PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1088 — 2005
Inwestycje finansowe i ubezpieczenia – tendencje światowe a polski rynek

Viera Pacáková, Erik Šoltés
University of Economics in Bratislava

SIMULATIONS OF INSURANCE LOSSES USING PARETO QUANTILE FUNCTION

1. Introduction

The conditions under which claims in non-life insurance are performed allow us to consider the claim amounts to be samples from specific heavy-tailed probability distributions.

Pareto distribution is often used as a model for insurance losses. This paper describes its very good properties for modelling of loss distribution using quantile functions and simulations of the largest losses. Process of modelling and simulation has been illustrated on the sample of observed claim size data in accident insurance. We apply the process of modelling that was presented in our last year article [Pacáková 2004, pp. 91-99].

2. Pareto distribution of insurance losses

We will assume that individual claim amounts are drawn from a particular distribution, called a *loss distribution.* The aim is to describe the variation in claim amounts by finding a loss distribution that adequately describes the claims that actually occur.

The simplest of the loss distributions is exponential distribution with cumulative distribution function (CDF)

$$F(x) = 1 - e^{-\lambda x}, \qquad \lambda > 0.$$

The tail probability

$$P(X > x) = 1 - F(x) = e^{-\lambda x}$$

goes to zero exponentially in x. We would need a distribution with more weight in the upper tail to correct this problem. We can slow the tail down by asking that the tail go to zero as a power of x. For example we can set

$$P(X > x) = 1 - F(x) = \left(\frac{\lambda}{\lambda + x}\right)^{\alpha},$$

which is the upper tail of the Pareto distribution that is often used as a model for insurance losses when also a very large losses would be observed.

A random variable X has the Pareto distribution with parameters α and λ if CDF is

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

and probability density function

$$f(x) = \frac{dF(x)}{dx} = \frac{\alpha\lambda^{\alpha}}{(\lambda + x)^{\alpha+1}}, \ x > 0, \ \alpha > 0, \ \lambda > 0. \quad (2)$$

The quantile function QF, denoted by $Q(p)$, expresses p-quantile x_p as a function of p:

$x_p = Q(p)$ is the value of x for which $p = P(X \le x_p) = F(x_p)$.

The definitions of the QF and the CDF can by written for any pairs of values (x, p) as $x = Q(p)$ and $p = F(x)$. These functions are thus simple inverses of each other, provided that they are both continuous increasing functions. Thus we can also write $Q(p) = F^{-1}(p)$.

When X has a Pareto distribution, from (1) we can find

$$Q(p) = \lambda.(1 - p)^{-1/\alpha} - \lambda. \quad (3)$$

The basic statistics are

$$E(X) = \frac{\lambda}{\alpha - 1}, \ \alpha > 1$$

$$D(X) = \frac{\alpha\lambda^2}{(\alpha - 1)^2(\alpha - 2)}, \ \alpha > 2.$$

The Pareto distribution has two parameters α and λ. The method of moments of point estimation of parameters α and λ is very easy to apply. We equate the first two population and sample moments and we find

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

Eliminating λ we obtain

$$\tilde{\alpha} = \frac{2s^2}{s^2 - \bar{x}^2} \text{ and } \tilde{\lambda} = (\tilde{\alpha} - 1)\,\bar{x}. \quad (4)$$

The estimates obtained in this way will tend to have rather large standard errors, because s^2 has a very large variance. We will obtain estimates of α and λ using maximum likelihood method.

Let $\hat{\alpha}$ and $\hat{\lambda}$ be the maximum likelihood estimates given data $x_1, x_2, ..., x_n$ from the Pareto distribution $P(\alpha, \lambda)$. From $f(x; \alpha, \lambda)$ the likelihood is

$$L(\alpha, \lambda, \mathbf{x}) = \prod_{i=1}^{n} \frac{\alpha \lambda^{\alpha}}{(\lambda + x_i)^{\alpha+1}},$$

with *ln* likelihood

$$l(\alpha, \lambda, x) = \sum_{i=1}^{n} \left[\ln \alpha + \alpha \ln \lambda - (\alpha+1) \ln(\lambda + x_i) \right] =$$

$$= n \ln \alpha + n\alpha \ln \lambda - (\alpha+1) \sum \ln(\lambda + x_i).$$

Setting $\frac{\partial l}{\partial \alpha} = 0,\ \frac{\partial l}{\partial \lambda} = 0$ and solving them for α we find two expressions for $\hat{\alpha}$:

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \ln\left(1 + \frac{x_i}{\lambda}\right)} \quad \text{or} \quad \hat{\alpha} = \frac{\sum_{i=1}^{n} \frac{1}{\lambda + x_i}}{\sum_{i=1}^{n} \frac{x_i}{\lambda(\lambda + x_i)}}. \tag{5}$$

We equate the two expressions for $\hat{\alpha}$ and find that $\hat{\lambda}$ satisfies $f(\lambda) = 0$, where

$$f(\lambda) = \frac{\sum \frac{1}{\lambda + x_i}}{\sum \frac{x_i}{\lambda(\lambda + x_i)}} - \frac{n}{\sum \ln\left(1 + \frac{x_i}{\lambda}\right)}. \tag{6}$$

Illustrative example

We introduce the illustrative example of modelling losses by Pareto model. We have observed sample of values of 91 individual claims from accident insurance policies. Using χ^2 goodness-of-fit test we will test of H_0: the data come from a Pareto model $Pa(\alpha, \lambda)$.

We start by estimating of α and λ parameters using the method of moments. With help of the equations (4) the method of moments gives

$$\tilde{\alpha} = 2{,}617 \quad \text{and} \quad \tilde{\lambda} = 73\ 163{,}76.$$

The method of moments does provide initial estimates for more efficient maximum likelihood method. Substituting the initial estimate $\tilde{\lambda} = 73\ 163{,}76$ into nonlinear equation (6) and solving it with help of Solver function of table processor Excel we find the maximum likelihood estimator $\hat{\lambda} = 37\ 277{,}81375$. Substituting

$\hat{\lambda}$ into expressions (5) we find $\hat{\alpha} = 1{,}739399006$ in both cases. The value of $\hat{\alpha}$ may be surprising, since the population variance does not exist in this case because of $\hat{\alpha} < 2$. This simply says that the Pareto distribution selected by ML has very heavy tail.

Table 1 contains the procedure of χ^2 goodness-of-fit test for Pareto models with parameters chosen by two different methods. We will denote as O_i observed frequencies, as E_1 the expected frequencies under Pareto model with parameters estimated by the method of moments and as E_2 the excepted frequencies under Pareto model with parameters estimated by maximum likelihood method.

Table 1. Observed and fitted values for the Pareto model

Lower limit	Upper Limit	O_i	E_1	E_2
at or below	19 000	46	40	40
19 000	37 000	17	19	17
37 000	55 000	9	10	9
55 000	91 000	8	10	8
91 000	160 000	6	7	6
above	160 000	5	5	5
		91	**1,197859**	**0,058203**

Source: own calculations.

The χ^2 statistic is computed as usual:

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}.$$

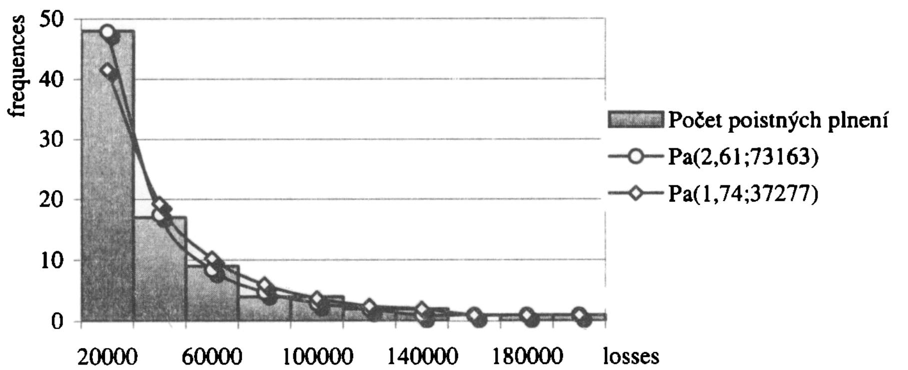

Fig. 1. Pareto loss distributions

We find $\chi^2 = 1{,}1978$ for method of moments and $\chi^2 = 0{,}582$ for ML method of parameters estimation. To compare computed values of χ^2 with quantile $\chi^2_{0,99;3} = 11{,}3449$, we can see the both of Pareto models give a good fit, although, as expected, the ML fit is better as assessed by the χ^2 value. The improvement in fit is evident also from Fig. 1.

3. The quantile models of the order statistics

We denoted a set of ordered sampling data of losses by

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

The corresponding random variables being denotes by

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

Thus $X_{(n)}$ for example is the random variable representing the largest observation of the sample of n. The n random variables are referred to as the *n order statistics*. These statistics play a major role in modeling with quantile-defined distributions.

Consider first the distribution of the largest observations $X_{(n)}$ with CDF denoted by $F_{(n)}(x) = p_{(n)}$. The probability

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

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

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

Inverting the $F(x)$ to get the quantile function, we have

$$Q_{(n)}\left(p_{(n)}\right) = Q\left(p_{(n)}^{1/n}\right). \quad (7)$$

If X has a Pareto distribution with CDF (1) then the quantile function of the largest observation is

$$Q_{(n)}(p) = \lambda\left(1 - p^{1/n}\right)^{-1/\alpha} - \lambda. \quad (8)$$

The distribution will have a median value of $Q_{(n)}(0{,}5) = \lambda\left(1 - 0{,}5^{1/n}\right)^{-1/\alpha} - \lambda$.

For the general *r-th* order statistic $X_{(r)}$ we explain in [Pacáková 2004, pp. 91-99] the order statistics distribution rule:

If a sample of n observations from a distribution with quantile function $Q(p)$ is ordered, then the quantile function of the distribution of the *r-th* order statistic is given by

$$Q_{(r)}\left(p_{(r)}\right) = Q\left(BETAINV\left(p_{(r)}, r, n - r + 1\right)\right). \quad (9)$$

$INVBETA(.)$ 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.

A particularly useful application of this result lies in evaluating the medians of the distributions of ordered data. Thus, the median M_r of the distribution of the *r-th* ordered data is $M_r = Q(INVBETA(0{,}5, r, n-r+1))$. This we will term the *median rankit.* A plot of $x_{(r)}$ versus M_r will give a straight line through the origin for a correct and well-fitted model. This plot can form the basis of model validation.

4. Demonstration of the simulation of the largest observations

Table 2 contain the results of simulation of the 20 largest observations in sample of 1000 Pareto distributed losses with quantile function (3) step by step by our last year article [Pacáková 2004].

Table 2. Simulation of 20 the largest from 1000 losses

v	n	$1/n$	$v^{1/n}$	u	$Q(u)$
0,135493	1000	0,0010000	0,9980032	**0,9980032**	**1291697,514**
0,331321	999	0,0010010	0,9988948	**0,9969002**	**994804,452**
0,253843	998	0,0010020	0,9986272	**0,9955316**	**799110,676**
0,993465	997	0,0010030	0,9999934	**0,9955251**	**798406,978**
0,180922	996	0,0010040	0,9982849	**0,9938177**	**656697,334**
0,997123	995	0,0010050	0,9999971	**0,9938148**	**656511,689**
0,855881	994	0,0010060	0,9998434	**0,9936592**	**646673,091**
0,919813	993	0,0010070	0,9999158	**0,9935756**	**641539,780**
0,943984	992	0,0010081	0,9999419	**0,9935178**	**638057,160**
0,761040	991	0,0010091	0,9997245	**0,9932441**	**622188,114**
0,865165	990	0,0010101	0,9998537	**0,9930988**	**614169,640**
0,561498	989	0,0010111	0,9994166	**0,9925194**	**584666,674**
0,436941	988	0,0010121	0,9991623	**0,9916880**	**548102,839**
0,068052	987	0,0010132	0,9972808	**0,9889915**	**460784,822**
0,198585	986	0,0010142	0,9983619	**0,9873713**	**422982,371**
0,905523	985	0,0010152	0,9998993	**0,9872719**	**420910,870**
0,130303	984	0,0010163	0,9979311	**0,9852293**	**383336,723**
0,624701	983	0,0010173	0,9995215	**0,9847579**	**375807,461**
0,648640	982	0,0010183	0,9995593	**0,9843239**	**369193,331**
0,554228	981	0,0010194	0,9993986	**0,9837319**	**360622,519**

Source: own calculations.

Column denoted by v contains 20 values v_i, generated from uniform distribution under interval $\langle 0;1 \rangle$ not ordered in any way. Column u includes values of transformed variables

$$u_{(n)} = v_n^{1/n},$$

$$u_{(n-1)} = \left(v_{n-1}\right)^{\frac{1}{n-1}} . u_{(n)},$$

$$u_{(n-2)} = \left(v_{n-2}\right)^{\frac{1}{n-2}} . u_{(n-1)},$$

$$\vdots$$

where the v_i, $i = n, n-1, n-2, \ldots$, are generated values from the column v. The values $u_{(i)}$ form an ordering sequence from a uniform distribution.

The order statistics for the largest observations on X in column denoted by $Q(u)$ are then simulated using the uniform transformation rule and Pareto quantile function (3) with maximum likelihood estimated parameters $\hat{\alpha} = 1{,}739399006$ and $\hat{\lambda} = 37277{,}81375$ as

$$x_{(n)} = Q\left(u_{(n)}\right),$$

$$x_{(n-1)} = Q\left(u_{(n-1)}\right),$$

$$x_{(n-2)} = Q\left(u_{(n-2)}\right),$$

$$\vdots$$

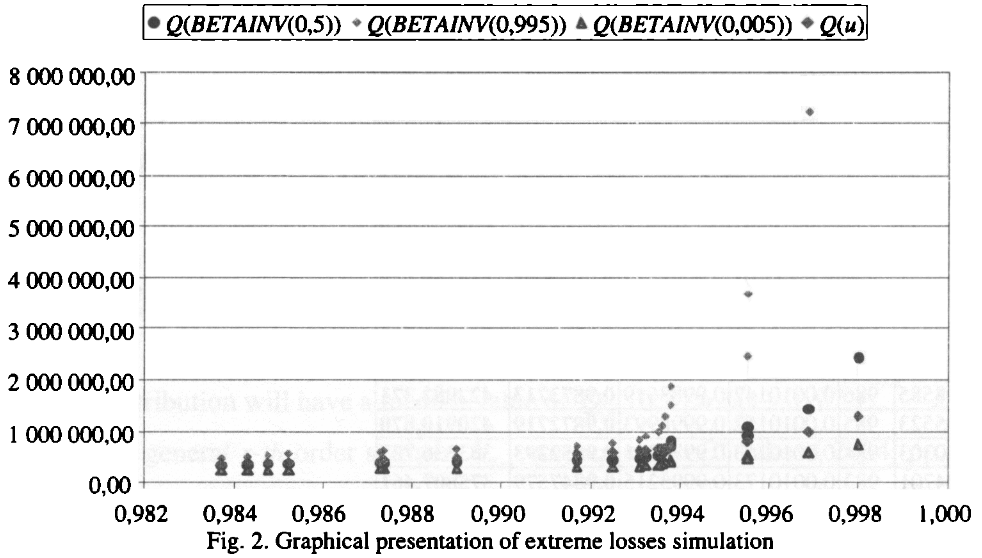

Fig. 2. Graphical presentation of extreme losses simulation

The quantile function thus provides the natural way to simulate values for those distributions for which it is an explicit function of p. As we have seen it is possible to simulate the observations in upper tail without simulating the central values.

On the Figure 2 we can see except simulated values of $Q(u)$ also the quantiles of the order statistics $X_{(981)}, X_{(982)}, ..., X_{(999)}, X_{(1000)}$ for probabilities $p = 0,5$ (median rankits), $p = 0,005$ and $p = 0,995$. Quantiles for $p = 0,005$ and $p = 0,995$ give the bounds which the 20 largest observations of 1000 Pareto distributed losses would exceed with probability only 0,01.

Literature

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

Gilchrist W.G., *Modelling with Quantile Distribution Functions*, „Journal of Applied Statistics" 1997, vol. 24, nr 1, pp. 113-122.

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

Pacáková V., Sodomová E., *Modelling with Quantile Distribution Functions,* „Ekonomika a Informatika" 2003, nr 1, pp. 30-44.

Pacáková V., Šoltés E., *Modelling and Simulations of the Extreme Losses Using Quantile Functions*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1037, AE, Wrocław 2004, pp. 91-99.

Sipková Ľ., *Zovšeobecnené lambda rozdelenie a odhad jeho parametrov*, „Ekonomika a Informatika" 2004, nr 1, pp. 107-128.

SYMULACJA STRAT W UBEZPIECZENIACH Z ZASTOSOWANIEM FUNKCJI KWANTYLOWYCH ROZKŁADU PARETO

Streszczenie

Artykuł dotyczy zastosowania rozkładu Pareto w modelowaniu strat w ubezpieczeniach. Autorzy opisują właściwości tego rozkładu, szczególnie na potrzeby modelowania strat z zastosowaniem funkcji kwantylowych. Rozważania teoretyczne są zilustrowane przykładem empirycznym z zakresu ubezpieczeń wypadkowych.