COMPARISON BETWEEN PARAMETRIC AND NON-PARAMETRIC METHODS OF ESTIMATING COMPREHENSIVE MOTOR CLAIM SEVERITY DISTRIBUTIONS

Ruth Wangari Kimani - 092833

Submitted in partial fulfillment of the requirements for the Degree of Bachelor's in business science in Actuarial Science at Strathmore University

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

Strathmore University

Nairobi, Kenya

November, 2019

DECLARATION

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

Ruth Wangari Kimani ..... [*Name of Candidate*]

..... [*Signature*]

28-11-2019 ..... [*Date*]

This Research Project has been submitted for examination with my approval as the Supervisor.

Dr Elphas Luchemo Okango: [*Name of Supervisor*]

[*Signature*]

28th November, 2019: [*Date*]

Strathmore Institute of Mathematical Sciences

Strathmore University

Table of Contents

DECLARATION ..... **i**

COMPARISON BETWEEN PARAMETRIC AND NON-PARAMETRIC METHODS OF ESTIMATING COMPREHENSIVE MOTOR CLAIM SEVERITY DISTRIBUTIONS. ..... **iv**

ABSTRACT iv

CHAPTER 1 ..... **1**

1.0 INTRODUCTION 1

1.1 BACKGROUND OF THE STUDY 1

1.2 Problem Statement 4

1.3 Objectives of the Study 5

1.3.1 General Objective 5

1.3.2 Specific Objectives 5

1.4 Research Questions 5

CHAPTER 2 ..... **6**

2.0 LITERATURE REVIEW 6

CHAPTER 3 ..... **15**

3.0 METHODOLOGY 15

PARAMETRIC METHODS 15

3.1 Introduction 15

3.2 Selection of Claim Distributions 15

3.2.1 Gamma Distribution 16

3.2.2 Log-Normal Distribution 16

3.3 Non-Parametric methods 17

3.3.1 Classical Kernel Density estimation 17

3.3.2 Data Transformation 18

3.3.3 Criteria for selection of the bandwidth and the transformation parameters 18

3.4 Checking Model Fit 19

CHAPTER 4 20

4.0 DATA ANALYSIS AND RESULTS 20

4.1 Descriptive Statistics and Parametric Fitting 20

4.1.1 Log-normal Distribution fitting 22

4.1.2 Gamma Distribution Fitting 23

4.1 Non-Parametric Fitting 25

4.2 Goodness of Fit Tests 26
CHAPTER 5 27
5.0 CONCLUSION 27
References 28

List of Figures

Figure 1 log-normal fitting to the original data 22
Figure 2 Log-normal fitting to the log-transformed data 22
Figure 3 Gamma fitting to the original data 23
Figure 4 Gamma Fitting for the log-transformed data 24
Figure 5 KDE Plot for the original data 25
Figure 6 KDE Plot for the log-transformed data 26

List of Tables

Table 1. Descriptive Statistics of the Original Data 20
Table 2 Descriptive Statistics of Log-Transformed Data 20
Table 3 Log-likelihood Estimates 26

COMPARISON BETWEEN PARAMETRIC AND NON-PARAMETRIC METHODS OF ESTIMATING COMPREHENSIVE MOTOR CLAIM SEVERITY DISTRIBUTIONS.

ABSTRACT

Claim severity is the amount of loss associated with an insurance claim. Insurers compensate policyholders who have suffered a loss from the occurrence of an insured risk. Insurance companies have been estimating claim severity by using normal distribution meaning; they assume an average cost of motor claims to estimate the total claims amount. However, this method is not very efficient because not all motor claims follow a normal distribution. To deal with this, there has been an introduction to using other parametric distributions such as the gamma and log-normal distribution. Parametric distributions do not consider the outlier claims that do not follow any of the parametric distributions and this is what led to using non-parametric distributions. The data used in this research study consisted of an auto-insurance portfolio of a company operating in Sweden, which was compiled by the Swedish Committee on the Analysis of Risk Premium in Motor Insurance, (Hallin & Ingenbleek, 1983). The motor insurance data is cross sectional and it involves the third-party liability auto-insurance claims for the year 1977. The only variable I worked with were the claim amounts. The main aim of this research study was to employ both parametric and non-parametric models in estimating the claim size distribution. From the data analysis that was carried out, it can be concluded that the non-parametric method is the most suitable one for estimating motor claims severity distributions.

CHAPTER 1

1.0 INTRODUCTION

1.1 BACKGROUND OF THE STUDY

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). Insurance is a process that entails the provision of a method of offsetting the risk of a likely future loss with a payment of a premium. The underlying concept 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. 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.

In the non-life insurance industry, there is an increased interest in the auto mobile industry because it entails the management of large numbers of risk events. These comprises of instances of theft and damage to vehicles due to accidents or other causes as well as the extent of damage to the parties involved. According to (Kingman, 2018), 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. Most non-life 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.

One of the main duties of an actuary has been to apply statistical techniques in the analysis and interpretation of data. In this research paper I intend to demonstrate the use of parametric and non-parametric modern statistical techniques in solving actuarial problems. In particular, I focused on a portfolio of motor insurance policies and, in analyzing the historical data drawn from this portfolio, we look at some of the classical challenges encountered by actuaries when dealing with insurance data.

Comparatively large insurance claim amounts, which may be unusual, necessitates the need to discover and employ specified statistical distributions with relatively fat tails and are highly skewed such as: the, gamma, pareto, weibull and log-normal. (Boland, 2006) These models are very useful and efficient to insurance companies in providing crucial information which enables them to make vital decisions such as: premium loading, future expected profits, adequate reserves which will assure the companies' profitability and the effect of reinsurance and expenses.

In analyzing the data, we focus on two concerns of the actuary. First, it is a consensus, at least in motor insurance, of the importance of identifying important 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, (W.Frees & Valdez, 2012). Some of these factors include; driver (age and/or gender) and vehicle (make, cubic capacity, kilometers driven) features.

The second concern is probably one of the most important aspects of the actuary's job: 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. In predicting claims distribution, at least for motor insurance, we often associate the cost of claims with two components: the event of an accident and the amount of claim, if an accident occurs. Actuaries express these two

components as the claim frequency and claim severity respectively. However, in this research paper we look at prediction models for claim severity only.

Modeling claim size of an insurance company is an important part of insurance price estimation and forecasting of future claims. The capability of forecasting future claims experience validates the insurance company to make suitable important arrangements to decrease the possibilities of making a loss. Such arrangements comprise of establishing a suitable premium for the policies and keeping money aside necessary for settling future reserves. An appropriate premium for an individual policy should be adequate cater for the expense cost of paying a claim. Sufficient reserves that are held should entitle the insurance company to remain solvent, such that it can be sufficient enough to pay claims when they arise. The number of claims in a discrete portfolio makes discrete standard distributions appropriate since their probabilities are explained on non-negative numbers. Furthermore, claims amount is supported on the positive integer line and seems to be positively skewed. A model on claim size was based on a non-negative continuous random variable according to their research. Many actuarial models for claims amounts are established on continuous distributions. The log-normal and gamma distributions fall mostly among the commonly used distributions for modeling claim amounts (Bahnemann, 2015). Various distributions consisting of claim size are the exponential, Weibull, and Pareto distributions. A model of claim amounts from First Assurance Company limited, Kenya for motor comprehensive policy consisted of the log-normal distribution which was chosen as the most suitable model that would provide a good fit to the motor insurance claims size data. The Pareto distribution has proved to belong to the tail-behavior of claim amounts and hence gives a good fit for claims amount data.

We will also look at non-parametric methods of estimating motor claims severity distributions. Non-parametric methods make no assumption on the population distribution or sample size. This is quite the opposite with most 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. Generally, conclusions drawn from non-parametric methods are not as powerful as the parametric ones. However, since non-parametric methods make fewer assumptions, they tend to be more flexible, more robust and applicable to non-quantitative data. (Hesse, J.B, & E.N., 2017)

It is popularly known that large claims are highly unpredictable and they result to financial instability and so, since solvency is a main concern for both insurance managers and insurance regulators, there is a need to estimate the density of claim cost amounts and to include the extremes in all the analyses.

In particular, I looked at the univariate kernel density estimation (KDE). This is a non-parametric method for estimating the probability density function f_x of a random variable X, it is a fundamental data smoothing problem where inferences about the population are made based on a finite data sample, (Guidoum, 2015) presented that a kernel is a weighting function used in non-parametric estimation methods, (Silverman, 1986).

1.2 Problem Statement

For actuaries, estimating claims severity distribution is very important since a good understanding and interpretation of loss distribution helps in making decisions in the insurance industry regarding premium loading, expected profits, reserves necessary to ensure profitability and the need for reinsurance. Motor claims increase every other year and there is a need to have an appropriate method of estimating motor claims severity distributions to help in the financial management of the company.

Most insurance companies use parametric distributions in estimating their claim size distributions. This limits those claims which are not represented by parametric distributions such as Gamma or the Log-normal distributions.

This brings about the need to include non-parametric distributions when estimating motor claim severity distributions. Since they make no assumptions about the data, they

let the data speak for itself and hence they may provide a better representation of motor claim severity distributions than parametric distributions do.

A proper understanding of statistical distributions is important in modeling claims data. By finding the appropriate statistical distributions which fit the claims severity data, it becomes easier to estimate the expected future claims based on these models, which aids in proper financial management in terms of holding reserves and other important actuarial decisions.

1.3 Objectives of the Study

1.3.1 General Objective

The general objective of this study was to employ both parametric and non-parametric models in estimating the claim size distribution.

1.3.2 Specific Objectives

2. To estimate the distribution for claim severity data using non-parametric technique i.e. fitting the KDE.
3. To measure the goodness of fit of the chosen models to determine which of the tested models fits best the given insurance motor data.
4. To use the chosen models to estimate the given claim amounts distributions.

1.4 Research Questions

1. How will the statistical distributions fit into the claims data?
2. Which statistical estimation method fits best into the given claims data after performing the goodness of fit tests?

CHAPTER 2

2.0 LITERATURE REVIEW

There is a rich literature on modeling the joint frequency and severity distribution of automobile insurance claims. To differentiate this modeling from the classical risk theory application (Klugman, Panjer, & Willmot), we pay attention to cases where explanatory variables, such as policyholder characteristics are available. There has been substantial interest in statistical modeling of claims frequency, see (Denuit & Boucher, 2015) for a recent example. However, the literature on modeling claims severity especially in conjunction with claims frequency is less extensive. One possible explanation, noted by (Coutts, 1984), is that most of the variation in overall claims experience may be attributed to claim frequency. He also remarks that the first paper to analyze claim frequency and severity separately seems to be (Levy, Kahane, & Haim, 1975).

An overview of how statistical modeling of claims severity can be helpful in pricing automobile coverage from the study of (Brockman & Wright, 1992). They narrowed down on categorical pricing variables to form cells that could be used with traditional insurance underwriting techniques. (Renshaw & Arthur, 1994) shows how generalized linear models can be used to analyze both the frequency and severity portions based on individual policyholder level data.

Another researcher provided a more modern statistical approach, fitting not only cross-sectional data but also following policyholders over time. (Pinquet, 1997) was interested in two lines of business, claims at fault and not at fault with respect to a third party. For each line, he also hypothesized a frequency and severity component that was- allowed to be correlated with one another. In particular, the claims frequency distribution was assumed to be bivariate Poisson. Severities were modeled using log-normal and gamma distributions. Also, at the individual policyholder level, (Frango & D.Vrontos, 2001) examined a claim frequency and severity model, using negative binomial and Pareto distributions, respectively. They used their statistical model to develop experience rated (bonus-malus) premiums.

Another research was done which applied the actuarial modeling steps to fit models to 490 claim amounts that were obtained from 7 consecutive years. Analytic loss distributions were fitted using maximum likelihood estimation for each of the 7 years, (Brockman & Wright, 1992). They applied P-P plots and the Kolmogorov-Smirnov tests (K-S test) to judge the perfect quality of fitted data. They used various statistical distributions which included the inverse Pareto, Pareto, burr, Pearson VI, inverse burr log-normal and restricted benktander families.

Another analysis applied an actuarial modeling to fit a statistical distribution of 250 claims, (Meyers, 2005). The statistical distributions that were tested were log-normal, gamma and Weibull distributions. He used maximum likelihood estimation to fit claims amounts into the distributions. Thereafter, he calculated the likelihoods and then applied them in obtaining the posterior probabilities of each model.

(Renshaw & Arthur, 1994) also came up with a useful approach in actuarial modeling for claim amounts for non-life insurance based on quasi-likelihood and extended quasi likelihood. They also applied maximum likelihood estimates to fit the claim amounts into the model since the quasi-likelihood parameter estimates according to them contain similar asymptotic properties as the maximum likelihood estimates.

Another research paper done by (Guiahi, 2000), presented the issues and methodologies necessary for fitting alternative statistical distributions to samples of different insurance data. His presentation was mainly focused on a sample of data with Log-normal as the main distribution. In his research he applied the method of maximum likelihood to estimate model parameters and, also his criteria for comparing which probability distribution fits the insurance data set best was focused on the value of Akaike's Information Criteria, AIC.

The AIC criterion is defined by:

AIC= -2(maximum log-likelihood) + 2(no. of parameters estimated)

In AIC when various models are being estimated, the model with the smallest AIC value is the most prudent one.

Another researcher, (Fiete, 2005), also presented actuarial modeling on his COTOR solution by applying gamma, lognormal, Weibull and Pareto distributions. He used similar steps of actuarial modeling as the previously discussed researchers and also used maximum likelihood estimation to obtain parameter values but because of the nature of his data, he determined the goodness of fit using P-P plots because they enabled him to compare the goodness of fit across the entire range of possible outcomes in order not to depend on one number from the log-likelihood to describe the goodness of fit.

A previous research study done by (Ismaili, 2018) about modelling claim severity in personal line general insurance, evaluating the suitable statistical model for modeling some general insurance policies such as motor, property, armed robbery plan, theft using data from the leading line of general insurance company in Nigeria for claim size and also predicting the risk premiums for each of the various policies are also determined. He illustrated that some of those policies are modelled more effectively using their own dissimilar distributions. The finding after the analysis was carried out, generally indicates that Gamma distribution was chosen as the best loss model for property and commercial insurance policies. The log normal distribution seemed best for the theft and motor insurance policies, showing that the Weibull distribution fitted best the armed robbery policy plan.

A study which was by the Claims Standards Council published 2010-2011 regarded the rise in the counts and size of personal injury claims to the introduction of conditional fee arrangements and also a large number of subsequent forms. These changes explained the notable extension in the frequency of claims. This tasked claim actuaries with realizing

more effective and efficient models for estimating these claims so that adequate reserves can be provided in order to meet the expected future claims.

A research paper focusing on the asymptotic behavior of the compound claim distribution, illustrated that under some certain conditions, if the distribution of the frequency of claims is negative binomial, then the distribution of the aggregate claim has asymptotic as a gamma-type distribution in its tail, (Yulia, 2010).

(Shi, 2011), stated that under specific given conditions, a negative binomial frequency gives rise to an aggregate distribution which is appropriate gamma. The skewness of gamma distribution is most often double its coefficient of variation since the loss distribution is often skewed to the right. To avoid having skewness that doubles the coefficient of variation, adding a third parameter to the gamma will be helpful. The most appropriate approach to test the fit of theoretical statistical distributions to a loss claim data is to contrast the theoretical distributions with the actual statistical family of distributions to the specific data.

(Yulia, 2010), worked on a study pertaining the modelling of claim severity in actuarial insurance practice and concluded that distributions that explain claim amounts are generally positively skewed, and the regression models of Gamma and Log-normal have been used previously by to estimate claim amounts. However, the fitting of claim sizes via regression models assumes that claim types are independent. In their study, they made an independent assumption between claim types and, also investigated three types of Malaysian motor insurance claims; Third Party Body Injury, Third Party Property Damage and Own Damage. They applied the normal (Clayton & Frank, 2011), used copulas for modeling dependence between these three insurance claim types. Their results showed that the Akaike's Information Criterion and Bayesian Information Criterion reveal that the Clayton is the most suitable fit model for modeling copula defense between own damage (OD) and Third Party Body Injury (TPBI) claims, whereas the t-copula is the most suitable copula for modeling dependence between TPBI and

TPPD claims. The main benefit of using copula is that each marginal distribution can be specified independently based on the distribution of individual variable and then linked by the copula which considers the dependence between these variables.

In another work done by (Forum, 2010) attempting to model claim severity brought about intriguing major terms when fitting probability distributions for the severity of random events. The important examples include events with negative impacts such as the distribution of insurance loss claimed under insurance policies, the severity of damages caused by other factors, and events with positive impact such as order sizes for products characterized demand. In the research, the severity process determines parameters of s continuous statistical distribution that are used to model the severity of a continuous-valued event of interest. In the research study, it was explained that the severity of an event does not follow classical distributions such as the normal distribution that are frequently assumed by standard statistical methods. It gives a default set of probability distribution models for various distributions that are used to model severity data.

Hierarchical Insurance Claims Modeling describes three components of a modeling process that are fitted to a model estimating a third-party insurance claim to find which model fits best the claim data, (W.Frees & Valdez, 2012). The three component (hierarchical) models consist of; claim frequency, type of claim and claim size and they observe that the claim severity likelihood depends majorly on the combination of the types of claims observed in the given data.

When analyzing the trends in Loss Frequency and Severity (Ethan, 2009), Hanover Insurance evaluates historical data to analyze trends in loss frequency and severity of claims. The trends are as a result of external factors, such as legislative, environmental and economic forces. The trends were analyzed using two methods; correlating the trends from prior data to external factors and comparing the impact of events to trends in the

data. The analysis quantified the effect of each external force and isolated factors which were most significant to trends.

(Mcguire, 2007) worked on a paper about Individual Claim Modeling of Compulsory Third Party (CTP) motor insurance data where he applied generalized linear model to extend the finalized claim amount. (Taylor & Mcguire, 2004) incorporated claim severity to enable the model to deal appropriately with the changing mix of claims.

(W.Frees & Valdez, 2012) in their hierarchical insurance claim modelling, explained the statistical modeling of detailed, micro-level auto-mobile insurance records when considering data from a leading insurance company. Detailed micro-level insurance records, infers to experience one vehicle level, including the make of the vehicle and driver characteristics, insurance coverage and claims experience, for each year. The claims experience comprises of detailed information on the type of insurance claim, such as whether the claim is due to an injury to a third party, comprehensive or claims for damage to the insured, and the corresponding claim amount. They suggested a hierarchical model for three components, corresponding to the frequency, type and severity of claims. They applied loss models for determining claim amounts and frequency. The driver's age, gender, past record and no claims discount as well as vehicle age and make seem to be key variables used to estimate the occurrence of a claim.

The other model is a multinomial logit model which estimates the type of insurance claim, whether it is a third-party injury, third party property damage, insured own damage or some combination. The current year, vehicle age and vehicle model are important predictors for this component. Their third model explains the severity component. In this model, they used a generalized beta of the second kind long tailed distribution for claim amounts and, to also incorporate predictor variables. The current year, driver's experience, vehicle age and the driver's age seem to be key predictors for this component. These variables imply that there is a notable dependence among the

various claim types, however, they went ahead to use a t-copula to account for this Dependence.

All the model components give a justification for assessing the importance of a rating variable. When used together, the integrated model enables an actuary to estimate motor insurance mobile claims more efficiently and accurately than when applying traditional insurance methods. By using simulation, they indicate this by coming up with predictive distributions and establishing premiums under alternative reinsurance coverage. (Gordon, 2006) fitted a tweediest compound Poisson model to a general insurance claims data, and he observed that the dependence of the likelihood function on p is as for a linear exponential family, so that modeling similar to that of generalized linear models is possible. Through the study they realized that when modeling the cost of insurance claims, it is generally necessary to model the variation of the costs together with their mean. In order to model the variation using the framework of double generalized linear models, (Gordon, 2006) discovered that variation increases the precision of the estimated tariffs. He also stated that the use of double generalized linear models also enables us to deal with the instance where only the total cost of claims has been recorded and estimated. (Meyers, 2005) came up with a classical non- Bayesian confidence interval of parameter selected distributions in modeling claim severity, by using a likelihood ratio test; he revealed a classic illustration that portrays how to apply likelihood function and Bayes' theorem to predict the high claim amounts of motor insurance.

(Chavez-Demoulin & Davison, 2005), have discussed some of the smooth extreme models in insurance. They pointed out on highlighting nonparametric trends, as a time dependence is present in many catastrophic risk events (for example storms, floods or natural disasters) and in the financial markets. A recent study done by (Cooray & Ananda, 2005) combines the lognormal and the Pareto distribution and obtains a distribution with an appropriate shape for small claims and can handle data which contains heavy tails.

(Wand & Jones, 1995) initiated the proposal of using the transformed kernel density estimation for asymmetrical variables and based on the shifted power transformation family. The original method is a good approximation for heavy-tailed distributions. The statistical features of the density estimators are also equally important when estimating the cumulative density function (CDF). (Bolance, Guillen, & Nielsen, 2003) the transformed kernel estimation seems to be an efficient approach in estimating quartiles near 1 and hence it can be used to estimate Value-at-Risk (VaR) in financial and insurance related industries.

(Buch-Larsen, Guillen, Nielsen, & Bolance, 2005) suggested an alternative transformation based on a generalization of the Champernowne distribution; simulation studies have shown that it is preferable to other transformation density estimation approaches for distributions that are Pareto-like in the tail. The existing distributions still experience a problem with the choice of the bandwidth parameter in transformation kernel density estimation. One way of solving this issue is to establish the transformation approach so that it leads to a beta distribution, then apply an existing theory to optimize the bandwidth parameter selection on beta distributed data and transform it back to it's original state. The main disadvantage with this method is that the beta distribution may be very steep in the domain boundary, resulting to numerical instability when the derivative of the inverse distribution function is needed for the backward transformation.

The work by (Chefd'hotel, 2003) specifically focuses on the contribution of non-parametric density estimation (such as kernel density estimation) to the group of linear shape deformations. The authors want to discover a better method for kernel density estimation than using standard kernels. They are focusing on this in order to understand how kernel estimators can be used in vision problems, which pertain to shape and texture transformations of an object's appearance.

A parametric model assumes that the density is known up to a finite number of parameters, while a nonparametric model allows great flexibility in the possible form, usually assuming that it belongs to some infinite collection of curves (differentiable

with square integrable second derivatives for example). The most used approach is kernel smoothing, which dates back to (Rosenblatt, 1956) and (Parzen, 1962).

A large extent of econometric research concerning estimate on of densities has shown that a well estimated density can be extremely useful for applied purposes. An interesting comprehensive review of kernel smoothing, and its applications can be found in (Bierens, 1987). (Silverman, 1986) and (Scott, 1981) discuss kernel density estimation thoroughly, giving details about assumptions on the kernel weight, properties of the estimator such as bias and variance, and discusses how to choose the smoothness of the estimate.

CHAPTER 3

3.0 METHODOLOGY

PARAMETRIC METHODS

3.1 Introduction

This section provides the methodology which was used in the study. This research paper used an auto-insurance portfolio of a company operating in Sweden, which was compiled by the Swedish Committee on the Analysis of Risk Premium in Motor Insurance, (Hallin & Ingenbleek, 1983). The motor insurance data is cross sectional and it involves the third-party liability auto-insurance claims for the year 1977, for which the insurance is covering the losses up-to the limits of the insured amount. It contains various variables such as: kilometers, zone, recent driver claim experience, vehicle make, frequency of claims and insured claims payment. The variable used during this study is the claims payment as the study is dealing with a univariate case. The statistical modeling of claims data involves the fitting of standard probability distributions to the observed claims data.

3.2 Selection of Claim Distributions

The initial selection of the models is based on prior knowledge on the nature and form of the claim's data. Claim severity is best modeled using non-zero continuous distributions which are positively skewed with heavy tails. This is due to the fact, that extremely large claim values often occur in the upper right tails of the distribution. Prior knowledge of claims experiences in non-life insurance together with descriptive analysis of major features of the claims data and graphical techniques are applied in the selection of the initial approximate probability distributions of claim amount. (Merz, Mario, & Wuthrich, 2008), noted that majority of claims data arising from the general insurance industry are positively skewed with heavy tails. They argued that statistical distributions which exhibit these characteristics may be suitable for modeling such claims.

3.4 Checking Model Fit

A goodness-of-fit (GoF) test is a statistical process that evaluates how well a distribution fits a set of given data. The test that I used to determine which method is the most suitable involved obtaining the log-likelihood function of the gamma, log-normal and the kernel density estimate.

The MLE usually results in better estimates when it is compared to other methods such as; the least-squares estimation (LSE), the method of quantile and method of moments (MME), especially when the given sample size is large. It was discovered that the MLE method completely utilizes all the information about the parameters that are provided in the data and results in a highly flexible estimator which has better asymptotic properties such as; efficiency, unbiasedness, asymptotic consistency, invariance and normality, (Denuit & Boucher, 2015).

Let Xi be the i^{th} claim amount, where n = $i_1; i_2; i_3, \ldots, n$

n represents the number of claims in the given data.

L represents the likelihood function of the distribution

θ represents the maximum likelihood estimator.

f(x) represents the probability distribution function of a specific distribution.

The likelihood function of the claims data is L(θ) which is provided as:

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

The maximum likelihood, determine on the equation above is given by:

$$L(\theta) = \log L(\theta) = \sum_{i=1}^{n} logf(x_i|\theta)$$

L(θ) can be differentiated with respect to θ, the MLE, express as θ^ and then equate it to zero.

$$\text{MLE} = \frac{dL}{d\theta}(\theta^{\wedge}) = 0$$

The estimation method with the highest log-likelihood estimate is deemed to be the most appropriate method of estimating univariate claim distributions

CHAPTER 4

4.0 DATA ANALYSIS AND RESULTS

4.1 Descriptive Statistics and Parametric Fitting

As mentioned earlier, the data belongs to use an auto-insurance portfolio of a company operating in Sweden. The motor insurance data is cross sectional and it involves the third-party liability auto-insurance claims for the year 1977, for which the insurance is covering the losses up-to the limits of the insured amount. The data consists of 2,182 settled motor insurance claim payments.

Mean	Std.Dev	Variance	Min	Max	Skewness	Kurtosis
257007.6	1017283	1.034864e+12	0	18245026	8.294279	92.87296

Table 1. Descriptive Statistics of the Original Data

The above statistics represent the original claims amount data. The average amount of claim payments is 257,007.6, with an average distance of 1,017,283 between each amoun t and the mean. The variance, which is 1.034864e+12, gives the square of the average distance between each amount and the mean. The minimum amount, of settled claims is 0 while the maximum amount of settled claims amount is 18,245,026. The skewness, whi ch is the measure of the asymmetry of the probability distribution of a real-valued rand om variable about its mean is 8.294279. This data is highly skewed since the skewness is much greater than 1. Finally, the kurtosis of the original data is 92.87296. This value sho ws the degree of peaked-ness of the data. In our case, this degree of peaked-ness is ver y high indicating that the distribution of the data has heavy tails.

Mean	Std.Dev	Variance	Min	Max	Skewness	Kurtosis
10.67858	1.976245	3.905543	0	16.7194	0.1639466	2.830637

Table 2 Descriptive Statistics of Log-Transformed Data

The above statistics represent the log-transformed claims amount data. The average amount of claim payments is 10.67858, with an average distance of 1.976245 between each amount and the mean. The variance, which is 3.905543, gives the square of the average distance between each amount and the mean. The minimum amount, of settled claims is 0 while the maximum amount of settled claims amount is 16.7194.

The skewness, which is the measure of the asymmetry of the probability distribution of a real-valued random variable about its mean is 0.1639466. This data is moderately ske wed since the skewness is between 0 and 1. Finally, the kurtosis of the original data is 2.830637. This value shows the degree of peaked-ness of the data. In our case, this degree of peaked-ness is moderately high, since its greater than 0, indicating that the dis tribution of the data has heavy tails.

We provide univariate histograms for the individual claim amounts below. Our variable is payments under the *severity_only* data.

Sv represents the original data of payments that are greater than 0 and are in matrix form.

Svlog represents the log-transformed data of payments that are greater than 0 and are in matrix form.

4.1.1 Log-normal Distribution fitting

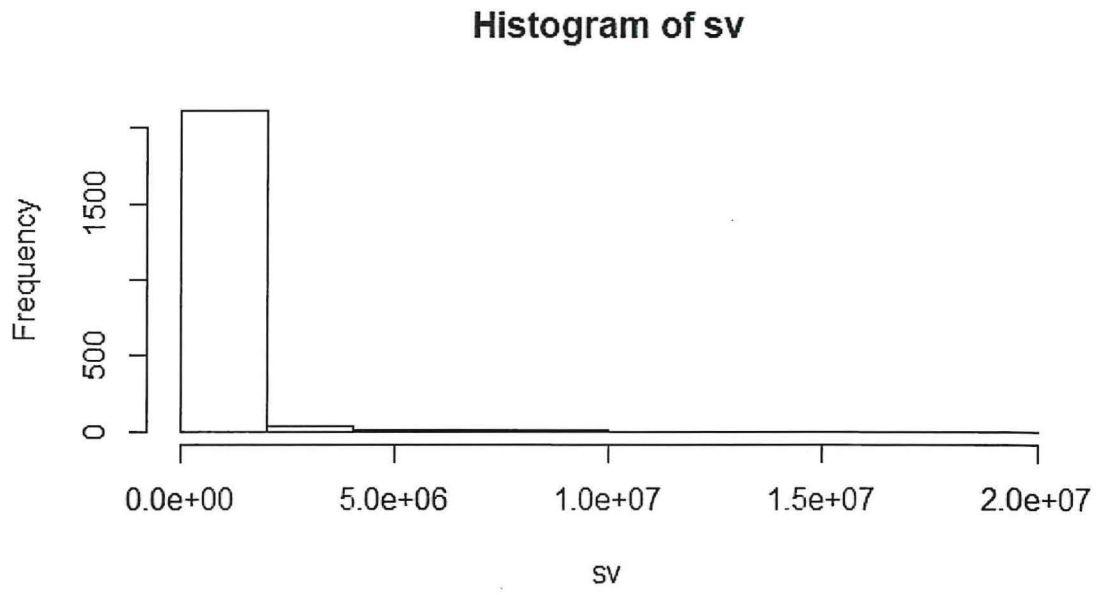

Figure 1 log-normal fitting to the original data

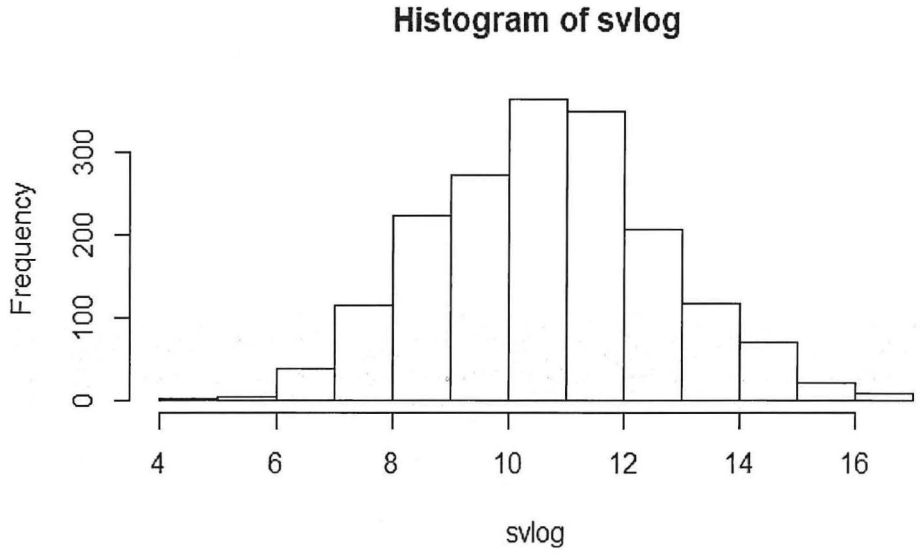

Figure 2 Log-normal fitting to the log-transformed data

From the above log-normal histogram plots, it is clear that the original claims amount data is highly right-skewed with majority of the claim amounts having little severity while only a few being of very high severity. The transformed data is showing a more symmetrical distribution which has little skewness.

4.1.2 Gamma Distribution Fitting

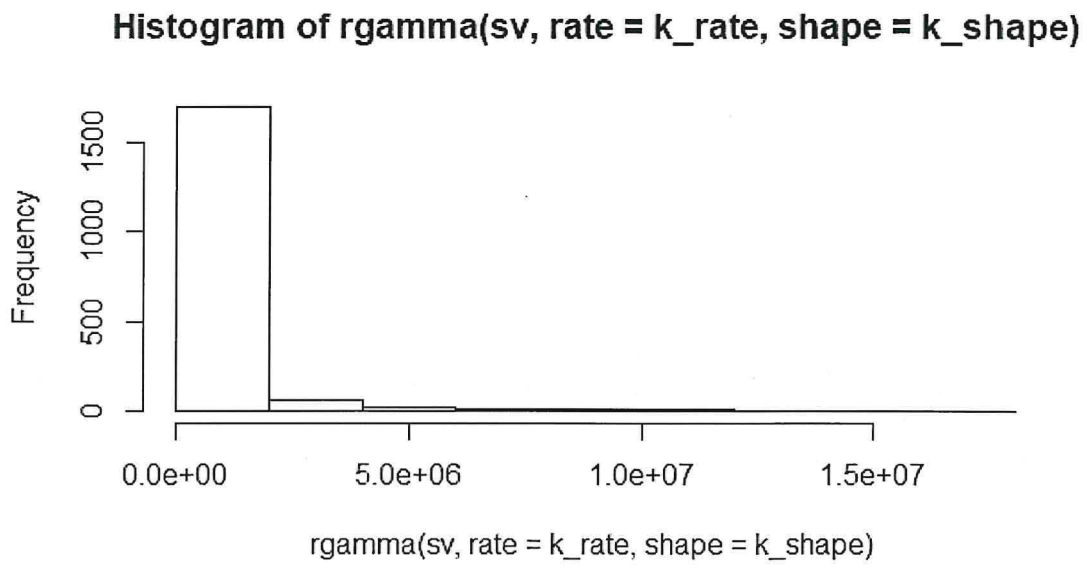

Figure 3 Gamma fitting to the original data

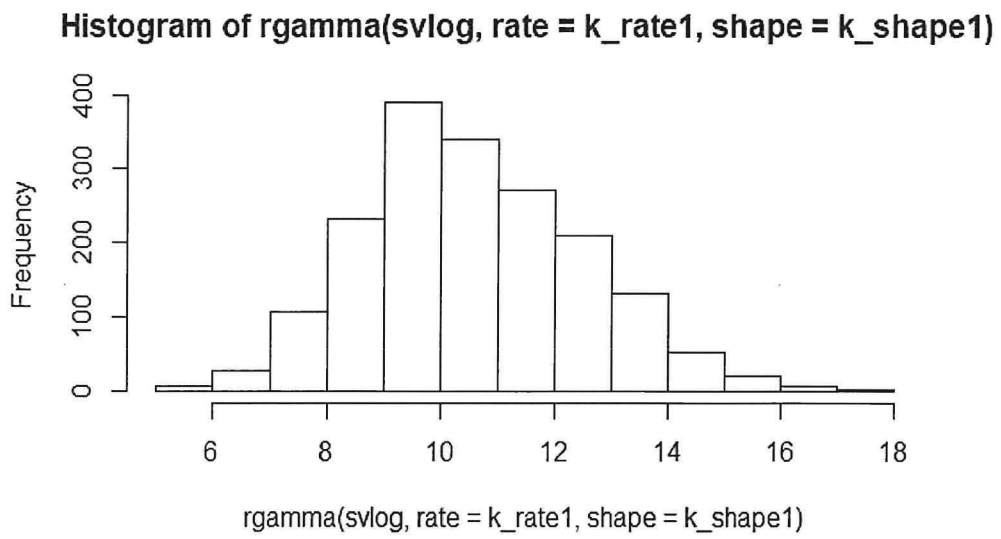

Figure 4 Gamma Fitting for the log-transformed data

Just like the log-normal fitting, from the above gamma histogram plots, it is clear that the original claims amount data is highly right-skewed with majority of the claim amounts having little severity while only a few being of very high severity. The transformed data is showing a more symmetrical distribution which has little skewness.

Both the lognormal and the gamma distributions provide a good fit to the claims amount data. We now compare how the non-parametric estimation method, in specific the Kernel Density Estimation (KDE) Method, fits into the claims amount data, then we compare which of the two estimation methods is better.

4.1 Non-Parametric Fitting

In this section, we show the results of the kernel density estimation, including both the results of the original univariate claims payment data and the log-transformed univariate claims payment data.

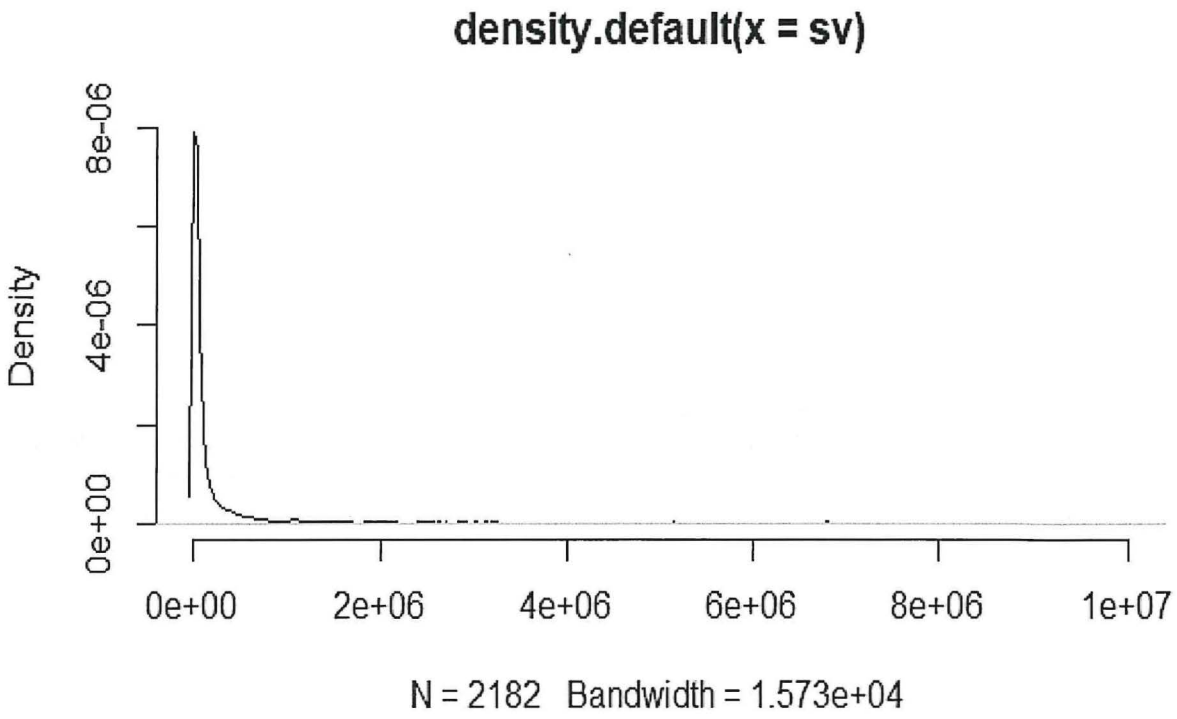

Figure 5 KDE Plot for the original data

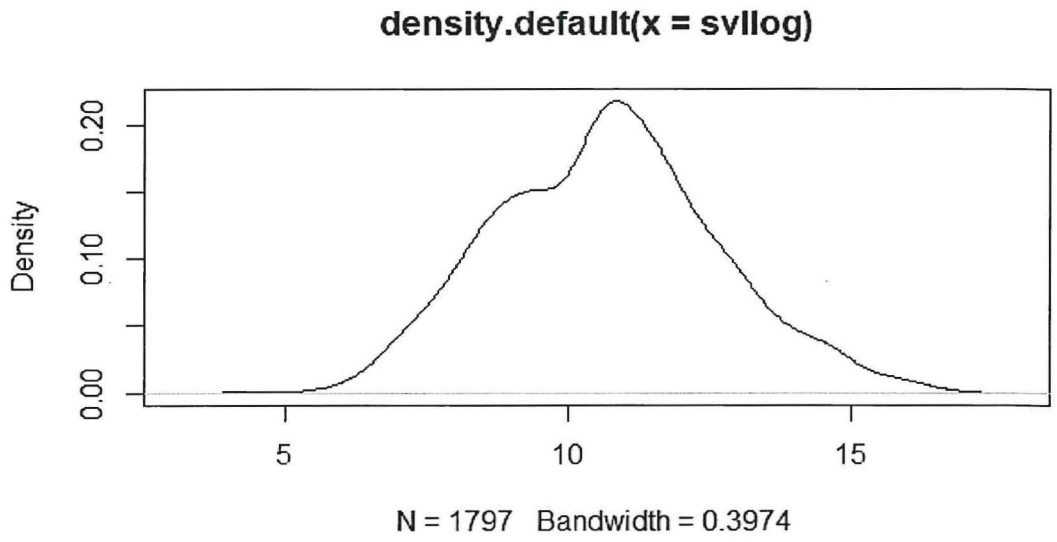

Figure 6 KDE Plot for the log-transformed data

From the above classical KDE plots, it can be observed that the KDE plot for the original univariate claims data does not have a smooth shape. However, we can see that there is a significant improvement in the kernel density estimate when it is applied to the log-transformed claim data compared to the fit obtained when applied to the untransformed original data.

4.2 Goodness of Fit Tests

For the goodness of fit test, we obtain the log-likelihood estimation of the gamma, log-normal and kernel density estimation methods as shown below.

Estimations	Log-Likelihood
Log-Normal	186,362.1
Gamma	142,820.6
Kernel density	196,990.5

Table 3 Log-likelihood Estimates

CHAPTER 5

5.0 CONCLUSION

From the above results, we can deduce that among the parametric and non-parametric estimation methods, the non-parametric method in specific the kernel density method is the most suitable method for estimating claim amounts since it has the largest log-likelihood estimate.

Non-parametric estimation methods are preferred to parametric estimation methods mainly because they don't make assumptions about the data distribution, since they let the data to speak for itself. Another reason is that non-parametric methods explain outliers found in observations which may not be represented by parametric estimation methods. Other advantages include; they can be used with smaller sample sizes, they have, in many cases, a high level of asymptotic relative efficiency compared to the classical parametric tests, they can be more easily understood intuitively, they can be used with more types of data and they can be applied to a large number of situations.

References

Achieng, M. O. (2012). *Actuarial Modeling for Insurance Claim Severity in Motor Comprehensive Policy Using Industrial Statistical Distributions.* Nairobi.

Bahnemann, D. (2015). *Distributions for Actuaries.* Fairfax: Casualty Actuarial Society.

Bierens, H. J. (1987). Kernel estimators of regression functions. *Advances in Econometrics*.

Bolance, C., Guillen, M., & Nielsen, J. (2003). Kernel density estimation of actuarialloss functions. *Journal of Insurance and Mathematics*, 19–36.

Boland, P. J. (2006). Statistical methods in general insurance.

Boleat. (2010).

Brockman, M., & Wright, T. (1992). Statistical Motor Rating: making effective use of your data. *Journal of Institute of Actuaries*.

Buch-Larsen, T., Guillen, M., Nielsen, J., & Bolance, C. (2005). Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. *Journal of Statistics*, 503–518.

Chavez-Demoulin, V., & Davison, P. (2005). *Generalized additive modelling of sample extremes.*

Chefd'hotel, E. G. (2003). Practical non-parametric density estimation on a transformation group for vision. In *Proceedings of the 2003 IEEE Computing Society Conference on Computer Vision and Pattern Recognition (CVPR'03* (pp. pp. 114–121).

Clayton, & Frank. (2011). Modeling dependence structure with Archimedan copuls and applications to the iTraxx CDS index. *Journal of Computational and Applied Mathematics*.

Cooray, K., & Ananda, M. (2005). Modeling actuarial data with a composite lognormal-Pareto. *Actuarial Journal*, 321–334.

Coutts, S. (1984). Motor Insurance Rating, an Actuarial Approach. *Journal of the Institute of Actuaries*.

Denuit, & Boucher. (2015). Modeling the Frequency of Auto Insurance Claims by Means of Poisson and Negative Binomial Models.

Ethan. (2009). *Loss Frequency and Severity* .

Fiete, S. (2005). *Benfield Analystics.*

Forum, S. G. (2010).

Frango, N., & D.Vrontos, S. (2001, May 22). Design of Optimal Bonus-Malus Systems with a Frequency and a Severity Component on an Individual Basis in Automobile Insurance. *Astin Bulletin*.

Gordon, S. K. (2006). Fitting Tweedies Compund Poisson Model to Insurance Claims Data:Dispersion Modeling.

Guiahi, F. (2000). *Fitting Loss Distributions with Emphasis on Rating Variables.*

Guidoum, A. C. (2015). Kernel Estimator and Bandwidth Selection for Density and its Derivatives.

Hallin, M., & Ingenbleek, J.-F. (1983). *Analysis of Risk Premium in Motor Insurance.*

Hesse, C., J.B, O., & E.N., N. (2017). *Introduction to Nonparametric Statistical Methods.* Accra: Akrong Publications Ltd.

Hsiao, R. (2002, January 05). A new perspective on the dynamics of IT-enabled strategic change.

Ismaili, A. (2018). *Creating a Scene for Property Claims Adjustment.*

Kingman, J. (2018). *Financial Reporting Council.* London: APS Group.

Klugman, Panjer, & Willmot. (2004). *Generalized Linear Models Beyond The Exponential Family With Loss Reserve Applications.* Society of Actuaries.

Levy, Kahane, Y., & Haim. (1975). Regulation in the Insurance Industry:Determination of Premiums in Automobile Insurance. *Journal of Risk and Insurance*.

Mcguire, G. (2007). Individual Claim Modeling of CTP Data. *Institute of Actuaries of Australia.*

Merz, M., Mario, V., & Wuthrich, V. (2008). *Modeling the Claims Development Result For Solvency Purposes.* Caualty Actuarial Society.

Meyers, G. (2005). *On Predictive Modeling for Claim Severity.*

Mihaela, D. (2015, 11). MODELING THE FREQUENCY OF AUTO INSURANCE CLAIMS BY MEANS OF POISSON AND NEGATIVE BINOMIAL MODELS. *Scientific Annals of the "Alexandru Ioan Cuza" University of Iaşi*, pp. 151-168.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 1065–1076.

Pinquet, J. (1997). Allowance for Cost of Claims in Bonus-Malus Systems. *Insurance Mathematics and Economics*.

Renshaw, E., & Arthur. (1994). Modeling the Claims Process In The Presence of Covariates.

Rosenblatt, M. (1956). Remarks on some nonparametrice stimates of a density function. *Annals of Mathematical Statistics*, 832–837.

Scott, D. W. (1981). Using computer-binned data for densityestimation. *Computer Science and Statistics*, 292–294.

Shi, P. (2011). Longitudinal Modeling of Insurance Claim Counts Using Jitters.

Silverman, S. J. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.

Taylor, G., & Mcguire, G. (2004, December). Individual Claim Loss Reserving.

Tomberlin, J., Weisberg, H., & Thomas. (1982). A Statistical Perspective on Actuarial Methods for Estimating Pure Premiums from Cross-Classified Data. *Journal of Risk and Insurance*.

W.Frees, E., & Valdez, E. A. (2012). Hierarchical Insurance Claims Modeling. *Journal of the American Statistical Association*.

Wand, M., & Jones, M. (1995). *Kernel Smoothing.* London: Chapman & Hall.

Yulia. (2010). *Asymptotic Behaviour of Compound Distributions.*