Central European Review of Economic Issues
EKONOMICKÁ REVUE

Fitting probability distributions to market risk and insurance risk

Kateřina ZELINKOVÁ*

Department of Management, Faculty of Economics, VSB – Technical University of Ostrava, Sokolská 33, Ostrava, Czech Republic.

Abstract

Determining the parametric VaR approach is very important in establishing the probability distribution of a risk factor. We assume that a normal distribution is symmetric; however, it has some limitations. This distribution is used for modelling asymmetric data or data that have only positive values, such as insurance claims. The aim of the paper is to find the best probability distribution for stock exchange index returns and for insurance claims. The paper is structured as follows. Firstly, we describe the typical probability distributions used in finance, namely normal, Student, logistic, gamma, exponential and lognormal distribution, and the methods of verification. Subsequently, parameters of the distribution types are estimated via the maximum likelihood method, and after that we calculate the value at risk. The VaR is calculated even though the time series do not correspond to the stated types of probability distribution; nevertheless, we calculate the value at risk for all the stated types of probability distribution because it is apparent that large mistake can arise if an incorrect type of probability distribution is used.

Keywords

Exponential distribution, Gamma distribution, Kolmogorov–Smirnov test, Logistic distribution, QQ plot, Goodness-of-fit tests.

JEL Classification: C19, G21, G22

* katerina.zelinkova@vsb.cz

This paper has been developed in the framework of the financial support of the student grant of the Faculty of Economics, Technical University of Ostrava, in project no. SP2015/93.

ISSN 1212-3951 (Print), 1805-9481 (Online)
ER-CEREI, Volume 18: 167–173 (2015).
doi: 10.7327/cerei.2015.12.03

Fitting probability distributions to market risk and insurance risk

Katerina ZELINKOVÁ

1. Introduction

In market risk management, it is frequently necessary to fit probability distributions to risk factors and portfolios for descriptive, predictive, explicative or simulation purposes. For example, risk measures such as value at risk require a probability model for the return distribution. The specified distribution is important because it describes the potential behaviour of the risk factor in the future. Crucial to the determination of the parametric value at risk is the probability distribution of returns. This requires the fitting of an appropriate probability distribution to the data. Usually, as directed by the Basel Committee, a normal distribution is assumed, but Fama (1965) stated that empirical revenues do not fit normal distribution.

The value at risk (VaR) is a risk measure representing a value of loss that will not be exceeded over a given risk horizon at a certain significance level. It is also possible to refer to the VaR as a methodology for managing risk that is applied widely to model credit, operational, market and insurance risk (Alexander, 2008b; Hull, 2007; Jorion, 2007; Morgan, 1996).

The aim of the paper is to find the best probability distribution for stock exchange index returns and insurance claims. We will use the daily returns data of the CAC 40 indices from 1 March 1990 to 31 December 2012, which have positive and negative random variables. We will use data containing the claims of individual policyholders within motor hull insurance during the year 2009.

2. Types of probability distribution and methods of verification

Distributions of probability have three basic characteristics. The first characteristic is location, which indicates where the distribution is on a line and consists of the mean, mode and median. The next is a scale that indicates how the scores are spread around the central point; this consists of variance and standard deviation. The last is shape, which shows us how the distribution is skewed and describes how peaked or flat the distribution is; this consists of skewness and kurtosis. Not every distribution of probability has all three basic characteristics. In this chapter, we present selected types of probability and methods of verification.

2.1 Types of probability distribution

The described types of probability (Alexander, 2008a; Beirlant et al., 2004; Lewis, 2003; McNeil et al., 2005) are used for random variables that have positive and negative values, for example revenues, or only positives values, for example insurance claims.

The normal distribution is well known and the most-used distribution of probability. One reason for its popularity is the central limit theorem, which states that, under mild conditions, the mean of a large number of random variables independently drawn from the same distribution is distributed approximately normally, irrespective of the form of the original distribution. Thus, physical quantities that are expected to be the sum of many independent processes (such as measurement errors) often have a distribution that is very close to normal.

For continuous random variables for which $-\infty \leq x \leq \infty$, the probability density function of the normal distribution is the following:

$$\phi(X) = \frac{1}{\sqrt{2\pi\sigma}} exp\left[-\frac{(x-\mu)}{2\sigma}\right], \qquad (1)$$

where μ is the mean and σ is the standard deviation. The basic characteristics of the normal distribution are as follows:

- Mean = μ,
- Standard deviation σ,
- Skewness = 0,
- Kurtosis = 3.

If $\mu = 0$ and $\sigma = 1$, the distribution is called standard normal or unit normal distribution.

In probability and statistics, *Student's t-distribution* (or simply the *t*-distribution) is a family of continuous probability distributions discovered by William S. Gosset in 1908. Gosset was a statistician employed by the Guinness brewing company, which had stipulated that he could not publish under his own name. He therefore wrote under the pen name *Student*. The Student distribution is closely related to the normal distribution; it is a symmetric curve that converges to the standard normal density as the degrees of freedom (v) increase. If $v \geq 30$, we consider it to be normal distribution. The degrees of freedom are the only parameter in the Student *t*-distribution and the lower the degrees of freedom, the lower the peak of the distribution and the

longer the tails. For a continuous random variable for which $-\infty \le x \le \infty$, the density function for the Student *t*-distribution with *v* degrees of freedom is

$$f_v(t) = (v\pi)^{\frac{-1}{2}} \Gamma\left(\frac{v}{2}\right)^{-1} \Gamma\left(\frac{v+1}{2}\right)\left(1+\frac{t^2}{v}\right)^{-\left(\frac{v+1}{2}\right)}, \quad (2)$$

where the gamma function Γ is an extension of the factorial function n! to non-integer values. The distribution has zero expectation and zero skewness, and for v > 2, the variance of a Student *t*-distributed variable is $\sigma = v(v-2)$.

For continuous random variables for which $-\infty \le x \le \infty$, the probability density function of the *logistic distribution* is given by

$$f(x) = \frac{\exp\left(\frac{x-\alpha}{c}\right)}{c\left[1+\exp\left(\frac{x-\alpha}{c}\right)\right]}, \quad (3)$$

where $c = \sqrt{3}\frac{\beta}{\pi}$. Note that $c > 0$ and we can interpret α as the mean and β as the standard deviation. The formulas for each characteristic of the logistic distribution are:

- Mean = α,
- Standard deviation = β,
- Skewness = 0,
- Kurtosis = 4,2.

The *gamma distribution* is a two-parameter family of continuous probability distributions. This distribution is for continuous random variables for which $0 \le x \le \infty$, and the probability density function of the gamma distribution is the following:

$$f(x) = \left(\frac{x}{\beta}\right) \cdot \frac{\exp\left(-\frac{x}{\beta}\right)}{\alpha\Gamma(\xi)}, \quad (4)$$

where β is a scale parameter and ξ is a shape parameter – $\Gamma(\xi)$ is the gamma function given by $\Gamma(\xi) = \int_0^\infty exp(-u)u^{\xi-1}du$. Both parameter β and parameter ξ will be positive values. The formulas for each characteristic of the gamma distribution are:

- Mean = $\beta \cdot \xi$,
- Standard deviation = $\sqrt{\beta^2 \xi}$,
- Skewness = $2\xi^{-\frac{1}{2}}$,
- Kurtosis = $3 + \frac{6}{\xi}$.

The *exponential distributions* have only one parameter. This type of probability is for continuous random variables for which $0 \le x \le \infty$, and the probability density function of the exponential distribution is given by

$$f(x) = \frac{1}{\beta}\exp\left(-\frac{x}{\beta}\right), \quad (5)$$

where β is a scale parameter:

- Mean = β,
- Standard deviation = β,
- Skewness = 2,
- Kurtosis = 9.

For random variables for which $0 \le x \le \infty$, the probability density function of the *lognormal distribution* is given by

$$f(x) = \frac{1}{x\xi\sqrt{2\pi}}\exp\left(\frac{-\left[\log\left(\frac{x}{\alpha}\right)\right]}{2\xi}\right). \quad (6)$$

In this case, we can directly interpret α as the median and ξ as the shape parameter. The formulas for each characteristic of the lognormal distribution are

- Mean = $\alpha exp\left(\frac{1}{2}\beta^2\right)$,
- Standard deviation = $\alpha\sqrt{\left(c^2 - c\right)}$,
- Skewness = $(c+2)\sqrt{(c-1)}$,
- Kurtosis = $c^4 + 2c^2 + 3c^2 - 3$.

Further details concerning lognormal distribution can be found in Aitchison and Brown (1957).

2.2 Methods of verification

We can determinate the suitable type of probability via two kinds of methods. The first methods are graphic, namely the *QQ* plot, *PP* plot and histogram. The next methods are goodness-of-fit tests, namely the Kolmogorov–Smirnov and the Shapiro–Wilk test, which measure the compatibility of a random sample with a theoretical probability distribution function. In other words, these tests show how well the distribution fits the data.

The Kolmogorov–Smirnov test (Hendl, 2004), also referred to as the Kolmogorov–Smirnov *D* test or Kolmogorov–Smirnov Z test, is a non-parametric test for the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution. The Kolmogorov–Smirnov statistic quantifies the distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distri-

bution. The null distribution of this statistic is calculated under the null hypothesis that the sample is drawn from the reference distribution. In each case, the distributions considered under the null hypothesis are continuous distributions but are otherwise unrestricted. We use this test if the sample is greater than 2 000.

H_0: The sample has distribution function $F_0(x)$;

H_1: The sample does not have distribution function $F_0(x)$.

The test statistic is the following:

$$KS = \max\left|F(x) - G(x)\right|, \quad (7)$$

where $F(x)$ is the empirical distribution function of the sample and $G(x)$ is the distribution function consideration of the probability distribution of the population (when comparing the two samples $G(x)$ is the empirical distribution function of the second sample).

The null hypothesis is valid if the *p*-value is greater than α, and we reject H_0 and accept the alternative hypothesis H_1 if the *p*-value is less than α. The *p*-value is a number between 0 and 1 and is defined as the probability of obtaining a result equal to or *more extreme* than what was actually observed. Further details of the *K-S* test can be found in Paramasamy (1992).

A *Q-Q plot* is a plot of the percentiles (or quantiles) of a standard normal distribution (or any other specific distribution) against the corresponding percentiles of the observed data. If the observations follow approximately a normal distribution, the resulting plot should be roughly a straight line with a positive slope. If the *QQ* plot has a normal distribution, we call it a rank graph.

3. Results

In this part, we will determinate the suitable type of probability distribution and subsequently we will estimate the value at risk for the stated type distribution of probability.

We use data of daily CAC 40 log returns and data containing the claims of individual policyholders within motor hull insurance during the year 2009. The CAC 40 log returns are from 1 March 1993 to 31 December 2012.

Firstly, the goodness-of-fit test, namely the Kolmogorov–Smirnov test, and the *QQ* plot will be provided.

Subsequently, parameters will be estimated via maximum likelihood in SPSS, and the value at risk for the CAC 40 index will be determined. These values are positive and negative and therefore we will use types of probability distribution for continuous random variables for which $-\infty \le x \le \infty$, namely normal, Student and logistic distribution.

After that, the value at risk for the sample that has only positives values will be determined. Therefore, we will use probability distribution types for continuous random variables for which $0 \le x \le \infty$, namely gamma, lognormal and exponential distribution.

The estimation of the value at risk will be performed by Monte Carlo simulation with 10 000 scenarios. The process for estimating VaR by Monte Carlo simulation is as follows. First, on the basis of the estimated parameters' stated distribution functions, random numbers in the interval (0–1) are generated. These random numbers are transformed via appropriate inverse distribution functions of marginal distributions into random returns of individual assets. The simulated returns of portfolios are sorted into ascending order from the smallest to the largest loss and the VaR is determined as a quantile of probability distribution (i.e. as a loss in the order of $10\,000\alpha$). The VaR is computed for one day and α = 1%. Further details of the Monte Carlo simulation method can be found in Deepak and Ramanathan (2009).

Some statistical characteristics of our data sample, namely the mean, standard deviation, skewness and kurtosis, are shown in Table 1.

Table 1 Basic statistical characteristics

	Index CAC 40	*Insurance claims*
Mean	0.0122%	56 441 CZK
Standard deviation	1.4319 %	96 534 CZK
Skewness	4.45913	1 403
Kurtosis	–0.02292	29

3.1 Goodness-of-fit tests

In this part, the *K-S* test is performed for all the stated types of distribution. We try to ascertain whether random variables belong to the distribution containing the determined distribution function. The results are presented in Table 2.

Table 2 Kolmogorov–Smirnov test

Type of distribution	*P-value*
Normal	0.0000
Student	0.0650
Logistic	0.0210
Gamma	0.0580
Lognormal	0.0187
Exponential	0.0000

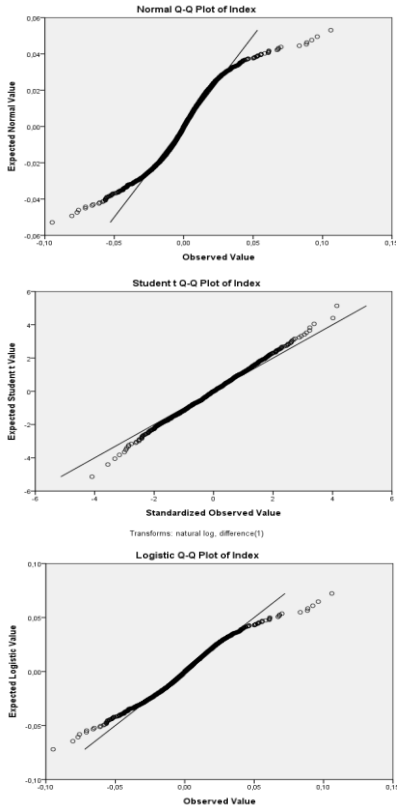

Figure 1 *QQ* plots of distributions for a random variable with $-\infty \le x \le \infty$

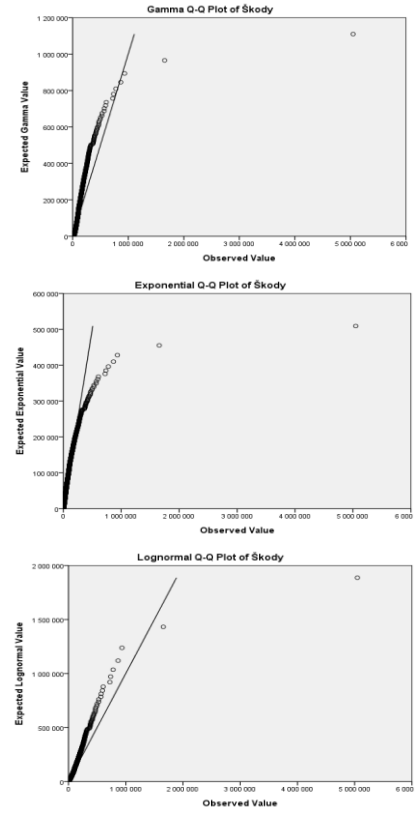

Figure 2 *QQ* plots of distributions for a random variable with $0 \le x \le \infty$

On the basis of the *K-S* test, the Student and gamma distributions are statistically significant because the *p*-value is greater than 0.05. The logistic and lognormal distributions are statistically significant for $\alpha = 0.01$ too. Thus, we accept the null hypothesis by assuming $\alpha = 0.05$ for the Student and gamma distributions and $\alpha = 0.01$ for the logistic and lognormal distributions. This means that the distribution function fits the Student and gamma empirical distribution functions for 5% probability and the logistic and lognormal distributions for 1% probability.

Figure 1 presents the *QQ* plots for the particular probabilities for the normal, Student and logistic distributions. We can see that the normal distribution does not correspond to the empirical data and the most suitable distribution is the Student distribution, because the plotted points for the Student distribution are markedly more linear than those for the normal *QQ* plot and logistic *QQ* plot.

Figure 2 provides the *QQ* plots for particular probabilities for gamma, lognormal and exponential distribution. On the basis of Figure 2, we can conclude that the data sample is characterized by a heavier fat tail than the theoretical distributions. The most suitable is gamma distribution, because the plotted points are markedly more linear than those for the lognormal *QQ* plot or the exponential *QQ* plot.

From Figure 2 we can conclude that the data sample has the characteristic of a heavier fat tail than the theoretical distributions. The most suitable distribution is gamma distribution because the plotted points are markedly more linear than those in the lognormal *QQ* plot or exponential *QQ* plot.

3.2 Results for market risk

The log returns have positive and negative values; therefore, we use the probability distribution types for a random variable for which $-\infty \le x \le \infty$, namely the normal, Student *t* and logistic distributions. The parameter estimation for the given types of probability are presented in Table 3.

Now we estimate the value at risk for market risk for the 1% significance level and a one-day risk horizon. The value of the first percentile of the data is 4.07%, which is a non-parametric estimate.

We can see in Table 4 that the value at risk is different for the given distributions. In the normal QQ plot of the normal distribution or in the result of the K-

Table 3 Parameter estimation of the normal, Student and logistic distributions

Type of distribution	*Location*	*Scale*	*Degree of freedom*
Normal	0.000123	0.01432	–
Student – *t*	0.000123	0.01432	8.17
Logistic	0.000123	0.008	–

S test, it is apparent that insufficient probability is given to extreme events and thus the value at risk may be underestimated. This is indeed the case, as the first percentile of the fitted normal distribution is 3.19%, which is less than the non-parametric estimation. The logistic and Student *t*-distributions offer a fatter-tailed alternative to the normal distribution. The Student *t*-distribution appears to fit the body and tail of the data much better than the normal distribution.

Table 4 Value at risk for market risk

	$VaR_{1\%}$
Normal	3.19 %
Student – *t*	4.45 %
Logistic	4.84 %

3.3 Results for insurance risk

The insurance claims have only positives values; therefore, we use the probability distribution types for a random variable for which $0 \le x \le \infty$, namely gamma, lognormal and exponential distribution. The parameter estimation for the given distributions of probability are shown in Table 5.

Table 5 Parameter estimation of gamma, lognormal and exponential distributions

Type of distribution	*Location*	*Shape*
Gamma	6.06E–06	0.342
Lognormal	33792.76	1.096
Exponential	56 440.97	

Now we estimate the value at risk for insurance risk for the 0.5% significance level and a one-day risk horizon. The non-parametric estimate is 279 051.2 CZK. The value at risk for the difference distribution is presented in Table 6.

Table 6 Value at risk for insurance risk in CZK

	$VaR_{1\%}$
Gamma	282 385.96
Lognormal	302 618.71
Exponential	299 042.18

The amount of capital required to cover unexpected losses at the 0.5% significance level is 282 385 CZK in the case of gamma distribution. The gamma distribution estimate is close to but higher than the non-parametric estimate.

4 Conclusion

The aim of paper was to find the best probability distribution for stock exchange index returns and insurance claims.

In this paper, the basic types of probability distribution, namely the normal, Student, logistic, gamma, lognormal and exponential distributions, and the methods of fitting probability (Kolmogorov–Smirnov test and *QQ* plot) were introduced. In the application part, we estimated the parameters of given distributions, stated *QQ* plots and determined the value at risk. We used data of CAC 40 log returns and data containing the claims of individual policyholders within motor hull insurance.

These distributions are frequently used for financial data. In market risk management, it is necessary to fit probability distributions to risk factors and portfolios, for example to determine capital requirements in banks or insurance institutions. For the determination of parametric measurement, value at risk is a crucial probability distribution of data. In most cases, we assume the normal distribution, but this does not take into account the fat tails of distribution. This assumption can lead to underestimation or overestimation of the value at risk. Thus, it is important to deal with fitting probability.

References

AITCHISON, J., BROWN, J. A. C. (1957). The lognormal distribution. *The Economic Journal* 67(268): 713–715. http://dx.doi.org/10.2307/2227716

ALEXANDER, C. (2008a). *Quantitave Methods in Finance*. Chichester: Wiley.

ALEXANDER, C. (2008b). *Value at Risk Models*. Chichester: Wiley.

BEIRLANT, J., GOEGEBEUR, Y. SEGERS J., TEUGELS J. (2004). *Statistics of Extremes: Theory and Applications.* Chichester: Wiley. http://dx.doi.org/10.1002/0470012382

DEEPAK, J., RAMANATHAN, T. (2009). Parametric and nonparametric estimation of value at risk. *The Journal of Risk Model Validation* 3(1): 51–71.

FAMA, E. F. (1965). The behavior of stock-market prices. *The Journal of Business* 38(1): 34–105. http://dx.doi.org/10.1086/294743

HENDL, J. (2004). *Přehled statistických metod*. Praha: Portál.

HULL, J. (2007). *Risk Management and Financial Institutions*. New Jersey: Pearson Education.

JORION, P. (2007). *Value at Risk: The New Benchmark for Managing Financial Risk*. 2. ed. New York: McGraw-Hill.

LEWIS, N. D. C. (2003). *Market Risk Modelling*. London: Risk Books.

McNEIL, A., J., FREDY, R. EMBRECHTS, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools.* Princeton: Princeton University Press.

PARAMASAMY, S. (1992). On the multivariate Kolmogorov-Smirnov distribution. *Statistics and Probability Letters* 15(2): 149–155. http://dx.doi.org/10.1016/0167-7152(92)90128-R

Additional sources

MORGAN, J. P. (1996). *RiskMeasures^{TM} Technical Document*. [Online] Available at: http://yats.free.fr/ papers/td4e.pdf [cited 2015-12-12].