Loss Distributions in Insurance Risk Management

V. Packová, D. Brebera

Abstract—Probability modelling has a wide range of applications in the field of insurance. An improvement of methods for reducing of actuarial risk in insurance company is effective tool for insurance risk management. While the risk assessment of insurance company in connection with her solvency is a complex and comprehensive problem, its solution starts with statistical modelling of number and amounts of individual claims. The objective of this article is to present possibilities how to obtain appropriate probability model that adequately describe the insurance losses and how to use such the model for the purposes of risk management. Modern computer techniques and statistical software open up a wide field of practical applications for this aim. The article includes application of presented methods based on data of claim amounts in motor third-party liability insurance.

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

I. Introduction

Although the empirical distribution functions can be useful tools in understanding claims data, 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 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 [2]. 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. [3, 4, 5, 6].

The conditions under which claims are performed (and data are collected) allow us to consider the claim amounts in non-life insurance branches to be samples from specific, very often heavy-tailed probability distributions. As a probability models for clam sizes we will understand probability models of the financial losses which can be suffered by individuals and disbursed 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" [6]. 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.

II. Claim Amounts Modelling Process

We will concerned 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 package.

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 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 is 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 salient features. This involved finding the mean, median, standard deviation, coefficient of variance, skewnes and kurtosis. This was done using Statgraphics Centurion XV package.

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).

ISBN: 978-1-61804-293-4

The Distribution Fitting procedure of this software fits any of 45 probability distributions (7 for discrete and 38 for continuous random variables) to a column of numeric data represented random sample from the selected distribution. Distributions selected for our analysis aαe defined in Statgraphics Centurion as follow [7].

Gamma Distribution
Probability density function (PDF)

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

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

Lognormal Distribution
Probability density function (PDF)

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

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

Weibull Distribution
Probability density function (PDF) α β

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

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

A good tool when selecting 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 latter selection usually gives a smoother result, with the desirable value of h depending on the size of the data sample. Therefore in the application we will use Cosine function

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

B. Parameters Estimation

We will use the method of Maximum Likelihood (ML) to estimate the parameters of the selected loss distribution. 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: Let $\mathbf{x}=(x_1, x_2, ..., x_n)$ be a vector of n independent observations taken from a population with PDF $f(x;\boldsymbol{\Theta})$, where $\boldsymbol{\Theta}'=(\Theta_1, \Theta_2, ..., \Theta_p)$ is a vector of p unknown parameters. Define the likelihood function $L(\boldsymbol{\Theta};\mathbf{x})$ by

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

The ML estimate $\hat{\boldsymbol{\Theta}}=\hat{\boldsymbol{\Theta}}(\mathbf{x})$ is that value of $\boldsymbol{\Theta}$ which maximises $L(\boldsymbol{\Theta};\mathbf{x})$.

C. *Goodness of Fit Tests*

Various tests may be used to assess the fit of a proposed model. For all tests, the hypotheses of interest are:

H_0: data are independent samples from the specified distribution,

H_1: data are not independent samples from the specified distribution.

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

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:

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

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 in Reinsurance

Modelling of the tail of the loss distributions in non-life 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.

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 [8].

The Pareto distribution function of the losses X_a that exceed known deductible a is

$$F_a(x) = 1 - \left(\frac{a}{x}\right)^b, \quad x \ge a \tag{10}$$

The density function can be written

$$f_a(x) = \frac{b \cdot a^b}{x^{b+1}}, \quad x \ge a \tag{11}$$

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 [9, 10].

Let losses above this *OP*

$$X_{OP,1}, X_{OP,2}, ..., X_{OP,n}$$

be independent identically Pareto distributed random variables with distribution function

$$F_{OP}(x) = 1 - \left(\frac{OP}{x}\right)^b, \qquad x \ge OP \tag{12}$$

The maximum likelihood estimation of Pareto parameter b is given by formula

$$\frac{n}{\sum_{i=1}^{n} \ln\left(\frac{X_{OP,i}}{OP}\right)} \tag{13}$$

The Pareto distribution expressed by (10) 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

$$LF(a) = LF(OP) \cdot P(X_{OP} \rangle\, a) = LF(OP) \cdot \left(\frac{OP}{a}\right)^b \tag{14}$$

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

$$RP = LF\left(a\right) \cdot EXL \tag{15}$$

where

$$EXL = E\left(X_a\right) = \frac{a \cdot b}{b-1}, \qquad b > 1 \tag{16}$$

IV. Application of the Theoretical Results

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 1352 claim amounts (in thousands of Czech crowns - CZK) from the portfolio of 26 125 policyholders in compulsory motor third-party liability insurance.

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

Table 1 Summary statistics for X

Count	1352
Average	1376,29
Median	996,0
Standard deviation	1705,32
Coeff. of variation	123,907%
Minimum	1,0
Maximum	24986,7
Skewness	5,0977
Kurtosis	42,7794

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 for *X* in Fig. 2 can be concluded that loss distribution in our case is skew and long or heavy tailed.

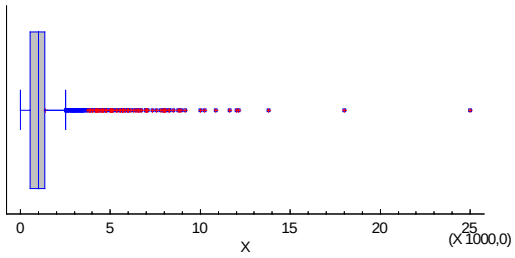

Fig. 1 Box-and-Whisker plot of claim amounts data

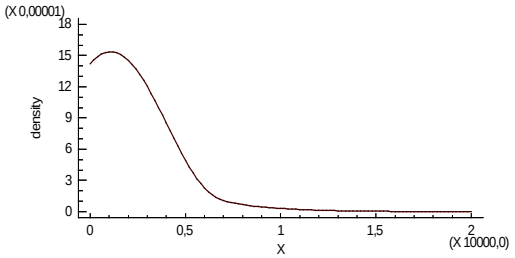

Fig. 2 Density Trace for *X*

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.

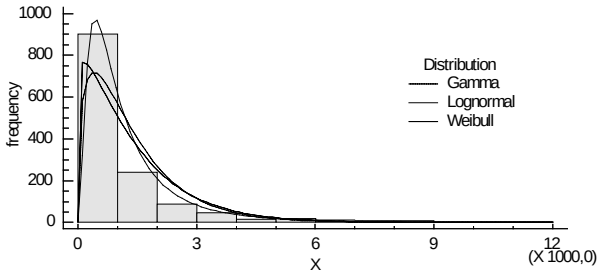

Fig. 3 Histogram and estimated loss distributions

From Fig. 3 it seems that lognormal distribution follows the data best, it is also suitable for both small and large claims. It is hard to compare the tail fit, but clearly the all distributions have high discrepancies at middle claims intervals.

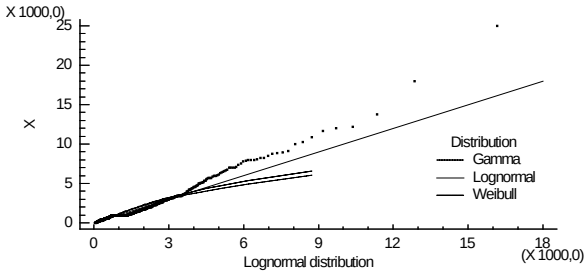

Fig. 4 Quantile-Quantile plot of selected distributions

The Quantile-Quantile (Q-Q) plot shows the fraction of observations at or below X plotted versus the equivalent percentiles of the fitted distributions. One selected distribution, in our case lognormal, is used to define the X-axis and is represented by the diagonal line. The others are represented by curves.

In Fig. 4 the fitted lognormal distribution has been used to define the X-axis. 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. Unfortunately, all selected 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.

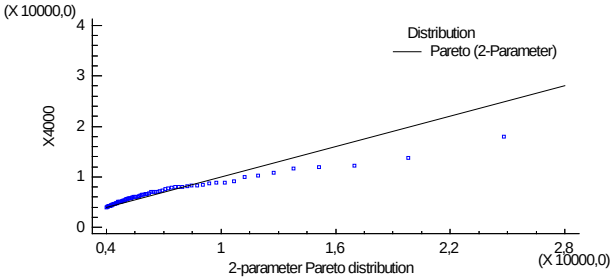

Fig. 5 Quantile-Quantile plot for Pareto distribution of X_{4000}

In the Fig. 5 the fitted Pareto distribution has been used to define the X-axis. The fact that the points lie close to the diagonal line confirms the fact that this distribution provides a good model for the clam amounts data above 4 million CZK.

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.

Table 2 Estimated parameters of the fitted distributions

Gamma	*Lognormal*	*Weibull*
shape = 1,41869	mean = 1355,69	shape = 1,0931
scale = 0,001031	Std. Dev. = 1438,37	scale = 1433,38
	Log mean = 6,83502	
	Log std. dev. = 0,868387	

The estimated parameters of the fitted distributions are shown in Table 2.

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

	Gamma	*Lognormal*	*Weibull*
A^2	45,5961	21,8625	53,1055
Modified Form	45,5961	21,8625	53,1055
P-Value	<0.01	<0.01	<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

ISBN: 978-1-61804-293-4

lognormal distribution with parameters estimated by ML. Since the smallest P-value is less than 0,01, we can reject the hypothesis that X comes from a lognormal distribution with 99% confidence.

Table 4 Chi-Squared test with lognormal distribution

	Lower Limit	*Upper Limit*	*Observed Frequency*	*Expected Frequency*	*Chi-Squared*
below		500,0	303	321,07	1,02
	500,0	3000,0	930	911,03	0,40
	3000,0	5500,0	72	92,41	4,51
	5500,0	8000,0	30	18,57	7,03
	8000,0	10500,0	10	5,38	3,98
above	10500,0		7	3,55	3,36

Chi-Squared = 20,2961 with 3 d.f. P-Value = 0,000147369

Table 4 confirmed the poor fit with the lognormal distribution especially for the claim amounts more than 3 million KCZ and Fig. 4 for the claim amounts more than 4 million KCZ. By Fig. 5 we can assume that a good model for losses above 4 million KCZ can be Pareto distribution with PDF expressed by the formula (11).

Table 5 Estimated parameters by ML

Pareto (2-Parameter)
shape = 2,07701
lower threshold = 4000,0

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

	Pareto (2-Parameter)
DPLUS	0,0589548
DMINUS	0,128705
DN	0,128705
P-Value	0,172365

There is 74 values ranging from 4000 to 24986,7 thousand KCZ. Table 5 and Table 6 show the results of fitting a 2-parameter Pareto distribution to the data on X_{4000}. The estimated parameters of the fitted distribution are shownin Table 5. The results of K-S test whether the 2-parameter Pareto distribution fits the data adequately contain Table 6. Since the P-value = 0,172365 is greater than 0,05, we cannot reject the hypothesis that sampling values of the variable X_{4000} comes from a 2-parameter Pareto distribution with 95% confidence.

Suppose the insurance company wants to reduce technical risk by non-proportional XL reinsurance with priority (deductible) a = 10 000 (thousand KCZ). Pareto distribution for variable X_{4000} with parameters from Table 5 we will use to determine reinsurance risk premium by (15). As OP we put value 4000. To calculate $LF(a)$ by (14) we need to know $P(X_{OP} \rangle a)$. We can use Tail Areas pane for the fitted 2-parameter Pareto distribution. It 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 10000 for the fitted 2-parameter Pareto distribution of X_{4000} is 0,149099, as we can see in Table 7. The value of $LF(OP)$ we can estimate by relative frequency of the losses above 4000:

$$LF(OP) = \frac{74}{1352} = 0,054734$$

Table 7 Tail Area for X_{4000} Pareto (2-Parameter) distribution

X	*Lower Tail Area (<)*	*Upper Tail Area (>)*
10000,0	0,850901	**0,149099**

Then by (14) we get

$$LF(a) = LF(OP) \cdot P(X_{OP} \rangle a) = 0,0547 \cdot 0,1491 = 0,007712$$

So that we can use the formulas (15) and (16) to calculate the reinsurance premium RP we will fit the 9 values ranging from 10000,0 to 24986,7 of the variable X10000 by Pareto (2-Parameter) distribution using Statgraphics Centurion Goodnes of fit procedure. The rsults contain the Table 8 and Table 9.

Table 8 Estimated parameters by ML

Pareto (2-Parameter)
shape = 3,65624
lower threshold = 10000,0

Table 9 K-S goodness-of-fit tests for X_{10000}

	Pareto (2-Parameter)
DPLUS	0,157301
DMINUS	0,105537
DN	0,157301
P-Value	0,979142

Table 9 shows the results of tests run to determine whether X_{10000} can be adequately modelled by a 2-parameter Pareto distribution with ML estimated parameters in Table 8. Since the P-value = 0,979142 is greater than 0,05, we cannot reject the hypothesis that values of X_{10000} comes from the 2-parameter Pareto distribution with 95% confidence.
You can also assess visually how well the 2-parameter Pareto distribution fits by selecting Q-Q graph on Fig. 6.

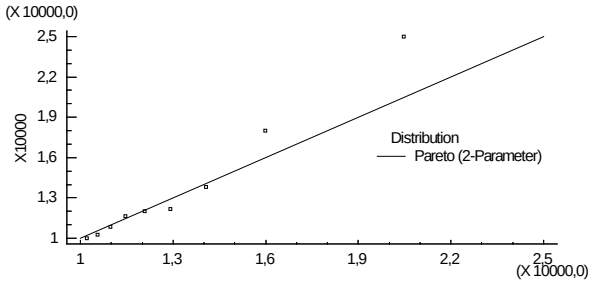

Fig. 6 Quantile-Quantile plot for Pareto distribution of X_{10000}

By values of estimated parameters in Table 8 we can calculate

$$E(X_a) = \frac{a \cdot b}{b-1} = \frac{1000 \cdot 3,65624}{2,65624} = 13764,72$$

Then by formula (15) we get reinsurance premium in thousand KCZ:

$$RP = LF(a) \cdot EXL = 0,007712 \cdot 13764,72 = 106,16.$$

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

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

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

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

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

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

[8] 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

[9] 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.

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

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 top Mathematician in Insurance Company of the Czech Savings Bank since 1999 to 2006. At the present she has been working at Faculty of Economics and Administration in Pardubice since 2005. She has been concentrated on statistics, actuarial science and management of financial risks.