KWAME NKRUMAH UNIVERSITY OF SCIENCE AND
TECHNOLOGY, KUMASI

A SUITABLE CLAIM SEVERITY MODEL OF COMPREHENSIVE INSURANCE POLICY

BY

ANYANUMEH HERBERT TETTEH

(Bsc Actuarial Science)

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS,
KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN
PARTIAL FUFILLMENT OF THE REQUIREMENT FOR THE DEGREE
OF Msc. ACTUARIAL SCIENCE

April 2016

DECLARATION

I hereby declare that this submission is my own work towards the award of the M. Phil degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgment had been made in the text.

Anyanumeh Herbert Tetteh

Student(PG8730912) ....

Student Signature Date

Certified by:

Mr NANA KENA FREMPONG ....

Supervisor Signature Date

Certified by:

Prof. S. K. AMPONSAH	....	....
Head of Department	Signature	Date

DEDICATION

I dedicate this work to the Almighty God and my family especially my wife and my kids, Duncan and Emmanuella and Desmond for their prayers and support.

ABSTRACT

This study was base on a claim data collected from a recognized motor insurance companies in Ghana. The data from this non-life insurance company was to find the appropriate statistical distributions that best fit the claims data. In the insurance industry two major problems exist; how to get an appropriate model for a large claim data and how to fit and test the appropriateness of the model identified to make future decision. In this study, a clear methodology was set up to deal with the identified problems. Specific distributions were chosen from a group of family of exponential distribution. The study began with a brief introduction for the modeling process as practice by Actuaries. The purpose of the study was to get the best statistical distribution that best fit the claim amount from the insurance company data and fit the data using Quantile–Quantile plot (Q-Q plots) for all the tested distributions. The modeling process established one of the statistical distributions efficiently fitted the claim data. The study went ahead to test for the goodness of fit test using (AIC) that is Akaike's Information Criterion and a graphical presentation was done using the Q-Q plot. Finally the findings were discussed and conclusions, was made to the work and recommendations were given to guide for further study on similar or same topic.

ACKNOWLEDGMENTS

It take vision, enthusiasm passion and faith to accomplished a mission but support and encourage brings success. Many individuals have supported me to turn the wheel to success both in knowledge and other diverse ways. First of all am grateful to my maker the Almighty God for His care protection and wisdom. Secondly, the man who had turn things run, my noble hard working kind hearted supervisor Nana Kena Frempong for his time, piece of advice, encouragement and constructive criticism throughout the research period. I own it again to Dr A.Y Omari- Sasu, and Derrick Owusu Asamoah for their advice and encouragement. I would also like to express my profound gratitude to my entire course mates for their support and encouragement. Also thank Dafour Lambert and Francis Berbiye. Additionally I would like to thank Lawrence of NTHC, Accra, Vincent Anyomi, Adamkie Rebecca, my mum and all my sibling, Finally to all friends, Kingsly, Victor and all love ones.

CONTENTS

LIST OF ABBREVIATION

AIC ....Akaike's Information Criterion

COMP ....Comprehensive Insurance OD

....Own Damage TPBI

....Third Party Body Injury

TPPD .... Third Party Property Injury BIC ....

Bayesian Information Criterion

GLM .... Generalized Linear Model

COTOR .... Commision on Risk Theory

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

In Ghana, and many other developing countries, even though the general insurance industry is growing steadily, most of these industries either see it important to have proper data base for more actuarial work to be done for the industry and even when the data is available, they do not find it very crucial to study and analyzed this existing data beyond the simple descriptive statistics, (Ghana Insurance Commission annual survey 2013). This probably might explain why general insurance companies in Ghanaian economy are either under or not maximizing their profit . Claims, more importantly large ones, had been the major problem on the heels of these companies as past occurrences of repudiations of claims had problem for the industry. Again, company's in ability to settle claim payment as a result of the time, size and the frequency at which this claims come posses a treat. Severity of Claim analysis is crucial in determining actual and fair premiums. If this is not done, an insurance company might risk solvency. Manager of this claims data need both past and present data to model the behaviour of the claims and the probability of their occurrence but most of these companies sometimes do not understand how to incorporate this in the insurance business. Such records on the past claim will serve as spring board and source of information for their business to grow. (Adeleke, 2011)

The underwriter and the policy manager, for that reason there is needs to ensure that the premium charged to individual members of the policy holders proportionate the level of insurance claim been paid by the insurer or is equitable, compared with the premium to keep the business in good standing to others, having in mind the probability of frequency and severity of loss claims that may be made by that policy holder. In the absence of the actuarial work about the occurrence of claims, premiums are only determined arbitrarily, or using an adjusted parameter from other developed country which might be in variance to the local market such as the rates for other developed economies are adopted for developing economies. Unfortunately in many cases, the alterations is inaccurate and inappropriate as there might by different economic market conditions of the developed to the developing economies. This continue with the practice of inaccurate and inappropriate premium adjustment would do little to mitigate the poor image problems from which the insurance companies had been struggling with (Osokka, 1992).

All over the world and in particular Ghana, insurance companies faces the problem of claim payment especially when the claim payment involved comprehensive insurance. This piece of work would like to look at how claims involving large payments are handled using statistical distributions in some of the motor insurance companies in Ghana. The work consists of various chapters. This chapter consist of the background to the research, statement to the problem, objectives, significance of the study, organization of the study among others.

1.2 Background of Study

(Boland, 2006) Insurance industry is a data-driven industry, and insurance companies employ Actuarial analysts to understand the claims data. No one likes to lose, not even to break-even and an actuary in particular needs to model both the claim frequency and severity the claim and premiums, losses and claims. Techniques in exploratory data analysis such as histograms, percentile plots(P-P), quantile-quantile (Q-Q), the various summary statistics including mean, variance,sample estimates of skewness and kurtosis are very important tools in obtaining a feeling for the typical claim size. Relatively large claims, which may be infrequent, are of particular concern and hence the need to find and use distributions with relatively fat tails likes the normal distribution, Pareto distribution, Weibull, distribution and log normal distributions. Although the empirical distribution function can be a useful tool in understanding claims data, there is often a natural desire to “fit” probability distribution with reasonably tractable mathematical properties to claims data. Actuarial scientist are always interested to use both statistical and actuarial software that attempt to fit different types of distributions to a data set through different types of fitness test such as the Kolmogorov Smirnoff and Anderson Darling tests, Chi-square Goodness of Fit tests and the AIC (Akaike’s Information Criteria). Fitting those distributions most of the time statistical and actuarial softwares, such as the R or stata make things easier but quiet technical, but these softwares are able to find “the distribution curve which best describes your data sets. The expect in claim modeling will like to consider the impact of loss payments, deductibles, interest rates, reinsurance arrangements and inflation on the part of a claim that will be handled by the base insurance company. This practicably needs a good appreciation of conditional probabilities and distributions.

Loss severity modeling is very crucial component in general insurance more importantly the comprehensive insurance to avert the collapse of the industry. Losses will be observed over a time period, the policy provider can then be able to better the inherent risk and the probability loss characterized by policyholders and to adjust premiums and the amount of reserves where applicable. The data type considered is purely an individualized claim data with the relevance of being able to plan a model that will capture the temporal inter dependence in the pattern of claims amount. The study will investigate the importance of various types of probability models for the insurance claim data that will provide better overview in understanding all the possible problems associated with insurance claim amount observed over a time period.

Insurance is described as the protection of oneself or property against risk or uncertainty by purchasing an insurance policy or assured, willing to protect against unforeseen future risk. This strategy for protecting oneself against future uncertainty is the best management strategy since the future is on certain and can not be predicted. Car usage has embedded mortality and morbidity risk and the loss that may arise as a result of fire and theft in nations. In pursuance of that the general insurance products were developed to take care of the possibility of losses that might occur from the use of motor vehicles and other non-life properties. Motor insurance is a vital form of general insurance that involves the use of motor car or motorbike. This is a way of transferring the risk of uncertainty of or loses, from the owner of a vehicle to the insurer in exchange for a premium, and can be a form of a guaranteed for devastating future loss. Motor insurance, is increasingly becoming the commonest insurance industry all over the world.

Motor or automobile insurance is in operation to protect road users and other motorist from potential financial loss from operating a vehicle. Therefore, insurers and policy enforcers require motorist to buy an insurance cover to guide other third parties holder and the motor itself in the case of comprehensive insurance holders as well as reducing the impart of loss. Basically there are two form of vehicle insurance policy in Ghana; these are: The comprehensive motor insurance(COMP), and the third party motor insurance.

Comprehensive insurance (COMP) covers provides protection against physical damages or losses resulting from incidents apart from collision. This sometimes includes coverage for the Vandalized glass (such as a broken windshield) damage sustained from hitting an animal or bird, damage from falling objects or missile fire but for the purpose of this study all definition is limited to motor accident.

In Ghana and most part of the world, comprehensive motor insurance is not compulsory. But a car or property owner can guard against the risk of financial loss by purchasing insurance cover, comprehensive motor insurance is the best way because it covers compensation for car accidents and other kinds of loss in the cause of doing business. In case there is a damage to another party or injury to a third person or damage to a property, the protection policy will protect the policy holder in that kind of person by indemnifying the person. Moreover the cover also takes care of the

policy holder's own medical, vehicle and other expenses. Comprehensive motor insurance covers the policy holder's policy documents covers when the loss occurs, this might come in many form such as
vandalism, fire, theft or other perils.

One fundamental work of an actuarial Scientist consists of using statistical models to predict forecast, building and analyzing statistical models to serve as a framework for describing the processes by which an insurance company's cash flows. This task uses both statistical and mathematical and various quantitative skills to enable one "make financial meaning of the future". Operations of non-life (motor) insurance company should be as profitable than any other business, the challenges of moral hazards might be address challenge might be moral hazard. Insurance companies need to guide against future uncertainties and the risk of large claims payment.

At a fundamental level, insurance companies accept premiums in exchange for promises to indemnify a policyholder on the event of uncertain occurrence of the insured. This indemnification is known as a claim. A positive amount, also known as the severity, of the claim, is a key financial expenditure for an insurer. One can also think about a zero claim as equivalent to the insured event not occurring. Knowing only the claim amount summarizes the reimbursement to the policyholder. Ignoring expenses, an insurer that examines only amounts paid would be indifferent with regard to two claims of when compared to one claim of events though the number of claims will be distinct.

Most of the time, Many actuaries are not interested in the occurrence of the claims but rather in the consequences of the random amount, they are concern with the quantum of claim the insurance company would have to bare rather than the peril in particular or the circumstances which give rise to the claim amount to be paid or the number of claims involved. (Denuit, 2007) the actuary in the field of general insurance needs to have an understanding of the various actuarial models for the risk consisting of the total or aggregate amount of claims payable by an insurance company over a fixed time period.

Globally, general insurance is one of the fastest growing in the actuarial industry. This includes personal or property Insurance such as motor and home insurance, health insurance, as well as large liability insurance and many other commercial risks. For any economy, the insurance industry is very critical because it is a major source for which economies and government all over the world depends on for emergency funding over a longer time period. Insurance Industries must be research oriented institution with the main aim to find and protect the company being claim payments, Insurance companies involves large number expects and analysts with key analytical skill, risk mangers who understand and appreciate the claims data and how it could be applied to reduce the probability of loss (Boland, 2006).

Data of Insurance contains relatively large claim amounts, which may be infrequent, hence the need to find and use appropriate to model data with statistical distributions such the gamma, Pareto, Weibull, log-normal and Burr, exponential, log-normal acceptable methodology (Boland, 2006).

These statistical models provide more reliable and accurate results, they also provides information to the company and they enable companies make an informed, reliable and decisive decisions on amongst certain important insurance parameters including: expected profits, premium, reserves which are necessary to ensure(high probability) profitability and the impact of reinsurance and deductibles. Comprehensive motor insurance in developing country like Ghana is fast growing but one basic problem for motor insurance companies is public perception on claim payment.

Even though theoretical and empirical distribution functions are very important tools in understanding and interpretation of claims data for comprehensive motor insurance policy, there is always the zeal to “fit” statistical and probability distribution with much reasonable and reliable and accurate properties to the insurance claims data. This study applied all the steps requires in actuarial modeling to find a suitable probability distribution for the claims data and testing for the goodness of fit for the distributions. "Constructing models for insurance comprehensive claims severity can often give one a much added insight into the complexity of the large amounts of claims that may often be hidden in a huge amount of data. (Raz and Shaw, 2000)"

1.3 Problem Statement

The problem of most Motor insurance companies and especially for claim actuaries is how much more claim payment or loss amount is expected in the coming year. The issue of whether claim paid this year will be higher than last year or lower, and how best to find the probability of loss by using probability functions. The problem of most insurers is how to have a foresight of the likelihood of claim on how much more or less claim payment to be expected. In Ghana, motor insurance companies record an over-whelming number of claims in any given period, some of the claim sizes are large and are unbearable for these insurance companies because, either they don't have the financial muscles to pay such a claim or they were unexpected especially the claims that involve luxury motor and the amount involved are large.

Knowledge on how likely the claim amount occurs is important for insures' expectation into the future. The most thing for an insurance company is to get an appropriate distribution claim data, this will help you to estimate both premiums and claims. One major problem is how the insurer will calculate the expected number and amount of future claim.

1.4 Objectives of the Study

The main objectives of the study are;

- To fit loss models.
- To measure the goodness of fit among loss models
- To use the chosen model to estimate the expected claim amount per risk in the coming year.

1.5 Justification of study

There is a growing need for an appropriate model for estimating an insurance claim with the non- life (motor) insurance circle. The survival of any insurance company is to be able to estimate adequate reserve and make provision for future claims.

This study targeted to recognized the importance of understanding statistical and probability distributions for use by motor insurance companies and their under writers. General insurance companies need to have an understanding of the various statistical model for estimating individual and aggregate claims payable by an insurance company over a time period. This technical knowledge enables the company makes an informed financial decisions to ensure high profitability, expected profits and the impact of re-insurance.

The paper analyses the theoretical back-ground of the modeling process which takes place with insurance claims data. This is because insurance companies receive an overwhelming number of claims within a particular period, especially in motor policy; therefore an accurate modeling process to evaluate these huge amounts of claims would clearly be of assistance in managing the data and the amount expected to pay. Applying statistical distributions to model claim severity, The company added knowledge into the complexity of the large insurance claim amount of claim that may often be hidden in a huge amount of data. The company will able to project base on this evidences attain from this models and finally the company will apply this model on their future claim data.

1.6 Overview of Research Methodology

This research will use secondary sources of data from a general insurance company in Ghana. The researcher will use some statistical tools in analyzing the data. The population under study will be from one motor insurance company in General insurance providers in Ghana. The various statistical models that would be will use are Weibull distribution, normal distribution, lognormal distribution, and gamma distribution the various continuous statistical distributions are the main models that will be apply to the data.

The variables applies in the study were the loss amounts incurred by an automobile insurance company reported to National insurance commission. All the random variables in this study were independent. Various statistical tools will aid this study and histograms, tables and other harts will be use. Basic statistical packages will be use.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this section there is a review of the work of several authors regarding definition, concept claim severity modes and various studies done to discover suitable claim severity model of comprehensive insurance policy.

2.2 Key Terms in Insurance

2.2.1 Rate Making

The process of calculating a premium to charge a customer for an insurance policy is known as rate making. In this process, loss frequency and severity are analyses in order to predict how much money the company must make to break even, and then adjust the price to make a profit. Pure premium is also reflected in the premium price to account for commissions for insurance salespeople, company expenses, and other miscellaneous expenses. Premium figure that is created through this process reflects a group of policy buyers who share a similar expectation of loss. To create a different premium for each policyholder would be impractical. The data that is examined is normally recorded on a quarterly basis. (McClenahan, 2001).

2.2.2 Exposure

Exposure is the name for the basic rating unit that affects the premium. The unit varies based on the type of coverage that is being provided by the insurance company. For example, a car year is considered one automobile insured for a period of twelve months. A policy covering three cars for a six month term involves 1.5 car years. There are several exposure statistics examined: written exposure, which are the units of exposure from policies that were written during the period earned exposure, which are the exposure units that experienced loss during the period; in-force exposures, which are exposure units that experienced loss at a certain point in time. The units of exposure that this project uses in calculations are the earned exposures. (McClenahan, 2001).

2.2.3 Claims

The demand for payment by a policyholder or by an injured third party is considered a claim. The claims are organized by accident date-the date on which the accident occurred, leading to the claim-and by report date-the date on which the insurer is notified of the claim. The claims are recorded as “feature-paid” in the historical data for Hanover, which is used in calculations.(McClenahan, 2001)

Claim Frequency $= \frac{ClaimCount}{Exposure}$

2.2.4 Losses

Losses are the amounts paid or to be paid to the claimants under their insurance policy contracts. There are several divisions of losses that are recorded: paid losses, the losses of a period that have been paid to the claimant; case reserve, the amount that is expected to be paid for a claim in the future; accident yearcase incurred losses, the sum of paid losses and case reserve for a specific year; ultimate incurred losses, the accident year-case incurred losses plus the losses that have not yet been reported. For this project, paid losses are used in calculations with claims and exposure. (McClenahan, 2001) Loss Cost = Claim Frequency $*$

Claim Severity

2.2.5 Severity

Severity is the average loss per claim on a policy. It can be calculated using any recorded type of losses, such as paid loss, case reserve, etc. In this case, paid losses are used, and the formula for this calculation is: Claim Severity =$\frac{Loss}{ClaimCount}$

$$S = \frac{L}{C} \tag{2.1}$$

Where S is the severity, L equals paid losses, and C equals the amount of claims for the period. The severity is already calculated on a per-unit basis, so there is no scale factor that needs to be eliminated. (McClenahan, 2001)

2.2.6 Pure Premium

Pure premium is the amount of money needed to pay the amount of losses over the entire exposure. The formula for this quantity is:

$$P = \frac{L}{E} \tag{2.2}$$

Where P is the pure premium, L is the paid losses, and E is the number of exposure units. The pure premium can also be written as:

$$P = \frac{C}{E} * \frac{L}{C} \tag{2.3}$$

With C equalling claim count, which is the same as:

$$P = F * S \tag{2.4}$$

Therefore, when frequency is calculated on a per-unit basis, pure premium is the product of frequency and severity. Since the pure premium is based on both frequency and severity, and is more volatile than the other factors, it was not examined as closely as frequency and severity were in this project. (McClenahan, 2001) Motor insurance has gradually grown to be become the most prevalent insurance industry all over the world. The literature review present a comprehensive overview on actuarial modelling as applied in insurance claim severity using statistical distributions as viewed and present by other research studies. The section is base on published materials, reports journals, and opinions prior collected by other researchers. It is believed that by presenting this study and the scope is widened and thus revealing the gap in informations that has
previously presented.

All over the world, Motor insurance premiums both third party and comprehensive are steady rise, despite the long-term trend towards being fewer casualties from road accidents. Young drivers particularly must pay large insurance premiums because of their higher risk. The high cost of motor insurance can cause a number of policy problems for motor insurers when it's come to claim payment.

In a recent study published by (Ismail *et al* 2011) on modelling claim size in personal line non-life insurance, testing which statistical model best fitted for some non-life insurance policy such as motor, property, Armed robbery plan, Theft using data from the most prominent line of non-life insurance company in Nigeria for claim amount and also estimating the risk premiums for each class of policy are also estimated. His study shows that some of the line policies are indeed better modelled with different distribution than the other. Their finding in summary revels that Gamma distribution was chosen as the best loss model, property and commercial policy. Log normal was best for the theft and motor, it is also reveal that Weibull was best fit for the armed robbery policy plan, There was an omission or detailed information of the impact of the other used distribution on other policies.

In UK, it is compulsory for car owners to purchase an insurance package cars against third party, the risk of damaging another person's property or causing bodily injury to others or death. Most drivers purchase comprehensive insurance, to cover risks to themselves as well, and some buy extra cover so that can recouped excess payments when involved in accidents for which it were not at fault of theirs or pursue personal injury claims without risk of incurring any costs to themselves. Currently, there is a rush for comprehensive insurance for car owners to guide against future loss (House of Commons Transport Committee, 2010-11). One of their key recommendations was that key policy makers facilitates investigation of effective means of deploying and publicizing new technology which can assess how cars are driven by young drivers and sponsor research into international experience in restraining the number of personal injury claims relating to motor accident insurance.

In a recent publication by (Boleat, 2010) and the Claims Standards Council published 2010-2011 attributed the increase in the frequency and severity of personal injury claims to the introduction of conditional fee arrangements and a number of subsequent reforms. These changes would account for the significant and continuing growth in the number of claims. The Claims Standards Council said that the introduction of conditional fee arrangements had led to the introduction of claims management firms, which "promoted themselves to consumers offering advice to people wishing to make a claim and referring them on to a solicitor. This is evident that claim actuaries must find a better and efficient model for modeling these claims in order that efficient reserve can be made available to meet the future.

Again, (Peng Shi *et al*, 2011)" in their study to mode longitudinal insurance claim count using Jitters he touches on very vital aspects in the rate making process for property and casualty insurance. His study involves the characterization of copulas to model the number of insurance claims for an individual policyholder within a longitudinal context. In order to take care of the problems associated with copulas, commonly attributed to multivariate discrete data, they adopted the “jittering” method to the heterogeneity on the frequency of claims. Cogently subject-specific effects are accounted in the model by using available covariate information through a regression model.

The predictive distribution together with the corresponding credibility of claim frequency can be derived from the model claim counts which have the effect of continualization of his the data. They proposed Elliptical copulas to accommodate the inter temporal nature of the “jittered” claim counts and the unobservable subject-specific for rate making and risk classification purposes. For empirical illustration, they analysed an unbalanced longitudinal dataset of claim counts observed from a portfolio of auto-mobile insurance policies of a general insurer in Singapore. They further establish the validity of the calibrated copula model, and demonstrated that the copula with jittering” method outperforms standard count regression models when modelling claim. The only shortfall of their modelling process was that they combine both individual and aggregate claim in their process".

Loss distributions are used extensively in actuarial practices, both in rate making, reserving and premium adjustment. A number of approaches have been developed to calculate individual and aggregate loss distribution, including the Heckman-Meyer mathod, Penjer method, Fast Fouvier transform, and stochastic simulation. All these methodes are base on the assumption that separate loss frequency and loss severity distribution are available.

Sometimes, however,it is not practical to obtain frequency and not severity distribution separately and only aggregate information is available for analysis. In this case the assumption about the shape of claim of individual loss distribution becomes very important experientially in "tail" of distribution. This study will address the question of what type of probability distribution is the most appropriate to use to approximate an individual loss data.

(Yulia, *et al*, 2010) published a paper on the asymptotic behavior of the compound claim distribution. He shown that under some special conditions, if the distribution of the number of claims negative binomial, then the distribution of the aggregate claim have asymptotic as a gamma-type distributions in its tail.

A similar result is describe in (Peng Shi, 2011). The theorem states that under certain condition a negative binomial frequency leads to an aggregate distribution which is appropriate gamma. The skewness of gamma distribution is always twice its coefficient of variation since the loss distribution is usually positively skewed, but not always have skewness double it coefficient of variation adding a third parameter to the gamma was suggested. The ideal method to test the fit of a theoretical distributions to a loss claim is to compare the theoretical distribution with the actual statistical family of distributions to the data.

(Yulia *et al*, 2010) in their paper Handling the Dependence of claim severities with Copula Models did a study on the modelling of claim severity data in actuarial literature as well as in insurance practice agreed that claim cost distributions generally have positive support and are positively skewed, the regression models of Gamma and Log normal have been used by practitioners for modelling claim severities. However, the fitting of claim severities via regression models assumes that the claim types are independent. In their study, they made an independent assumption between claim types and also investigated consider other categories of insurance three types of Malaysian motor insurance claims namely Third Party Body Injury, Third Party Property Damage and Own Damage and applied the normal,(Frank and Clayton, 2010) used copulas for modelling dependence structures between these insurance claim types. Their findings were that the Akaike's Information Criterion and Bayesian Information Criterion reveal that the Clayton is the best fit model for modelling copula defence between own damage (OD) and third Party Body Injury (TPBI) claims and between Third party property damage and own damage claims, whereas the t-copula is the best copula for modelling dependence between TPBI and TPPD claims. This study modelled the dependence between insurance claim types using copulas on the Malaysian motor insurance claim severity data. The main advantage of using copula is that each marginal distribution can be specified independently based on the distribution of individual variable and then joined by the copula which takes into account the dependence between these variables. Based on the results, the estimated of copula parameter for claim severities indicate that the dependence between claim types is significant but have failed to do further test on the other statistical distributions mention earlier in the other papers.

In another work done by (SAS Global Forum, 2010) attempt of modelling claim severity came out with an interesting key terms when fitting probability distributions for the severity of random events. Vital examples include events with negative impact 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 study, the severity procedure estimates parameters of any continuous statistical distribution that is used to model the severity of a continuous-valued event of interest. In the study severity was explained as the severity of an event does not follow typical distributions such as the normal distribution that are often assumed by standard statistical methods. It provides a default set of probability distribution models for various distributions that are used for severity modelling. In the simplest form, you can estimate the parameters of any of these distributions by using a list of severity values that are recorded in data set.

(Emiliano A., 2009) Hierarchical Insurance Claims Modelling outline three component of his modelling process where he fit model of a third party insurance claim to find which model fit his data outline Three component (hierarchical) models: claim frequency, type of claim and claim severity observe that The severity likelihood clearly depends on the combination of the types of claims observed. He also notes the additional complication of observing claims for “own damages” type for only above the applicable excess.

In analysing trends in Loss Frequency and Severity (Ethan *et al*, 2009) Hanover Insurance evaluates historical data to analyse trends in loss frequency and severity of claims. The trends are caused by external factors, such as legislative, environmental and economic forces. Trends were analysed using two different approaches, one correlating the trends from prior data to external factors, and another comparing the impact of events to trends in the data. The analysis mathematically quantified the effect of each external force and isolated factors which were most significant to the trends.

(Grainne, 2007) presented a paper on Individual claim modelling of Compulsory Third Party(CTP) motor insurance data where he use generalized linear model to further extend the finalized claim size generalized linear model in Taylor and (McGuire, 2004) to incorporate claim severity to enable the model to deal appropriately with the changing mix of claims. His finding presents a similar realistic case study of the modelling of a compulsory third party insurance data set; in this paper.

The modelling process for (CTP). This paper has discussed the fitting of a severity based claim size model for CTP data. It has covered such issues as how to use interactions to produce a model that takes into account different features of different severity classes and how premiums could be adjusted to take care of future claims.

(Edward w. *et al*, 2007) in their hierarchical insurance claim modelling, "they explained statistical modeling of detailed, micro-level auto-mobile insurance records when considered data from a major insurance company". By detailed micro-level records, insurance they refer to experience an individual vehicle level, including vehicle and driver characteristics, insurance coverage and claims experience, by year. The claims experience consists of detailed information on the type of insurance claim, such as whether the claim is due to injury to a third party, comprehensive or claims for damage to the insured, as well as the corresponding claim amount they propose a hierarchical model for three components, corresponding to the frequency, type and severity of claims. "They use loss models for assessing claim severity and frequency. The driver's age, gender, past record and no claims discount as well as vehicle age and type turn out to be important variables for predicting the event of a claim.

The second is a multinomial logit model to predict the type of insurance claim, whether it is third party injury, third party property damage, insured own damage or some combination. Year, vehicle age and vehicle type turn out to be important predictors for this component their third model is for the severity component". Here, they used a generalized beta of the second kind long-tailed distribution for claim amounts and also incorporate predictor variables. Year, diver's experience, vehicle age and driver's age turn out to be important predictors for this component. Not surprisingly, they show that there is a significant dependence among the different claim types; they went ahead to use a t-copula to account for this Dependence.

All the models components model provides justification for assessing the importance of a rating variable. When taken together, the integrated model allows an actuary to predict auto-mobile claims more efficiently than traditional methods. Using simulation, they demonstrate this by developing predictive distributions and calculating premiums under alternative reinsurance coverage. (Gordon, 2006) fitted tweediest compound Poison model to a general insurance claims data, he observed that the dependence of the likelihood function on p is as for a linear exponential family, so that modelling similar to that of generalized linear models is possible. The study, they found that, when modelling the cost of insurance claims, it is generally necessary to model the dispersion of the costs as well as their mean. In order to model the dispersion using the framework of double generalized linear models, (Gordon, 2006) found that dispersion increases the precision of the estimated tariffs. Gordon also stated that the use of double generalized linear models also allows us to handle the case where only the total cost of claims and not the number of claims has been recorded and estimated.

(Glenn, 2005) constructing a classical non- Bayesian confidence interval of parameter selected distributions in modeling claim severity, using a likelihood ratio test; he revel some classic example showing how to use the likelihood function and Bayes' theorem to estimate the costs of high claim of motor insurance.

Most of his assumptions are arguable and regardless of how the debate surrounding his work, it is resolved that much is work is still needed to be done on his work. He hope his paper provides strong evidence that such an approach can succeed and provide a sound basis and methodology for insures to use in premium setting when the exposures are high risk to the insurer and determining claim amount within a certain period.

(Keiding *et al*, 1998) also had a study on how to modify cox regression model commonly used survival function to study the hazard occurrence of claim on motor, property and household insurance, His finding has attributed certain factor that bring about claims as preventable.

(Cramer, 1930) stated that "the objective of the theory of risk is to give a statistical and mathematical interpretations of the random fluctuations in the general insurance industry and to analyze the various ways of protection against their inconvenient effects." The actuary in the general insurance industry needs to have a wider understanding of various types of probability models for the risk consisting of the individual and aggregate amount of claims payable by a company over a fixed period of time. Such models will inform the company and enable it to make decisions on amongst other things: expected profits, premium loadings, reserves necessary to ensure (with high probability) profitability, and the impact of reinsurance when necessary and deductibles.

2.3 Actuarial Approach to Modeling

The brain behind any insurance work, is base on the power of actuarial projections and analysis. The survival of the company rest on the shoulder of the actuary. the Actuary is the key behind premium loading, all underwriting work, claim estimation profit testing and reserving.

Series of actuarial studies have been carried out on the modelling of claim frequency and severity data in actuarial literature as well as in general insurance practice. It has well been established that the claim cost distributions generally have positive support and are heavy tailed and skewed, the regression models of weibull, Gamma and Log normal have been used by practitioners all the years for modelling claim severities. It is always assume that, fitting of claim severities via regression models assumes that the claim types are independent.

The early 1980's and early 1990's there was various effort by actuaries to combine financial theory with stochastic approach into their modelling processes to improve the quality of their finding and providing a wider range of models for establishments. (D'arcy, 1989). In resent ages the actuarial profession, both in practice and in research, many actuarial organizations combines life tables, claim models, stochastic models and loss models, methods and financial theory in taking decisions.

(Feldblum, 2001) Unfortunately in Africa and specifically Ghana there are limited articles written about actuarial modelling for insurance claim severity from different perspective. (Miguel *et al* , 2009) used inflated generalized Poison regression model to model disability severity in motor insurance claims, it observes that bodily injury claims have the greatest impact on the claim costs of motor insurance companies. The disability severity of motor claims is assessed in numerous European countries by means of score systems. The study estimates the disability severity score of victims involved in motor. He show that the injury severity estimates may be automatically converted into financial terms by insurers at any point of the claim handling process. As such, the methodology described may be used by motor insurers operating in the

Spanish market to monitor the size of bodily injury claims. By using insurance data, various applications are presented in which the score estimate of disability severity is of value to insurers, either for computing the claim compensation or for claim reserve purposes.

(Wright, 2005) used the actuarial modelling steps to fit models in data drawn several institution in consecutive years. The study fitted loss amount on statistical distributions using maximum likelihood estimation for the of year of all the observations. The study used q-q plots and (K-S) test to assess the absolute quality of fit. For this study, in the study Q-Q plots was use to confirm the quality of fitness as its advantages was cite for a given response. Wright tried several statistical distributions which included the inverse fetched distribution, Pareto distribution, gamma distribution burr distribution, inverse burr log-normal,restricted benktander including many other families of distribution families. The benktander family has the property (like Pareto and exponential families) that left truncation gives another distribution in the same family. While this research used almost similar distributions like the above, it did not use the benktander family even though it's a statistical distribution because the study did not focus on the excess loss amount but the whole claims data set. (Meyers, 2005) In actuarial modelling to fit a data on some family of selected distributions, the log-normal, gamma, Weibull and pareto distribution were tested. The study uses used maximum likelihood estimation to fit the distributions, that idea was extended to many of his research papers, only that he later applied the idea based on Bayesian solution. That is, after he calculated the likelihoods, he used them to get the posterior probabilities of each model. For this research however, the likelihoods were used to calculate the Akaike's Information Criterion so as to choose the best best probability distribution that fits the claims data best.

(Yulia *at el*, 2010) A research published by yuli 2010 on the dependence of thirty party insurance claim severities with Copula models, he made independent assumption between claim types will be investigated as we will consider three types of Malaysian motor insurance claims namely Third Party Body Injury Third Party Property Damage and Own Damage and applied the normal, Frank and Clayton copulas for modelling dependence structures between these claim types. The Akaike's Information Criterion and Bayesian Information find that the Clayton is the best copula for modelling dependence between TPBI and OD claims and between TPPD and OD claims, whereas the t-copula is the best copula for modelling dependence between TPBI and TPPD claims. This study modelled the dependence between insurance claim types using copulas on the Malaysian motor insurance claim severity data. The main advantage of using copula is that each marginal distribution can be specified independently based on the distribution of individual variable and then joined by the copula which takes into account the dependence between these variables. Based on the results, the estimated of copula parameter for claim severities indicate that the dependence between claim types is significant.

(Noiszura *at el*, 2006) did a comparison of risk classification methods for claims severity date. They compare several risk classification methods for claim severity data by using weighted equation which is written as a weighted difference between the observed and fitted values. The weighted equation was being applied to estimate claim severities which is equivalent to the total claim costs divided by the number of claims. From their data they 2 observed that, the classical and regression fitting procedures give equal values for parameter estimates. However, the regression procedure provides a faster convergence. The multiplicative and additive models give similar parameter estimates. The smallest chi-squares are given by the minimum chi-squares model. Except for the exponential model, all models provide similar values for absolute difference.

(Renshaw, 2004) Introduces another useful approach in actuarial modelling loss amounts for non-life insurance basing on the concept of quasi-likelihood and extended quasi likelihood. he employed the use of the maximum likelihood estimates to fit the model structure; this is because the quasi-likelihood parameter estimates according to him have similar asymptotic properties to the maximum likelihood parameter estimates. This research however did not touch on the quasi likelihood approach but used the maximum likelihood estimates and the steps followed by renshaw in his modelling procedure goes in line with the procedures involved in all actuarial modelling process.

(Fiete, 2005) commission on risk theory COTOR) said, using Pareto, gamma, lognormal, Weibull to revise a paper on 4050 accident claims in Basi insurance. He used similar steps in modelling process and also used maximum likelihood estimation to obtain parameter values but due to the nature of his data he evaluated the goodness of fit using P-P plots because they allowed him to examine goodness of fit across the entire range of possible outcomes so as not to depend on a single number from the log-likelihood to describe goodness of fit but have to apply other method for validation.

(Luyan, 2004) also presented a paper on Severity Distributions for generalized linear model; Gamma Log normal, and Weibull. The was basically on how motor insurance claims severity can be fitted to a profitability distribution, with the assumption that there has been a claim through accident year. The study has clearly spelt out all the necessary steps involved in Actuarial modelling process, giving more detailed on the criteria and his methodology.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter consists of the methodology. Claims data considered were comprehensive motor insurance policies of a major motor insurer in Ghana for claims that paid in the year 2014 from January 2014- December 2014. Data correspond to a cost of claims, expressed in Ghana cedis in a random sample of all comprehensive motor insurance claims. Certain basic assumptions were made. The analysis used followed assumptions that attributed to the claim data under consideration, these included:

1. All the motor insurance claims came from one distribution distribution (they were independent and identically distributed).
2. There exist no zero insurance claim for any motor under the policy.
3. There exist no catastrophic claim.

3.2 Exponential Family of Distribution

Many distributions used by actuaries share a common structure and can be grouped into an exponential family. This has made it possible to construct the common framework for analysis referred to as generalized linear models. For GLMs, response variable is assumed to have a probability distribution function that can be written as;

$$f(y,\theta,\phi) = exp[\frac{y\theta - b(\theta)}{\alpha(\phi)} + c(y,\phi)] \tag{3.1}$$

Where: φ is the dispersion parameter or, sometimes, nuisance parameter θ is often referred to as the canonical parameter, natural parameter, or parameter of interest. $c(y,\varphi)$ determine the type of distribution; for example, normal, Poison, and so on.

3.2.1 Processes in Actuarial Modelling

In Actuarial modelling, there are basic standard that will be followed in fitting the various probability distribution under consideration. (Kaishev, 2001) These steps are:

- choosing family for the a model distribution
- Specifying of the criteria to choose one model from the family of distributions
- Checking for all the model fit
- revising and validating model fit if necessary

3.2.2 Choosing Model for the Family of distribution

These are the fundamental step in the modelling process. considerations were made of a number of parametric probability distributions as potential candidates for the data generating mechanism of the claim amounts. Most data in general insurance data is skewed to the right and therefore most distributions that exhibit this characteristic can be used to model the claim severity, (Fiete, 2005). However, the list of potential probability distributions is enormous and it is worth emphasizing that the choice of distributions is to some extend subjective. For this study the choice of the sample distributions was with regard to:

- Having Prior knowledge and experience in curve fitting
- considering time constraint
- Availability of computer soft-ware to facilitate the study
- The volume and quality of data

Therefore four statistical distributions were used, these included: normal, gamma, log normal and Weibull.

Let X be the claim amount.

Weibull Distribution

The Weibull is a very general and popular failure distribution that has been shown to apply to a large number of diverse situations. Weibull distribution is right skewed and it is use to represent nonnegative task times. This distribution is named after WaloddiWeibull, a Swedish physicist, who used it in 1939 to represent the distribution of the breaking strength of materials. The distribution has also been used in reliability and quality control.
(Formulae and Table FOA, 2002).

X:Weibull(c,γ) parameter:

c>0,γ >0

The probability distribution function is given by;

$$f(x) = c\gamma x^{\gamma-1}e^{-cx^{\gamma}}, x > 0 \tag{3.2}$$

$$F_x X = 1 - e^{-cx^{\gamma}} \tag{3.3}$$

3.2.3 Gamma Distribution

The Gamma distribution is a two-parameter family of continuous probability distributions and has a right skewed distribution. It is very useful for risk analysis modeling, particularly, for claims size modeling. The Gamma distribution is given by The probability distribution function(PDF) is given by; (Formulea and Tables, FOA, 2002)

$$f(x,\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}exp(-\lambda x), x \geq 0, \alpha, \lambda > 0 \tag{3.4}$$

The cumulative distribution function (CDF) is given by:

$$F(x,\alpha,\beta) = 1 - \sum_{i=0}^{\alpha-1} \frac{(\beta x)^i}{i!} e^{-\beta x} = \sum_{i=\alpha}^{\infty} \frac{(\beta x)^i}{i!} \quad (3.5)$$

The gamma distribution has a mean, variance and skewness as:

$$Var(X) = \frac{\alpha\lambda^2}{(\alpha-1)(\alpha)-2}, (\alpha>2) \quad (3.6)$$

$$Var(X) = \frac{\alpha}{\lambda^2} \quad (3.7)$$

$$Skew(X) = \frac{2}{\sqrt{\alpha}} \quad (3.8)$$

Log-normal Distribution

The lognormal distribution is applicable to random variables that are constrained by zero but have a few very large values. The resulting distribution is asymmetrical and positively skewed. The application of a logarithmic transformation to the data can allow the data to be approximated by the symmetrical normal distribution, although the absence of negative values may limit the validity of this procedure.

(formulae and Table, FOA, 2002) Log normal distribution ($log \sim N(\mu,\sigma^2)$)

Parameter: $-\infty < \mu\infty, \sigma > 0$ The probability distribution function is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{logx-\mu}{\sigma}\right)^2\right\}, x > 0 \quad (3.9)$$

The Cumulative distribution function is given by;

$$\phi\left(\frac{lnx-\mu}{\sigma}\right), N(0,1), x > 0, \sigma > 0, \mu \in R \quad (3.10)$$

The log normal has mean, variance and skewness as follows:

$$E(X) = e^{r\mu + \frac{1}{2}\sigma^2} \quad (3.11)$$

$$Var(X) = e^{2\mu+\sigma^2}\left(e^{\sigma^2-1}\right) \quad (3.12)$$

$$Skewness(X) = \left(e^{\sigma^2+2}\right)\sqrt{e^{\sigma^2}-1} \quad (3.13)$$

Normal Distribution

The normal distribution is important in insurance and finance since it appears like limiting distribution in many cases. The normal distribution is applicable to a very wide range of phenomena and is the most widely used distribution in statistics. It was originally developed as an approximation to the binomial distribution when the number of trials is large and the Bernoulli probability p is not close to 0 or 1. It is also the asymptotic form of the sum of random variables under a wide range of conditions. Parameter: $\mu,\sigma(\sigma > 0)$ The probability distribution function is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\frac{1}{x}exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, -\infty < x < \infty \quad (3.14)$$

The normal distribution has mean and variance as follows:

$$E(X) = \mu \quad (3.15)$$

$$Var(X) = \sigma^2 \quad (3.16)$$

3.2.4 Parameters Estimators

The study concerned mainly interested with probability theory: which events may occur and the probability of occurrence, given a model family and choices for the parameters. This is useful only in the situation where we know the precise model family and parameter values for the situation of interest. These are the exceptions, not basically a rule, for both scientific inquiry and for assumptions learning inference. In many occasions, we are not certain with the processes of data which are generative from source we are not rarely certain estimation of the parameters for each of the above sampled probability distributions using the insurance claims data. Once the parameters of a given distribution were estimated, then a fitted statistical

distributions was available for further analysis. The maximum likelihood estimation approach shall be use to estimate the
parameters.

3.2.5 Maximum Likelihood Estimation

The method of Maximum likelihood was used to fit unknown parameters of the distribution to the data set. The likelihood function of a variable, x, is the probability of observing what was observed given a hypothetical value of the parameter. The maximum likelihood estimate (MLE) is the one that yields the highest probability that is that which maximizes the likelihood function. A parameter is a descriptor of the model. A familiar model might be the normal distribution of a population with two parameters: the mean and variance.(ActEd Study Materials, 2013) The idea of likelihood was encountered as a basic measure of the quality of a set of predictions with respect to observed data. In the context of parameter estimation, the likelihood is naturally observed as a function of the parameters θ to be estimated. The Maximum Likelihood estimates will be use because it provides several acceptable properties which include: efficiency, unbiased, asymptotic consistency, invariance normality and . The merit of considering the Maximum Likelihood estimation is that, it fully uses all the available information about the known parameters contained in the data and that it is highly flexible, and the method is statistically well understood. Let X_i be the ith claim amount , where $n = i_1, i_2, i_3, \ldots, n$ n is the number of claims in the data. L is the likelihood function θ is the maximum likelihood estimator.

f(x) is the probability distribution function of a specific distribution. (ActEd Study Materials: CT6-PC-13) The likelihood function of the claims data is L(θ) given by:

$$L(\theta) = \prod_{i=1}^{n} f(x_i/\theta) \tag{3.17}$$

The maximum likelihood, determine on the equation(3.17) above is given by:

$$L(\theta) = logL(\theta) = \sum_{i=1}^{n} logf(X_i/\theta) \tag{3.18}$$

$L(\theta)$ can be differentiated with respect to θ, the MLE, express as $\hat{\theta}$ and equate it to zero.

$$MLE = \frac{dL}{d\theta}(\hat{\theta}) = 0 \quad (3.19)$$

The only thing needed was to derive the maximum likelihood estimate for the set family of the selected statistical models in the form of a likelihood function as a probability of getting the data readily available. All that was needed to estimate the parameters was the R software. All the parameter were estimated using software and no manual work were done. The researcher will fit values for AIC and BIC. The AIC and the BIC are two most commonly use to compare the likelihood mode. Both AIC and BIC are use to check consistency.

They are simply define as; The Akaike's Information Criteria (A.I.C) The A.I.C is a type of criteria used in selecting the best model for making inference from a sampled group of models. It is an estimation of kullback-leibler information or distance and attempts to select a good approximating model for inference based on the principle of parsimony. (Anderson & Burnham, 2004) This criterion was derived based on the concept that truth is very complex and that no "true model" exists. Therefore in A.I.C, the model with the smallest value of A.I.C is selected because this model is estimated to be closest to the unknown truth among then candidate models considered.

The AIC criterion is defined by:

$$AIC = -2 * ln(likelihood) + 2k \quad (3.20)$$

$$BIC = -2 * ln(likelihood) + ln(N) * K \quad (3.21)$$

Where K is the number of parameters and N is the number of observations.

CHAPTER 4

ANALYSIS

4.1 DATA ANALYSIS

Introduction

This chapter basically discusses the various distributions of claim data from a general assurance company in Ghana from January 2014 to November 2014. The chapter used exploratory data analysis (histogram, mean, variance skewness, kurtosis maximum value, minimum value, standard deviation, and 1st and 3rd quartile) to assist in the identification of the family of distribution which the data might follow. The diagnostics test Probability plot was used to graphically demonstrate goodness of fit to the Log-normal distribution, the

Weibull distribution, the Gamma distribution, and the normal distribution. The Goodness-of-Fit tests were used to test fitness of the distributions.

4.2 Descriptive Data Analysis for the Claims

The exploratory data analysis gave detail accounts preliminary analysis of the findings of the study. The sections below explain them.

4.2.1 Descriptive Statistics

A total of 1,050 claims data on motor comprehensive insurance for 2014 was used for the modeling. Table 4.1 below summarizes the result of the descriptive data analysis of the claims amount. The data was truncated in order to allow values that are above 354,200 and those below 5.86, fit into the data set so that the data set will be equalized to the underlying sample with values outside the bound entirely omitted. This is to allow some level of homogeneity in the modelling process and to permit actuarial modelling. The average of claims was computed. The sum claim, standard deviation, skewness, minimum and maximum value as well as the quantiles are also shown. This summary was

necessary because it helps to identify essential features of the data.

Table 4.1: Descriptive statistics for the motor claim data

Statistics	Value
Minimum(GHS)	5.86
Maximum(GHS)	354,200
1^{th} Quatile (GHS)	10,09
2^{nd} Quantile(GHS)	5,972.25
Mean(GHS)	8,271 .63
Sum(GHS)	8685246.00
Variance	75317440
Standard Deviation(GHS)	2,3941.06
Skewness	8.63
Kurtosis	96.45

From table 4.1 above summarizes the result of the descriptive data analysis of the claim data. The GHS 5.86 is the minimum loss amount that was paid to policy holders. The maximum loss amount of claim was GHS 354,200.00. This mean that within that period, the highest claim amount paid by the insurer to a policy holder was GHS 354,200.00. The 25^{th} quartile of loss paid was GHs 10,092 and that of the 75^{th} quartile was GHS 5,972.25, the mean claim loss payment was GHS 8,271.63 and the variance was GHS 57,317,440.00 The standard deviation is 12394106 and that of the coefficient of skewness is 8.63. The skweness was measuring the symmetric nature of the claim. The value 8.63 informs how the claim amount was positively skewed. Kurtosis has a value of

96.447, kurtosis measures whether the data is heavy-tailed or light tailed. The value 96.447 indicates that the data is heavy tailed the sum claim payment is GHS 8,685,246.00.

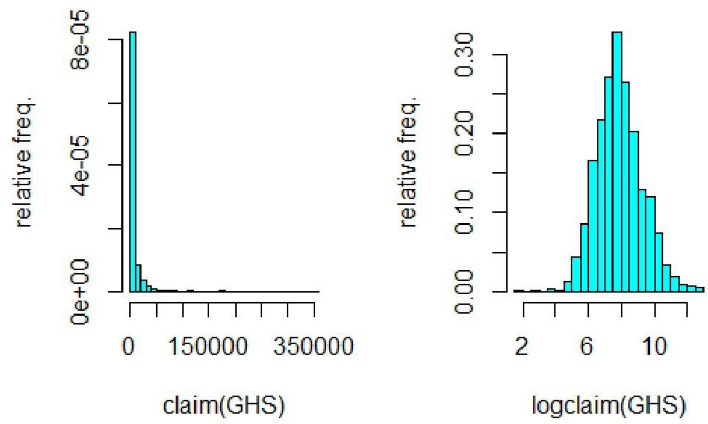

Figure 4.1: Histogram of claim amount and log claim amount respectively

Figure 4.1 above shows the relative frequency graph of the loss amount is high for less claims and positively skewed. As stated by (Oyugi,2012) "This curve shows the skewness of the claims data. From the figure 4.1, it can be seen that the original claims data has skewed positive. This was interpreted to mean that the claims data had less claim amounts which occurs frequently while large claim amounts rarely occurs " this mean that a larger claim amount has a smaller relative frequency and a smaller claim amount have larger relative frequencies (Oyugi, 2012). This is an underlying characteristics of claim severity. The histogram was to visualize the shape on the claim amount. It was also observed that the claim data has less claim amounts but very high values because of the high variability in the claim data set even though the claim was only one year. The claim amount was transformed to log claim amount to improve symmetry . After the claim data was transformed, a new histogram, log claim amount was plotted (ie from figure 4.1) this is an attempt to reduce the level of skewness of the data and make it quite symmetric.

The figure 4.2 shows the histogram of the transformed claim amount superimposed with the kernel density. The kernal density is a non parametric technique which was used to visualized the underlying distribution for claim amount. Essentially, the kernel density has converge all the loss claim amount to the center. This ensures that the kernel is symmetric about all the claim. The kernel density estimates converge the data set faster to the true underlying density for continuous claim amount. The clumsy and unpronounced nature of the histogram from the raw claim amount has been overcome (partially) using the nonparametric kernel density graph. In principle, the kernel estimate was used in averaging and smoothing the histogram. The interval for the relative frequency axis and the claim (GHS) axis looks consistent than that if the histogram. This has guided the researcher to select a set of distribution for the study. Before choosing one or more models for the data, it is generally, necessary to choose good model among a predetermine set of models. This choice may be guided by the knowledge or the behavior the modeled variable, or, in the absence of knowledge regarding the underlying process, by the observation of the empirical distribution.

The Gamma distribution, Normal distribution, Log normal distribution and Weibull distribution were chosen as a family of models for the study. This was base on the nature of the kernel density

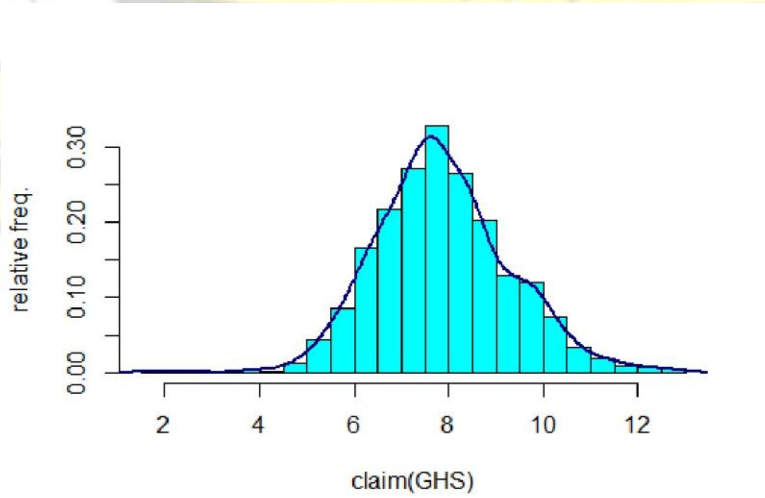

Figure 4.2: Histogram of log claim amount with a kernel density estimate.

4.2.2 Selection of Candidate Models

The descriptive data analysis results helped in the in the selection of the following distribution for the modelling process. Due to the symmetric nature of the data set the Normal distribution, the Log-normal distribution, the Weibull distribution, the Gamma distribution, were selected as the candidate models after testing a wide range of distributions.

4.2.3 Estimation of Model Parameters for the Claim Data

Given any model, there exists a great deal of theories for making estimates of the model parameters based on the empirical data (Achieng, 2010), in this case, the claims were rounded up to compute the maximum likelihood estimates of the several distributions and were fitted whiles the first step used in fitting the model to the claims data was done by finding the parameter estimates of their statistical distribution. When the parameters of any distribution have been obtained using the claims data, then literally, the statistical distribution has been fitted to the claims data (Achieng, 2010). Tables 4.2 show the parameter estimates from the models considered: the normal (μ and σ parameters) the Log-normal distribution (μ and σparameters); the Weibull distribution (α and β parameters); and the Gamma distribution (α and β parameters); were fitted to the claims data obtained from the motor comprehensive insurance data.

Table 4.2: Parameter estimates from maximum likelihood

Probability Distribution	Parameter Estimate	Value
Normal	$\hat{\mu}$	7.872044
	$\hat{\beta}$	1.415005
Lognormal	$\hat{\mu}$	2.0464112
	$\hat{\beta}$	0.1877511
Gamma	$\hat{\alpha}$	29.74034

	$\hat{\lambda}$	3.77800
Weibull	$\hat{c}$	5.799700
	$\hat{\lambda}$	8.465012

From table 4.2, the estimated parameters for the family of distributions normal, log normal, Gamma and Weibull were; normal has a mean of 7.872044 and a variance of 1.415005, log-normal has a mean of 2.0464112 and a variance of 0.1877511. The gamma distribution with a location of 29.74034 and a scale estimate of 3.77800, and the Weibull has location of 5.799700 and and scale estimate of 8.465012.

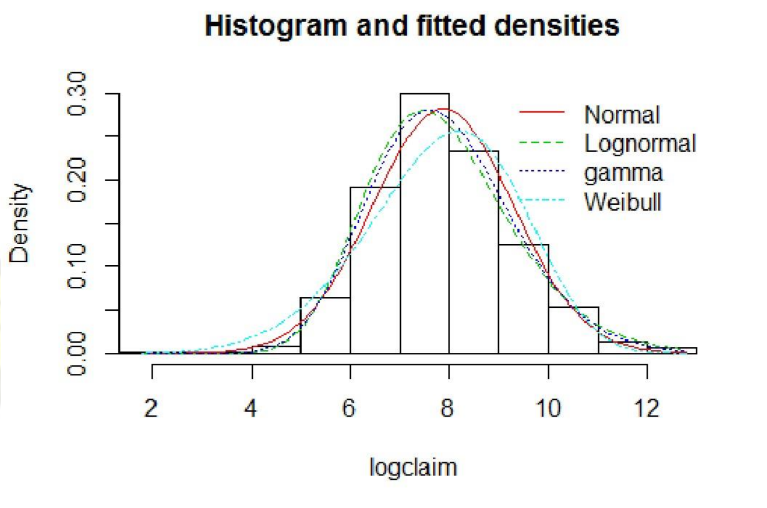

Figure 4.3: Histogram and fitted densities of claim amount

Figure 4.3 shows the fitted density plot for log claim amount with Log-normal, Normal, Gamma and the Weibull models. The models were fitted to loss claim data within period January 2014 to November 2015 covering about 1,050 beneficiaries. To visualize the diagram of which of the family distribution is accurately fitted to the log claim data. The pattern of diagram is sufficiently regular, approximate it with a smooth curve. It was observed that the pattern the histogram superimposed by the densities of the four models under consideration are sufficiently regular and approximate with he smooth curve. Some of the densities or curves are lying above or underneath the histogram. The area under the curves indicates the proportion of range of values within the claim range.

From the histogram and the fitted densities, it was observed that the normal and the lognormal are well superimposed to the histogram bars. The two were chosen to further test for their empirical and the gamma and the Weibull as the theoretical distributions.

Additional test was conducted. This test was to visualize the behavior of the four loss model to the loss data. This was done by using the quantile- quantile plot. The q-q plot as an exploratory visual devise was used to check appropriateness of the model. It was use to compute the theoretical model to the expected amount to see if the plot will fall appropriately with the reference line Figure 4.4, show how close the fitted model are to the claim data. The Normal and Log-normal were paired as empirical and the Gamma and Weibull were also paired as theoretical quantiles for a visual representation plot. A careful observation reveals that, the Normal and Log-normal lie on the 45- degree reference line with just few outliers . This is a clear indication that the normal and the Log-normal we choose to the distribution of claim amount. This is because almost all the points fell approximately along the reference line with only some few outliers below and above the plot. This is a greater evidence that the Log-normal and the Normal fit the claim data. The Gamma and the Weibull shows a deviation at both the bottom and upward the 45-degree reference line with a lot of outliers quiet fall away from the 45-degree line. This is an indication that they might not be fitting well to the claim amount. In furtherance to the above explanation, the log-normal and the normal can be said as models from the claim data because they are linear to the reference line. Again, there is some systematic deviation of the Weibull and the Gamma from the ideal reference line observing from the rare.

Figure 4.5 shows the fitted plot for the Normal and the Log-normal model. The empirical and theoretical CDFs plot was used to asses whether the sample claims data observed came from the Normal or Log-normal. Comparing the general trend of both the Normal and the Log-normal CDFs.The Normal curve look more flatten than the Log-normal. Comparing the two models, the normal looks more appropriate than the Log-normal. Even though the S shape indicating that both distributions are better choices. But critical observation shown that the normal has shown a more flattens shape than the Log-normal.

The Log-normal has shown some level of skewness indicating that it has some level of heavier tail than the log-normal.

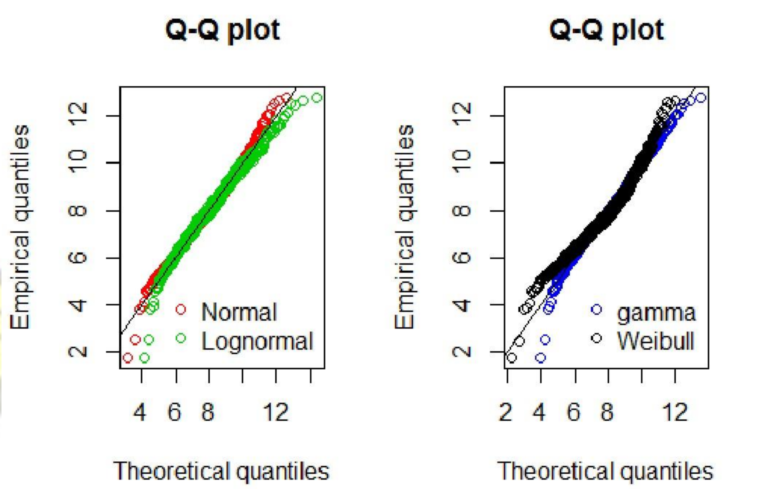

Figure 4.4: Empirical Quantile plot and theoretical Quantile plot

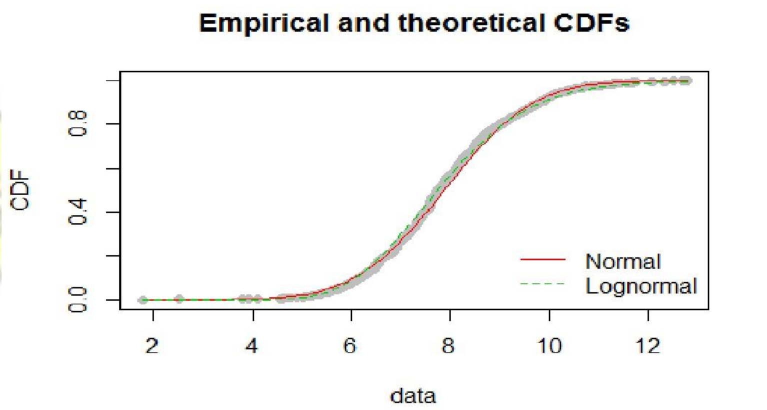

Figure 4.5: Empirical and theoretical CDFs for distributions

The parameters obtained were then used in the computation of the log-likelihoods AIC and BIC statistics of the four distributions as The maximum likelihood estimation method was used to estimate the individual parameters because of its numerous advantages as discussed in Chapter three over the other methods of parameter estimation. With the aid of R software statistical software the parameters were computed and displayed in the table below.

Table 4.3: Log likelihood and AIC and BIC statistics

Distribution	AIC	BIC
Normal	1856.599	3717.197
Lognormal	1882.437	3738.695
Gamma	3740.88	3740.793
Weibul	3800.634	3810.547

Table 4.3 shows Akaike's Information Criterion (AIC) and Baysian Information CriterionBIC statistics. The values computed were used as a basis for choosing the best model that fit the data. From the table above the normal distribution has AIC of 1856.99 and BIC of 3717.197. The lognormal AIC of 1882.437 and BIC of 3768.695. The gamma distribution AIC of 3740.88 and BIC of 3740.793 and finally the Weibull distribution AIC of 3800.634 and BIC of 3810.547. From the

AIC and BIC is the best distribution to model the claim amount is the normal distribution followed by the lognormal. The other two models do not best fit claim data.

4.3 Normal distribution

This section is interested in the post model selection fit to affirm the selected models. (Achieng 2010) argued that the central problem in analysis is the kind of model one needs to use for making inferences from the claim dataset. This is known as the model selection problem (Achieng, 2010). As already shown and proven in both the selection criterion and the goodness of fit test in the earlier sections the normal emerged as the best fit since it has least AIC followed by the logormal and with the BIC it was the Normal followed by the Log-normal distribution. It was necessary to use goodness of fit criterion in order to select the loss model that best fits the claim data. This stage has really stated the need for such work. The expectation of the claim amount using the best fit model that is the Normal model is given by:

$$E(X) = e^{\mu} \tag{4.1}$$

where μ is 7.872044 Finally the expected claim amount was estimated as GHS 2,622.92

CHAPTER 5

CONCLUSION

5.1 Introduction

This chapter is the final stage of the study process; it includes the summary, conclusion to the finding and recommendations future research.

5.2 Summary

The objectives set out for this research was to investigate which best actuarial model will fit for the comprehensive insurance claims data. It was also targeted to see how well this model will fits the claims data so that the model will be recommended to the company for modeling the claim amounts. In a very important sense, the study was not interested with the steps of modeling the data instead; the study tries to model the information in the data to fit a
particular distribution.

An attempt was made to come out with the best loss model that best fits the claims amount data using numerical computations and graphical representation. From the rigorous analysis carried out coupled with the results displayed in table 4.2 and 4.3 and figure 4.4 and 4.5 it can concluded that the insurance company can use Normal model to model motor comprehensive severity claims data since that is the ultimate loss model among the four fitted models.

From the results presented in table 4.3, Normal distribution has highest likelihood function of 3778.608 followed by the Log-normal model with 3727.11, which making those two distributions to pass through their fitting, with the data.

Out of the four selected statistical distributions, testing which one is providing a good fit for the claims data. The Weibull and Gamma are not good quite for the claim amount. As discussed earlier, the theorem of AIC and BIC states that the model with the smallest AIC and BIC is the best model. The Normal has AIC of 18562.599 followed by the Log-normal with AIC with 1882.437.608. Even though Gamma model and Weibull model has some nice fitness it is not as appropriate like the normal and the subnormal model.

Finally, Q-Q plots for the selected statistical distribution were plotted. To graphically re-affirm the goodness of fit test computed by the A.I.C. 4.2 indicates that the Q-Q plot of the normal distribution gives the best fit to the insurance claims data since almost all points plotted on lies directly on the reference line and just few of these points scattered along the line. Again, Looking at the Q-Q plot from the empirical quantile for the Log-normal model, it has also shown some level of fitness but the better choice is the normal. The Gamma and the Weibull not given enough evidence as choices from the family of distribution as good fit for the purpose of this study even though they have also exhibited that they have some level of fitness for a number of points on the extreme ends of the theoretical plot, plotting far from the reference line.

After careful and systematic study and going through all the necessary process in the statistical modeling processes with the above summary indicates that the normal distribution would provide a good fit to the motor insurance comprehensive claims data. Therefore if the company is to model the motor comprehensive policy claim amounts experienced in January, 2014 to November 2014, then the appropriate statistical distribution to use to give the appropriate or reliable claim forecasts would be the normal distribution.

Having tested the goodness of fit of the log-normal distribution taking into consideration all the parameter generated from the likelihood and graphical representations using the histogram, the Q-Q plots, from Akaike' Informative Criterion (AIC) and Bayesian Informative Criterion comparing values and outcome of all other distribution value, it is quite clear and evident that the procedural and steps followed in the actuarial modeling process is capable of giving reliable outcome that can be used to make inferences and take useful for decision in the general motor insurance industry and general insurance industry as a whole.

The study has reveal that the assumptions made before the analysis of motor comprehensive insurance claims data may greatly impart the final results as the assumptions made led to the choice a family of distribution consisting of the Normal, Gamma, Weibull, distribution of which the Normal distribution has proved to providing a good fit for the claims amounts from the company. From the modeling process carried out, it can only be concluded that more heavily right tailed statistical distributions would have been included in the study by using more advanced software so as to increase the sample distributions used in the study for accuracy of findings since from statistical point of view the higher the sample size the accuracy the results or findings.

In conclusion, the modeling process is an important step before any decision can be made with regard to future policies in the insurance industry, therefore more effort must be dedicated to ensure that the process adopted yields accurate and reliable result.

The study has shown that the assumptions made before the analysis of insurance claims data may greatly affect the final results as the assumptions made in section 3 led to the choice a family of distribution consisting of the Normal, the lognormal, the gamma and the weibull of which the Normal distribution was proved to be capable of providing a good fit for the claims amounts. From the modeling carried out one can conclude that more right hand tailed statistical distributions would have been included in the study by using more advanced software so as to increase the sample distributions used in the study for accuracy of findings.

5.3 Conclusion

The Normal model was found to be the best fit model for the claim data among the four models tested.

The Q-Q plots, AIC and BIC, Empirical CDFs was used to measure the fitness, the three tools confirmed, affirmed and reaffirmed that the normal model is the best fit model among the Gamma, Log normal and the Weibull model. The expected claim amount per risk to be paid in the coming year was estimated to be GHs 2,622.92.

After a systematic study and going through actuarial modeling process by using comprehensive claim amount paid to policy holder, the research concludes that Normal model is the appropriate model for modeling comprehensive insurance claim severity with a heavier tail. Finally, the expected claim amount to be paid per risk during the period was estimated to be GHS 2,622.92.

5.4 Limitation of the Study

Going through modeling steps stated from the initial stage of this work, the following are likely limitations encountered when applying these results on future study,

- The choice of the size of data set to be used was limited by the software used in this study. This means that, the data size was reduced in order for the software to take of the data size.
- Some few assumption were considered on the absence of zero claim amounts on the data which might reality not exists in comprehensive insurance policy. This may assumption introduces a bias to the findings of the of the research.
- This researcher assumed that the entire claim amounts experience between January 2014 to November, 2014 was all reported but by the time the data was assembled, some of the claims had been reported but the claim amounts has not been reported not included in the study.
- The researcher failed to consider the likelihood values in checking the model

fit.

- The research duration was very short and this could not allow the research to do more thorough studies.

5.5 Recommendation

In the cause of this paper some few recommendations to be consider by the motor insurance service providers in order to get very reliable result have been noted and these could greatly improve the outcome yielded by the actuarial modeling process to be able to address all claim amounts in comprehensive insurance policies. Some of these are highlighted below:

There is need for insurance companies in the general insurance industry to carry out more research about the expected future claim amounts in insurance industry. This is because the study has acknowledged that future claim amounts to help the in their premium loading:

This research did not take into account whether there exists any relationship between the claim amounts and the claim frequencies. However there is need for insurance companies to establish this relationship and the strength of the relationship between the two variables such that the modeling of one variable would be sufficient in estimating the other variable.

Implementation of the results of this research can help general motor insurance companies build their solvency states and therefore amend their policies regulations to avoid ruin before it is too late.

The researcher also recommends that data storage should be a very critical issue to consider since the insurance industry is heavily dependent on data so that the future can be predicted.

More exponential distribution with characteristics exhibited by this data can be added to the set of selected family of models.

The data size could be increase by future researcher and could also consider about four or more year to validate the data.

Reference

Achieng, O. M., & NO, I. (2010). Actuarial Modeling for Insurance Claim Severity in Motor Comprehensive Policy Using Industrial Statistical Distributions. In International Congress of Actuaries (pp. 7-12).

Achieng, O. M., & NO, I. (2010). Actuarial Modeling for Insurance Claim Severity in Motor Comprehensive Policy Using Industrial Statistical Distributions. In International Congress of Actuaries (pp. 7-12).

Adeleke, I. A.,&Ibiwoye, A. (2011). Modeling Claim Sizes In Personal Line Non-Life Insurance. International Business & Economics Research Journal (IBER),

Ammour, B. S. (2010). Introduction des effets fixes dansunmodèle binomial négatif: application à la consommation de médicaments au Canada.

Berke, E. M., & Shi, X. (2009). Computing travel time when the exact address is unknown: a comparison of point and polygon ZIP code approximation methods. International Journal of Health Geographics, 8(1), 23.

Boland, N., Domínguez-Marín, P., Nickel, S., & Puerto, J. (2006). Exact procedures for solving the discrete ordered median problem. Computers & operations research, 33(11), 3270-3300.

Budtz-Jørgensen, E., Keiding, N., &Grandjean, P. (2001). Benchmark dose calculation from epidemiological data. Biometrics, 698-706.

Cassidy, J. D., Carroll, L. J., Cote, P., Lemstra, M., Berglund, A., &Nygren, Å. (2000). Effect of eliminating compensation for pain and suffering on the outcome of insurance claims for whiplash injury. New England Journal of Medicine, 342(16), 1179-1186.

Cullen, A. C., & Frey, H. C. 1999 techniques in exposure assessment: a handbook for dealing with variability 1999 techniques in exposure assessment: a handbook for dealing with variability).

D'Arcy, S. P., &Garven, J. R. (1990). Property-liability insurance pricing models: an empirical evaluation. Journal of Risk and Insurance, 391-430.

Davies, K. F., Melbourne, B. A., McClenahan, J. L., & Tuff, T. (2011
Davies, K. F., Melbourne, B. A., McClenahan, J. L., & Tuff, T. Polar biology, 34(1), 119-125. Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. Delignette-Muller, M. L., Dutang, C., Pouillot, R., Denis, J. B., &Delignette-Muller, M. M. L. (2015). Package 'fitdistrplus'.

Dunn, P. K.,& Smyth, G. K. (2008). Statistics and Computing, 18(1), 73-86.

Efron, B., &Tibshirani, R. J. (1994). An introduction to the bootstrap (chapman & hall/crc monographs on statistics & applied probability).

Feldblum, S. (2001, November). Underwriting cycles and business strategies. In Proceedings of the Casualty Actuarial, pp. 175-235). Society (Vol. 88)

Frees, E. W., Shi, P., & Valdez, E. A. (2009). Actuarial applications of a hierarchical insurance claims model. Astin Bulletin, 39(01), 165-197.
Hossain, M. I., Samir, B. B., El-Harbawi, M., Masri, A. N., Mutalib, M. A., Hefter, G., & Yin, C. Y. (2011). Development of a novel mathematical model using a group contribution method for prediction of ionic liquid toxicities. Chemosphere, 85(6), 990-994

House of Commons Transport Select Committee. (2010). The performance of the Department for Transport: Government response to the Committee's Fourth Report of Session 2009–10: Third Special Report of Session 2010–11.

Kallah-Dagadu, G. A. B. R. I. E. L. (2013 Kallah-Dagadu, G. A. B. R. I. E. L.)

Lowther, M. W., & McMillan, W. J. PRINCIPLES-BASED CPD FOR SOUTH AFRICAN ACTUARIES.

Meyers, G. (2005). On Predictive Modeling for Claim Severity. In Casualty Actuarial Society Forum (pp. 215-253).

Meyers, G. (2005). On Predictive Modeling for Claim Severity. In Casualty Actuarial Society Forum (pp. 215-253).

Nolan, S. J., Nespor, J. S., Harris Jr, G. W., Lane, J. P., Merrell, A. R., LagueuxJr, R. A., &Panas, M. G. (2003). U.S. Patent No. 6,640,278. Washington, DC: U.S. Patent and Trademark Office.

Nourse, V., & Shaffer, G. (2009). Varieties of New Legal Realism: Can a New World Order Prompt a New Legal Theory?. Cornell Law Review, 95, 61. Shaw, A., & Hurst, J. A. (2009). 'I don't see any point in telling them': attitudes to sharing genetic information in the family and carrier testing of relatives among British Pakistani adults referred to a genetics clinic. Ethnicity & health, 14(2), 205-224.

Silverman, B. W. (1986). Density estimation for statistics and data analysis (Vol. 26). CRC press.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical therapy, 85(3), 257-268.

So, Y. (2010). Analyzing Interval-Censored Survival Data with... In SAS GLOBAL FORUM 2010.

Stapley, J., Reger, J., Feulner, P. G., Smadja, C., Galindo, J., Ekblom, R., ...& Slate, J. (2010). Adaptation genomics: the next generation. Trends in ecology & evolution, 25(12), 705-712.

Strandström, H., Sandelin, K., &Oker-Blom, N. (1962). Inhibitory effect of Coxsackie virus, influenza virus, and interferon on Rous sarcoma virus. Virology, 16(4), 384-391.

Taylor, G., McGuire, G., & Sullivan, J. (2008). Individual claim loss reserving conditioned by case estimates. Annals of Actuarial Science, 3(1-2), 215-256.

Yang, G. Z., Hawkes, D. J., Rueckert, D., Noble, A., & Taylor, C. (Eds.). (2009). Medical Image Computing and Computer-Assisted Intervention– MICCAI 2009: 12th International Conference, London, UK, September 20-24, 2009, Proceedings (Vol. 5761). Springer.

Wu, L., Su, X., Shi, P., &Qiu, J. (2011). A new approach to stability analysis and stabilization of discrete-time TS fuzzy time-varying delay systems. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 41(1), 273-286.

APPENDIX A

claim.amt$claims

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75
1050	0	915	1	8272	300.0	479.4	1009.2	2300.0	5972.2 1

lowest : 5.86 12.21 45.00 50.02 60.00

highest: 176614.43 222640.00 273700.00 322429.91 354200.00

	X..claim.amt.claims
nobs	1.050000e+03
NAs	0.000000e+00
Minimum	5.860000e+00
Maximum	3.542000e+05
1. Quartile	1.009237e+03
3. Quartile	5.972245e+03
Mean	8.271663e+03

Median 2.300000e+03 Sum

8.685246e+06 SE Mean 7.388372e+02

LCL Mean	6.821896e+03
UCL Mean	9.721430e+03
Variance	5.731744e+08

Stdev 2.394106e+04 Skewness

8.683519e+00 Kurtosis 9.644714e+01

Normal

Fitting of the distribution ’ norm ’ by maximum likelihood

Parameters :

estimate Std. Error mean

7.872044 0.04376057 sd

1.418005 0.03094333

Loglikelihood: -1856.599 AIC: 3717.197 BIC: 3727.11

Correlation matrix:

	mean	sd
mean	1	0
sd	0	1

Log-normal

Fitting of the distribution ' lnorm ' by maximum likelihood

Parameters :

estimate Std. Error meanlog

2.0464112 0.005794125 sdlog 0.1877511

0.004096542

Loglikelihood: -1882.347 AIC: 3768.695 BIC: 3778.608

Correlation matrix:

	meanlog	sdlog
meanlog	1	0
sdlog	0	1

Gamma

Fitting of the distribution ' gamma ' by maximum likelihood

Parameters :

estimate Std. Error shape

29.74034 1.2907596 rate 3.77800

0.1653567

Loglikelihood: -1863.44 AIC: 3730.88 BIC: 3740.793 Correlation matrix:

shape rate

shape 1.0000000 0.9916065 rate

0.9916065 1.0000000

Weibull

Fitting of the distribution ’ weibull ’ by maximum likelihood

Parameters :

estimate Std. Error shape

5.799770 0.12987642 scale 8.465012

0.04770304

Loglikelihood: -1898.317 AIC: 3800.634 BIC: 3810.547

Correlation matrix:

shape scale

shape 1.0000000 0.3294148 scale

0.3294148 1.0000000