

# ACTUARIAL MODELING FOR INSURANCE CLAIM SEVERITY IN MOTOR COMPREHENSIVE POLICY USING INDUSTRIAL STATISTICAL DISTRIBUTIONS

OYUGI MARGARET ACHIENG  
BOSOM INSURANCE BROKERS LTD  
P.O.BOX 74547-00200  
NAIROBI, KENYA  
PHONE: +254 724 163 221  
[margaretactuary@yahoo.com](mailto:margaretactuary@yahoo.com)  
ICA REF NO 22  
TRACK: ASTIN

---

## ABSTRACT

General Insurance companies typically face two major problems when they want to use past or present claim amounts in forecasting future claim severity. First, they have to find an appropriate statistical distribution for their large volumes of claim amounts. Then test how well this statistical distribution fits their claims data. This paper sets forth a methodology for dealing with these problems.

The paper starts with introducing how actuarial modeling comes into practice in insurance claims data. Then an exclusive summary of what is entailed in the actuarial modeling process is outlined. The objective of the paper is to obtain an appropriate statistical distribution for the insurance claim amounts and to test how well the chosen statistical distribution fits the claims data.

The variable modeled is claim amounts from First Assurance Company limited, Kenya. The modeling process will establish one statistical distribution that could efficiently model the claim amounts, and then the goodness of fit test will be done mathematically using the Akaike's Information Criterion (A.I.C) and graphically using the Quantile-Quantile Plots (Q-Q plots)

Finally the study gives a summary, a conclusion and recommendations that can be used by insurance companies to improve their results concerning future claim inferences.

**Keywords:** Statistical distributions, Actuarial Modeling, Insurance Claim Amount

---

**ACRONYMS**

C.D.F	Cumulative Distribution Function
COTOR	Committee on Theory of Risks
K-S Test	Kolmogorov-Smirnov test
MATLAB	Matrix Laboratory
MLE	Maximum Likelihood Estimator
P.D.F	Probability Distribution Function
P-P Plot	Probability- Probability Plot
Q-Q Plot	Quantile – Quantile Plot
SPSS	Statistical Package for Social Scientists
TPO	Third Party Only

## KEY DEFINITIONS

Claim	- An accident in which an insured suffers damages which are potentially covered by their insurance contract
Frequency Distribution	-The distribution of the number of claims in one period
Loss	-The monetary amount of damage suffered by a policy-holder as a result of a loss event (claim)
Loss Distribution	-The probability distribution associated with either the loss or the amount paid due to the loss
M.L.E	-This is the value of the parameter that makes the observed data most likely to have occurred given the data generating process assumed to have produced the observations
Parameter	- A non-random Variable in a model that once chosen remains constant
Parametric	-The probabilities are known functions depending on a finite number of (real-valued) parameters
Parametric Estimation	-It assumes the distribution belongs to a known parametric family (e.g. Normal) and use the data to guide the selection of the form of the distribution and to calibrate its parameters
Premiums	-The amount of money the insurer needs to collect from the policy-holder in order to cover the expected losses, expenses and a provision for profit
Probability Distribution	-For a discrete random variable, it is a list of probabilities associated with each of its possible values
Random Variable -	A function that associates a unique numerical value with every outcome of an experiment
Reserve -	Provisions for future liabilities which indicate how much money should be set aside now to reasonably provide for future pay-outs
Severity	-This can either be the amount paid due to a loss or the size of a loss event

## Table of Contents

ABSTRACT.....	i
ACRONYMS.....	i
KEY DEFINITIONS.....	iii
1.0 INTRODUCTION.....	2
1.1 Back Ground of the Study.....	2
1.2 Executive Summary.....	3
1.3 Problem Statement.....	3
1.4 Justification of the Study.....	4
1.5 Objective of the Study.....	4
1.5.1 General Objective.....	4
1.5.2 Specific Objective.....	4
1.6 Hypothesis.....	4
1.7 Variables.....	4
2.0 LITERATURE REVIEW.....	5
2.1 Actuarial Modeling.....	5
3.0 METHODOLOGY.....	7
3.1 Scope of Data.....	7
3.2 Actuarial Modeling Process.....	7
3.2.1 Selecting a Model Family.....	7
3.2.2 Estimation of Parameters.....	8
3.2.3 Criteria for Choosing One Distribution for the Claims Data.....	9
3.2.4 Checking Model Fit.....	9
3.2.5 Revising the Model If Necessary.....	10
3.3 Data Processing.....	10

4.0 COMPUTATION AND INTERPRETATION .....	11
4.1 Summary Statistics .....	11
4.2 Interpretation Of Histograms.....	11
4.2.1 Histogram Of The Claim Amounts .....	11
4.2.2 Histogram Of The Log Of Claim Amounts .....	12
4.2.3 Histogram Of Triple Log Of The Claim Amounts.....	12
4.3 Maximum Likelihood Estimates.....	13
4.4 The Log-Likelihoods .....	13
4.5 Probability Density Functions .....	14
4.5.1 Graphical Implication.....	15
4.6 Goodness of Fit Test.....	15
4.6.1 The Akaike's Information Criteria Interpretation .....	15
5.0 SUMMARY, CONCLUSION AND RECOMMENDATION.....	19
5.1 Summary .....	19
5.2 Conclusion.....	20
5.3 Limitations and Problems Encountered .....	21
5.4 Recommendations.....	21
APPENDIX I .....	23
REFERENCES.....	25

## 1.0 INTRODUCTION

### 1.1 Back Ground of the Study

Actuarial science consists of building and analyzing mathematical models to describe the processes by which money flows into and out of an insurance company. It is a combination of diverse quantitative skills to enable one “make financial sense of the future”(www.actuaries.org )

General insurance is perhaps the fastest growing area for actuaries. It includes Health insurance, personal/property Insurance such as home and motor insurance as well as large commercial risks and liability insurance. (Boland, 2006) For any country, the insurance industry is of great importance because it is a form of economic remediation. It provides a means of reducing financial loss due to the consequences of risks by spreading or pooling the risk over a large number of people. Insurance being a data-driven industry with the main cash out-flow being claim payments, Insurance companies employ large numbers of analysts, including actuaries, to understand the claims data.

Claim actuaries are not interested in the occurrence of the claims themselves but rather in the consequences of its random out-come. That is, they are concern with the amount the insurance company will have to pay than with the particular circumstances which give rise to the claim numbers. (Denuit, marechal, pitrebois & walin, 2007) The general insurance actuary needs to have an understanding of the various models for the risk consisting of the total or aggregate amount of claims payable by an insurance company over a fixed period of time.

Insurance data contains relatively large claim amounts, which may be infrequent, hence there is need to find and use statistical distributions with relatively heavy tails and highly skewed like the exponential, gamma, pareto, weibull and log-normal. (Boland, 2006)

These models are informative to the company and they enable it make decisions on amongst other things: premium loading, expected profits, reserves necessary to ensure (with high probability) profitability and the impact of reinsurance and deductibles.

In view of the economic importance of motor comprehensive insurance in developing countries, many attempts have been made in the actuarial literature to find a probabilistic model for the distribution of the claim amounts reported by insured drivers. (Denuit, marechal, pitrebois & walin, 2007)

Although the empirical distribution functions can be useful tools in understanding claims data for motor policy, there is always a desire to “fit” a probability distribution with reasonably tractable mathematical properties to the claims data. Therefore this paper involves the steps taken in actuarial modeling to find a suitable probability distribution for the claims data and testing for the goodness of fit of the supposed distribution.

Finally, constructing interpretable models for 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 & Shaw, 2000)

## 1.2 Executive Summary

The paper deals with industrial statistical distributions used in actuarial analysis of insurance claim amounts and more specifically in motor policy

First, the variables and the hypothesis which govern the modeling process are introduced. The description of the sampled probability distributions used in modeling insurance claim amounts follows. The description made on these distributions entails the nature and property of the distribution, its PDF, its histogram plots, mean, variance, skewness and kurtosis. The distributions used in this study include: The Exponential, Gamma, Log-normal and Weibull.

From there, steps followed in the actuarial modeling process follow as below;

- First, a model family is selected from which one distribution is to emerge as the best fitted distribution for the claims data
- Parameters are estimated using the maximum likelihood estimation method. After which the log likelihood estimates are computed.
- Testing of the goodness of fit is then done using the A.I.C and the Q-Q plots.
- If the selected model does not fit the claims data, another distribution is to be chosen and the process started again.

This modeling process was aided by the MATLAB soft ware. The method that was used to test the goodness of fit was the Quantile-Quantile (Q-Q) Plots as its advantages will be sited in section 3. This was confirmed by the Akaike's information criterion (A.I.C). The Q-Q plots were used to provide more insight into the nature of the distributions after which a conclusion was made.

## 1.3 Problem Statement

For claim actuaries, claim modeling is very crucial since a good understanding and Interpretation of loss distribution is the back-bone of all the decisions made in the Insurance industry regarding, premium loading, expected profits, reserves necessary to Ensure profitability and the impact of re-insurance. (Boland, 2006)

At First Assurance Company limited, motor policy records an over-whelming number of claims in any given period. For this same reason the study dealt with motor Comprehensive claims because the claim amounts are larger compared to the third party claims, which makes motor comprehensive more sensitive to the insurance fund managers. Choosing the most suitable loss distribution therefore is important especially for the large amounts of claims settled by insurance companies.

An understanding of the probability and statistical distribution is vital for the general Insurance actuary to be able to summarize and model the large amounts of claims data and give timely outcomes. These distributions are "must-have" tools for any actuarial assessment, that's why the study went further to get their specific descriptions to emphasize their different properties and how they are useful in insurance claims Data. (Raz & Shaw, 2000)

There is need to cite the different methods used to test for the goodness of fit of the Supposed probability distribution chosen for the claims data. This is paramount because it is only by choosing the best method amongst the sampled methods that the actuaries attain accuracy and consistency in making financial sense of the future.

#### **1.4 Justification of the Study.**

This study was geared to demonstrate the importance of understanding probability and statistical distributions for use by general insurance actuaries. They also need to have an understanding of various models for the risk consisting of the aggregate claims payable by an insurance company over a period of time. This information enables the company make important decisions regarding premium loading, reserves necessary to ensure high profitability, expected profits and the impact of re-insurance.

The paper analyzes 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. By using statistical distributions to model claim severity, one has a much added insight into the complexity of the large amounts of claim that may often be hidden in a huge amount of data.

#### **1.5 Objective of the Study**

##### **1.5.1 General Objective**

The general objective was to test for an appropriate statistical distribution for the insurance claim amounts and to test how well the chosen statistical distribution fits the claims data.

##### **1.5.2 Specific Objective**

1. Testing for the appropriate statistical distribution for the claim amounts
2. Test the goodness of fit of the chosen statistical distribution.

#### **1.6 Hypothesis**

##### **Null Hypothesis**

1. HO: The exponential distribution provides the correct statistical model for the claim amounts.
2. HO: The gamma distribution provides the correct statistical model for the claim amounts.
3. HO: The log-normal distribution provides the correct statistical model for the claim amounts.
4. HO: The weibull distribution provides the correct statistical model for the claim amounts.

#### **1.7 Variables**

The random variables used in the study were the claim amounts associated with the occurrence of automobile accidents reported to First Assurance Company Limited on motor comprehensive covers .The random variables in this study were independent.

## 2.0 LITERATURE REVIEW

This section presents a comprehensive overview on actuarial modeling as applied in insurance claim severity using probability distributions as viewed and present by other research studies. This section is based on published books, journals, reports and opinions prior collected. It is hoped that by presenting this, the study and the scope is widened and thus revealing the information gap.

### 2.1 Actuarial Modeling

In the late 1980's and early 1990's there was a distinct effort for actuaries to combine financial theory with stochastic methods into their established models (D'arcy, 1989). Today, the profession, both in practice and in the education syllabi of many actuarial organizations combines life tables, claim models, loss models, stochastic methods and financial theory. (Feldblum 2001). Therefore currently there is a hand full of articles written about actuarial modeling for insurance claim severity from different angles.

Wright (2005) used the steps involved in actuarial modeling to fit models to 490 claim amounts drawn from 7 consecutive years. He fitted analytic loss distributions using maximum likelihood estimation for each of the 7 years. He used P-P plots and the kolmogorov-smirnov tests (K-S test) to assess the absolute quality of fit. For this study, the Q-Q plots were used to assess the quality of fit as its advantages will be sited in section 3 below. Wright used several statistical distributions which included the inverse Pareto, Pareto, burr, Pearson VI, inverse burr log-normal and restricted benktander 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 claim amounts but the whole claims data set.

Meyers (2005) used actuarial modeling to fit a statistical distribution to 250 claims. The statistical distributions he tested were log-normal, gamma and weibull distribution. Meyers used maximum likelihood estimation to fit the distributions, an idea that was extended to this research, only that he based his methodology on the 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 compute A.I.C so as to choose the probability distribution that fits the claims data best.

Renshaw (2004) also brought an approach useful in actuarial modeling in claim amounts for non-life insurance basing on the concept of quasi-likelihood and extended quasi likelihood. He used 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 rensshaw in his modeling procedure goes in line with the procedures involved in all actuarial modeling process.

Guiahi (2000) presented a paper on issues and methodologies for fitting alternative statistical distributions to samples of insurance data. His illustration was based on a sample of data with log-normal as the underlying distribution. Like for this research he used the method of maximum likelihood to estimate model parameters and also his criteria for comparing which probability distribution fits the data set best was based upon the value of akaike's information criteria, AIC.

The AIC criterion is defined by:

$$\text{AIC} = -2(\text{maximized log-likelihood}) + 2(\text{no. of parameters estimated})$$

In A.I.C when two models are being compared, the model with the smaller AIC value is the more desirable one.

Fiete (2005) also presented actuarial modeling on his COTOR answer using gamma, lognormal, weibull and Pareto. He used similar steps of actuarial modeling 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 rely on a single number from the log-likelihood to describe goodness of fit.

Therefore, this study was based on how insurance claims severity can be fitted to a statistical distribution, assuming no zero claim. The study brought forth the steps involved in the modeling process, hence explaining the criteria and the methodology discussed in the above literature.

### **3.0 METHODOLOGY**

This section presents the methodology which was used in the study. It explains in details the steps that were encountered in the modeling process which includes the data processing and data analysis that were used.

#### **3.1 Scope of Data**

Secondary data was used from First Assurance Company limited (K) regarding their motor comprehensive policy (June 2006-June 2007). Three assumptions were made on the data before use, these included:

1. All the claims came from the same distribution (they were independent and identically distributed).
2. There was no zero claim for any motor vehicle registered under the policy
3. All future claims were to be generated from the same distribution

#### **3.2 Actuarial Modeling Process**

This section will describe the steps that were followed in fitting a statistical distribution to the claim severity, that is, the steps that were taken in the actuarial modeling process.

(Kaishev, 2001) These steps are:

- Selecting a model family
- Estimating model parameters
- Specification of the criteria to choose one model from the family of distributions
- Check model fit
- Revise model fit if necessary

##### **3.2.1 Selecting a Model Family**

This is the first step in the modeling process. Here 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 extent subjective.

For this study the choice of the sample distributions was with regard to:

- Prior knowledge and experience in curve fitting
- 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: gamma, exponential, log-normal and weibull. Still in this step, it was necessary to do some descriptive analysis of the data to obtain its salient features. This involved finding the mean, median, mode, standard deviation, skewness and kurtosis. This was done using SPSS. Histograms were plotted using SPSS to show the graphical representation of the data.

## PROPERTIES OF THE SAMPLED PROBABILITY DISTRIBUTIONS (kaishev, 2001)

### i. GAMMA DISTRIBUTION ( $\alpha, \lambda$ )

P.d.f:  $f(X) = \alpha\lambda / \Gamma(\lambda) X^{\lambda-1} e^{-\alpha X}$

If  $\lambda$  is an integer then,

C.D.F:  $F(X) = 1 - \sum_{j=0}^{\infty} e^{-\alpha X} (\alpha X)^j / j!$

Where  $X > 0$  and  $\alpha, \lambda > 0$

### ii. EXPONENTIAL DISTRIBUTION ( $\alpha$ )

This is the gamma distribution with  $\lambda = 1$  such that:

P.d.f:  $f(X) = \alpha e^{-\alpha X}$

C.D.F:  $F(X) = 1 - e^{-\alpha X}$  where  $X > 0$  and  $\alpha > 0$

### iv. LOG-NORMAL DISTRIBUTION ( $\mu, \delta$ )

p.d.f:  $f(X) = 1/\delta X (2\pi)^{-0.5} \exp[-(\log X - \mu)^2 / 2\delta^2]$

C.D.F:  $F(X) = \Phi[(\log X - \mu) / \delta]$

Where  $\Phi$  is the  $N(0, 1)$  c.d.f and  $X > 0, \delta > 0$  while  $\mu$  is a real number

### v. WEIBULL DISTRIBUTION ( $C, \gamma$ )

p.d.f:  $f(X) = C\gamma X^{\gamma-1} \exp\{-CX^\gamma\}$

C.D.F:  $F(X) = 1 - \exp\{-CX^\gamma\}$

Where  $X > 0, C, \gamma > 0$

## 3.2.2 Estimation of Parameters

It involved estimation of the parameter(s) for each of the above sampled probability distributions using the claims data. Once the parameter(s) of a given distribution were estimated, then a fitted distribution was available for further analysis. The maximum likelihood estimation method was used to estimate the parameters

### 3.2.2.1 Maximum Likelihood Estimator

The Maximum Likelihood estimates were used because they have several desirable properties which include: consistency, efficiency, asymptotic normality and invariance. The advantage of using Maximum Likelihood Estimation is that it fully uses all the information about the parameters contained in the data and that it is highly flexible. (Deniut, 2007)

Let  $X_i$  be the  $i$ th claim amount, where  $1 \leq i \leq n$ .

$n$  is the number of claims in the data set

$L$  is the likelihood function

$\theta$  is the parameter

$f(x)$  is the probability distribution function of a specific distribution

The likelihood function of the claims data is given by:

$$L = \prod f(X) \dots\dots\dots(1)$$

To get maximum likelihood, differentiate equation (1) above

$$M.L.E = dL/d\theta \dots\dots\dots(2)$$

Therefore to solve the value of the parameter, equate equation (2) to zero:

$$dL/d\theta = 0 \dots\dots\dots(3)$$

However all that was needed to derive the maximum likelihood estimators was to formulate statistical models in the form of a likelihood function as a probability of getting the data at hand. With the help of MATLAB statistical package this likelihood estimates were derived for each of the four distributions according to the data set.

### 3.2.3 Criteria for Choosing One Distribution for the Claims Data

Out of the four probability distributions which were sampled, only one which appeared to fit the data set more than the rest had to be chosen. Since the parameters were obtained using maximum likelihood, the criteria for choosing one distribution out of the four was also based on the values of the estimated maximum likelihood estimates, the larger the likelihood, the better the model. ( Fiete, 2005)

### 3.2.4 Checking Model Fit

It was assumed that no model in the set of models considered was true; hence, selection of a best approximating model was the main goal. (Anderson & Burnham, 2004) Just because a distribution got the highest log-likelihood out of the four distributions, this was not sufficient evidence to show that it is the right distribution for the claims data set. Therefore an assessment was made on how good this distribution fitted the claims data using The Q-Q Plots and the A.I.C. (Level of significance:  $\alpha=0.01$ )

#### 3.2.4.1 The Quantile-Quantile (Q-Q) Plots.

The Quantile-Quantile (Q-Q) plots are graphical techniques used to check whether or not a sampled data set could have come from some specific target distribution i.e. to determine how well a theoretical distribution models the set of sampled data provided. This study used the Q-Q plots to check for goodness of fit of the chosen distribution to the sampled claim severity. The Q-Q plots were chosen because of their multiple functions while analyzing data sets and also because of their advantages as sited below.

(<http://maths.murdoch.edu.au/qq/qqnormal>)

#### Using Q-Q Plot to Check Goodness-Of-Fit of a Probability Distribution

The first Q stands for the quantiles of the sampled data set and the second Q stands for the quantile of the distribution being checked whether the data fits.

In this case, The Q-Q plot is a plot of the target population quantile(y) against the respective sample quantile (x) (<http://stats.gla.ac.uk/steps/grossary/qqplot>)

If the sample data follows the distribution suspected, then the quantiles from the sample data would lie close to where they might be expected and the points on the plot would straggle about the line  $y=x$

Theoretically, in order to calculate the quantiles of the distribution, this target distribution must first be specified, i.e. its population mean and standard deviation but in practice, the sample estimates are used, therefore sample mean and standard deviation of the distribution were estimated to be same as the ones of the sampled data set. MATLAB soft-ware was used to develop the Q-Q plot for measuring the goodness of fit of the chosen probability distributions.

### Advantages of Q-Q Plots

1. The sample sizes do not need to be equal
2. Many distributional aspects can be simultaneously tested for example shifts in locations, shifts from scale, changes in symmetry and the presence of outliers. This is important because if the claim amounts and the claim counts of the data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line. ([http://en.wikipedia.org/wiki/Q-Q plot](http://en.wikipedia.org/wiki/Q-Q_plot))

#### 3.2.4.2 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 the candidate models considered.

The AIC criterion is defined by:

$$AIC = -2(\text{maximized log-likelihood}) + 2(\text{no. of parameters estimated})$$

For this research, the A.I.C was used when testing for the goodness of fit after computing the likelihood function.

#### 3.2.5 Revising the Model If Necessary

This was the final step in the modeling process. If the Q-Q plot between the quantiles of the claim amounts (y) against the respective sample quantile (x) of the chosen distribution had not lied close to where it was expected such that the points on the plot were not along the line  $y=x$  then another probability distribution would have been chosen. This would have meant the modeling process was to be started again so as to choose another probability distribution to fit the claim amounts, this process would have been repeated until an appropriate statistical distribution was fitted to the claims amount. Luckily the model did not need to be revised as will be evident in section 4

### 3.3 Data Processing

The study mostly involved the use of computer statistical packages to perform the tests and to plot the graphs. MS-Excel was used for representation of the findings, MATLAB was the main toolbox used for the parameter estimation, computation of likelihood values, Probability distribution functions (p.d.f) and in plotting the Q-Q plots. SPSS was used in getting the summary statistics and plotting the histograms. The MATLAB codes are provided in the appendix.

## 4.0 COMPUTATION AND INTERPRETATION

This section gives a practical illustration of the methodology given in section three. Here, the computations that were involved in the actuarial modeling process during the study will be encountered and the results displayed both numerically and graphically. The logic behind the selection criteria will be established basing on the results. And finally a conclusion will be made on the findings of this study so as to know which statistical distribution best fits the claims data.

### 4.1 Summary Statistics

The actuarial modeling process started with the computation of the summary statistics of the claim amounts. These are presented in table 4.1 below. This summary was necessary in pointing out the salient features of the data. From table 4.1, most of the statistics computed depended on the sample size, N. The data is heavily skewed which was an important feature in selecting the family of distributions to use.

**Table 4.1 Statistics Summary**

N	Mean	Median	Mode	Standard Dev.	Skewness	Kurtosis	Sum
125	148002.8	41978	11020	315150.8	4.1	19.8	18500349

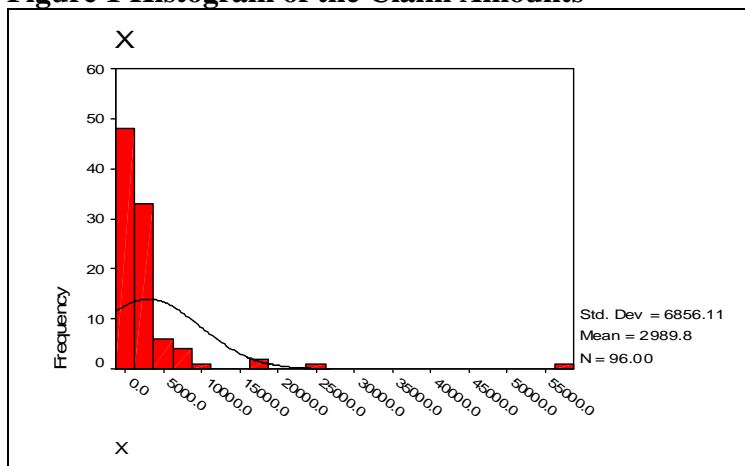
*Source: Computed by author*

In the same concept, the histogram of the claims data is plotted to identify the shape of the data

## 4.2 INTERPRETATION OF HISTOGRAMS

### 4.2.1 Histogram of the Claim Amounts

**Figure 1 Histogram of the Claim Amounts**



*Source: Plotted by Author*

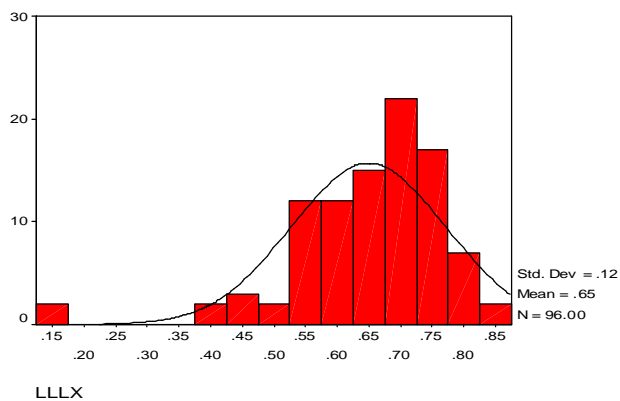
The bars of this histogram are not well pronounced and it has small widths such that it may not have been practical to begin fitting a distribution to the data in its original form. The histogram in figure 1 also has a normal curve superimposed on it. This curve shows the skewness of the claims data. From the diagram, it can be seen that the original claims data has a heavy right-hand

tail. This was interpreted to mean that the claims data had few claim amounts with very high values while most of the claim amounts were of low values.

#### 4.2.2 Histogram of the Log of Claim Amounts

In an attempt to reduce the skewness of the data, log transformation was used. With this, the log of the claims data was computed using MATLAB. The histogram of the log transformation was plotted as shown in figure 2 below. The bars of this histogram are well pronounced and the normal curve is less skewed. This log transformation could have been used for modeling the claims data before having a look at the histogram of the triple log of the data

**Figure 2 Histogram of the Log of Claims Amounts**

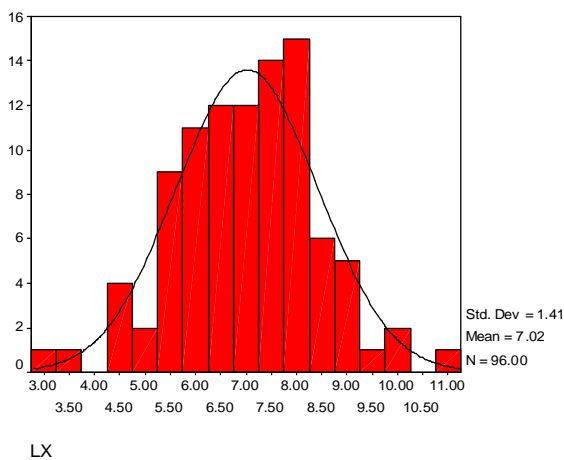


*Source: Plotted by author*

#### 4.2.3 Histogram of Triple Log of The Claim Amounts

In further attempt to reduce the skewness of the data, the triple log was obtained. This yielded the histogram in figure 3 below. The histogram of the triple log of the claims data has bars with even broader widths and the normal curve is relatively normally distributed.

**Figure 3 Histogram of the Triple Log of the Claims**



*Source: Plotted by Author*

Therefore, the triple log of the claims data was taken and maximum likelihood fits for the, four distributions begun. From figure 3, it is clear that the triple log of the claims data could be used to fit the statistical distributions. As a result, the parameter estimates for the four distributions were computed at this point.

### 4.3 Maximum Likelihood Estimates

Given any model, there exists a great deal of theories for making estimates of the model parameters based on the empirical data, in this case, the triple log of the claims data was to be used to compute the maximum likelihood estimates of the four sampled distributions. The first step in fitting a model to a claims data is finding the parameter estimates of the particular statistical distribution. When the parameters of any distribution have been obtained using the triple log of the claims data, then literally, the statistical distribution has been fitted to the claims data. With this regard, table 4.2 below gives the parameters of the four distributions having been fitted to the claims data. Where  $\mu$  was taken to be the value of the first parameter and  $\delta$  was taken to be the second parameter value. Table 4.2 also shows the confidence interval within which the parameters lie at 99% level of confidence. The parameters obtained were then used in the computation of the log-likelihoods of the four distributions.

**Table 4.2 Summary of the Estimation Results**

Distributions	Log-Likelihood	A.I.C	$\mu$	$\delta$	Parameters Confidence Interval
Exponential	-109.6749	221.3500	0.8609		$0.6913 \leq \mu \leq 1.0972$
Lognormal	182.0632	-360.1300	-0.1517	0.0628	$-0.1664 \leq \mu \leq -0.3677$ $0.0539 \leq \delta \leq 0.0749$
Gamma	-6460.9000	12925.8000	255.8749	0.0034	$164.6777 \leq \mu \leq 347.0722$ $0.0022 \leq \delta \leq 0.0046$
Weibull	174.8366	-345.6700	0.8865	16.6898	$0.8736 \leq \mu \leq 0.8996$ $14.0754 \leq \delta \leq 19.7897$

*Source: Computed by Author*

### 4.4 The Log-Likelihoods

The log-likelihood theory provides rigorous and omnibus inference methods if the model is given, that is, after the parameters of a distribution have been obtained. The log-likelihoods form the basis of the selection of the distribution that fits the data. It was the first tool used in the primary elimination stage. Table 4.2 shows the computed log-likelihoods. MATLAB was used to obtain the values. From the tabulated statistics, the log-normal distribution, with log-likelihood value of 182.0632 has the highest log-likelihood value amongst the sampled distributions; hence the log-normal distribution was the best fit of the four distributions. The weibull distribution

followed with log-likelihood of 174.8366; therefore it was relatively not a bad fit. The log likelihoods of the exponential and the gamma distributions are -109.6749 and -6460.9 respectively. These are far from the log-likelihood of the log-normal distribution.

Therefore the exponential and the gamma distribution could not provide a good fit to the claims data. On the other hand, in the quest to reduce the skewness of the data, it would have been possible to compute a quadruple log of the data since the triple logs were still positives. But that was as far as the log transformation could have been applied on to the claims data set since most values of the triple log were less than one, actually most of the values were nearing zero such that the quadruple log values may have ended up as negatives!

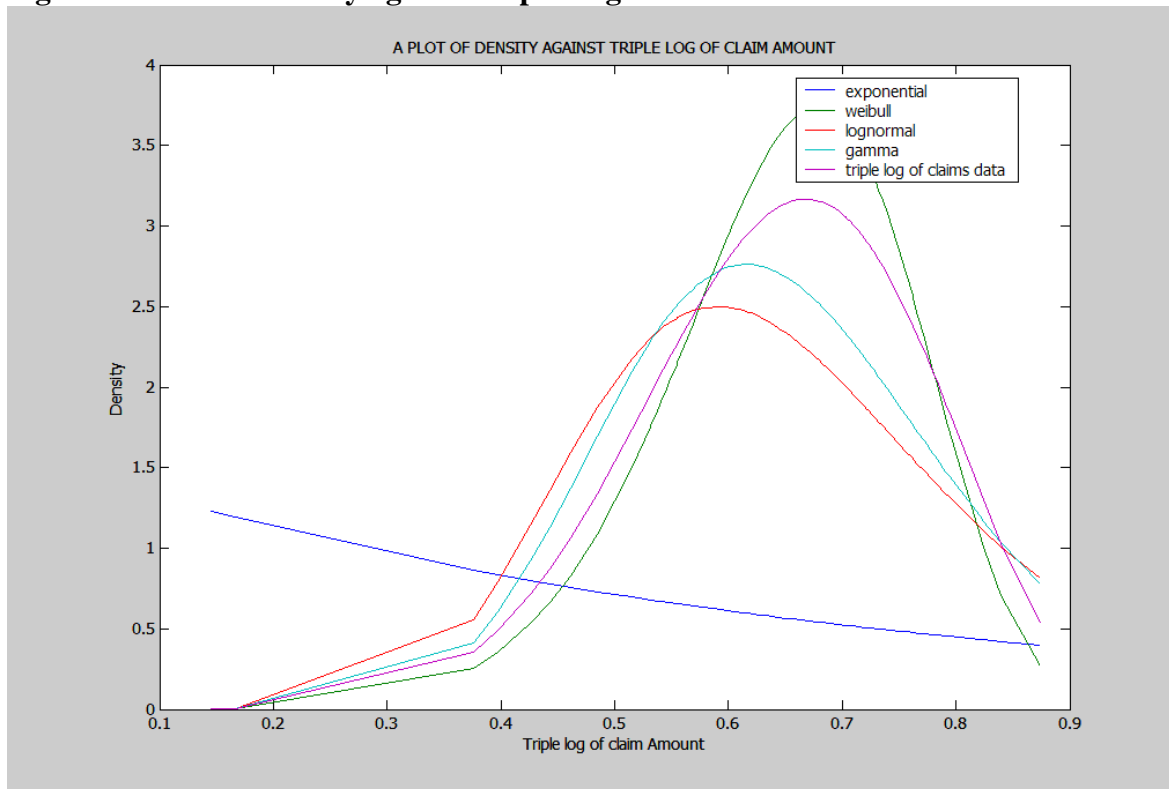
With this, the log-normal distribution was selected as the statistical distribution that gave a relatively good fit for the claims data as compared to the weibull, the exponential and the gamma distribution.

#### 4.5 Probability Density Functions

A plot of the Probability Density Functions (p.d.f) of the each of the four distributions as compared to the plot of the triple log of the claims data was also used to give a clear visual indication of how the distributions follow the trend of the claims data.

This is shown in below.

**Figure 4 A Plot of Density against Triple Log of First Assurance Claim Amounts**



Source: Plotted by Author

### 4.5.1 Graphical Implication

Figure 4 shows the plot of the p.d.f for the four statistical distributions in comparison to the plot of the triple log of the claims data. From the diagram, the gamma plot cannot provide a good fit to the claims data as its tail does not in any way correspond with the one observed from the triple log of the claims data. The exponential and the weibull plots are definitely not good fits for the data as they have much lower density values. The log-normal plot was relatively concluded to have the best fit.

### 4.6 Goodness of Fit Test

This section is interested in the post Model selection fit to affirm the selected model.

The central problem in analysis is which model to use for making inferences from the data- this is the model selection problem. As highlighted in section 3, just because the log-normal distribution emerged as having the highest log-likelihood value amongst the four distributions could not have meant that it was the best statistical distribution to model the claims data.

With this argument, it was necessary to carry out a goodness of fit test in order to select a statistical distribution that best fits the data. In this study, the goodness of fit test was done both mathematically and affirmed graphically. This called for computation of the Akaike's Information Criterion (A.I.C) and plotting of the Q-Q plots. Since it was necessary to ascertain how well the distribution fits the data, the A.I.C was computed as presented in table 4.2 above. Graphically, the goodness of fit was established using the Q-Q plots of the four distributions as fitted on to the triple log of the claims data.

#### 4.6.1 The Akaike's Information Criteria Interpretation

This criterion was derived based on the concept that the truth is very complex and that no "true model" exists for any sampled data set. Therefore given the four statistical distributions, it was possible to estimate which distribution was closest to the unknown true model.

The formulae for AIC was cited in section 3, AIC is used in selecting the model for making inferences for the claims data. In this study the AIC was manually computed after the values of the log likelihoods for the individual distributions obtained. These results are tabulated in table 4.2. In AIC one should select the model with the smallest value of AIC. With this regard the log-normal distribution had the smallest value of AIC. It was therefore concluded that the log-normal distribution was the best fitted distribution among the chosen distributions for the claims data since it had the smallest value of AIC. That is, the log-normal distribution was estimated to be the closest to the unknown true distribution among the four candidate distributions considered.

Before drawing a final conclusion with regard to how 'good' the log-normal distribution fits the claims data, it is prudent to consider the graphical affirmation of the 'goodness of fit'. In this order, hypothesis of the four distributions were formulated and their Q-Q plots plotted and interpreted as follows.

#### 4.6.2 The Quantile-Quantile (Q-Q) Plots Interpretation

The Q-Q plots for each of the four distributions were constructed using MATLAB, and the selection criterion was based on the specific hypothesis for each specified model.

That is;

H0: The statistical distribution provides the correct statistical model for the claims data

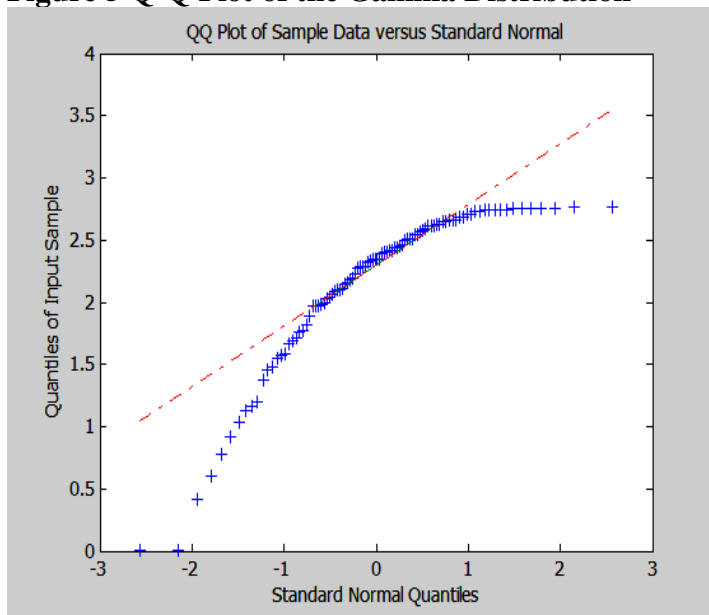
H1: The statistical distribution does not provide the correct statistical model for the Claims data

##### 4.6.2.1 The Gamma Distribution

H0: The gamma distribution provides the correct statistical model for the claims data

H1: The gamma distribution does not provide the correct statistical model

**Figure 5 Q-Q Plot of the Gamma Distribution**



*Source: Plotted by Author*

This plot depicts the “worst” fit because the tails of the diagram vary far away from the reference line while there are very few data points falling onto or around the reference line.

#### **Conclusion**

The null hypothesis was therefore rejected and a conclusion was made that at 99% level of confidence, the gamma distribution does not provide the correct statistical model for the claims data

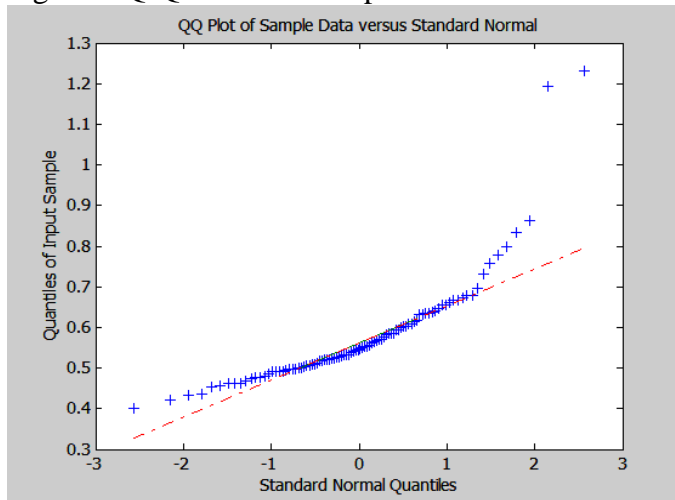
##### 4.6.2.2 The Exponential Distribution

H0: The exponential distribution provides the correct statistical model for the claims data

H1: The exponential distribution does not provide the correct statistical model

Figure 6 below shows the Q-Q plot. This plot shows that there is random variation in the data set. Unlike the Q-Q plot of the gamma distribution, several points plot on the reference line though the plot depicts that the exponential distribution has heavy tails on both ends as most points not falling on the reference line are at the extreme ends of the reference line and not in the middle

Figure 6 Q-Q Plot of the Exponential Distribution



Source: Plotted by Author

### Conclusion

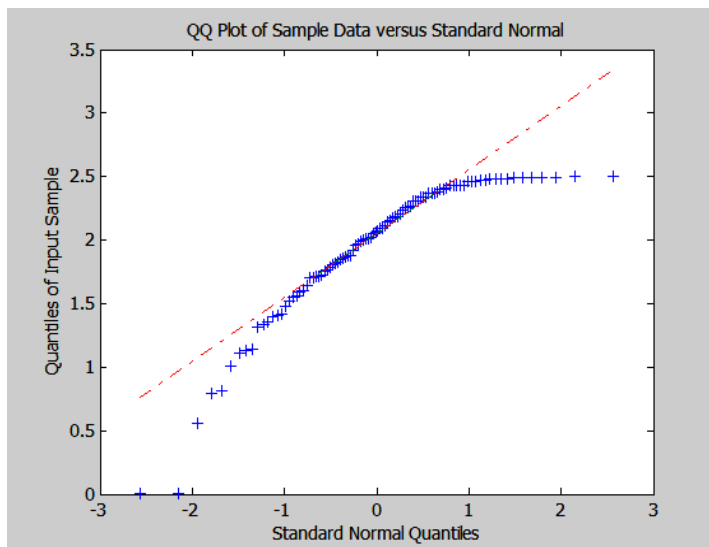
The null hypothesis was therefore rejected and a conclusion was made that at 99% level of confidence, the exponential distribution does not provide the correct statistical model for the claims data.

#### 4.6.2.3 The Weibull Distribution

H<sub>0</sub>: The weibull distribution provides the correct statistical model for the claims data

H<sub>1</sub>: The weibull distribution does not provide the correct statistical model

Figure 7 Q-Q plot of the Weibull Distribution



Source: Plotted by Author

This plot shows that there is less random variation in the data sets. The Q-Q plot indicates solid agreement in the body of the distribution as most of the points plotted fall on the reference line

and only a few points at both ends of the reference line fall far from the line. This plot shows that the claims data has heavy tails on both ends.

### Conclusion

The null hypothesis was therefore rejected and a conclusion was made that at 99% level of confidence, the weibull distribution does not provide the correct statistical model for the claims data.

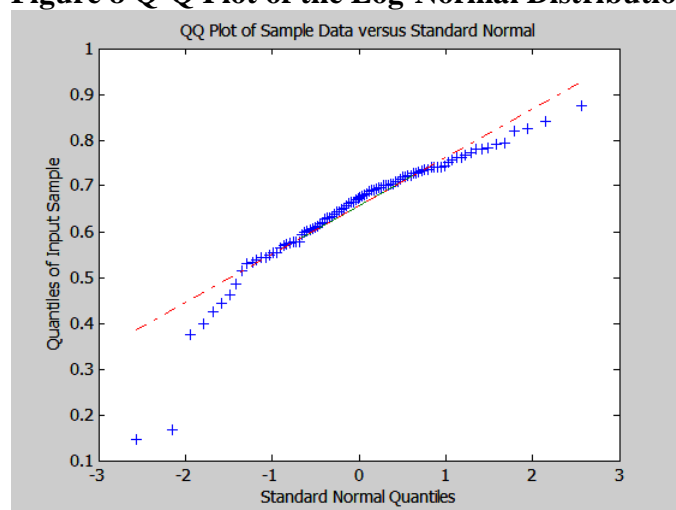
#### 4.6.2.4 The Log-Normal Distribution

H0: The log-normal distribution provides the correct statistical model for the claims data

H1: The log-normal distribution does not provide the correct statistical model

As seen in figure 8 below this plot shows the best fit not only as compare to the previous Q-Q plots but it is evident that there is solid agreement in the body of the distribution with very minimal fluctuations at the tails. From the plot, almost all the points fall on the reference line and for the few that are not on the line; they are very close to the line.

**Figure 8 Q-Q Plot of the Log-Normal Distribution**



*Source: Plotted by Author*

### Conclusion

Therefore the null hypothesis was not rejected and a conclusion was made that at 99% level of confidence, the log-normal distribution provides the correct statistical model for the claims data.

Hence, this brought the modeling process to an end as it was confirmed, affirmed and reaffirmed that in a family of distribution consisting of the exponential, the gamma, the lognormal and the weibull distribution, the log-normal distribution would be the best statistical distribution to model the claim amounts of First Assurance Company limited at 99% level of confidence.

Finally, this conclusion meant that the log-normal distribution could effectively be relied upon to model the claim severity such that future claim amounts could be forecasted for decision making by the company managers regarding future premium rates, future reserves and expected profits as sighted in section 1

## 5.0 SUMMARY, CONCLUSION AND RECOMMENDATION

### 5.1 Summary

The major objective of this study was to come up with one statistical distribution for the insurance claims data and to test how well this statistical distribution fits the claims data so that this distribution can be used for modeling the claim amounts. In a very important sense, the study was not concern with the steps of modeling the data; instead, the study tries to model the information in the data to fit a particular distribution.

Therefore, an attempt was made to establish an appropriate statistical distribution that best fits the insurance claims data using numerical computations and graphical implications using MATLAB and SPSS software's. From the analysis carried out coupled with the results displayed in table 4.2 it is revealed that the insurance claims data for First Assurance company can best be modeled using the log-normal distribution.

According to table 4.2, the log-normal distribution has the highest log-likelihood function of 182.0632, which implies that among the chosen statistical distributions it stands a better chance in providing a good fit for the claims data. The log-normal distribution is then followed by the weibull distribution which has a log-likelihood of 174.8366, and then the exponential distribution follows with a log-likelihood of -109.6749. Finally the gamma distribution trails with a log-likelihood of -6460.9. The results in this log-likelihood column imply that the gamma distribution is far from giving an appropriate model for the claims data, followed by the exponential distribution. The weibull distribution can be considered in some cases but in this case the log-normal distribution is present displaying even a higher log-likelihood function. In figure 4 the P.d.f of the four distributions have been plotted in comparison with the plot of the triple log of the claims amounts. This figure illustrates that gamma and the exponential distributions plots have shapes which don't much the shape of the claims amounts. The weibull distribution does not display a shape far from the one displayed by the claims data however the log-normal distribution's shape matches the shape of the claims amount indicating that it can actually be used to model the claims amount. To test whether the log-normal distribution provides a good fit to the claims data, the A.I.C is computed in the third column of table 4.2. With regard to the A.I.C, the lognormal distribution has the least A.I.C value of -360.13 indicating that it provides a good fit for the claims data. This log-normal value is followed closely by the weibull's A.I.C value of -345.67 then the exponential distribution has a high A.I.C value of 221.35 showing that it would not be regarded as a good fit to the claims data in this case. The A.I.C value of the gamma distribution is even higher, with a value of 12,925.80 this can be considered a bad fit but this would still mean the gamma distribution cannot in any case model this claims amounts.

Finally, Q-Q plots for each of the four statistical distributions was plotted on figure 5 to figure 8 to graphically re-affirm the goodness of fit test computed by the A.I.C. Figure 8 shows that the Q-Q plot of the log-normal distribution provides the best fit to the claims data as almost all points plot on the reference line and only a few points plot near the line. The Q-Q plot for the weibull distribution almost indicates a good fit only for a number of points on the extreme ends

plotting far from the reference line while Q-Q plots of the exponential and the gamma distributions indicate that these two distributions could not be used to model the claims amounts to provide a reliable model for forecasting.

## 5.2 Conclusion

After carrying out each step in the actuarial modeling processes with diligence and accuracy, the above summary clearly indicates that the log-normal distribution would provide a good fit to the claims data. Therefore if First Assurance Company limited is to model the motor comprehensive policy claim amounts experienced in June 2006- June 2007, the appropriate statistical distribution to use to yield reliable claim forecasts would be the log-normal distribution.

Having tested the goodness of fit of the log-normal distribution both graphically using the Q-Q plots and mathematically using the A.I.C value, it is evident that the steps followed in the actuarial modeling process is capable of yielding reliable results that can be used to make inferences useful for decision making in the general insurance industry.

This 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 log-normal, the exponential, the gamma and the weibull. Of which the log-normal distribution was proved to be capable of providing a good fit for the claims amounts. From the modeling carried out using the MATLAB software, 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.

However the results of this study are dependent on a number of factors outside the modeling process. This means that the insurance company has to acknowledge these factors before using the results of this study in making future inferences.

These factors include:

- Future claim amounts to be experienced by the insurance company may depend on an increase in recklessness of drivers
- Change of road traffic policies may reduce future claims amounts. This maybe as a result of more traffic police-men being employed along the roads
- Increase in awareness of the general public on the advantages of comprehensive insurance may lead to more claims amounts as more people would be moving from Third party (TPO) Insurance to Comprehensive insurance. This means that the reserves computed using the forecasted claims from the modeling process would not meet all the future expenses.
- As years go by, more people are buying cars and those with cars are buying more cars. This can cause an increase in future claim amounts such that the forecasts made earlier on would be lower than the actual values. Despite the dependence of the results of this study, the analysis has yielded results which can be used to amend insurance policies with regard to the level of premiums charged in view of expected claim amounts and this will go a long way in improving the profits registered by the insurance company.

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

### **5.3 Limitations Encountered**

In adopting the actuarial modeling process displayed in this study, the following are some of the likely limitations to be encountered when applying this results:

- I. In section 3, an assumption was made on the absence of zero claim amounts which in reality exists in comprehensive insurance policy. This assumption introduces a bias to the findings of the study.
- II. The choice of the statistical distributions to be used was limited by the software used in this research. This means that, other potential distributions such as the Pareto distribution could not be included as part of the family of distribution in this study due to limitations of the in-built functions of the MATLAB software.
- III. This research assumed that all the claim amounts experiences between June 2006 and June 2007 were all reported. Yet by the time the data was assembled, some of the claims had been reported but the claim amounts had not yet been established therefore they were not included in the study.

### **5.4 Recommendations**

Throughout this research some suggestions to be followed by the insurance companies in order to attain reliable forecasts have been noted and these could greatly improve the results yielded by the actuarial modeling process to be able to address all claim amounts in comprehensive insurance policies. Some of these are highlighted below:

Improving the level of implementation of the results of this study can help insurance companies establish their solvency states and therefore amend their policies before it's too late. This is because the study revealed that, for the insurance companies to be able to reliably estimate their future profits or losses they have to first accurately forecast future claim amounts then be able to set premium amounts, such that they can estimate required future reserves hence compute expected profits. But just knowing the results doesn't change anything; they have to take a further step in implementing the result findings.

There is need for insurance companies to carry out more research about the expected future claim amounts in their regions of location. This is because the study has acknowledged that future claim amounts are dependent on a number of factors experienced by the policyholder and not just by computational analysis. It should also be observed that the assumptions made on the claim amounts used in this study are not universal assumptions.

Insurance companies should make appropriate assumptions with regards to the claim amounts at hand. This would be appropriate in yielding unbiased estimates and reliable results. Different policies have different trends in the claim amounts and therefore different assumptions are tailored for different policies.

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.

## APPENDIX I

### COMPUTATION CODES

- Histogram of claims data is got from SPSS.
- Summary statistics of claims is got from SPSS.

### MATLAB CODES

- ❖ Get triple log of claims data
- ❖  $Lx = \text{Log}(x)$
- ❖  $LLx = \text{Log}(Lx)$
- ❖  $LLLx = \text{Log}(LLx)$

after getting the triple log of claims data, the 4 distributions are then fitted to the claims data by calculating the parameters.

Confidence interval for the parameters is calculated.

- ❖  $\text{Alpha} = 0.01;$
- ❖  $[\text{expparms}, \text{expci}] = \text{expfit}(LLLx, \text{alpha});$
- ❖  $[\text{lognparms}, \text{lognci}] = \text{lognfit}(LLLx, \text{alpha});$
- ❖  $[\text{gamparms}, \text{gamci}] = \text{gamfit}(LLLx, \text{alpha});$
- ❖  $[\text{wblparms}, \text{wblci}] = \text{wblfit}(LLLx, \text{alpha});$

the log-likelihood function is then calculated;

- ❖  $\text{expmudelta} = (\text{expci}(2, 1) - \text{expci}(1, 1));$
- ❖  $\text{temp}(1) = \text{expci}(1, 1) + \text{expmudelta};$
- ❖  $-\text{explike}(\text{temp}, LLLx)$
- ❖  $\text{lognmudelta} = (\text{lognci}(2, 1) - \text{lognci}(1, 1));$
- ❖  $\text{lognsigdelta} = (\text{lognci}(2, 2) - \text{lognci}(2, 1));$
- ❖  $\text{temp}(1) = \text{lognci}(1, 1) + \text{lognmudelta};$
- ❖  $\text{temp}(2) = \text{lognci}(1, 2) + \text{lognsigdelta};$
- ❖  $-\text{lognlike}(\text{temp}, LLx)$
- ❖  $\text{gammudelta} = (\text{gamci}(2, 1) - \text{gamci}(1, 1));$
- ❖  $\text{gamsigdelta} = (\text{gamci}(2, 2) - \text{gamci}(1, 2));$
- ❖  $\text{temp}(1) = \text{gamci}(2, 1) + \text{gammudelta};$
- ❖  $\text{temp}(2) = \text{gamci}(2, 1) + \text{gamsigdelta};$
- ❖  $-\text{gamlike}(\text{temp}, LLLx)$
- ❖  $\text{Wblmudelta} = (\text{wblci}(2, 1) - \text{wblci}(1, 1));$
- ❖  $\text{wblsigdelta} = (\text{wblci}(2, 2) - \text{wblci}(1, 2));$
- ❖  $\text{temp}(1) = \text{wblci}(1, 1) + \text{wblmudelta};$
- ❖  $\text{temp}(2) = \text{wblci}(1, 2) + \text{wblsigdelta};$
- ❖  $-\text{wbllike}(\text{temp}, LLLx)$

A graphical representation of the 4 distributions is then plotted, where parameter for exponential is E, for lognormal is L1, L2, for gamma, G1, G2 & for Weibull W1, W2.

Such that:

- ❖  $E1 = \text{exp pdf}(LLLx, E);$
- ❖  $L3 = \text{lognpd f}(LLLx, L1, L2);$
- ❖  $G3 = \text{Gampdf}(LLLx, G1, G2);$
- ❖  $W3 = \text{wblpdf}(LLLx, W1, W2);$
- ❖  $\text{plot}(y, E1, y, L3, y, G3, y, W3)$

- ❖ legend ('exponential', 'log-normal', 'Gamma', 'Weibull');
- goodness of fit is tested using Q-Q plots:
- ❖ qq plot (LLx, E)
  - ❖ qqplot (LLx, L1, L2)
  - ❖ qqplot (LLx, G1, G2)
  - ❖ qqplot (LLx, W1, W2)

*END*

## REFERENCES

1. Anderson D.R & Burnham K.P (2004) “MultiModel inference: Understanding A.I.C & B.I.C in model selection” Colorado Cooperative Fish & wildlife Research Unit (USGS-BRD)
2. Boland P. J. (2006) “Statistical Methods in General Insurance”
3. D’arcy, Stephen P. (1989) “On becoming an actuary of the third kind” (PDF) Proceedings of the casualty Actuarial Society LXXVI (145): 45-76
4. Denuit .M, Marechale .X, Pitrebois .S. and Walhin J.F (2007) “Actuarial Modeling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems” John Wiley & sons Ltd
5. Feldblum S. (2001) “Introduction” in Robert F. Lowe: Foundation of CAS, 4th Arlington, Virginia. CAS ISBN 0-9624762-2-6
6. Fiete S. (2004) COTOR Challenge Round 3, available at [www.casact.org/cotor/Fiete.doc](http://www.casact.org/cotor/Fiete.doc)
7. Guiahi F. (2000) “Fitting to loss distributions with emphasis on rating variables”. available at [www.casact.org/pubs/forum/01wforum/01wfl33.pdf](http://www.casact.org/pubs/forum/01wforum/01wfl33.pdf)
8. Kaishev V. (2001) SPMI, Cass Business School Hand Out notes.
9. Kitungulu K.S. (2007), Hand outs for notes on “Actuarial Data Processing”
10. Meyers.(2004) COTOR Challenge Round 2, available at [www.casact.org/cotor](http://www.casact.org/cotor)
11. Raz .O. and Shaw .M. (2000) “An Approach to preserving sufficient correctness in open resource coalitions, 10th International Work-shop on software specification & Design (IWSSD-10) IEEE Computer Society, November 2000.”
12. Renshaw A.E (2004) “Modeling the claim process in the presence of covariates”
13. Smyth G.K and Jørgensen B. (2004) “Fitting Tweedie’s compound poisson model to insurance claims data: dispersion modeling. ASTIN bulletin, 32(1), 143-157
14. Wright T. (2005) FIA, Deloitte, UK. COTOR Challenge Round 3, available at [www.casact.org/cotor/](http://www.casact.org/cotor/)
15. [www.actuaries.org.uk](http://www.actuaries.org.uk)
16. [www.en.wikipedia.org/wiki/qqplot](http://www.en.wikipedia.org/wiki/qqplot)
17. [www.maths.murdoch.edu.au/qq/qqnormal](http://www.maths.murdoch.edu.au/qq/qqnormal)
18. [www.wiki.math.yorku.ca/index.php/statistics](http://www.wiki.math.yorku.ca/index.php/statistics)