Technical University of Varna Annual Journal

https://doi.org/10.29114/ajtuv.vol5.iss1.226

Vol.5 Issue 1 (2021)

Published: 2021-06-09

ISSN 2603-316X (Online)

Identification of Statistical Distribution Type by Sample Data

Ventsislav Nikolov, *Aleksandar Krastev*, and *Snezhina Yanakieva*

1 – Technical University of Varna, Computer Science and Engineering Department, 9010, 1 Studentska Street, Varna, Bulgaria

Corresponding author contact: v.nikolov@tu-varna.bg

Abstract. *The present paper provides a full description of a software implementation of a system for statistical distributions. Such a system is almost indispensable in many simulation applications where the factors incorporated adhere to a specific non-normal distribution. The realization is developed as a software library that can be integrated in different other applications. There is also the possibility for additional theoretical distribution types to be added.*

Keywords: statistical distribution, software library, histogram, simulation

1 Purpose

The main goal of the statistical distributions identification software modules is an automatic determination of theoretical distribution type and its parameters for a given data sample. This is done by a software module developed by the authors of the paper hereto proposed and which is currently being used in Monte Carlo simulations with generation of future data with distribution and parameters according to the available historical data (Robert, 1999). Thus, the most suitable modeling of empirical histogram is achieved by exact reproduction of the identified theoretical distribution type.

The existing software products have the disadvantage of being mainly mathematical general purpose systems (Ricci, 2005). The system presented here is developed as a library in Java programming language and as such it can be integrated in different types of other software systems that can call its functionalities either directly or as services (Josuttis, 2007).

2 Implementation

The input data for the software module involves the data sample whose distribution is to be determined and settings like a list of standard distributions types that will be checked against the selected data sample. Examples of such distribution types are: Beta, Cauchy, Student, Weibull, etc. (Krishnamoorthy, 2006). The result of the module as output data is identified distribution type that best fits the presented data sample and the specific distribution parameters. Computed, additionally, is the numerical distance between the histograms of the empirical data sample and theoretical distributions, so that the sample can be assigned to some distribution type but taking into account the observations or measurements in the sample, another similar distribution types might also to be perceived.

Some of the most often used statistical distribution is the Normal distribution and often data generation in different simulation tasks is done using this distribution type. However, this assumption is not always correct since the best fitting theoretical distributions are likely to be skewed which, in turn, leads to the underestimation of the simulation accuracy. To improve the simulation accuracy the best fitting distribution identification must be done by identification and usage of non-normal distributions. Presented in fig.1 is an example of distribution where a given confidence level must be found and the distribution type in case of simulation is very important.

Technical University of Varna
Annual Journal
https://doi.org/10.29114/ajtuv.vol5.iss1.226
Vol.5 Issue 1 (2021)
Published: 2021-06-09
ISSN 2603-316X (Online)

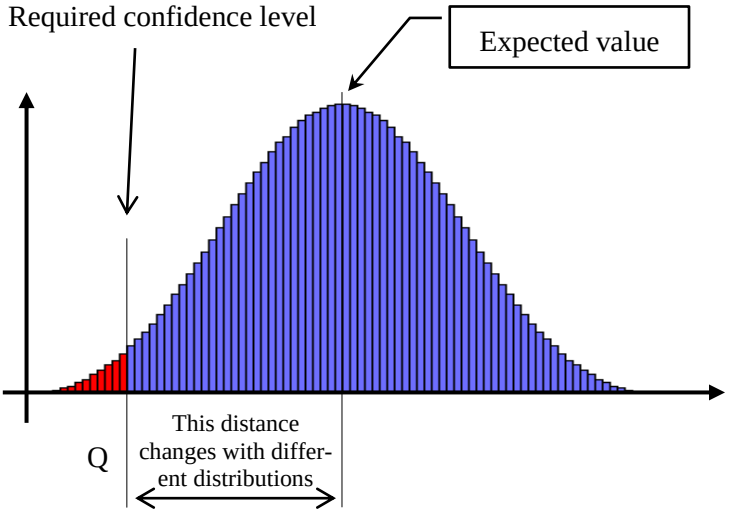

Fig. 1 An example of normal distribution with changes in its confidence levels when other distribution type is in use.

The simulation accuracy is even more important when the data sample is ordered time series and correlations between a set of such series are taken into account. Situation like this arises in value-at-risk (VaR) estimation by Monte Carlo simulation with numerous factors exerting an enormous influence over the result (Mun, 2006). Correct distribution type that can be non-normal significantly improves the accuracy of the results.

In order to identify the most suitable distribution type, the following steps are performed:

1) The histogram of the empirical data sample is calculated;
2) Theoretical probability density function is computed in the following way:
   2.1) First the specific distribution parameters are estimated from the data sample;
   2.2) The theoretical distribution is built using estimated parameters;
   2.3) The distance between the empirical histogram and theoretical density function is measured. This distance is largely determined by the chosen criterion which, in the present case, is the Euclidean distance between the distribution histogram bins.
3) The selected theoretical distribution types are ordered ascending according to the distances the best fitting distribution is considered and stored for later usage in the simulation.

Illustrated in fig. 2 is an example of a non-normal distribution that fits the sample data histogram better than the normal distribution.

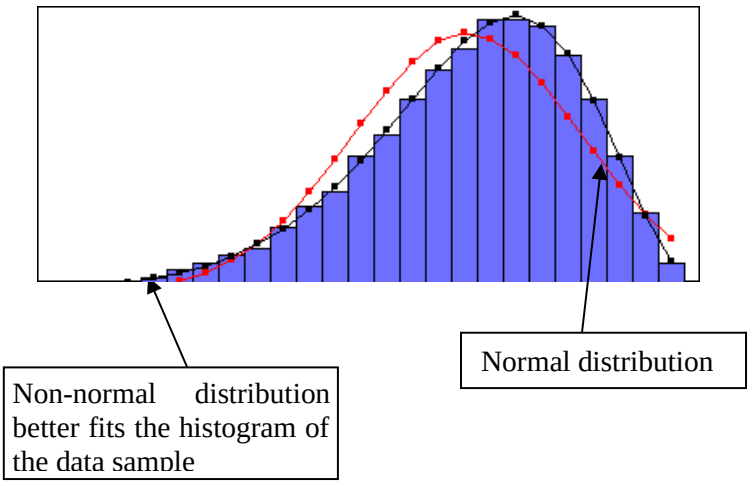

Fig. 2 Comparison of normal and non-normal distributions to an empirical histogram.

Technical University of Varna
Annual Journal
https://doi.org/10.29114/ajtuv.vol5.iss1.226
Vol.5 Issue 1 (2021)
Published: 2021-06-09
ISSN 2603-316X (Online)

.

Goodness-of-fit test are designed to determine whether a given data sample follows a specific distribution type. Employed, to this effect, is a technique like "chi squared", Anderson-Darling, Kolmogorov-Smirnov, etc. (Cameron, 1998; Draper, 1998). Here the best fitting distribution type to a given sample data is determined by comparing the histogram of the sample data (empirical histogram) and the probability density function (PDF) values (theoretical histogram) for all the selected theoretical distribution types – fig. 3. The distance is calculated as average squared distance between the histogram bin frequencies of empirical and theoretical histograms:

$$d^2 = \frac{1}{k}\sum_{i=1}^{k}(O_i - E_i)^2 \quad (1)$$

where k is the number of the histogram bins; O_i is the value of the theoretical PDF corresponding to the bin i from the data sample histogram; E_i is the number of values in bin *i* from the data sample histogram.

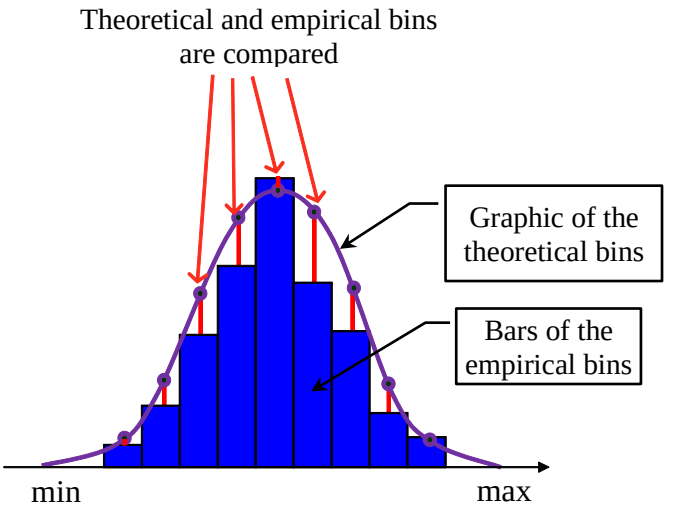

Fig. 3 Comparing sample data histogram with the theoretical distribution histogram.

The suggested technique can be considered as modified "chi squared" goodness-of-fit test (Bury, 1999). Accordingly, the empirical data sample histogram should be calculated as well as the theoretical distribution histograms. The empirical histogram is obtained from the available sample data and the theoretical histogram is calculated by the Cumulative Distribution Function (CDF) formula for every chosen theoretical distribution type. Therefore, for all the selected theoretical distribution types, the distribution parameters should be preliminary estimated in order for the CDF formula to be applied. Table 1 lists the theoretical distribution types built upon in our module along with their parameters (Balakrishnan, 2003).

The distribution parameters can be classified into two categories:

- **Specific parameters related to the distribution**. For each distribution type specific parameters are estimated using different techniques according to their computational effectiveness.
- **Parameters for sample transformation**. It is possible for the sample data values to be such so that they do not allow for the distribution formulas to be defined. In such a case the sample is transformed by shift and/or scale.

Technical University of Varna
Annual Journal
https://doi.org/10.29114/ajtuv.vol5.iss1.226
Vol.5 Issue 1 (2021)
Published: 2021-06-09
ISSN 2603-316X (Online)

Table 1. Distribution parameters and transformation.

Distribution	*Specific parameters*		*Transformation*	
	First parameper	*Second parameter*	*Shift*	*Scale*
Beta	α	β	yes	yes
Cauchy	x_0	γ	no	no
Exponential	λ	N/A	yes	no
Gumbel	μ	β	no	no
Gamma	k	θ	yes	no
Inverse Normal	μ	λ	yes	no
Log Normal	m	σ	yes	no
Logistic	μ	s	no	no
Maxwell-Boltzmann	a	N/A	yes	no
Normal	μ	σ	yes	no
Pareto	x_m	α	no	no
Pearson type VII	m	α	no	no
Rayleigh	σ	N/A	yes	no
Student	ν	N/A	yes	yes
Weibull	λ	k	yes	no

When all the parameters are estimated, the theoretical histogram is calculated. In this regard, use is made of the known CDF (Vallentin, 2011):

$$F_x(x) = P(X <= x) \tag{2}$$

where P is the probability for the variable X to contain a value that is smaller than or equal to x.

The probability density function (PDF) considered as a theoretical histogram is obtained from the distribution with known CDF – fig. 4.

$$P(y < X <= z) = F_x(y) - F_x(z) \tag{3}$$

where $F_x(x)$ is the CDF of the selected distribution.

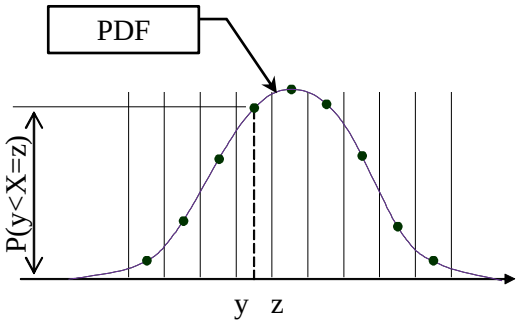

Fig. 4 Probability density function of a theoretical distribution represented as a histogram with bins.

The values of the data sample are considered one by one and they are classified to the histogram bins. The number of values in a given bin determines the bins height. Obtained, thereby, is the empirical histogram. The values that are on a border between two bins are half referred to the bins from the

Technical University of Varna
Annual Journal
https://doi.org/10.29114/ajtuv.vol5.iss1.226
Vol.5 Issue 1 (2021)
Published: 2021-06-09
ISSN 2603-316X (Online)

both sides of the border. The number of bins is chosen by a setting in the module and it can be either fixed or varying according to the size of the sample data.

For validation purposes a sample generator is also developed which generates values with a chosen distribution type. This is done by generating values with equal probability in interval from 0 to 1 either random or equally distanced one from the other.

3 Generating values in the simulation stage

In the simulation stage, the correlated factors are identified and for each factor the respective distribution type is identified and its CDF is saved. Generated, then, for every such factor is a huge number of values obeying the corresponding best fitting distribution by means of the CDF of the identified distribution. This allows for the accuracy to be significantly improved compared to the hypothesis that all factors are normally distributed. The process of sample values generation in the simulation stage using stored CDF is shown in fig. 5.

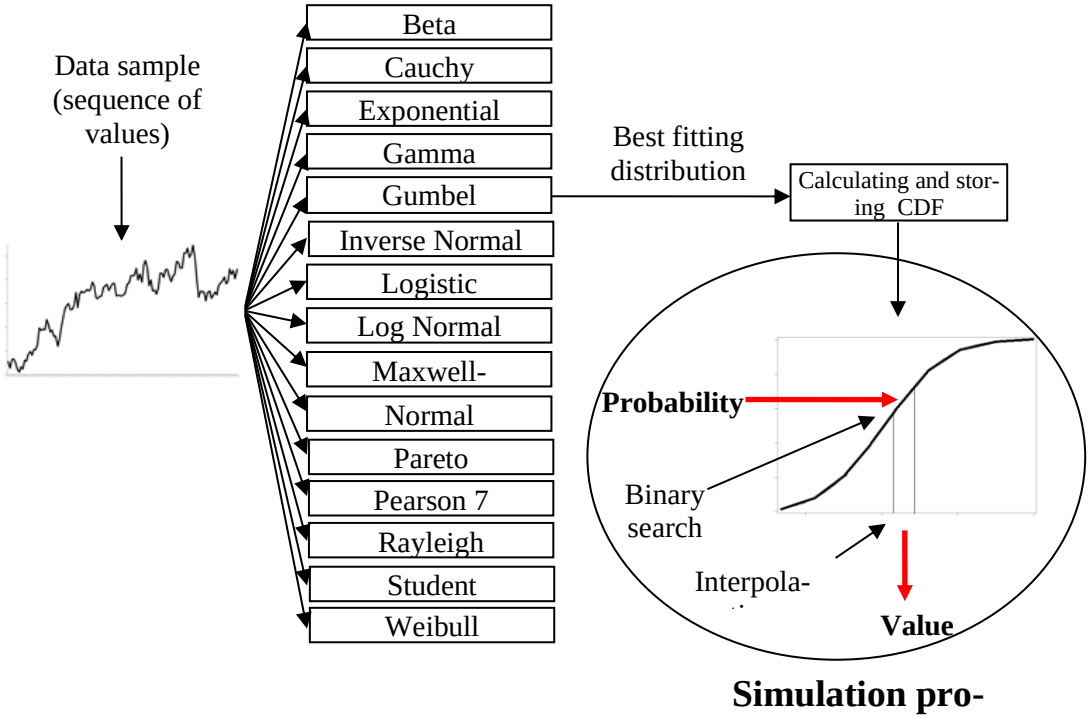

Fig. 5 Using the best fitting distribution type in the process of simulation.

Keeping the CDF of the best fitting distribution enables its use in the simulation stage. As the CDF values for the bins are sorted ascending, the probability is transformed into a sample value by a binary search (Sedgewick, 1998) in the CDF line. When the bin is established, the exact value is calculated by interpolation between its left and right border.

4 Conclusion

The developed software module was tested with both randomly generated sample data with different size and by uniformly generated values. Sometimes the suggested distribution type in not correctly identified as this happens more often if the data sample is relatively small. Such phenomenon occurs in the other known statistical distribution identification systems as well. The module is usable and integrated in other systems for Monte Carlo simulation even more when the distribution type is important for accuracy improvement. Identified, currently, are 15 theoretical distribution types but in view of the scalable software architecture of the module supporting various practical design patterns, other theoretical distribution types can be added as well. The developed software module performs:

Technical University of Varna
Annual Journal
https://doi.org/10.29114/ajtuv.vol5.iss1.226
Vol.5 Issue 1 (2021)
Published: 2021-06-09
ISSN 2603-316X (Online)

- Identification of the best fitting theoretical distribution and the values of its parameters;
- Comparison of empirical and theoretical histograms;
- Implementation of the best distribution for simulation purposes.

In fig. 6 part of the graphical user interface of the prototype system is shown, with built-in software library, where the generated empirical and several theoretical histograms can be seen.

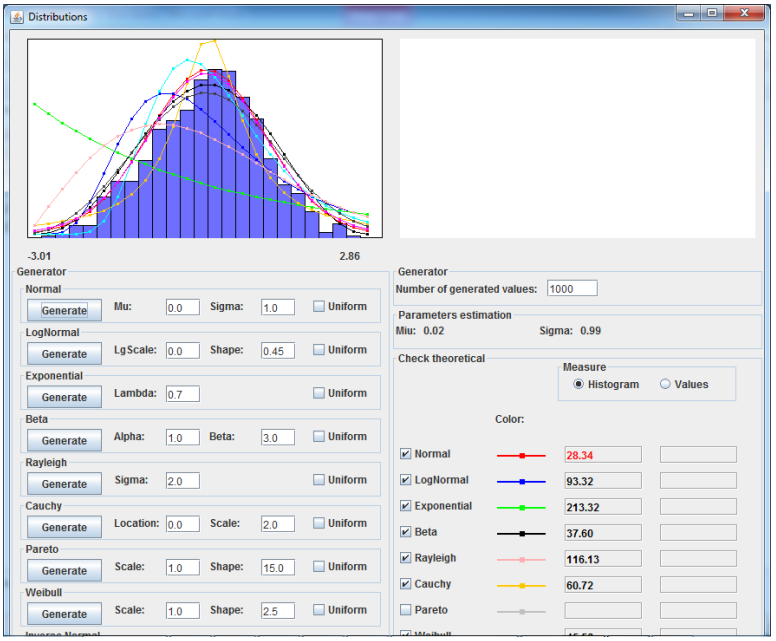

Fig. 6 Realized prototype using the distribution type identification software module

References

Balakrishnan, N., & Nevzorov, V. B. (2003). A primer on statistical distributions. Hoboken, New Jersey: A John Wiley & Sons.

Bury, K. (1999) Statistical Distributions in engineering. Cambridge University Press.

Cameron, A. C., & Trivedi, P. K. (2013). Regression analysis of count data (Vol. 53). Cambridge university press.

Draper, N. R., & Smith, H. (1998). Applied regression analysis (Vol. 326). John Wiley & Sons.

Josuttis, N. M. (2007). SOA in practice: the art of distributed system design. " O'Reilly Media, Inc.".

Krishnamoorthy, K. (2006) Handbook of statistical distributions with applications. Chapman & Hall.

Mun, J. (2006). Modeling risk: applying Monte Carlo simulation, real options analysis, forecasting, and optimization techniques (Vol. 347). John Wiley & Sons.

Ricci, V. (2005). Fitting distributions with R. Contributed Documentation available on CRAN, 96.

Robert, C., G. Casella (1999) Monte Carlo statistical methods. Springer-Verlag.

Sedgewick, R. (1998) Algorithms in C++, Parts 1-4: Fundamentals, Data Structure, Sorting, Searching, (3rd Ed.), Addison-Wesley.

Vallentin, M. (2011) Probability and Statistics Cookbook. Matthias Vallentin. Retrieved from Link