International Journal of Mathematics and Statistics Invention (IJMSI)
E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759
www.ijmsi.org Volume 10 Issue 6 || Nov.-Dec., 2022 || PP-01-04

Statistical Modelling for Insurance Data

B Navatha

ABSTRACT
Uncertainty refers to randomness and is different from a lack of predictability, or market inefficiency. An emergent research view holds that financial markets are both uncertain and predictable. Also, markets can be efficient but also uncertain. Insurance companies typically face two major problems when they want to forecast future premiums paid by using past or present behavior of premiums paid. For this, one has to find an appropriate statistical Probability distribution for the premiums paid. Then after test how well this statistical distribution fits the claims data. In modeling insurance claims, when there are extreme observations in the data, the commonly used loss distributions often are able to fit the bulk of the data well but fail to do so at the tail. One approach to overcome this problem is to focus on the extreme observations only and model them with the generalized Pareto distribution, supported by extreme value theory. The objective of this paper is to obtain an appropriate statistical Probability distribution for the insurance premium amounts and to test how well the chosen statistical distribution fits the premiums data. The modeling process will ascertain a statistical distribution that could capable model the claim amounts, and then the goodness of fit test was done mathematically using graphically using the Probability-Probability Plots (P-P plots) and Quantile-Quantile Plots(Q-Q plots). Finally, the study gives a summary, conclusion and recommendations that can be used by insurance companies to improve their results concerning future premium inferences..

***Key words**: premiums, extreme value theory, generalized Pareto distribution,generalized extreme value distribution, P-P plot, Q-Q Plot.*

Date of Submission: 05-12-2022 Date of Acceptance: 19-12-2022

I. Introduction

This Paper basically discusses the various distributions of premiums paid from an insurance company in India for the year 2010-2011. Here the data contains 48,000 observations and 36 variables like age, gender, Type of policy, Date of birth, Type of disease, Bonus, Floater amount, sum insured, premium, claims paid etc. we are interested in fitting of distribution for the variable Premiums paid .The paper used exploratory data analysis (histogram, mean, variance skewness, kurtosis maximum value, minimum value, standard deviation, and 1st and 3rd quartile) to help in the identification of the family of distribution which the data might follow. Probability plot was used to graphically demonstrate goodness of fit to different distributions. Different Goodness-of-Fit tests were used to test fitness of the distributions.

II. Descriptive Statistics

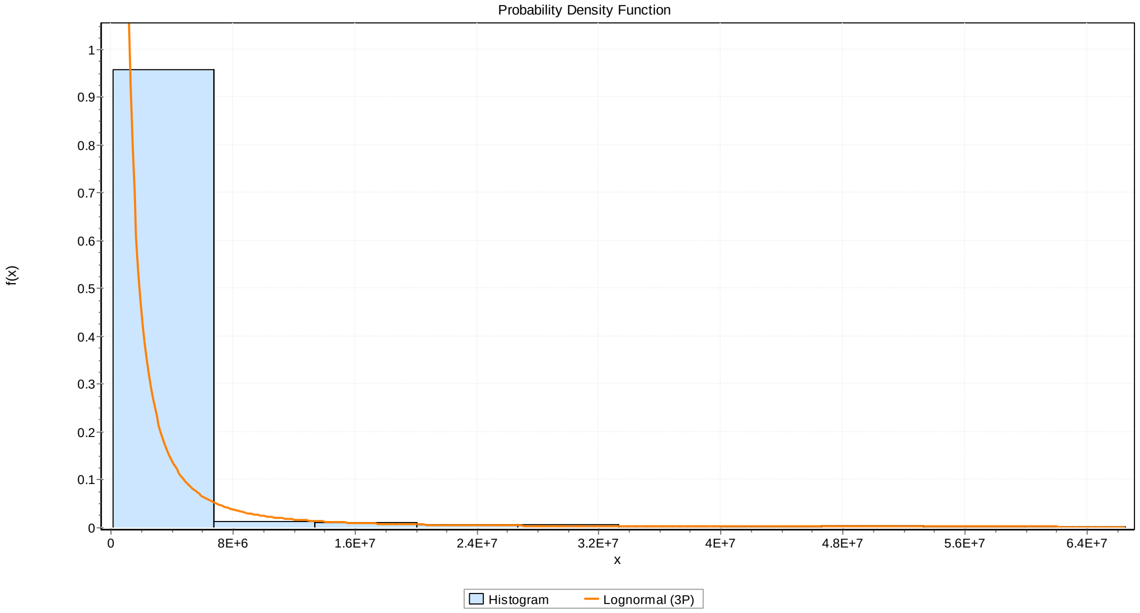
A total of 600 Premiums paid data on Health insurance for 2010-2011 was used for the modelling. below summarizes the result of the descriptive data analysis of the premiums paid.

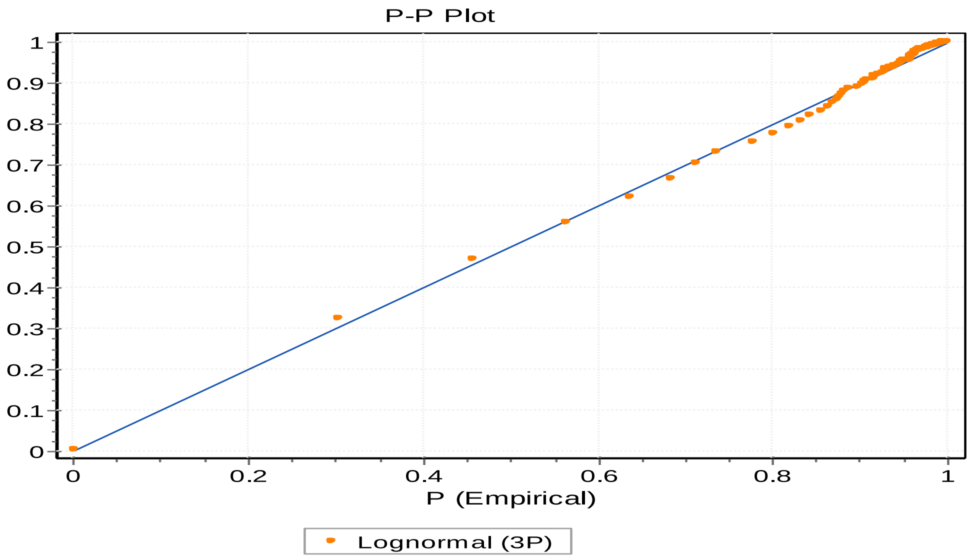
Statistic	**Value**	**Percentile**	**Value**
Sample Size	600	Min	1.0006E+5
Range	6.6474E+7	5%	1.1505E+5
Mean	1.7981E+6	10%	1.3035E+5
Variance	3.3726E+13	25% (Q1)	1.9517E+5
Std. Deviation	5.8074E+6	50% (Median)	4.1097E+5
Coef. of Variation	3.2298	75% (Q3)	1.0812E+6
Std. Error	2.3709E+5	90%	3.0155E+6
Skewness	7.0791	95%	6.3466E+6
Excess Kurtosis	58.567	Max	6.6574E+7

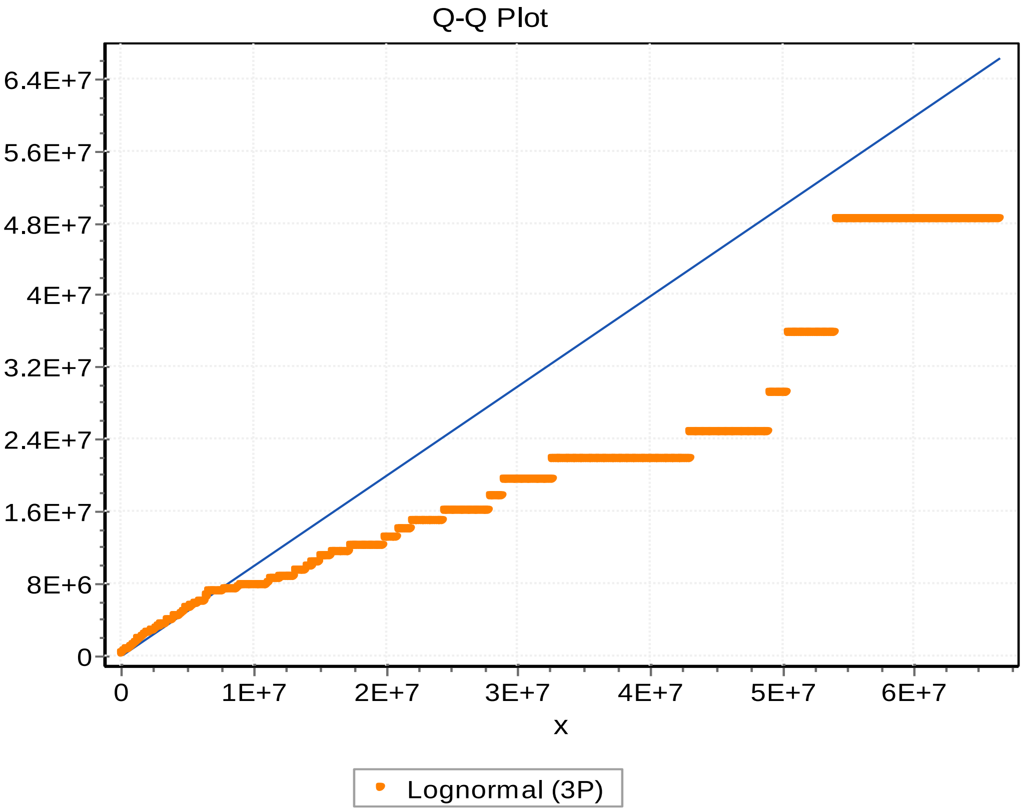
The average of premium was computed. The standard deviation, skewness, minimum and maximum value as well as the quantiles are also shown. This summary was necessary because it helps to identify key features of the data.: Descriptive statistics for the Health claim data from table above summarizes the result of the descriptive data analysis of the Premium data. 1,00,000 is the minimum Premium amount that was paid. The maximum Premium amount is 6,645,74,000. This indicates that within that period, the highest premium amount paid by the policy holder to the insurer. The 25th quartile of premium paid was 1,95,170 and that of the 75th quartile was 1,08,12,00The mean premium paid was 1,79,81,00 and the standard deviation is 5,80,7400 and that of the coefficient of skewness is 7. 0791.The skewness was measuring the symmetric nature of the claim. The value 7.0791 informs how the premium amount was positively skewed. Kurtosis has a value of 58.567, kurtosis measures whether the data is heavy-tailed or light tailed. The value 32.055 indicates that the data heavy tailed. The histogram was to visualize the shape on the premium amount. It was also observed that the premium data has less premium amounts but very high values because of the high variability in the premium data set even though the claim was only one year. The premium amount was transformed to log premium amount to improve symmetry. After the premium data was transformed, a new histogram, log premium amount was plotted this is an attempt to reduce the level of skewness of the data and make it quite symmetric. Before choosing one or more models for the data, it is necessary to choose good model among a predetermine set of models. This can be done with the help of fitting different distributions such as Gamma, Normal distribution, Log Normal , Weibull, other distributions were chosen as a family of models for the study.

Before fitting any distribution one has to know the parameters. These are estimated by Minitab Software.After fitting distributions, goodness of fit by different tests. with this one can know that which distribution fits well for the premiums paid data. Not only manual we can test graphically also, that is done with of P-P plots and also Q-Q plots.

Histogram P-P plot and Q-Q plot for Lognormal distribution is shown below.







3.The Parameters of all the distributions fitted to the data are listed in the are listed in the following Table

Sno	**Distribution**	**Parameters**
1	Exponential	λ=5.5615E-7
2	Exponential (2P)	λ=5.8892E-7 g=1.0006E+5
3	Fatigue Life	a=1.6945 b=8.4593E+5
4	Fatigue Life (3P)	a=2.3941 b=4.9326E+5 g=88577.0
5	Gamma	a=0.09586 b=1.8756E+7
6	Gamma (3P)	a=0.34631 b=4.9388E+6 g=1.0006E+5
7	Gen. Gamma	k=1.1061 a=0.26705 b=1.8756E+7

8	Gen. Gamma (4P)	k=1.5957 a=0.15411
		b=1.6335E+7 g=1.0006E+5

9	Gen. Pareto	k=0.74662 s=4.4157E+5 m=55327.0
10	Log-Gamma	a=107.42 b=0.12264
11	Lognormal	s=1.2701 m=13.174
12	Lognormal (3P)	s=1.8001 m=12.639 g=98427.0
13	Normal	s=5.8074E+6 m=1.7981E+6
14	Pareto	a=0.60207 b=1.0006E+5
15	Pareto 2	a=1.4472 b=7.7219E+5

k-threshold,a-Shape parameter and b-scale parameter.

4. Goodness of Fit – Summary

After treating the data fits with various distributions, we now test these with KS-test, Anderson Darling and Chi-squared test and found that, the Log Normal fits well with all the three tests for the claim data which is listed under Table ,Serial Number 12 in the table below which ranks "1" under the three tests. Therefore one can conclude that, Premiums paid follows Log Normal Distribution with authencity and while Generalized Pareto Distribution and Log Gamma falls under second and so on.

SNo	Distribution	Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
		Statistic	Rank	Statistic	Rank	Statistic	Rank
1	Exponential	0.34897	12	154.35	12	401.53	11
2	Exponential (2P)	0.37595	13	226.94	14	609.3	14
3	Fatigue Life	0.21556	9	54.995	9	132.95	8
4	Fatigue Life (3P)	0.15565	6	20.177	7	77.976	7
5	Gamma	0.63504	15	275.1	15	1166.3	15
6	Gamma (3P)	0.15742	7	31.22	8	181.7	9
7	Gen. Gamma	0.31341	11	106.64	11	455.15	13
8	Gen. Gamma (4P)	0.24693	10	64.965	10	407.9	12
9	Gen. Pareto	0.09303	3	6.115	2	25.873	2
10	Log-Gamma	0.09118	2	9.215	3	30.127	3
11	Lognormal	0.09548	4	11.843	4	33.113	4
12	**Lognormal (3P)**	**0.03249**	**1**	**0.77271**	**1**	**12.884**	**1**
13	Normal	0.38499	14	159.29	13	309.64	10
14	Pareto	0.10416	5	12.623	5	43.251	5
15	Pareto 2	0.16166	8	15.547	6	55.2	6

III. CONCLUSIONS AND FUTURE STUDY

With this empirical study of the claim data suggest that the premium paid for the beneficiaries follows the Log Normal Distribution, Thus, with this we can estimate the number of persons paying premium above any premiums paid by using the above Log Normal Distribution. Also, from this we know that estimation is possible by the Log Normal Distribution that how many are paying premium below or above 3 lakhs as compensation from the data on premium paid ,This study can further be studied for various data sets on actuaries which will be helpful to process the premium and other parameters of the data. Thus, we can explore this procedure of fitting and estimating the parameters for other than premiums paid data for different set of variables such as claimspaid ,floater amount ,Sum assured ,etc.