15}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

It's log-likelihood function is:

$$\log L = n \log(\lambda) - \lambda n \bar{x} \tag{16}$$

4.3.2 Gamma Distribution

The probability density function of gamma distribution is given by:

$$f(x) = \frac{\lambda \alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\lambda x},\ x > 0 \tag{17}$$

The mean and variance are given below:

$$E(x) = \frac{\alpha}{\lambda} \text{ and } var(x) = \frac{\alpha}{\lambda^2} \tag{18}$$

4.3.3 Lognormal Distribution

The probability density function is given by:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2},\ \text{for } 0 < x < \infty \tag{19}$$

The mean and variance are given by:

$$E(x) = e^{\mu + \frac{1}{2}\sigma^2} \text{ and } var(x) = e^{2\mu + \sigma^2 (e^{\sigma^2} - 1)} \tag{20}$$

M and σ^2 may be estimated using the log-transformed data hence easy to estimate the MLEs.

We let $x_1, x_2, ..., x_n$ be the observed values and, therefore MLE is given by:

$$\bar{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{21}$$

We also have $\hat{\sigma}^2 = s_y^2$ where subscript y is the sample variance computed on the y-values.

$$\hat{\sigma}^2 = \frac{1}{\eta}\sum_{i=1}^{n} (\bar{y_i} - y)^2 = \frac{1}{n}\sum_{i=1}^{n} (ln(\hat{x}_i) - \mu)^2 \tag{22}$$

4.3.4 Weibull Distribution

The distribution function is given by:

$$F(x) = 1 - e^{-cx^{\gamma}} \tag{23}$$

Its probability density function is given as:

$$f(x) = c\gamma x^{\gamma-1} e^{-cx^{\gamma}}, \text{ where } x > 0 \tag{24}$$

When c and γ are unknown, it is not easy to apply the method of Maximum Likelihood.

However, the equations are elementary when we use a computer.

Where γ has a known value, Maximum Likelihood is now easy.

4.3.5 Pareto Distribution

The distribution function of Pareto distribution is given by:

$$F(x) = 1 - \left(\frac{\lambda}{\lambda + x}\right)^{\alpha}, \ x > 0 \tag{25}$$

The probability density function is given below:

$$f(x) = \frac{\alpha\lambda^{\alpha}}{(\lambda + x)^{\alpha+1}}, x > 0 \tag{26}$$

The mean and the variance are given by:

$$E(x) = \frac{\lambda}{\alpha - 1},\ (a > 1);\ \text{and}\ var(x) = \frac{\alpha\lambda^2}{(\alpha - 1)^2\alpha - 2},\ (\alpha > 0) \tag{27}$$

The likelihood function is:

$$L = \prod_{i=1}^{n} \frac{\alpha\lambda^{\alpha}}{x_i\alpha + 1},\ 0 < \lambda \leq \min(x_i), \alpha > 0 \tag{28}$$

Its log-likelihood function is given by:

$$\log L = n \times \log(\alpha) + \alpha n \times \log(\lambda) - (\alpha + 1)\sum_{i=1}^{n} \log(x_i) \tag{29}$$

4.4 Goodness Of Fit Test

The purpose of goodness-of-fit tests is typically to measure the distance between the fitted parametric distribution and the empirical distribution. To select the most appropriate continuous distributions for the claims severity we apply both the Kolmogorov-Smirnov and Anderson-Darling tests since they are suitable for performing an exact test on continuous distribution.

For both the Goodness of Fit tests, our hypotheses of interest are:

- H_0: The claims data sample follows a particular distribution.
- H_1: The claims data samples do not follow the particular distribution.

Smaller K-S and A-D test statistic values will indicate that the distribution fits the data better.

4.5 Information Criteria

Both the information criteria, AIC and BIC are utilized for the selection of the most appropriate distribution for the claims data for all the selected claims frequency and claims severity distributions that pass the goodness of fit tests. The lower the value of these two criteria the better a model is.

4.6 Model Selection

The likelihood function (LLF), AIC and BIC criteria are employed for purposes of selecting the appropriate distribution among the fitted distributions. The distribution function with the maximum LLF value subject to passing the goodness of fit tests and minimum AIC or BIC values is selected as the most appropriate model.

5 Data Analysis

5.1 Introduction

The motor insurance claim severity data used was obtained from the annual reports of Insurance Regulatory Authority. The data was from $2016 - 2020$ and contained 37 insurance companies, all licensed and regulated by `IRA`. Incurred claims for both motor commercial and motor private classes were analyzed side by side to obtain suitable models for each category. All the analysis was done using the R Programming Language.

5.2 Descriptive Statistics

Getting a summarised overview of the data was critical in discovering patterns, especially on skewness.

Table 2: Descriptive statistics for incurred claims (2016-2020)

Stat	Motor Commercial	Motor Private
No. Of Observations	185	185
Mean	296940.7	408137.7
Standard Error	23745.16	30003.67
Median	173695	271419
Standard Deviation	322969.1	408094.0
Kurtosis	5.937997	5.777819
Skewness	1.741602	1.683955
Minimum	0	-24861
Maximum	1470770	1992246
Sum	54934035	75505472

It is clearly evident that incurred claims data for both classes of motor insurance are positively skewed, with motor private having a higher mean than motor commercial.

The next step was to get a visual representation of the distribution of the data.

From the histograms we can affirm that the data from both classes is not only skewed to the right, but also long-tailed. This gives a good hint of the kind of distributions most appropriate to model the data.

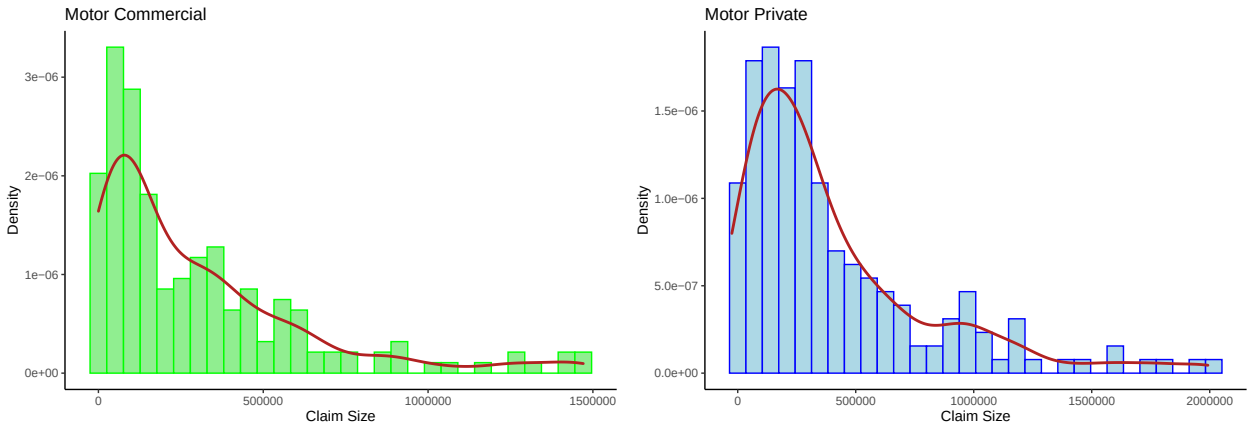

Figure 1: Histograms of original datasets

The Normal QQ-plots in Figure 2 show that the datasets don't match the normal standard distribution.

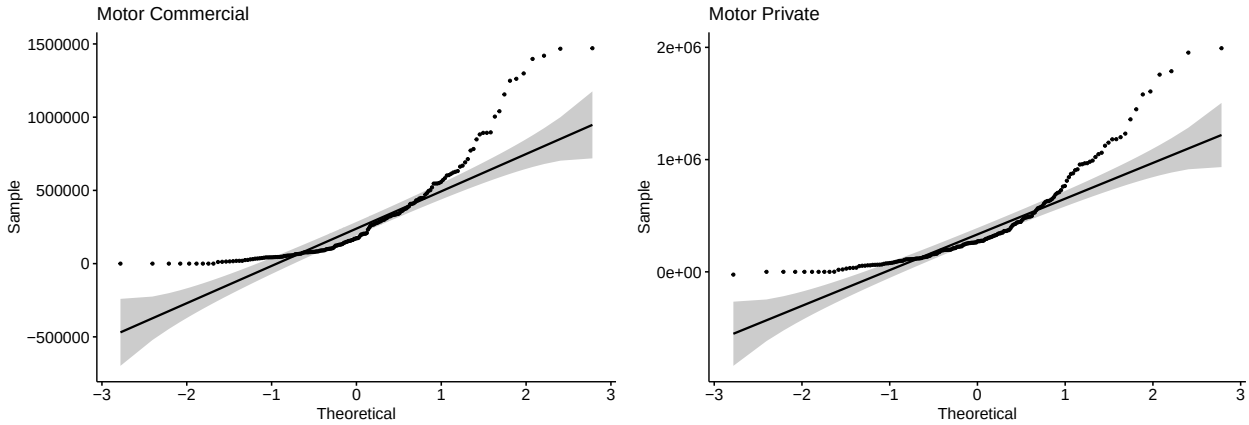

Figure 2: Normal QQ-plots of original data

This is a good indication that after fitting the distributions, non-parametric tests would be applied as opposed to parametric tests.

To make it simpler to work with, we transformed the data using the cube root function. QQ-Plots of the transformed data are shown in Figure 3.

The cube root transformed data seemed to be more suitable for fitting the distributions compared to the original data.

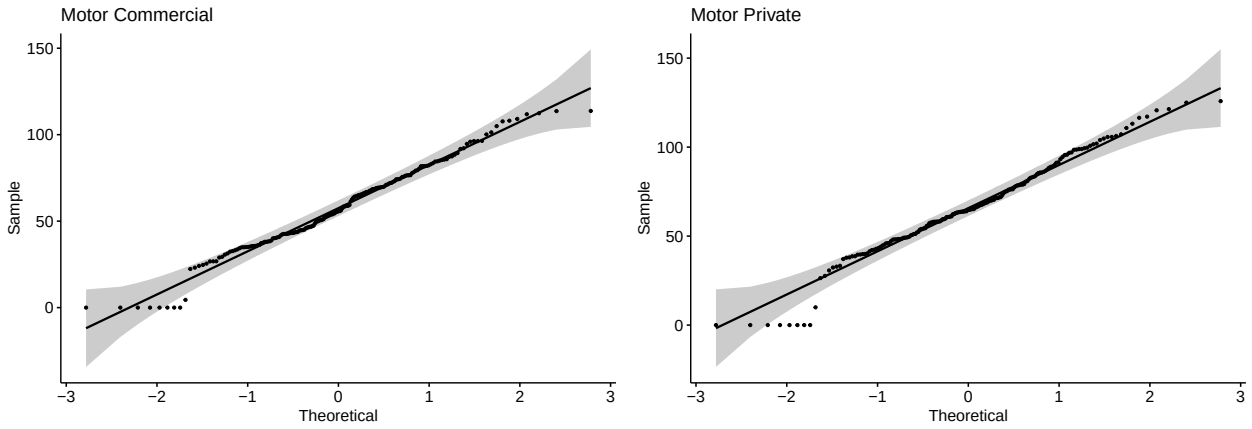

Figure 3: QQ-Plots of transformed data

5.3 Parameter Estimation

We used the Maximum Likelihood Estimation (`MLE`) method to obtain the various fitted distributions. Consequently, the most appropriate distribution is the one with the highest log-likelihood function (`LLF`).

Table 3: Estimated Parameters For Fitted Distributions

Distribution	Parameter	Motor Commercial	Motor Private
Exponential	Rate	0.02	0.01
	Std. Error	0.00	0.00
	LLF	-901.93	-920.29
Gamma	Shape	6.54	8.90
	Shape Std. Error	0.68	0.93
	Rate	0.11	0.13
	Rate Std. Error	0.01	0.01
	LLF	-800.47	-794.92
Log Normal	Mean Log	4.02	4.17
	Mean Log Std. Error	0.03	0.03
	SD Log	0.42	0.35
	SD Log Std. Error	0.02	0.02
	LLF	-810.28	-801.37
Weibull	Mean Log	2.91	3.33
	Mean Log Std. Error	0.17	0.19
	SD Log	67.40	76.47
	SD Log Std. Error	1.84	1.83
	LLF	-798.55	-795.27

Under the motor commercial class, Weibull distribution has the highest log-likelihood function (-798.55), followed by the Gamma distribution (-800.47), then Log-Normal (-810.28) and finally Exponential distribution (-901.93).

For motor private, Gamma distribution has the highest LLF (−794.92) followed closely by Weibull (−795.27), then Log-Normal (−801.37) and finally the Exponential distribution (−920.29).

From the LLF values, the Weibull distribution is the most appropriate for the motor commercial class and the Gamma distribution for motor private.

5.4 Goodness-Of-Fit Test

Typically, measures of goodness-of-fit summarize the discrepancy between observed values and the values expected under the model in question.

Two non-parametric tests were performed:

- Kolmogorov-Smirnov (K-S) test
- Anderson-Darling (A-D) test

Those two were used to determine the appropriateness of the fitted distributions to the incurred claims data.

Table 4: K-S and A-D test statistic values for fitted distributions

Test Statistic	Distribution	Motor Commercial	Motor Private
K-S	Exponential	0.3428	0.3785
	Gamma	0.0693	0.0395
	Log Normal	0.0846	0.0622
	Weibull	0.0655	0.0679
A-D	Exponential	31.8254	37.2804
	Gamma	0.6340	0.2429
	Log Normal	1.1405	0.6198
	Weibull	0.8083	0.9116

To determine the most suitable continuous distribution for the incurred claims, we fish for smaller K-S and A-D test statistic values.

For the motor commercial class, Weibull distribution had the smallest K-S statistic (0.0655), followed very closely by Gamma distribution (0.0693). But under the A-D statistic, Gamma (0.6340) beats Weibull (0.8083) by a considerable margin. The Gamma distribution would be the best fit for this class.

As expected for the motor private class, the Gamma distribution has the least K-S statistic (0.0395) as well as A-D statistic (0.2429) which again makes it the most appropriate fit for this class.

From the K-S and A-D tests, the Gamma distribution seems to be the best fit for both auto-insurance classes.

5.5 Information Criteria

Two approaches were used here:

- Akaike's Information Criteria (AIC)
- Bayesian Information Criteria (BIC)

Lower values for both AIC and BIC indicate a more appropriate distribution.

Table 5: AIC and BIC values for fitted distributions

Information Criterion	Distribution	Motor Commercial	Motor Private
AIC	Exponential	1805.85	1842.59
	Gamma	1604.95	1593.83
	Log Normal	1624.55	1606.74
	Weibull	1601.10	1594.54
BIC	Exponential	1809.03	1845.76
	Gamma	1611.30	1600.17
	Log Normal	1630.91	1613.08
	Weibull	1607.45	1600.88

For the motor commercial class, Weibull had the lowest AIC and BIC values (1601.10, 1607.45).

For motor private, Gamma distribution had the minimum AIC and BIC values (1593.83, 1600.17).

From the AIC and BIC values, Weibull was the most appropriate distribution for motor commercial and Gamma for motor private. In both auto-insurance classes, the Exponential distribution seems to have the worst fit.

6 Conclusion

Our research entailed determining the statistical distribution that best fits claim severity of motor insurance in Kenya. It could then be used to predict future claim incurrence.

After selecting a family of continuous, positively skewed distributions (Exponential, Gamma, Lognormal and Weibull), their parameters were estimated using the MLE method, K-S and A-D tests applied to test their goodness-of-fit, and finally AIC and BIC criterion used to determine the most appropriate distribution among the chosen ones.

The most suitable distribution is the one with:

- Maximum LLF
- Minimum K-S and A-D test statistic values
- Minimum AIC and BIC values

According to our study, the Weibull distribution is the best fit for modelling claim severity in motor commercial class, while the Gamma distribution is best for the motor private class.

7 References

[1] Achieng, O.M. (2010). Actuarial Modeling for Insurance Claim Severity in Motor Comprehensive Policy Using Industrial Statistical Distributions. International Congress of Actuaries, Cape Town, Vol. 712. 7-12.

[2] Ahmad, Z., Mahmoudi, E., Sanku, D. and Saima, K. K. (2020). Modeling Vehicle Insurance Loss Data Using a New Member of T-X Family of Distributions. Journal of Statistical Theory and Applications, 19, 133-147.

[3] Akaike, H. (1974). A New Look at Statistical Model Identification. IEEE Transactions on Automatic Control, 19, 716-723.

[4] Bahnemann, D. (2015). Distributions for Actuaries. CAS Monograph Series No. 2, Casualty Actuarial Society, Arlington.

[5] Boland, P. J. (2006). Statistical Methods in General Insurance. Unpublished M. Ed. Dissertation, National University of Ireland, Dublin.

[6] Boucher, J.P., Denuit, M. and Guillén, M. (2007). Risk Classification for Claim Counts: A Comparative Analysis of Various Zero-Inflated Mixed Poisson and Hurdle Models. North American Actuarial Journal, 11, 110-131.

[7] Burnecki, K., Misiorek, A. and Weron, R. (2010). Loss Distributions. Unpublished M. Ed. Dissertation, Wroclaw University of Technology, 50-370 Wroclaw, Poland.

[8] Dodge, Y. (2008). The concise encyclopedia of statistics: With 247 tables. New York, NY: Springer.

[9] Dutta K. and Perry J., (2006). A Tale of Tails: An Empirical Analysis of Loss Distribution Models for Estimating Operational Risk Capital. Federal Reserve Bank of Boston, MA, USA. Working Papers, No. 06-13.

[10] Frees, W. E. and Valdez, E. A. (2012). Hierarchical Insurance Claims Modeling. Journal of the American Statistical Association, 103, 484, 1457-1469.

[11] Hakim, A. R., Fithriani, I. and Novita M. (2021). Properties of Burr distribution and its application to heavy tailed survival time data. Journal of Physics: Conference Series, 1, 1, 6-8.

[12] Hesse, C., Ofosu, J. B. and Nortey, E. N. (2017). Introduction to Nonparametric Statistical Methods. Akrong Publications Ltd, Accra.

[13] Hogg R. V. and Klugman S. A. (1984). Loss Distributions. New York: John Wiley & Sons.

[14] lgnatov, Z. G., Kaishev, V. K. and Krachunov, R. S. (2001). An improved finite time ruin probability formula and its Mathematica implementation. Journal of Insurance: Mathematics and Economics, 29, 3, 375-386.

[15] Mazviona, B. W., & Chiduza, T. (2013). The Use of Statistical Distributions to Model Claims in Motor Insurance. International Journal of Business, Economics and Law, 3, 1, 44-57.

[16] Merz, M. and Wüthrich, M. V. (2008). Modelling the Claims Development Result for Solvency Purposes. Casualty Actuarial Society E-Forum Fall 2008, 542-568.

[17] Mihaela, D. and Jemna Dănuţ-Vasile (2015). Modeling the Frequency of Auto insurance claims by Means of Poisson and Negative Binomial models. Journal of Scientific Annals of Economics and Business of the "Alexandru Ioan Cuza" University of laşi, Romania, 62, 2, 151-168.

[18] Nduwayezu, F. (2016). Finding appropriate Loss Distributions to insurance data: Case study of Kenya (2010-2014) Unpublished M. Ed. Dissertation, Strathmore University, Nairobi, Kenya.

[19] Omari, C., Nyambura, S. and Mwangi, J. (2018). Modeling the Frequency and Severity of Auto Insurance Claims Using Statistical Distributions. Journal of Mathematical Finance, 8, 137-160.

[20] Packova, V., and Brebera, D. (2015). Loss Distribution in Insurance Risk Management. Proceedings of the International Conference on Economics and Statistics (ES 2015). Vienna, Austria, March 15-17, 2015, 17-22.

[21] Selvakumar, V., Satpathi, D. K., Praveen Kumar, P. T. V. and Haragopal, V. V. (2022). Modeling of Motor Insurance Extreme Claims through Appropriate Statistical Distributions. Unpublished M. Ed. Dissertation, Birla Institute of Technology and Science – Pilani, Hyderabad, India.

[22] Vaughan, E. J. (1997). Risk Management. New York: John Wiley.

8 Appendix: R Code For This Project

```
library(readxl)
library(tidyverse)
library(moments)
library(ggpubr)
library(actuar)
library(fitdistrplus)
library(kableExtra)
library(glue)

industry <- readxl::read_xlsx(
  path = "data/insurance_industry_stats_2016-2020.xlsx",
  skip = 1
) |>
  dplyr::select(
    `Class Name`, `2016`:`2020`
  )

industry_long <- industry |>
  tidyr::pivot_longer(
    cols = !`Class Name`,
    names_to = "Year",
    values_to = "Amount"
  )

# ----desc stats----
# standard error:
se <- function(x) {
  sqrt(var(x) / length(x))
}

# descriptive stats:
desc_stats <- industry_long |>
  group_by(`Class Name`) |>
  summarise(
```

```
    `No. Of Observations` = n(),
    Mean = mean(Amount),
    `Standard Error` = se(Amount),
    Median = median(Amount),
    `Standard Deviation` = sd(Amount),
    Kurtosis = kurtosis(Amount),
    Skewness = skewness(Amount),
    Minimum = min(Amount),
    Maximum = max(Amount),
    Sum = sum(Amount)
  ) |>
  t() |>
  as.data.frame()

# set column names:
colnames(desc_stats) <- desc_stats[1, 1:2] |>
  gsub(pattern = "_", replacement = " ") |>
  stringr::str_to_title()

# remove first row:
desc_stats <- desc_stats[-1, ]

# Make rownames the `Stat` column:
desc_stats["Stat"] <- rownames(desc_stats)

# remove rownames:
rownames(desc_stats) <- NULL

# make stat the first column:
desc_stats <- desc_stats |>
  dplyr::relocate(Stat)

# ----hist----
mc_hist <- industry_long |>
  dplyr::filter(`Class Name` == "motor_commercial") |>
  ggplot(
```

```
    aes(x = Amount)
  ) +
  geom_histogram(
    mapping = aes(y = ..density..),
    color = "green",
    fill = "lightgreen"
  ) +
  ylab(label = "Density") +
  xlab("Claim Size") +
  ggtitle(label = "Motor Commercial") +
  geom_density(
    color = "firebrick",
    lwd = 1
  ) +
  theme_classic()

mp_hist <- industry_long |>
  dplyr::filter(`Class Name` == "motor_private") |>
  ggplot(
    aes(x = Amount)
  ) +
  geom_histogram(
    mapping = aes(y = ..density..),
    color = "blue",
    fill = "lightblue"
  ) +
  ylab(label = "Density") +
  xlab("Claim Size") +
  ggtitle(label = "Motor Private") +
  geom_density(
    color = "firebrick",
    lwd = 1
  ) +
  theme_classic()

# ----qqplots----
```

```r
# From the descriptive stats and now from the histograms,
# we can affirm that the data is positively skewed.

# This implies the need to use continuous distributions that
# are +vely skewed to fit the data

mc_qqplot <- industry_long |>
  dplyr::filter(`Class Name` == "motor_commercial") |>
  ggqqplot(
    x = "Amount",
    title = "Motor Commercial",
    size = 0.8
  )

mp_qqplot <- industry_long |>
  dplyr::filter(`Class Name` == "motor_private") |>
  ggqqplot(
    x = "Amount",
    title = "Motor Private",
    size = 0.8
  )

# We used the cube root function to transform the data and
# make the claim sizes become closer to normally distributed

# transform data:
industry_long_trans <- industry_long |>
  dplyr::mutate(
    Amount = Amount ^ (1 / 3)
  )



# qqplots of transformed data:
mc_trans_qqplot <- industry_long_trans |>
  dplyr::filter(`Class Name` == "motor_commercial") |>
  ggqqplot(
```

```
    x = "Amount",
    title = "Motor Commercial",
    size = 0.8
  )

mp_trans_qqplot <- industry_long_trans |>
  dplyr::filter(`Class Name` == "motor_private") |>
  ggqqplot(
    x = "Amount",
    title = "Motor Private",
    size = 0.8
  )

# ----fit distrs----
# Extract positive values for fitting models, x > 0, and remove
# missing values:
industry_long_trans <- industry_long_trans |>
  dplyr::filter(Amount > 0, !is.na(Amount))

# positive data:
positive_data <- industry_long_trans |>
  dplyr::filter(Amount > 0, !is.na(Amount)) |>
  dplyr::group_by(`Class Name`) |>
  dplyr::mutate(indices = 1:n()) |>
  dplyr::ungroup() |>
  tidyr::pivot_wider(
    id_cols = indices,
    names_from = `Class Name`,
    values_from = Amount
  ) |>
  dplyr::select(-indices)

# |- exp----
exp_model <- purrr::map(
  .x = positive_data,
  .f = ~ fitdist(
```

```
    data = na.omit(.x) |> as.vector(),
    distr = "exp"
  )
)

exp_gof <- purrr::map(.x = exp_model, .f = gofstat)
# extract K-S, A-D, AIC, BIC of the model

# exp model data.frame:
exp_model_df <- exp_model |>
  purrr::imap(
    .f = ~ tibble::tibble(
      Distribution = "Exponential",
      Parameter = c("Rate", "Std. Error", "LLF")
    ) |>
      dplyr::mutate(
        "{.y}" := c(.x$estimate, .x$sd, .x$loglik)
      )
  ) |>
  Reduce(
    f = function(...) {
      dplyr::full_join(..., by = c("Distribution", "Parameter"))
    }
  )


#' Goodness-Of-Fit data.frame
#'
#' @param distr_gof object of class "gofstat.fitdist"
#' @param distr_name name of the distribution
#'
#' @return a data.frame obj
#' @export
#'
gof_df <- function(distr_gof, distr_name) {
  distr_gof |>
```

```
    purrr::imap(
      .f = ~ tibble::tibble(
        `Test Statistic` = c("K-S", "A-D"),
        Distribution = distr_name
      ) |>
        dplyr::mutate(
          "{.y}" := c(.x$ks, .x$ad)
        )
    ) |>
    Reduce(
      f = function(...) {
        dplyr::full_join(
          ..., by = c("Test Statistic", "Distribution")
        )
      }
    )
}

#' Information Criterion
#'
#' AIC and BIC values.
#'
#' @param distr_gof object of class "gofstat.fitdist"
#' @param distr_name name of the distribution
#'
#' @return a data.frame obj
#' @export
#'
aic_bic_df <- function(distr_gof, distr_name) {
  distr_gof |>
    purrr::imap(
      .f = ~ tibble::tibble(
        `Information Criterion` = c("AIC", "BIC"),
        Distribution = distr_name
      ) |>
        dplyr::mutate(
```

```
          "{.y}" := c(.x$aic, .x$bic)
        )
    ) |>
    Reduce(
      f = function(...) {
        dplyr::full_join(
          ..., by = c("Information Criterion", "Distribution")
        )
      }
    )
}


exp_gof_df <- gof_df(
  distr_gof = exp_gof, distr_name = "Exponential"
)

# |- gamma----
gamma_model <- purrr::map(
  .x = positive_data,
  .f = ~ fitdist(
    data = na.omit(.x) |> as.vector(),
    distr = "gamma"
  )
)

gamma_gof <- purrr::map(
  .x = gamma_model,
  .f = gofstat
)

gamma_model_df <- gamma_model |>
  purrr::imap(
    .f = ~ tibble::tibble(
      Distribution = "Gamma",
      Parameter = c(
```

```
        "Shape", "Shape Std. Error", "Rate",
        "Rate Std. Error", "LLF"
      )
    ) |>
      dplyr::mutate(
        "{.y}" := c(
          .x$estimate[["shape"]], .x$sd[["shape"]],
          .x$estimate[["rate"]], .x$sd[["rate"]],
          .x$loglik
        )
      )
  ) |>
  Reduce(
    f = function(...) {
      dplyr::full_join(..., by = c("Distribution", "Parameter"))
    }
  )

gamma_gof_df <- gof_df(
  distr_gof = gamma_gof, distr_name = "Gamma"
)

# |- lognormal----
lnorm_model <- purrr::map(
  .x = positive_data,
  .f = ~ fitdist(
    data = na.omit(.x) |> as.vector(),
    distr = "lnorm"
  )
)

lnorm_gof <- purrr::map(
  .x = lnorm_model,
  .f = gofstat
)
```

```r
lnorm_model_df <- lnorm_model |>
  purrr::imap(
    .f = ~ tibble::tibble(
      Distribution = "Log Normal",
      Parameter = c(
        "Mean Log", "Mean Log Std. Error", "SD Log",
        "SD Log Std. Error", "LLF"
      )
    ) |>
      dplyr::mutate(
        "{.y}" := c(
          .x$estimate[["meanlog"]], .x$sd[["meanlog"]],
          .x$estimate[["sdlog"]], .x$sd[["sdlog"]],
          .x$loglik
        )
      )
  ) |>
  Reduce(
    f = function(...) {
      dplyr::full_join(..., by = c("Distribution", "Parameter"))
    }
  )

lnorm_gof_df <- gof_df(
  distr_gof = lnorm_gof, distr_name = "Log Normal"
)

# |- weibull----
weibull_model <- purrr::map(
  .x = positive_data,
  .f = ~ fitdist(
    data = na.omit(.x) |> as.vector(),
    distr = "weibull"
  )
)
```

```r
weibull_gof <- purrr::map(
  .x = weibull_model,
  .f = gofstat
)

weibull_model_df <- weibull_model |>
  purrr::imap(
    .f = ~ tibble::tibble(
      Distribution = "Weibull",
      Parameter = c(
        "Mean Log", "Mean Log Std. Error", "SD Log",
        "SD Log Std. Error", "LLF"
      )
    ) |>
      dplyr::mutate(
        "{.y}" := c(
          .x$estimate[["shape"]], .x$sd[["shape"]],
          .x$estimate[["scale"]], .x$sd[["scale"]],
          .x$loglik
        )
      )
  ) |>
  Reduce(
    f = function(...) {
      dplyr::full_join(..., by = c("Distribution", "Parameter"))
    }
  )

weibull_gof_df <- gof_df(
  distr_gof = weibull_gof, distr_name = "Weibull"
)

# |- pareto----

# !!! NOT working: error code 100 !!!
# scale_data <- function(x) {
```

```
#   (x - min(x) + 0.01) / (max(x) - min(x) + 0.02)
# }
#
# pareto_model <- purrr::map(
#   .x = positive_data,
#   .f = ~ fitdist(
#     data = na.omit(.x) |> as.vector() |> scale_data(),
#     distr = "pareto"
#   )
# )
#
# pareto_gof <- purrr::map(
#   .x = pareto_model,
#   .f = gofstat
# )

# ----distrs-df----
distrs_df <- dplyr::bind_rows(
  exp_model_df,
  gamma_model_df,
  lnorm_model_df,
  weibull_model_df
)

# ----distrs-gof-df----
distrs_gof_df <- dplyr::bind_rows(
  exp_gof_df,
  gamma_gof_df,
  lnorm_gof_df,
  weibull_gof_df
) |>
  dplyr::arrange(
    dplyr::desc(`Test Statistic`)
  )

# ----aic-bic-df----
```

```
info_df <- dplyr::bind_rows(
  aic_bic_df(distr_gof = exp_gof, distr_name = "Exponential"),
  aic_bic_df(distr_gof = gamma_gof, distr_name = "Gamma"),
  aic_bic_df(distr_gof = lnorm_gof, distr_name = "Log Normal"),
  aic_bic_df(distr_gof = weibull_gof, distr_name = "Weibull")
) |>
  dplyr::arrange(
    `Information Criterion`
  )
```