Loss Distributions and Simulations in General Insurance and Reinsurance

V. Pacáková, D. Brebera

Abstract**—**The article presents the techniques suitable for practical purposes in general insurance and reinsurance and demonstrates their application**.** It focuses mainly on modelling and simulations of claim amounts by means of mathematical and statistical methods using statistical software products and contains examples of their applications. Article also emphasizes the importance and practical use of probability models of individual claim amounts and simulation of extreme losses in insurance practice.

Keywords**—**Goodness of fit tests, loss distributions, Pareto distribution, reinsurance, risk premium, simulation.

I. Introduction

ALTHOGH the empirical distribution functions can be useful tools in understanding claims data, there is always a desire to “fit” a probability distribution with reasonably good mathematical properties to the claims data.

Therefore this paper involves the steps taken in actuarial modelling to find a suitable probability distribution for the claims data and testing for the goodness of fit of the supposed distribution [1].

A good introduction to the subject of fitting distributions to losses is given by Hogg and Klugman (1984) and Currie (1993). Emphasis is on the distribution of single losses related to claims made against various types of insurance policies. 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 [1]. View of the importance of probability modelling of claim amounts for insurance practice several actuarial book publications dealing with these issues, e.g. [2], [7], [12].

The conditions under which claims are performed (and data are collected) allow us to consider the claim amounts in general insurance branches to be samples from specific, very often heavy-tailed probability distributions. As a probability models for claim sizes we will understand probability models of the financial losses which can be suffered by individuals and paid under the contract by non-life insurance companies as a result of insurable events. Distributions used to model these costs are often called “loss distributions” [7]. Such distributions are positively skewed and very often they have relatively high probabilities in the right-hand tails. So they are described as long tailed or heavy tailed distributions.

The distributions used in this article include gamma, Weibull, lognormal and Pareto, which are particularly appropriate for modelling of insurance losses. The Pareto distribution is often used as a model for claim amounts needed to obtain well-fitted tails. This distribution plays a central role in this matter and an important role in quotation in non-proportional reinsurance [15].

II. Claim Amounts Modelling Process

We will concern with modelling claim amounts by fitting probability distributions from selected families to set on observed claim sizes. This modeling process will be aided by the STATGRAPHICS Centurion XV statistical analytical system.

Steps of modelling process follow as below:

1. We will assume that the claims arise as realizations from a certain family of distributions after an exploratory analysis and graphical techniques.
2. We will estimate the parameters of the selected parametric distribution using maximum likelihood method based the claim amount records.
3. We will test whether the selected distribution provides an adequate fit to the data using Kolmogorov-Smirnov, Anderson-Darling or χ^2 test.

A. Selecting Loss Distribution

Most data in general insurance are skewed to the right and therefore most distributions that exhibit this characteristic can be used to model the claim amounts. For this article the choice of the loss distributions was with regard to prior knowledge and experience in curve fitting, availability of computer software and exploratory descriptive analysis of the data to obtain its key features. This involved finding the mean, median, standard deviation, coefficient of variance, skewness and kurtosis. This was done using Statgraphics Centurion XV package.

The Distribution Fitting procedure of this software fits any

V. Pacáková is with Institute of Mathematics and Quantitative Methods, Faculty of Economics and Administration, University of Pardubice, Pardubice, Studentská 84, 532 10 Pardubice, Czech Republic (e-mail: Viera.Pacakova@upce.cz).

D. Brebera with Institute of Mathematics and Quantitative Methods, Faculty of Economics and Administration, University of Pardubice, Pardubice, Studentská 84, 532 10 Pardubice, Czech Republic (e-mail: David.Brebera@upce.cz).

of 45 probability distributions (7 for discrete and 38 for continuous random variables) to a column of numeric data that represents random sample from the selected distribution. Distributions selected for our analysis are defined in Statgraphics Centurion as follow [7].

Gamma Distribution

Probability density function (PDF) is in the form

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

with parameters: shape $\alpha>0$ and scale $\lambda>0$.

Lognormal Distribution

Probability density function (PDF)) is in the form

$$f(x)=\frac{1}{\sigma\, x\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}\text{, } x>0 \tag{2}$$

with parameters: location μ, scale $\sigma>0$.

Weibull Distribution

Probability density function (PDF)) is in the form

$$f(x)=\frac{\alpha}{\beta^{\alpha}}x^{\alpha-1}e^{-(x/\beta)^{\alpha}},\ x>0 \tag{3}$$

with parameters: shape $\alpha>0$ and scale $\beta>0$.

A good tool to select a distribution for a set of data in Statgraphics Centurion is procedure *Density Trace*. This procedure provides a nonparametric estimate of the probability density function of the population from which the data were sampled. It is created by counting the number of observations that fall within a window of fixed width moved across the range of the data. The estimated density function is given by

$$f(x)=\frac{1}{hn}\sum_{i=1}^{n}W\left(\frac{x-x_i}{h}\right) \tag{4}$$

where h is the width of the window in units of X and $W(u)$ is a weighting function. Two forms of weighting function are offered: *Boxcar function* and *Cosine function*.

The *Cosine function* usually gives a smoother result, with the desirable value of h depending on the size of the data sample. Therefore in the application we use Cosine function

$$W(u)=\begin{matrix}1+\cos(2\pi u) & if\ |u|<0,5\\ 0 & otherwise\end{matrix} \tag{5}$$

B. Parameters Estimation

We will use the method of Maximum Likelihood (ML) to estimate the parameters of the selected loss distributions. This method can be applied in a very wide variety of situations and the estimated obtained using ML generally have very good properties compared to estimates obtained by other methods (e. g. method of moments, method of quantile). Estimates are obtained using ML estimation in procedure *Distribution Fitting* in Statgraphics Centurion XV package.

The basis for ML estimation is Maximum Likelihood Theorem [5]: Let $\mathbf{x}=(x_1,x_2,...,x_n)$ be a vector of n independent observations taken from a population with PDF $f(x;\mathbf{\Theta})$, where $\mathbf{\Theta}'=(\Theta_1,\Theta_2,...,\Theta_p)$ is a vector of p unknown parameters. Define the likelihood function $L(\mathbf{\Theta};\mathbf{x})$ by

$$L(\mathbf{\Theta};\mathbf{x})=\prod_{i=1}^{n}f(x_i;\mathbf{\Theta}) \tag{6}$$

The ML estimate $\hat{\mathbf{\Theta}}=\hat{\mathbf{\Theta}}(\mathbf{x})$ are that values of $\mathbf{\Theta}$ which maximizes $L(\mathbf{\Theta};\mathbf{x})$.

C. Goodness of Fit Tests

Various statistical tests may be used to check the fit of a proposed model. For all tests, the hypotheses tested are:

H_0: The selected distribution provides the correct statistical model for the claims,

H_1: The selected distribution does not provide the correct statistical model for the claims.

From the seven different tests that offer the procedure *Distribution Fitting* of package Statgraphics Centurion XV we will use the next three [18]:

Chi-Squared test divides the range of X into k intervals and compares the observed counts O_i (number of data values observed in interval i) to the number expected given the fitted distribution E_i (number of data values expected in interval i).

Test statistics is given by

$$\chi^2=\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i} \tag{7}$$

which is compared to a chi-squared distribution with $k-p-1$ degrees of freedom, where p is the number of parameters estimated when fitting the selected distribution.

Kolmogorov-Smirnov test (K-S test) compares the empirical cumulative distribution of the data to the fitted cumulative distribution. The test statistic is given by formula

$$d_n=\sup_x\left|F_n(x)-F(x)\right| \tag{8}$$

The empirical CDF $F_n(x)$ is expressed as follows:

$$F_n(x)=\begin{cases}0 & x\le x_{(1)}\\ \dfrac{j}{n} & x_{(j)}<x\le x_{(j+1)} \qquad j=1,2,\dots,n-1\\ 1 & x>x_{(n)}\end{cases} \tag{9}$$

where data are sorted from smallest to largest in sequence $x_{(1)}\le x_{(2)}\le ......\le x_{(n)}$.

Anderson-Darling test is one of the modifications of K-S test. The test statistic is a weighted measure of the area between the empirical and fitted CDF's. It is calculated according to:

$$A^2=-n-\frac{\sum_{i=1}^{n}\left((2i-1)\cdot\ln\left(z_{(i)}\right)+(2n+1-2i)\cdot\ln\left(1-z_{(i)}\right)\right)}{n}$$

where $z_{(i)}=F_n\left(x_{(i)}\right)$.

In all above mentioned goodness of fit tests the small P-value leads to a rejection of the hypothesis H_0.

III. Pareto model of Claim Amounts

Pareto distribution is commonly used to model claim-size distribution in insurance for its convenient properties.

Pareto random variable *X* has distribution function

$$F(x)=1-\left(\frac{\lambda}{\lambda+x}\right)^{\alpha} \tag{10}$$

with positive parameters α and λ and density function

$$f(x)=\frac{\alpha\lambda^{\alpha}}{(\lambda+x)^{\alpha+1}}. \tag{11}$$

When *X* is by Pareto distributed, it is readily determine the mean

$$E(X)=\frac{\lambda}{\alpha-1} \qquad \text{for } \alpha>1 \tag{12}$$

and variance

$$D(X)=\frac{\alpha\lambda^2}{(\alpha-1)^2(\alpha-2)} \qquad \text{for } \alpha>2. \tag{13}$$

Then method of moments to estimate parameters α, λ is easy to apply. To equate the first two population and sample moments we find estimates

$$\tilde{\alpha}=\frac{2s^2}{s^2-\bar{x}^2} \qquad \tilde{\lambda}=(\tilde{\alpha}-1)\bar{x} \tag{14}$$

The estimates $\tilde{\alpha},\tilde{\lambda}$ obtained by this way tend to have rather large standard errors, mainly because s^2 could has a very large variance for high probability of extreme values of claim amounts. We rather prefer estimates of *α* and *λ* using maximum likelihood method.

We denote as $\hat{\alpha},\hat{\tilde{\lambda}}$ the maximum likelihood estimates given data $x_1,x_2,\dots,x_n$ from the Pareto distribution [12, 13]. Solving equation $f(\lambda)=0$ using the initial estimate $\tilde{\lambda}$, where

$$f(\lambda)=A-B=\frac{\sum_{i=1}^{n}\frac{1}{\lambda+x_i}}{\sum_{i=1}^{n}\frac{x_i}{\lambda(\lambda+x_i)}}-\frac{n}{\sum_{i=1}^{n}\ln\left(1+\frac{x_i}{\lambda}\right)} \tag{15}$$

we obtain $\hat{\lambda}$. Substituting $\hat{\lambda}$ in *A* or *B* we find $\hat{\alpha}$.

IV. Pareto Model in Reinsurance

Modelling of the tail of the loss distributions in general insurance is one of the problem areas, where obtaining a good fit to the extreme tails is of major importance. Thus is of particular relevance in non-proportional reinsurance if we are required to choose or price a high-excess layer [13].

The Pareto model is often used to estimate risk premiums for excess of loss treaties with high deductibles, where loss experience is insufficient and could therefore be misleading. This model is likely to remain the most important mathematical model for calculating excess of loss premiums for some years to come [19].

The Pareto distribution function of the losses X_a that exceed known deductible *a* is [4], [15].

$$F_a(x)=1-\left(\frac{a}{x}\right)^{b}, \quad x\ge a \tag{16}$$

The density function can be written

$$f_a(x)=\frac{b\cdot a^b}{x^{b+1}}, \quad x\ge a \tag{17}$$

Through this paper we will assume that the lower limit *a* is known as very often will be the case in practice when the reinsurer receives information about all losses exceeding *a* certain limit.

The parameter *b* is the Pareto parameter and we need it estimate. Let us consider the single losses in a given portfolio during a given period, usually one year. As we want to calculate premiums for *XL* treaties, we may limit our attention to the losses above a certain amount, the "observation point" *OP*. Of course, the *OP* must be lower than the deductible of the layer for which we wish to calculate the premium [4], [15].

Let losses above this *OP* $X_{OP,1}, X_{OP,2}, ..., X_{OP,n}$ be independent identically Pareto distributed random variables with distribution function

$$F_{OP}(x) = 1 - \left(\frac{OP}{x}\right)^{b}, \qquad x \geq OP \qquad (18)$$

The maximum likelihood estimation of Pareto parameter *b* is given by formula [4].

$$\frac{n}{\sum_{i=1}^{n} \ln\left(\frac{X_{OP,i}}{OP}\right)} \qquad (19)$$

The Pareto distribution expressed by (16) is part of the *Distribution Fitting* procedure in Statgraphics Centurion XV package. This allows us to use the Pareto distribution to calculate the reinsurance risk premium. Risk premiums are usually calculated using the following equation:

risk premium = expected frequency × expected loss

The expected frequency is the average number of losses paid by reinsurer per year. For a given portfolio we should set *OP* low enough to have a sufficient number of losses to give a reasonable estimation of the frequency *LF*(*OP*).

If the frequency at the observation point *OP* is known than it is possible to estimate the unknown frequency of losses exceeding any given high deductible *a* as

$$LF(a) = LF(OP) \cdot P(X_{OP} \rangle\, a) = LF(OP) \cdot \left(\frac{OP}{a}\right)^{b} \qquad (20)$$

The reinsurance risk premium *RP* can now be calculated as follows:

$$RP = LF\left(a\right) \cdot EXL \qquad (21)$$

where

$$EXL = E\left(X_{a}\right) = \frac{a \cdot b}{b-1}, \qquad b > 1 \qquad (22)$$

V. Simulations Using Quantile Function

The *Quantile Function*, QF, denoted by $Q\left(p\right)$, expresses the *p*-quantile x_{p} as a function of *p*: $x_{p} = Q\left(p\right)$, the value of *x* for which $p = P\left(X \leq x_{p}\right) = F\left(x_{p}\right)$.

The definitions of the QF and the CDF can by written for any pairs of values $\left(x,\, p\right)$ as $x = Q\left(p\right)$ and $p = F\left(x\right)$. These functions are simple inverses of each other, provided that they are both continuous increasing functions. Thus, we can also write $Q\left(p\right) = F^{-1}\left(p\right)$, and $F\left(x\right) = Q^{-1}\left(x\right)$. For sample data, the plot of $Q\left(p\right)$ corresponds to the plot of *x* against *p* [6], [9].

We denoted a set of ordered sampling data of losses by

$$x_{(1)},\, x_{(2)},\, ...,\, x_{(r)},\, ...,\, x_{(n-1)},\, x_{(n)}\, .$$

The corresponding random variables are being denoted by

$$X_{(1)},\, X_{(2)},\, ...,\, X_{(r)},\, ...,\, X_{(n-1)},\, X_{(n)}.$$

Thus $X_{(n)}$ for example is the random variable representing the largest observation of the sample of *n*. The *n* random variables are referred as the *n* order statistics. These statistics play a major role in modelling with quantile distribution $Q\left(p\right)$.

Consider first the distribution of the largest observations on $X_{(n)}$ with distribution function denoted $F_{(n)}\left(x\right) = p_{(n)}$. The probability

$$F_{(n)}\left(x\right) = p_{(n)} = P\left(X_{(n)} \leq x\right)$$

is also probability that all *n* independent observations on *X* are less than or equal to this value *x*, which for each one is *p*. By the multiplication law of probability

$$p_{(n)} = p^{n} \text{ so } p = p_{(n)}^{1/n} \text{ and } F\left(x\right) = p = p_{(n)}^{1/n}.$$

Inverting $F\left(x\right)$ to get the quantile function we have

$$Q_{(n)}\left(p_{(n)}\right) = Q\left(p_{(n)}^{1/n}\right) \qquad (23)$$

For the general *r*-th order statistic $X_{(r)}$ calculation becomes more difficult. The probability that the *r-th* larges observations is less than some value *z* is equal

$$p_{(r)} = F_{(r)}\left(z\right) = P\left(X_{(r)} \leq z\right)$$

This is also probability that at least *r* of the *n* independent observations is less or equal to *z*. The probability of *s* observations being less than or equal to *z* is p^{s}, where $p = F\left(z\right)$ is given by the binomial expression

$$P\left(s \;\, observations \leq z\right) = \binom{n}{s} p^{s} \left(1-p\right)^{(n-s)}$$

and

$$p_{(r)} = \sum_{s=r}^{n} \binom{n}{s} p^{s} \left(1-p\right)^{(n-s)}$$

If it can be inverted, then we can write

$$p = BETAINV\left(p_{(r)},\, r,\, n-r+1\right).$$

From the last two expressions we get

$$Q_{(r)}\left(p_{(r)}\right) = Q\left(BETAINV\left(p_{(r)}, r, n-r+1\right)\right) \qquad (24)$$

$BETAINV\left(\cdot\right)$ is a standard function in packages such as Excel. Thus, the quantiles of the order statistics can be evaluated directly from the distribution $Q(p)$ of the data. The quantile function thus provides the natural way to simulate values for those distributions for which it is an explicit function of p [13].

VI. Simulation of Extreme Losses

In a number of applications of quantile functions in general insurance interest focuses particularly on the extreme observations in the tails of the data. Fortunately it is possible to simulate the observations in one tail without simulating the central values. We will present here how to do this.

Consider the right-hand tail. The distribution of the largest observation has been shown to be $Q\left(p^{1/n}\right)$. Thus the largest observation can be simulated as $x_{(n)} = Q\left(u_{(n)}\right)$, where $u_{(n)} = v_n^{1/n}$ and v_n is a random number from interval [0, 1]. If we now generate a set of transformed variables by

$$\begin{aligned} u_{(n)} &= v_n^{1/n} \\ u_{(n-1)} &= \left(v_{n-1}\right)^{\frac{1}{n-1}} \cdot u_{(n)} \\ u_{(n-2)} &= \left(v_{n-2}\right)^{\frac{1}{n-2}} \cdot u_{(n-1)} \end{aligned} \qquad (25)$$

where v_i, $i = n, n-1, n-2, \ldots$ is simply simulated set of independent random uniform variables, not ordered in any way. It will be seen from their definitions that $u_{(i)}$, $i = n, n-1, n-2, \ldots$ form a decreasing series of values with $u_{(i-1)} < u_{(i)}$.

In fact, values $u_{(i)}$ form an ordering sequence from a uniform distribution. Notice that once $u_{(n)}$ is obtained, the relations have the general form

$$u_{(m)} = \left(v_m\right)^{\frac{1}{m}} \cdot u_{(m+1)}, \qquad m = n-1, n-2, \ldots$$

The order statistics for the largest observations on X are then simulated by

$$\begin{aligned} x_{(n)} &= Q\left(u_{(n)}\right) \\ x_{(n-1)} &= Q\left(u_{(n-1)}\right) \\ x_{(n-2)} &= Q\left(u_{(n-2)}\right) \\ &\vdots \end{aligned} \qquad (26)$$

In most simulation studies of n observations are generated and the sample analyses m times to give an overall view of their behavior. A technique that is sometimes used as an alternative to such simulation is to use a simple of ideal observations, sometimes called a *profile*. Such a set of ideal observations could be the medians M_r, $r = 1, 2, \ldots, n.$.

VII. Application of the Theoretical Results

A. Data

Practical application of theoretical results mentioned in previous chapters we will performed based on data obtained from unnamed Czech insurance company. We will use the data set contains 745 claim amounts (in thousands of Czech crowns - CZK) from the portfolio of 26 125 policyholders in compulsory motor third-party liability insurance. In the article [16] we considered 1352 claim amounts, including the reserve estimates. Those estimates caused in the file were high number of the same estimated values, which distorted results of the analysis. Data file in this article shall include only actually paid out claim amounts.

B. Exploratory analysis

We will start by descriptive analysis of sample data of the variable X, which represents the claim amounts in the whole portfolio of policies.

Table 1 Summary statistics for *X*

Count	745
Average	1220,95
Median	739,84
Standard deviation	1521,07
Coeff. of variation	124,581%
Minimum	24,42
Maximum	22820,0
Skewness	5,72352
Kurtosis	60,6561

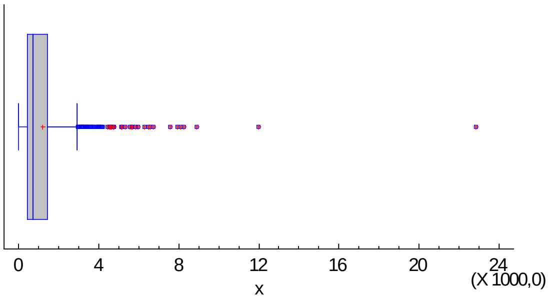

Fig. 1 Box-and-Whisker plot of claim amounts data

Tab.1 shows summary statistics for *X*. These statistics and Box-and-Whisker plot confirm the skew nature of the claims data. Also by density trace by (5) for *X* in Fig. 2 can be concluded that loss distribution in our case is skew and long or heavy tailed.

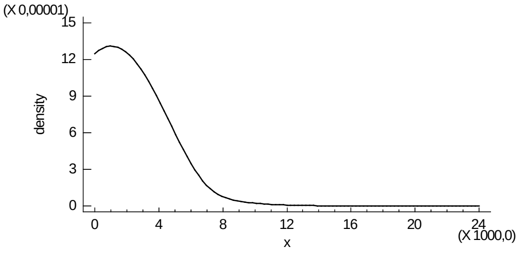

Fig. 2 Density Trace for *X*

C. Selected loss distributions

The results of exploratory analysis justify us to assume that gamma, lognormal or Weibull distributions would give a suitable model for the underlying claims distribution. We will now start to compare how well different distributions fit to our claims data. The best way to view the fitted distributions is through the Frequency Histogram. Fig. 3 shows a histogram of the data as a set of vertical bars, together with the estimated probability density functions.

From Fig. 3 it seems that lognormal distribution follows the data best. It is hard to compare the tail fit, but clearly the all distributions have discrepancies at middle claims intervals.

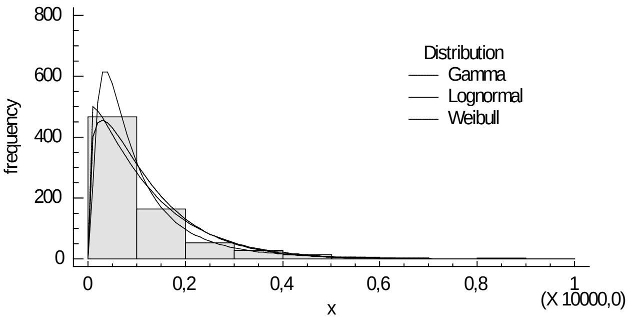

Fig. 3 Histogram and estimated loss distributions

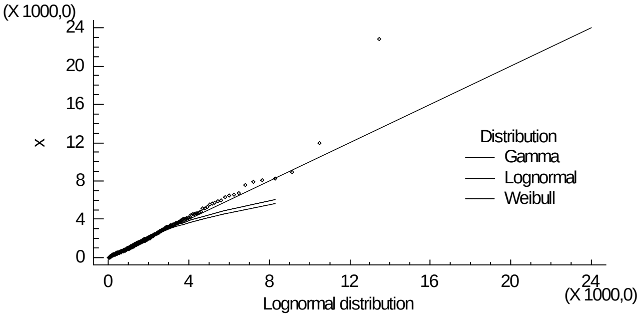

Fig. 4 Quantile-Quantile plot of selected distributions

The Quantile-Quantile (Q-Q) plot shows the fraction of observations at or below *X* plotted versus the equivalent percentiles of the fitted distributions. In Fig. 4 the fitted lognormal distribution has been used to define the X-axis and is represented by the diagonal line. The fact that the points lay the most close to the diagonal line confirms the fact that the lognormal distribution provides the best model for the data in comparison with other two distributions. Gamma and lognormal distributions deviates away from the data at higher values of *X*, greater than 4000 CZK of *X*. Evidently, the tails of these distributions are not fat enough.

Despite the adverse graphic results we will test whether the selected distributions fit the data adequately by using *Goodness-of-Fit Tests* of Statgraphics Centurion XV.

The ML parameters estimation of the fitted distributions by minimizing (6) is shown in Table 2.

Table 2 Estimated parameters of the fitted distributions

Gamma	*Lognormal*	*Weibull*
shape = 1,30953	mean = 1201,77	shape = 1,0647
scale = 0,001073	standard deviation = 1360,07	scale = 1257,15
	Log scale: mean = 6,67928	
	Log scale: std. dev. = 0,9080	

Table 3 Anderson-Darling Goodness-of-Fit Tests for *X*

	Gamma	*Lognormal*	*Weibull*
A^2	13,4802	1,58638	15,1049
Modified Form	13,4802	1,58638	15,1049
P-Value	<0.01	>=0.10	<0.01

Table 3 shows the results of tests run to determine whether *X* can be adequately modelled by gamma, lognormal or Weibull distributions. P-values less than 0,01 would indicate that *X* does not come from the selected distributions with 99% confidence.

Table 4 shows the results of chi-squared test by (7) run to determine whether *X* can be adequately modelled by lognormal distribution with parameters estimated by ML. Since the smallest *p*-value = 0,520495 amongst the tests performed is greater than or equal to 0,05, we cannot reject the hypothesis that *x* comes from a lognormal distribution with 95% confidence, contrary to the result of the chi-squared test in article [16].

Table 4 Chi-Squared test with lognormal distribution

	Upper Limit	*Observed Frequency*	*Expected Frequency*	*Chi-Squared*
at or below	1000,0	466	446,50	0,85
	2000,0	165	182,98	1,77
	3000,0	53	61,93	1,29
	4000,0	28	25,53	0,24
	5000,0	14	12,06	0,31
	6000,0	8	6,29	0,47
	7000,0	4	3,52	0,07
	8000,0	2	2,09	0,00
above	8000,0	5	4,11	0,19

Chi-Squared = 5,18357 with 6 d.f. P-Value = **0,520495**

Figure 5 and Tab. 4 lead to the conclusion that the fitted lognormal model undervalues the losses which exceed the 4000 CZK.. We need to find a distribution with rather more weight in upper tail than the lognormal distribution provides.

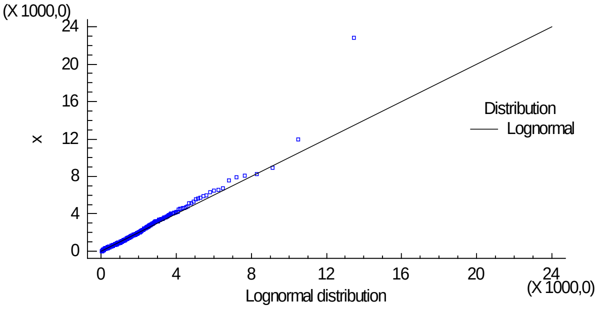

Fig. 5 Quantile-Quantile plot for lognormal distribution of *X*

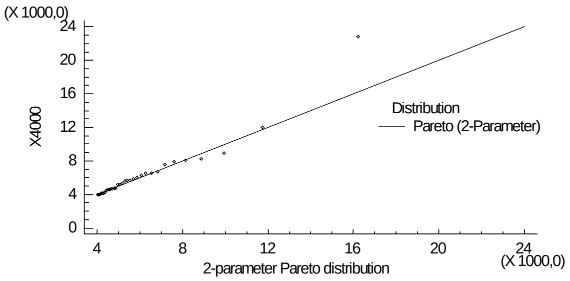

Fig. 6 Quantile-Quantile plot for Pareto distribution of X_{4000}

In the Fig. 6 we can see that the Pareto model (16) for the variable X_{4000} gives an better fit and is a considerable improvement over lognormal model. Pareto distribution has been used to define the *X*-axis at Fig. 6. The fact that the points lay close to the diagonal line confirms the fact that this distribution provides a good model for the claim amounts data above 4 million CZK and we can assume that a good model for losses above 4 million CZK can be Pareto distribution with PDF expressed by the formula (17).

Table 5 Estimated parameters by ML

Pareto (2-Parameter)
shape = 2,80078
lower threshold = 4000,0

Table 6 K-S goodness-of-fit tests for X_{4000}

	Pareto (2-Parameter)
DPLUS	0,0985791
DMINUS	0,0758859
DN	0,0985791
P-Value	**0,895778**

There are 34 values ranging from 4000 to 22 820 thousand CZK. Table 5 and Table 6 show the results of fitting a 2-parameters Pareto distribution (16) to the data on X_{4000}. The estimated parameters of the fitted distribution using maximum likelihood method are shown in Table 5. The results of K-S test whether the 2-parameter Pareto distribution fits the data adequately is shown in Table 6. Since the *p*-value = 0,895778 is greater than 0,05, we cannot reject the null hypothesis that sampling values of the variable X_{4000} comes from a 2-parameters Pareto distribution with 95% confidence.

We check yet hypothesis that the distribution with a good fit on the claim amounts empirical data is a Pareto distribution in the form (5). Since this probability model is not part of the offer distributions for Distribution Fitting procedure of Statgraphics Centurion XV package, we will perform a χ2 goodness of fit test on Pareto distribution in the form (16) in the usual way using Excel spreadsheet.

We suppose the random variable *X*, with values of the claim amounts in the whole portfolio of policies, is Pareto distributed with distribution function (16). Using the method of moments to estimate the parameters α, λ of a Pareto distribution to solve the equations (14) we find estimators by method of moments that are $\tilde{\alpha} = 5,6229,\ \tilde{\lambda} = 5644,402.$ Of course, maximum likelihood estimators we have preferred. We will obtain estimates of parameters α and λ using more efficient maximum likelihood method. Using the initial estimate $\tilde{\lambda} = 5644,402$ we find non-linear equation (15) as $f(5644,402) = 0,03533$ and using tool Solver of Excel we find $f(14745,8) = 8,2746 \cdot 10^{-10}$, hence the maximum likelihood estimator $\hat{\lambda} = 14745,8.$ Substituting $\hat{\lambda}$ into A or B in equation (15) we find $\hat{\alpha} = 13,19543$.

Now we can check whether the Pareto distribution with maximum likelihood estimators provides an adequate fit to the data using χ^2 goodness-of-fit test. The χ^2 statistic is computed by (7). The result of goodness of fit test presents tab. 7.

Table 7 Chi-Squared test for Pareto distribution

	Upper Limit	*Observed Frequency*	*Expected Frequency*	*Chi-Squared*
at or below	1000	466	431,5745	2,746026
	2000	165	174,3427	0,500663
	3000	53	74,38566	6,148315
	4000	28	33,31204	0,847073
	5000	14	15,5765	0,159557
	6000	8	7,571378	0,024265
	7000	4	3,811341	0,009338
above	7000	7	4,425835	1,497192
Total				**11,93243**

By comparing the calculated value $\chi^2 = 11,93243$ with quantile $\chi^2_{0,95} = 11,0705$ on 5 degrees of freedom (since there are two estimated parameters), we have to reject null hypothesis that the Pareto distribution with parameters estimated by maximum likelihood provides an adequate fit to the data.

Paradoxically, if we repeat χ^2 goodness-of-fit test for Pareto model with parameters estimated by method of moments, we get result $\chi^2 = 4,69 < \chi^2_{0,95} = 11,0705.$

Therefore Pareto distribution define by (10) with parameters $\tilde{\alpha} = 5,6229,\ \tilde{\lambda} = 5644,402$ gives an excellent fit and is a considerable improvement over lognormal and Pareto model with parameters estimated by the maximum likelihood method.

D. Using results in non-proportional reinsurance

Let's now suppose the insurance company wants to reduce technical risk by non-proportional XL reinsurance with priority (deductible) *a* = 8 000 (thousand CZK). We will use the Pareto distribution for variable X_{4000} with parameters from Table 5 to determine reinsurance risk premium by (21). As *OP* we put value 4000. To calculate *LF*(*a*) by (20) we need to know $P(X_{OP} \rangle\ a)$. We can use *Tail Areas* pane for the fitted 2-parameters Pareto distribution of the Statgraphics Centurion XV package. The software will calculate the tail areas for up to 5 critical values, which we may specify. The output indicates that the probability of obtaining a value above 8 000 for the fitted 2-parameter Pareto distribution of X_{4000} is

0,143509, as we can see in Table 8. The value of $LF(OP)$ we can estimate by relative frequency of the losses above 4000:

$$LF(OP) = \frac{33}{745} = 0,0443$$

Table 8 Tail Area for X_{4000} Pareto (2-Parameter) distribution

X	*Lower Tail Area (<)*	*Upper Tail Area (>)*
8000,0	0,856491	0,143509

Then by (20) we get

$$LF(a) = LF(OP) \cdot P(X_{OP} \rangle\, a) = 0,0443 \cdot 0,143509 = 0,006357$$

So that we can use the formulas (21) and (22) to calculate the reinsurance premium RP we will fit the 6 values ranging from 8 000,0 to 22 820,0 of the variable X_{8000} by Pareto (2-Parameters) distribution using Statgraphics Centurion XV *Goodnes of fit* procedure. The results are listed in the Table 9 and Table 10.

Table 9 Estimated Pareto parameters by ML

Pareto (2-Parameter)
shape = 3,74093
lower threshold = 8000,0

Table 10 *K-S* goodness-of-fit tests for X_{8000}

	Pareto (2-Parameter)
DPLUS	0,384835
DMINUS	0,14685
DN	0,384835
P-Value	0,339047

Table 10 shows the results of tests run to determine whether X_{8000} can be adequately modelled by a 2-parameter Pareto distribution (16) with ML estimated parameters from Table 9. Since the *p*-value = 0,339047 is greater than 0,05, we cannot reject the null hypothesis that values of X_{8000} come from the 2-parameter Pareto distribution with 95% confidence.

By the values of estimated parameters in Table 9 we can calculate

$$E(X_a) = \frac{a \cdot b}{b-1} = \frac{8000 \cdot 3,74093}{2,74093} = 10918,72$$

Then by formula (21) we get reinsurance premium in thousand CZK:

$$RP = LF(a) \cdot EXL = 0,006357 \cdot 10918,72 = 69,41.$$

E. Simulation of extreme losses

By the result in part VI-*C* Pareto model define by (10) with parameters $\tilde{\alpha} = 5,6229,\ \tilde{\lambda} = 5644,402$ fit well to the claim amounts data. Pareto quantile function $Q(p)$ or the inverse of the Pareto distribution function (10) has the form

$$Q(p) = F^{-1}(x) = \lambda\left[(1-p)^{-1/\alpha} - 1\right], \qquad 0 < p < 1. \tag{27}$$

That is in our case

$$Q(p) = F^{-1}(x) = 5644,402\left[(1-p)^{-1/5,6229} - 1\right] \tag{28}$$

We use results of parts V. and VI. to simulate 10 the largest claim amounts by (28) in case of number 1000 claim amounts. Table 11 contains the results of simulation step by step by part VI using formulas (25) and (26).

Table 11 Steps of simulation of 10 the largest claim amounts

v	*n*	*u*	*Q(u)*	*Q(BETAINV(0,5))*
0,235493	1000	0,99855	12 415,58	14 937,50
0,331321	999	0,99745	10 682,07	11 942,23
0,743843	998	0,99716	10 366,40	10 544,10
0,993465	997	0,99715	10 359,85	9 656,27
0,493922	996	0,99644	9 742,37	9 015,37
0,997123	995	0,99644	9 740,15	8 518,55
0,665588	994	0,99603	9 446,11	8 115,47
0,503882	993	0,99535	9 023,41	7 777,94
0,943984	992	0,99529	8 991,23	7 488,61
0,761040	991	0,99501	8 844,77	7 236,14

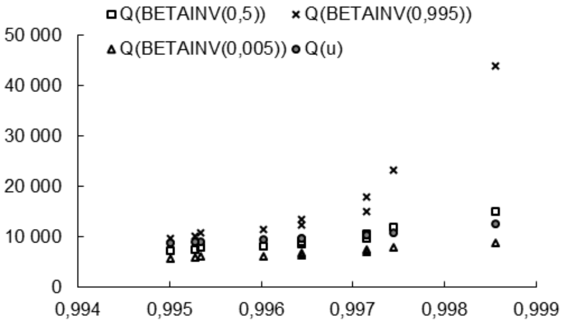

Fig. 7 Graphical result of simulation of 10 the largest claim amounts

On the Fig. 7 we can see simulated 10 values of $x = Q(u)$ and the quantiles $x_{0,5}$, $x_{0,995}$, $x_{0,005}$ of the order statistics $X_{(1000)}, X_{(999)}, \ldots, X_{(991)}$. Quantiles $x_{0,005}$ and $x_{0,995}$ give the bounds for 10 largest values which the Pareto distributed claim amounts would exceed with probability only 0,01.

Simulation of *p* the largest claim amounts in non-life insurance portfolio is useful in non-proportional reinsurance of the types of LCR(p), when insurance company cedes *p* the largest claim amounts to reinsurer, and ECOMOR, when reinsurance company pay losses that exceed *p-th* largest value in decreasing sequence of the claim amounts [4].

VIII. Conclusion

Probability models of claim amounts create the basis for solving of many substantial problems in general insurance practice. When trying to fit a distribution to claims data using traditional classic methods and classic loss distribution such as gamma, lognormal, Weibull and Pareto, as in this article, there are many situations where classical parametric distributions may not be appropriate to model claims. Mixture distributions [10], [20] or quantile models [9] might be good tools in such cases.

In the article we mentioned more examples of using probability models of claim amounts in general insurance. We consider should be emphasized that the individual amount is one of the components of the total claim amount [14]. The total amount of claims in a particular time period is a fundamental importance to the proper management of an insurance company [3], [11], [17]. The key assumption in various models for this total or aggregate claim amount is that we know distribution describing the claim amounts together with the model for events occurring in time.

References

[1] O. M. Achieng, Actuarial Modeling for Insurance Claim Severity in Motor Comprehensive Policy Using Industrial Statistical Distributions. [Online]. Available: http://www.actuaries.org/EVENTS/Congresses/Cape_Town/Papers/Non-Life%20Insurance%20(ASTIN)/22_final%20paper_Oyugi.pdf

[2] P. J Boland, *Statistical and Probabilistic Methods in Actuarial Science*. London: Chapman&Hall/CRC, 2007.

[3] A. Brdar Turk, A Quantitative Operational Risk Management Model, *Wseas Transactions on Business and Economics*, Issue 5, Volume 6, 2009, pp. 241-253.

[4] T. Cipra, *Reinsurance and Risk Transfer in Insurance* (Zajištění a přenos rizik v pojišťovnictví). Praha: Grada Publishing, 2004, ch. 11

[5] I. D. Currie, *Loss Distributions*. London and Edinburgh: Institute of Actuaries and Faculty of Actuaries, 1993.

[6] W. G. Gilchrist, *Statistical Modelling with Quantile Functions*, Chapman & Hall/CRC, London 2000.

[7] R. J. Gray, S. M. Pitts, *Riska Modelling in General Insurance*. Cambridge University Press, 2012, ch. 2.

[8] R. V. Hogg, S. A. Klugman, *Loss Distributions*. New York: John Wiley & Sons, 1984.

[9] P. Jindrová, Ľ. Sipková, Statistical Tools for Modeling Claim Severity. In: *European Financial Systems 2014*. Proceedings of the 11th International Scientific Conference. Lednice, June 12-13, 2014. Brno: Masaryk University, 2014, pp. 288-294.

[10] R. Kaas, M. Goovaerts, J. Dhaene, M. Denuit, *Modern Actuarial Risk Theory*. Boston: Kluwer Academic Publishers, 2001.

[11] M. Käärik, M. Umbleja, On claim size fitting and rough estimation of risk premiums based on Estonian traffic insurance example, *International Journal of Mathematical Models and Methods in Applied Sciences*, Issue 1, Volume 5, 2011, pp. 17-24.

[12] V. Pacáková, *Applied Insurance Statistics* (Aplikovaná poistná štatistika). Bratislava: Iura Edition, 2004.

[13] V. Pacáková, B. Linda, Simulations of Extreme Losses in Non-Life Insurance. *E+M Economics and Management*, ročník XII, 4/2009.

[14] V. Pacáková, Modelling and Simulation in Non-life Insurance. *Proceedings of the 5th International Conference on Applied Mathematics, Simulation, Modelling* (ASM'11), Corfu Island, Greece, July 14-16, 2011, Published by WSEAS Press, p. 129-133.

[15] V. Pacáková, J. Gogola, Pareto Distribution in Insurance and Reinsurance. Conference proceedings from 9th international scientific conference *Financial Management of Firms and Financial Institutions*, VŠB Ostrava, 2013. pp. 298-306.

[16] V. Pacáková, D. Brebera, Loss Distributions in Insurance Risk management. In: *Recent Advances on Economics and Business Administration*, Proceedings of the International Conference on Economics and Statistics (ES 2015), INASE Conference, Vienna, March 15-17, 2015. pp. 17-22.

[17] Přečková, L., Asymmetry of information during the application of the model for valuation the sum insured in case of business interruption in the Czech Republic, *International Journal of Mathematical Models and Methods in Applied Sciences*, Issue 1, Volume 5, 2011, pp. 212-219.

[18] Probability Distributions, On-line Manuals, StatPoint, Inc., 2005.

[19] H. Schmitter, Estimating property excess of loss risk premiums by means of Pareto model, Swiss Re, Zürich 1997. [Online]. Available http://www.kochpublishing.ch/data/2000_pareto_0007.pdf

[20] Y. K. Tse, *Nonlife Actuarial Models*. Cambridge: University Press, 2009.

Prof. RNDr. Viera Pacáková, Ph.D. graduated in Econometrics (1970) at Comenius University in Bratislava, 1978 - RNDr. in Probability and Mathematical Statistics at the same university, degree Ph.D at University of Economics in Bratislava in 1986, associate prof. in Quantitative Methods in Economics in 1998 and professor in Econometrics and Operation Research at University of Economics in Bratislava in 2006. She was working at Department of Statistics Faculty of Economic Informatics, University of Economics in Bratislava since 1970 to January 2011. At the present she has been working at Faculty of Economics and Administration in Pardubice since 2005. She has been concentrated on actuarial science and management of financial risks since 1994 in connection with the actuarial education in the Slovak and Czech Republic and she has been achieved considerable results in this area.

Mgr. David Brebera graduated in Mathematics in 1999 at Mathematics and Physics Faculty of Charles University. He was working as the senior actuary in Pojistovna Ceske sporitelny, a. s. (Insurance Company of the Czech Savings Bank), since 1999 to 2006. At the present he has been working at Faculty of Economics and Administration in Pardubice since 2005. He has been concentrated on statistics, actuarial science and management of financial risks.