A System for Approximating Stochastic Data ⋆

Zdena Rábová ⋆⋆
rabova@dcse.fee.vutbr.cz

Jaromír Bumbálek⋆⋆
bumbalek@dcse.fee.vutbr.cz

Abstract: The aim of this paper is to briefly characterise our computer tool for determining the type of probability distribution which is most suitable for an approximation of a given stochastic data file. This tool is also capable of deriving parameters of probability distributions, and check the hypothesis that given data correspond to the given probability distribution. Besides, the system removes values that were distorted during their collection from a given data file.

Key Words: Approximation of stochastic data, probability distribution (PD)

1 Introduction

When we model stochastic systems, we often deal with the problem of approximation of stochastic data (representing stochastic processes in real systems). An important task is to find the optimal type of PD for the approximation of a given stochastic process. If we perform this task successfully, we get an important qualitative information about the stochastic system we are modelling.

This paper gives an outline how the previously mentioned problem can be solved. Firstly, we describe basic requirements on a system for approximating stochastic data. Then, basic functions of the system we implemented are discussed. Finally, an example is mentioned.

2 Basic Requirements

The system for approximation of stochastic data system should be able:

– to remove values that are very likely distorted from a stochastic data file (this action is based on a statistical test),
– to find a PD type of a given stochastic data file (based on maximum information that can be gained from the given data),
– to determine PD parameters of a given stochastic data file,
– to check the hypothesis ”A given stochastic data file has a particular PD with certain parameters”,
– to create files of pseudo-random numbers using built-in tested generators.

⋆ This work was supported by GAČR under the contract No. 102/98/0552 “Research and Applications of Heterogenous Models”.

⋆⋆ Department of Computer Science and Engineering, Božetěchova 2, CZ-612 66 Brno

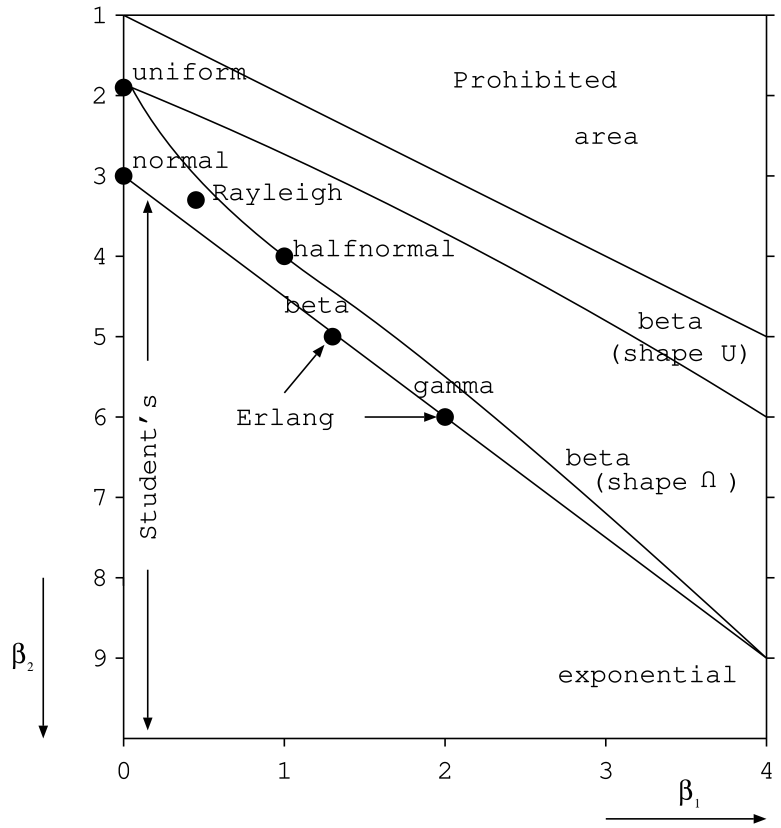

Figure 1: Space diagram (β_1, β_2).

3 Determining PD Type

When designing a suitable method, we start from basic characteristics of random magnitude which are defined by its moments. These characteristics are position, variability, skewness, and kurtosis.

For each PD type, an image is constructed. This image contains courses of the probability density function and the distribution function for chosen parameters. Moreover, an empirical histogram of 2000 values produced by random number generator of this PD type is created. Thus, images are graphical notations of Person's test results for a given file and PD type.

It is assumed that characteristic values β_1, the so-called skewness[***], and β_2, the so-called kurtosis[†], or the relation between them, are unambiguously assigned to the particular PD type. If the PD type is to be determined, the following steps have to be done. Firstly, a point estimator B_1 of the square of the skewness characteristic and a point estimator B_2 of the kurtosis characteristic of a given data sample are computed. Then, these point estimators are compared with values β_1 and β_2 of particular PD types. Selection of the optimal PD type is based on the position of the point $[B_1, B_2]$ in the space diagram (β_1, β_2). In the diagram, shown in figure 1, a considered PD is represented by:

1. a point, unless it has a shape parameter,

[***] third moment of PD
[†] fourth moment of PD

2. a countable set of points, if it has a discrete shape parameter,
3. a curve, if it has a real number parameter,
4. a surface, if it has two shape parameters.

From the previous contributions we see that those PD types which have more shape parameters are chosen more frequently. This is caused by the fact that their probability density function can be easily adjusted by suitable parameter values to empirical probability density function expressed by a histogram. The larger area PD contains in the space diagram (β_1, β_2), the bigger scale it provides.

Unfortunately, it is not possible to choose PD only according to strict equality of the value B_1 to the value β_1 and the value B_2 to the value β_2. This approach is not suitable because distributions, that are in the space diagram (β_1, β_2) assigned to points, would not be obviously chosen anytime.

According to Elderton [3], following equations are used for the calculation of the point estimator standard deviation:

$$\sigma\left(\sqrt{B_1}\right) = (6/n)^{1/2}$$

$$\sigma(B_2) = (24/n)^{1/2}$$

These equations hold exactly for normal PD. For other PDs, the accuracy increases with higher n.

Our system produces the name of the PD which could approximate a given stochastic data file with a precision specified by the user. Furthermore, the output contains the result of Pearson's test for the given PD and the processed file.

4 Calculation of PD Parameters

A few methods are available for calculating PD parameters. In our system, there is used a moment method. This method can be used for each PD that has finite moments. Appropriate equations are usually simple, only for Weibull PD it is necessary to use an iterative method.

In case that no PD type is selected, the user can exploit empirical percentiles (fifth, ninth, fifteenth).

If the user knows certain PD parameters but does not know the distribution type, he or she can enter these parameters. Sometimes this can help to calculate other PD parameters.

5 Evaluating Performed Approximation

The evaluation of statistic hypothesis that a given data file is driven by a particular PD is performed by Pearson's test. We select from following PD types: uniform, normal, half-normal, exponential, Weibull, Rayleigh, Pearson, Student's, beta, gamma, and Erlang.

6 Implementation

Our system was implemented two times, for the first time in Pascal under MS DOS [5], for the second time in C++ under MS Windows [6] using COM [4] to provide easier data transfers and possible extensions of the system. The system contains three parts: generators, testers, and approximators. The following example shows a practical usage of the system. The task was to find out a suitable PD type which approximate heights of trees within some forest area. Results are shown in table 1 and figure 2.

PD type	normal
shift	17.67 m
standard deviation	4.48 m
mean	17.62 m
skewness	2.15
kurtosis	4.01
samples	425

Table 1: Results of the approximation.

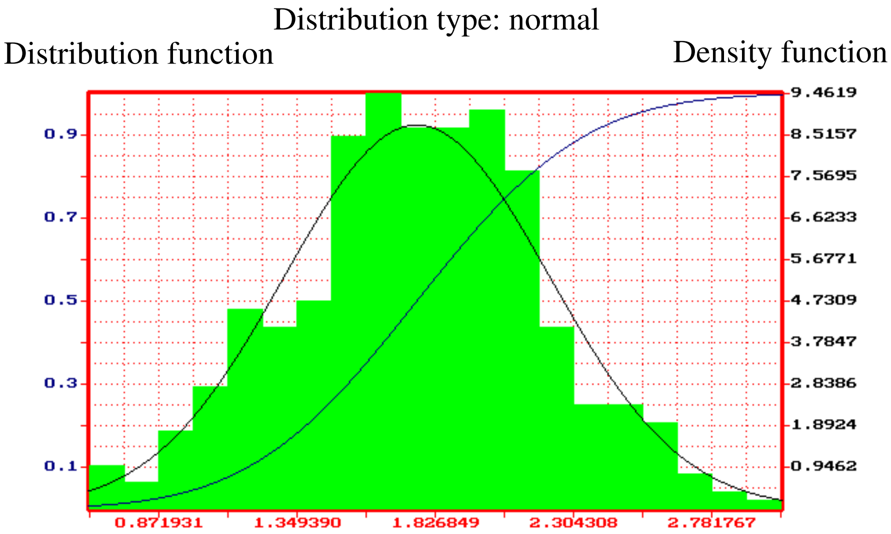

Figure 2: Graphical results of the approximation.

7 Conclusion

In this article, we present a solution of one particular problem of a stochastic data processing – the replacement of a stochastic data file by an appropriate probability distribution type including finding its parameters. The solution of this problem allow us to replace information represented with a large amount of data by a simple qualitative information about the probability distribution that drive this data. The main contribution is that this information can be easily manipulated.

References

1. Hahn, G., Shapiro, S.: Statistical Models in Engineering, John Wiley and Sons, New York, 1967
2. Johnson, N.: Systems of Frequency Curves Generated by Methods of Translations, Biometrika, vol. 36, 149
3. Elderton, W., Johnson, N.: Systems of Frequency Curves, Cambridge University Press, Cambridge, 1969
4. The Component Object Model: Technical Overview, Dr. Dobbs Journal, December, 1994
5. Ohnút, M.: Interaktivní systém pro zpracování stochastických dat, Master's thesis, TU Brno, 1991
6. Baláš, P.: Uživatelské rozhraní pro simulaci stochastických systémů, Master's thesis, TU Brno, 1995