International Journal of Applied Mathematics and Statistics,
Int. J. Appl. Math. Stat.; Vol. 53; Issue No. 4; Year 2015,
ISSN 0973-1377 (Print), ISSN 0973-7545 (Online)

The Modeling of Motor Insurance Claims with Infinite Mixture Distribution

S. Anantasopon [1], **P. Sattayatham** [1], **and T. Talangtam** [2]

[1] Institute of Science, Suranaree University of Technology,
Nakhon Ratchasima 30000, Thailand;
Email : s_anantasopon@hotmail.com
Email : psattayatham@gmail.com

[2] Faculty of Science, Khon Kaen University,
Khon Kaen 40002, Thailand;
Email : beesaporn@yahoo.com

ABSTRACT

This paper introduces an infinite mixture model which can be used to fit the real data set for 1,296 observations. We also compare the result of the fit for three classical models (i.e. exponential, inverse exponential and lognormal distributions) and that of infinite mixture distribution. The K-S test reveals that the infinite mixture distribution gives the best fit.

Keywords: claim severity, classical distributions, infinite mixture distribution, gamma distribution, inverse exponential distribution.

Mathematics Subject Classification: 62P05

1. INTRODUCTION

A primary attribute of the actuary is the ability to successfully apply mathematical and statistical techniques to insurance claim data for both analysis and interpretation. The modeling of claims is an important task for claim estimation, since a good estimation of a claim leads to good insurance pricing.

There are two kinds of claim modeling, i.e. modeling of claim frequency and claim severity. Claim severity refers to the monetary loss on an insurance claim and it is usually modeled as a non-negative continuous random variable which has a mixed distribution. This is due to the fact that, in general, the classical distributions can not be fitted to arbitrary claim data.

The mixture models can be classified into finite and infinite mixture distributions. In the 1960s and 1970s, finite mixture models appeared in the statistical literature and they are useful for modeling of unobserved heterogeneous populations. Many authors described the modeling of finite mixture models, such as Sattayatham and Talangtam (2012) described finite mixture lognormal distributions which can be applied to motor insurance claims data. Moreover, Erisoglu, Servi, Erisoglu and Calis (2013) use two mixture gamma distributions for estimation heterogenous wind data sets.

A finite mixture distribution has a limitation on the number of components (k), depending on mean clustering. In order to solve this problem, we are interested in a consideration of infinite mixture distributions. One reason for using an infinite mixture model is to obtain new probability distributions and work with unknown parameters which will be simpler than working with finite mixture distributions.

Several authors discussed claim severity and constructed new distributions by using infinite mixture distributions. Emilio, Jose, and Enrique (2006) proposed a negative binomial inverse Gaussian distribution (NBIG) which has been applied to compute automobile insurance premiums. Recently, Pacakova and Zapletal (2013) proposed the Pareto distribution which is derived from exponential and gamma distributions. The model provides a better fit to the claim amounts in compulsory third party liability for motor vehicles insured by some Czech companies.

The purpose of this study is to construct an infinite mixture model which can better fit a set of real data for some non-life insurance public companies in Thailand.

2. MATERIALS AND METHODS

We present some classical distributions and an infinite mixture distribution which is relevant to our research objective.

2.1. Classical distributions

We used a real data set of motor insurance claims from public non-life insurance companies in Thailand. There are 1,296 observations which will be fitted using three distributions, i.e., exponential, inverse exponential, and lognormal. The maximum likelihood estimation (MLE) is provided for an estimation of the parameters of those distributions. The statistical test for model fitting is the K-S test.

Statistical modeling: Let $X_i, i=1,2,\ldots$ be the amount of i^{th} claim. It is supposed that the random variables $X_1, X_2, \ldots$ are independent and identically distributed (iid).

The Model: (1) exponential distribution, $Exp(\theta)$.

(2) inverse exponential distribution, $IExp(\theta)$.

(3) lognormal distribution, $LN(\theta,\sigma)$.

Estimation for the model: Consider the amount $\{x_i\}, (i=1,2,\ldots,n)$, paid for the i^{th} contract. We shall fit the data set $\{x_i\}$ to the exponential, inverse exponential, and lognormal distributions. By MLE, we obtain estimatiors for the parameters θ, and σ as follows:

(1) The probability density function (pdf) for the exponential distribution is

$$f_X(x) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right);\ \theta \in R,\ x > 0. \tag{1}$$

The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta}\exp\left(-\frac{x_i}{\theta}\right);\ \theta \in R,\ x > 0.$$

Then

$$\ln L(\theta) = -n\ln\theta - \sum_{i=1}^{n}\left(\frac{x_i}{\theta}\right);\ \theta \in R,\ x > 0.$$

An estimator $\hat{\theta}$ for the parameter θ can be obtained by solving the equation $\frac{\partial \ln L(\theta)}{\partial \theta} = 0.$

$\hat{\theta}$ is given by: $\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}.$

(2) The pdf for the inverse exponential distribution is

$$f_X(x) = \frac{\theta}{x^2}\exp\left(-\frac{\theta}{x}\right);\ \theta > 0,\ x > 0. \tag{2}$$

The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{\theta}{x_i^2}\exp\left(-\frac{\theta}{x_i}\right);\ \theta > 0,\ x > 0.$$

Then

$$\ln L(\theta) = n\ln\theta - \sum_{i=1}^{n}\left[\frac{\theta}{x_i} + 2\ln x_i\right];\ \theta > 0,\ x > 0.$$

An estimator $\hat{\theta}$ for the parameter θ can be obtained by solving the equation $\frac{\partial \ln L(\theta)}{\partial \theta} = 0.$

$\hat{\theta}$ is given by: $\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$

(3) The pdf for the lognormal distribution is

$$f_X(x) = \frac{1}{\sqrt{2\pi}x\sigma}\exp\left(-\frac{(\ln x - \theta)^2}{2\sigma^2}\right);\ \theta \in R,\ \sigma > 0,\ x > 0. \tag{3}$$

The likelihood function is

$$L(\theta,\sigma)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}x_i\sigma}\exp\left(-\frac{(\ln x_i-\theta)^2}{2\sigma^2}\right);\ \theta\in R,\ \sigma>0,\ x>0.$$

Then

$$\ln L(\theta,\sigma)=-n\ln\sigma-\frac{n}{2}\ln 2\pi-\sum_{i=1}^{n}\left[\ln x_i+\frac{1}{2\sigma^2}(\ln x_i-\theta)^2\right];\ \theta\in R,\ \sigma>0,\ x>0.$$

The two estimators $\hat{\theta}$ and $\hat{\sigma}$ for the parameter θ and σ can be obtained by solving these two equations:

$$\frac{\partial \ln L(\theta,\sigma)}{\partial\theta}=0,\ \text{and}\ \frac{\partial \ln L(\theta,\sigma)}{\partial\sigma}=0.$$

The solutions are $\hat{\theta}=\dfrac{\sum_{i=1}^{n}\ln x_i}{n}$, and $\hat{\sigma}=\sqrt{\dfrac{\sum_{i=1}^{n}(\ln x_i-\hat{\theta})^2}{n}}$ respectively.

Goodness of fit test: The Goodness of Fit (GOF) test measures the compatibility of a random sample with a theoretical probability distribution function. We use the Kolmogorov-Smirnov test (K-S test).

The K-S test is used to decide whether a sample comes from a hypothesized continuous distribution. It is based on the Empirical Cumulative Distribution Function (ECDF)

$$F_n(x)=\frac{1}{n}[\text{Number of observations}\ \le x].$$

The K-S test statistic is defined by

$$D=\sup_x\ \left|F_n(x)-F_X^*(x)\right|,$$

where F_n^* is the theoretical cumulative distribution of the distribution being tested.

The K-S test is defined by:

H_0 : The data follow a specified distribution.

H_1 : The data do not follow a specified distribution.

Level of critical values: The hypothesis regarding the distributional form is rejected at the chosen singnificance level (alpha, α) if the test statistic D is greater than the critical value.

Those three classical distributions were applied to the real data set. An analysis involving some comparisons are presented from the results of the statistical tests.

Table 1 shows the statistical test value for fitting the classical distributions to the real data set. The results from the K-S test are as the follow. For the significance level $\alpha = 0.01$, we found that none of those classical distributions could be fitted to the real data set since the p Value is less than 0.01. Hence, we can reject the null hypothesis and conclude that the data set does not follow the classical distributions at a 99% confidence level.

Table 1: The fitting for classical distributions

Distribution	K-S tests		Estimated Parameters
	D Value	p Value	
exponential	0.1961	< 0.0100	$\hat{\theta} = 1.766 \times 10^4$
lognormal	0.0466	< 0.0100	$\hat{\theta} = 8.9672$
inverse exponential	0.0759	< 0.0100	$\hat{\sigma} = 1.1804$ $\hat{\theta} = 4.190 \times 10^3$

Therefore, we shall look for another distribution which can be fitted to our real data set. An infinite mixture distribution is of interest to us. The K-S test verifies that the lognormal distribution is superior to the exponential and inverse exponential distributions but we will use inverse exponential distribution for constructing the infinite mixture because of computational convenience. It will be used as a mixed distribution.

2.2. Infinite Mixture Model

This section describes the construction of infinite mixture distributions and an estimation of parameters using MLE.

We represent an insurance claim amount by the random variable X. Let $f_X\left(x|\theta\right)$ denote the pdf of the insurance claim amount if the risk parameter is known to be θ. The heterogeneity in the insurance portfolio is due to variability in the parameter θ.

Let $G\left(\theta\right) = P(\Theta \le \theta)$ be the cdf of Θ, where Θ is the risk parameter viewed as a random variable. $G\left(\theta\right)$ is called mixing distribution. Let $g\left(\theta\right)$ be the pdf of Θ. Then

$$h_X\left(x\right) = \int_{R^+} f_X\left(x|\theta\right) g\left(\theta\right) d\theta, \ \forall x \in R^+,$$

is the unconditional pdf of X.

The Model: An infinite mixture model is composed of gamma as mixing distribution and inverse exponential as mixed distribution. Let X be the inverse exponential random variable with parameter θ. We want to mix an infinite number of inverse exponential distributions, each with a different value of θ. We let the mixing distribution have a pdf of θ, namely, a gamma with parameters α and β.

We note that the pdf of the gamma and inverse exponential are

$$g(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta);\ \alpha,\beta>0,\ \theta>0,$$

and

$$f_X(x|\theta) = \frac{\theta}{x^2}\exp\left(-\frac{\theta}{x}\right);\ \theta>0,\ x>0,$$

respectively, Then the infinite mixture model is of the form:

$$\begin{aligned} h_X(x) &= \int_0^{\infty}\frac{\theta}{x^2}\exp\left(-\frac{\theta}{x}\right)\cdot\frac{\beta^{\alpha}}{\Gamma(\alpha)}\cdot\theta^{\alpha-1}\exp(-\beta\theta)\,d\theta \\ &= \int_0^{\infty}\frac{\theta^{1+\alpha-1}\beta^{\alpha}}{x^2\Gamma(\alpha)}\exp\left(-\frac{\theta}{x}-\beta\theta\right)d\theta \\ &= \int_0^{\infty}\frac{\theta^{(\alpha+1)-1}\left(\frac{1}{x}+\beta\right)^{\alpha+1}}{\Gamma(\alpha+1)}\exp\left[-\theta\left(\frac{1}{x}+\beta\right)\right]d\theta\cdot\frac{\beta^{\alpha}}{x^2\Gamma(\alpha)}\cdot\frac{\Gamma(\alpha+1)}{\left(\frac{1}{x}+\beta\right)^{\alpha+1}} \\ &= \frac{\beta^{\alpha}}{x^2\Gamma(\alpha)}\cdot\frac{\Gamma(\alpha+1)}{\left(\frac{1}{x}+\beta\right)^{\alpha+1}} \\ &= \frac{\alpha\beta^{\alpha}x^{\alpha-1}}{(1+\beta x)^{\alpha+1}};\ \alpha,\beta>0,\ x>0, \end{aligned} \tag{4}$$

which is the pdf of inverse Pareto distribution $IPa\left(\alpha,\frac{1}{\beta}\right)$.

Estimation for the model: Consider the amount x_i paid for the i^{th} contract. We fit the $IPa\left(\alpha,\frac{1}{\beta}\right)$ distribution in Eq.4 to the data set by MLE. The estimated value of parameters α and β can be obtained by the following method:

Assume that $\mathrm{X}\sim IPa\left(\alpha,\frac{1}{\beta}\right)$ with density

$$h_X(x) = \frac{\alpha\beta^{\alpha}x^{\alpha-1}}{(1+\beta x)^{\alpha+1}}; \quad \alpha,\beta>0,\ x>0.$$

The likelihood function can be written as

$$L(\alpha,\beta) = \prod_{i=1}^{n}\frac{\alpha\beta^{\alpha}x_i^{\alpha-1}}{(1+\beta x_i)^{\alpha+1}}; \ \alpha,\beta>0,\ x>0.$$

The log-likelihood function is in the from

$$\ln L(\alpha,\beta) = n\ln\alpha + n\alpha\ln\beta + (\alpha-1)\sum_{i=1}^{n}\ln x_i - (\alpha+1)\sum_{i=1}^{n}\ln(1+\beta x_i); \alpha,\beta>0,\ x>0.$$

Hence, the partial derivatives of the log-likelihood function are

$$\frac{\partial \ln L(\alpha,\beta)}{\partial\alpha} = \frac{n}{\alpha} + n\ln\beta + \sum_{i=1}^{n}\ln x_i - \sum_{i=1}^{n}\ln(1+\beta x_i),$$

$$\frac{\partial \ln L(\alpha,\beta)}{\partial\beta} = \frac{n\alpha}{\beta} - (\alpha+1)\sum_{i=1}^{n}\frac{x_i}{1+\beta x_i}.$$

The two estimations $\hat{\alpha}$ and $\hat{\beta}$ for parameters α and β can be obtained by solving these two equations.

$$\frac{n}{\alpha} + n\ln\beta + \sum_{i=1}^{n}\ln x_i - \sum_{i=1}^{n}\ln(1+\beta x_i) = 0, \tag{5}$$

$$\frac{n\alpha}{\beta} - (\alpha+1)\sum_{i=1}^{n}\frac{x_i}{1+\beta x_i} = 0. \tag{6}$$

Because of the difficulty of solving Eq. (5)-(6) algebraically, we preferred to solve the equations numerically by using the Newton-Raphson method to estimate parameters α and β. We used MATLAB to do this work.

3. RESULTS

Again, we tried to fit real data set of 1,296 observations to the inverse Pareto distribution $IPa\left(\alpha,\frac{1}{\beta}\right)$. The K-S test was used for testing of model fitting. The histogram for the real data set (log scale) is illustrated in Figure 1.

Table 2 shows the statistical test value for fitting of inverse Pareto distribution and the estimated parameters. The results of the K-S test reveal a p Value for inverse Pareto distribution of 0.0482 which is greater than 0.01. Hence, we can conclude that the real data set can be fitted by inverse Pareto distribution with a 99% confidence level. The estimated parameters for inverse Pareto distribution are $\hat{\alpha} = 4.7260$ and $\hat{\beta} = 8.787 \times 10^{-4}$.

Table 2: The fitting of Infinite Mixture distribution

Distribution	K-S tests		Estimated Parameter
	D Value	p Value	
inverse Pareto	0.0381	0.0482	$\hat{\alpha} = 4.7260$
			$\hat{\beta} = 8.787 \times 10^{-4}$

Figure 2, solid line shows Empirical Cumulative Distribution (ECDF) while the dashed line is the cdf of exponential distribution.

Figure 3, solid line shows ECDF while the dashed line is the cdf of inverse exponential distribution.

Figure 4, solid line shows ECDF while the dashed line is the cdf of lognormal distribution.

Figure 5, solid line shows ECDF while the dashed line is the cdf of inverse Pareto distribution.

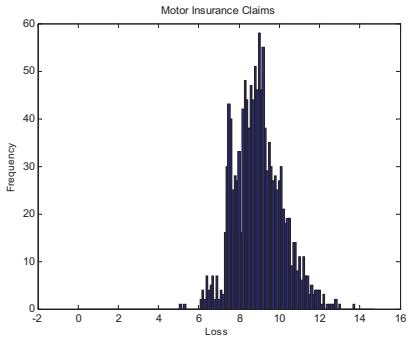

Figure 1. Histogram (log scale)

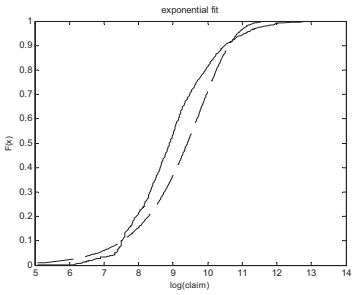

Figure 2. Model versus data cdf plot for the claim data set

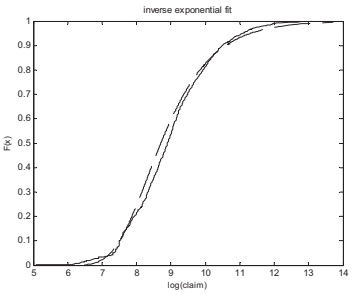

Figure 3. Model versus data cdf plot for the claim data set

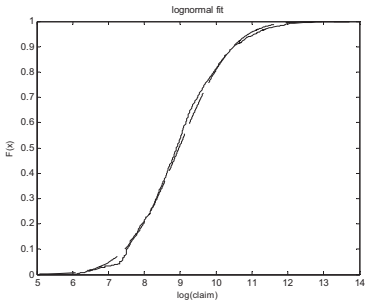

Figure 4. Model versus data cdf plot for the claim data set

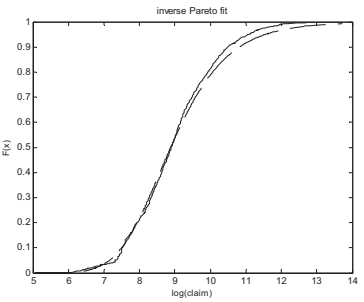

Figure 5. Model versus data cdf plot for the claim data set

Figures 6-9 show the P-P plot for exponential, inverse exponential, lognormal, and inverse Pareto distributions.

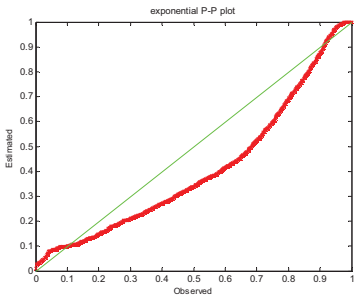

Figure 6. P-P plot for exponential distribution

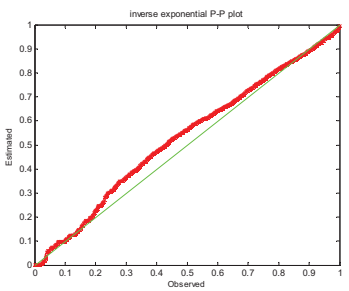

Figure 7. P-P plot for inverse exponential distribution

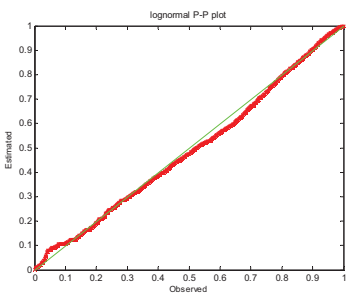

Figure 8. P-P plot for lognormal distribution

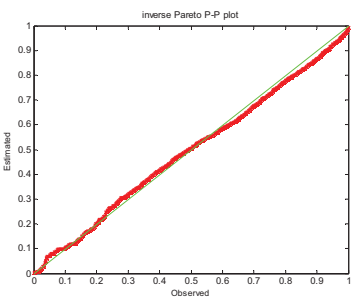

Figure 9. P-P plot for inverse Pareto distribution

4. DISCUSSION AND CONCLUSION

The new models can be constructed of infinite mixture distributions which are appropriated to our claim data set. They can be applied to any fields such as financial data, stock data and all practical purposes.

In this paper, we modelled the amounts for claims in heterogeneous portfolios of policies. Firstly, the classical models such as exponential, inverse exponential and lognormal have often been used by actuaries to fit real data sets. But these three classical models do not provide a good fit to a claim data set as determined by the K-S test. Then we attempted to find a solution by constructing an infinite mixture distribution which becomes an inverse Pareto distribution. Now, the real data set can be fitted to inverse Pareto distribution as shown by K-S test at a significance level of $\alpha = 0.01$. The classical distributions may not be appropriate to model claims, but an infinite mixture distribution does very well.

5. REFERENCES

Emilio, G.D., Jose, M.S., Enrique, C.O., 2006, Univariate and multivariate versions of the negative binomial-inverse Gaussian distributions with applications. Insurance : Mathematics and Economics. **42**, 39-49.

Erisoglu, M., Servi, T., Erisoglu, U., Calis, N., 2013, Mixture Gamma Distribution for Estimation of Wind Power Potential. Int. J. Appl. Math. Stat. **40**, 223-231.

Pacakova, V., Zapletal, D., 2013, Mixture Distributions in Modelling of Insurance Losses. Processing Conference on Applied Mathematics and Computational Methods in Engineeing.

Sattayatham, P., Talangtam, T., 2012, Fitting of Finite Mixture Distribution to More Insurance Claims. J. Math. Stat. **8**, 49-56.

Tosaporn, T., 2012, The Modeling of Loss For Non-life Insurance with Finite Mixture Models of Individual Data. Unpublished PhD Thesis, Suranaree University of Technology., 139 p.