BIOAUTOMATION, 2009, **13** (2), 27-44

Statistical Procedures for Finding Distribution Fits over Datasets with Applications in Biochemistry

Natalia D. Nikolova[1*], Daniela S. Toneva[2], Ana-Maria K. Tenekedjieva[3]

[1]*Technical University – Varna, 1 Studentska Str., 9010 Varna, Bulgaria,*
Phone /fax+359 52 383670, E-mail: natalia@dilogos.com

[2]*Technical University – Varna, 1 Studentska Str., 9010 Varna, Bulgaria,*
Phone+359 52 383725, E-mail: d_toneva@abv.bg

[3]*Duke University, North Carolina, USA*
E-mail: ana-maria.tenekedjieva@duke.edu

[*] *Corresponding author*

Received: November 12, 2009 **Accepted: February 27, 2009**

Published: August 5, 2009

Abstract: *A common problem in statistics is finding a distribution that fits to a certain dataset. Many theoretical distributions have been developed to give a good description of the empirical observations, and consequently, theory offers a variety of algorithms to test the quality of the resulting fits. It is reasonable to expect that each set of measurements should be described with the same theoretical distribution if one and the same experimental mechanism was applied. This paper presents procedures to find a theoretical distribution that best fits to several datasets. The procedure goes further, answering the questions of whether the given datasets come from the same general population, and assessing if the difference between the fitted distributions of two datasets are statistically significant. Kuiper test is used in all steps of the analysis. In two of those a Monte Carlo simulation procedure is elaborated to construct the Kuiper statistic's distribution. A platform with original program functions in MATLAB R2009a is created on the basis of the described procedures. It is applied to datasets from a biochemical experiment, which investigates the resulting density of fibrin network under different thrombin concentrations. The developed procedure has wide applications in different fields, as it models the behavior of datasets, generated through the same mechanism. The possibility to fit one type of distribution over different datasets allows comparing samples, performing interpolation and extrapolation procedures, and investigating the influence of the input conditions of an experiment over the parameters of the fitted distributions.*

Keywords: *Datasets, Distribution fits, Stair-case distributions, Kuiper test, Monte Carlo, MATLAB.*

Introduction

Statisticians often face a problem, where they have to analyze many datasets, derived in a similar way. It is possible to find a theoretical distribution that fits to each dataset. If the mechanism (experiment) to generate the samples was the same, then the distribution type that describes the datasets will also be the same. In that case, the difference between the sets will be captured not by changing the type of the distribution, but rather through change in its parameters. Finding a type of theoretical distribution that fits several datasets has several advantages. On one hand, it allows comparing the samples from the datasets and performing interpolation or extrapolation to generate new data without conducting experiments again. It is also possible to investigate how the variation of the input parameters influenced the parameters of the theoretical distribution. In some experiments it might be proven that the quantitative increase of the input parameters leads to qualitative changes in the output. In

other cases that variation may lead only to quantitative changes in the output (i.e. changes in the parameters of the distribution). Then it is of importance to investigate the statistical significance of the quantitative differences, i.e. to compare the statistical difference of the distribution parameters. In some cases it may not be possible to find a single type of distribution that fits all datasets. A possible option in these cases is to construct empirical distributions according to known techniques [8], and investigate whether the differences are statistically significant. In any case, proving that the observed difference between theoretical, or between empirical distributions, are not statistically significant allows uniting datasets and operating on larger amount of samples, which is a prerequisite for higher precision of the statistical results. This task is similar to testing for stability in regression analysis [10].

This paper approaches the problems of finding an appropriate distribution fit to datasets and testing the statistical significance of the observed differences. This problem shall be split into three tasks. The first task aims at identifying a theoretical distribution that fits the samples in all datasets by altering its parameters. The second task is to test the statistical significance of the difference between two empirical distributions. The third task is to test the statistical significance of the difference between two distribution fits over two arbitrary datasets.

Task 2 can be performed whether or not a suitable theoretical distribution fit was identified. Therefore, comparing and eventually uniting the samples will always be possible. This task requires comparing two independent discontinuous (stair-case) empirical CDFs. It is a standard problem and the approach here is based on the Kuiper two-sample test (a variation of the Kolmogorov-Smirnov test [15]). The last essentially does an estimate of the closeness of a pair of independent stair-case CDFs by finding the maximum deviation above and below the two [4].

Tasks 1 and 3 bring the novel elements of the paper. Task 1 searches for a type of theoretical distribution that fits multiple datasets by simply varying its specific parameter values. The performance of a distribution fit is assessed through four criteria, namely the Akaike Information Criterion (AIC) [1], the Bayesian Information Criterion (BIC) [5], the average and the minimal p_{value} of a distribution fit to all datasets. Since the datasets contain random measurements, the values of the parameters for each acquired fit in task 1 are random, too. That is why it is necessary to check whether the differences are statistically significant, for each pair of datasets. If not, then both theoretical fits are identical and the samples may be united. A distribution of the Kuiper statistic cannot be constructed in task 1, because the setup of the last compares a distribution with its own fit so independence is violated. A distribution of the Kuiper statistic cannot be constructed in task 3 either, because the last compares two distribution fits, and not stair-case CDFs. For that reason the distribution of the Kuiper statistic in tasks 1 and 3 is constructed via a Monte Carlo procedure [11].

The described statistical procedures are embodied into original program function in MATLAB R2009a environment. The platform tests 11 types of theoretical distributions in order to find the best fit. Additionally, new types of distributions can be easily added to this set. In addition to on-screen results, the programs also generate graphical output.

A biochemical problem is analyzed with the help of the statistical procedures and the platform. It concerns the analysis of the different density of the fibrin network under varying concentrations of thrombin. The 12 analyzed datasets contain measurements of the length of fibrin fibers in sectors between two nodes. Each dataset is collected under a given thrombin concentration and a given buffer. The measurements are taken from dry fibrin samples,

examined in Zeiss Evo40 scanning electron microscope. To enhance the data collection, an original program function in MATLAB R2009a is created to estimate lengths in electron microscope images.

Metods

Setup

Consider N datasets $\chi^i = \left(x_1^i, x_2^i, ..., x_{n_i}^i\right)$, for i = 1, 2, …, N. The data set χ^i contains $n_i > 65$ sorted positive samples $(0 < x_1^i \le x_2^i \le ... \le x_{n_i}^i)$ of a given random quantity under equal conditions. The datasets χ^i and χ^j, for $j \ne i$, contain samples of the same random quantity, but under slightly different conditions.

Assume that M types of theoretical distributions are analyzed. Each of them has a probability density function $PDF_j\left(x, \vec{p}_j\right)$, a cumulative distribution function $CDF_j\left(x, \vec{p}_j\right)$, and an inverse cumulative distribution function $invCDF_j\left(P, \vec{p}_j\right)$, for j = 1, 2, …, M. Each of these functions depend on n_j^p-dimensional parameter vectors $\vec{p}_j$ (for j = 1, 2, …, M), dependent on the theoretical distribution type.

There are three problems that have to be solved: 1) find the best theoretical distribution type, which fits the data in all datasets and identify the specific parameter values of this theoretical distribution type for each dataset; 2) check whether two different datasets are drawn from the same general population; 3) check the statistical significance of the difference between two theoretical distributions of one type fitted to two different arbitrary datasets.

First problem

The empirical cumulative distribution function $CDF_e^i(.)$ is initially linearly approximated over (n_i + 1) nodes as (n_i – 1) internal nodes $CDF_e^i\left(x_k^i/2 + x_{k+1}^i/2\right) = k/n_i$ for k = 1, 2, …, n_i–1 and 2 external nodes $CDF_e^i\left(x_1^i - \Delta_d^i\right) = 0$ and $CDF_e^i\left(x_{n_i}^i + \Delta_u^i\right) = 1$, where $\Delta_d^i = min\left(x_1^i, \left(x_{16}^i - x_1^i\right)/30\right)$ and $\Delta_u^i = \left(x_{n_i}^i - x_{n_i-15}^i\right)/30$ are the halves of mean inter-sample intervals in the lower and upper ends of the dataset χ^i, but the down external node is never with a negative abscissa $\left(x_1^i - \Delta_d^i\right) \ge 0$.

It is convenient to introduce "before-first" $x_0^i = x_1^i - 2\Delta_d^i$ and "after-last" $x_{n_i+1}^i = x_{n_i}^i + 2\Delta_u^i$ samples. When for some k = 1, 2, …, n_i and for p>1 it is true that $x_{k-1}^i < x_k^i = x_{k+1}^i = x_{k+2}^i = ... = x_{k+p}^i < x_{k+p+1}^i$ then the initial approximation of $CDF_e^i(.)$ contains a vertical segment of p nodes. In that case the p nodes on that segment are replaced by a single node in the middle of the vertical segment $CDF_e^i\left(x_k^i\right) = \left(k + p/2 - 1/2\right)/n_i$. The described two-step procedure [8] results in a strictly increasing function $CDF_e^i(.)$ in the closed interval $\left[x_1^i - \Delta_d^i; x_{n_i}^i + \Delta_u^i\right]$. That is why it is possible to introduce $invCDF_e^i(.)$ with the domain [0; 1] as the inverse function of $CDF_e^i(.)$ in $\left[x_1^i - \Delta_d^i; x_{n_i}^i + \Delta_u^i\right]$. The median and

the interquartile range of the empirical distribution can be estimated from $invCDF_e^i(.)$, whereas the mean and the standard deviation are easily estimated directly from the dataset χ^i:

- mean: $mean_e^i = \frac{1}{n_i}\sum_{k=1}^{n_i} x_k^i$;
- median: $med_e^i = invCDF_e^i(0.5)$;
- standard deviation: $std_e^i = \sqrt[2]{\frac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(x_k^i - mean_e^i\right)^2}$;
- interquartile range: $iqr_e^i = invCDF_e^i(0.75) - invCDF_e^i(0.25)$.

The non-zero part of the empirical density $PDF_e^i(.)$ is determined in the closed interval $\left[x_1^i - \Delta_d^i; x_{n_i}^i + \Delta_u^i\right]$ as a histogram with bins of equal area (each bin has equal product of density and span of data). The number of bins b_i is selected as the minimal from the Scott [6], Sturges [7] and Freedman-Diaconis [2] suggestions. The lower and upper margins of the k-th bin $m_{d,k}^i$ and $m_{u,k}^i$ are determined as quantiles $(k - 1)/b_i$ and k/b_i respectively: $m_{d,k}^i = invCDF_e^i(k/b_i - 1/b_i)$ and $m_{u,k}^i = invCDF_e^i(k/b_i)$. The density of the k-th bin is determined for $m_{d,k}^i \le x \le m_{u,k}^i$ as $PDF_e^i(x) = b_i^{-1}/\left(m_{u,k}^i - m_{d,k}^i\right)$. The described procedure [8] results in a histogram, where the relative error of the worst $PDF_e^i(.)$ estimate is minimal from all possible separation of the samples into b_i bins. The improper integral $\int_{-\infty}^{x} PDF_e^i(x)dx$ of the density is a smoothened version of $CDF_e^i(.)$ linearly approximated over (b_i + 1) nodes: $\left(invCDF_e^i(k/b_i); k/b_i\right)$ for $k = 0, 1, 2, \ldots, b_i$.

The likelihood of the dataset χ^i, if the samples are distributed with density $PDF_j(x, \vec{p}_j)$, is $L_j^i(\vec{p}_j) = \prod_{k=1}^{n_i} PDF_j(x_k^i, \vec{p}_j)$. The maximum likelihood estimates (MLEs) of $\vec{p}_j$ are determined as those $\vec{p}_j^i$, which maximize $L_j^i(\vec{p}_j)$, that is $\vec{p}_j^i = arg\left\{\max_{\vec{p}_j}\left[L_j^i(\vec{p}_j)\right]\right\}$. The numerical characteristics of the j-th theoretical distribution fitted to the dataset χ^i are calculated as:

- mean: $mean_j^i = \int_{-\infty}^{+\infty} x.PDF_j(x, \vec{p}_j^i)dx$;
- median: $med_j^i = invCDF_j(0.5, \vec{p}_j^i)$;
- mode: $mode_j^i = arg\left\{\max_x\left[PDF_j(x, \vec{p}_j)\right]\right\}$;
- standard deviation: $std_j^i = \sqrt[2]{\int_{-\infty}^{+\infty}\left(x - mean_j^i\right)^2.PDF_j(x, \vec{p}_j^i)dx}$;
- interquartile range: $iqr_j^i = invCDF_j(0.75, \vec{p}_j^i) - invCDF_j(0.25, \vec{p}_j^i)$.

The quality of the fit can be assessed using a statistical hypothesis test. The null hypothesis H_0 is that $CDF_e^i(x)$ is equal to $CDF_j\left(x, \vec{p}_j^{\,i}\right)$, which means that the sample χ^i is drawn from $CDFj\left(x, \vec{p}_j^{\,i}\right)$. The alternative hypothesis H_1 is that $CDF_e^i(x)$ is different from $CDF_j\left(x, \vec{p}_j^{\,i}\right)$, which means that the fit is not good. The Kuiper statistic V_j^i [3] is a suitable measure for the goodness-of-fit of the theoretical cumulative distribution functions $CDF_j\left(x, \vec{p}_j^{\,i}\right)$ to the dataset χ^i:

$$V_j^i = \max_x\left\{CDF_e^i(x) - CDF_j\left(x, \vec{p}_j^{\,i}\right)\right\} + \max_x\left\{CDF_j\left(x, \vec{p}_j^{\,i}\right) - CDF_e^i(x)\right\}. \quad (1)$$

The distribution of V, if H_0 is true, can be estimated by a Monte Carlo procedure, because the original Kuiper's distribution refers to two independent distributions, but not to the case when one is fitted to the data of the other [4]. In n^{MC} simulation cycles, n_i samples are drawn from the fitted distribution $CDF_j\left(x, \vec{p}_j^{\,i}\right)$, and n^{MC} synthetic datasets $\chi_r^{i,syn} = \left\{x_{1,r}^{i,syn}, x_{2,r}^{i,syn}, ..., x_{n_i,r}^{i,syn}\right\}$, for r = 1, 2,…, n^{MC} are formed. The dataset $\chi_r^{i,syn}$ contains n_i sorted positive samples ($0 < x_{1,r}^{i,syn} \le x_{2,r}^{i,syn} \le ... \le x_{n_i,r}^{i,syn}$). It is possible to do the following for each synthetic dataset $\chi_r^{i,syn}$ using the described algorithms:

1. Construct the synthetic empiric distribution $CDF_{e,r}^{i,syn}(.)$;
2. Find the synthetic maximum likelihood estimates $\vec{p}_{j,r}^{\,i,syn}$;
3. Fit the synthetic theoretical distribution function $CDF_{j,r}^{syn}\left(x, \vec{p}_{j,r}^{\,i,syn}\right)$ to $\chi_r^{i,syn}$;
4. Estimate the synthetic Kuiper statistic $V_{j,r}^{i,syn}$.

The p-value $P_{value,j}^{fit,i}$ of the statistical test (the probability to reject a true hypothesis H_0 that the j-th type theoretical distribution fits well to the samples in dataset χ^i) is estimated as the frequency of generating synthetic Kuiper statistic greater than the actual Kuiper statistic V_j^i:

$$P_{value,j}^{fit,i} = \frac{1}{n^{mc}} \sum_{\substack{r=1 \\ V_j^i < V_{j,r}^{i,syn}}}^{n^{mc}} 1 \quad (2)$$

In fact, (2) is the sum of the indicator function of the crisp set, defined as all synthetic datasets with a Kuiper statistic greater than V_j^i.

The performance of each theoretical distribution should be assessed according to its goodness-of-fit measures to the N datasets simultaneously. If a given theoretical distribution cannot be fitted even to one of the datasets, then that theoretical distribution has to be discarded from further consideration. The other theoretical distributions have to be ranked according to their ability to describe all datasets. One basic and three auxiliary criteria are useful in the required ranking.

The basic criterion is the minimal p-value of the theoretical distribution fits to the N datasets:

$$minP_{value,j}^{fit} = min\left\{P_{value,j}^{fit,1}, P_{value,j}^{fit,2}, ..., P_{value,j}^{fit,N}\right\}, \text{ for } j = 1, 2, ..., M. \quad (3)$$

The first auxiliary criterion is the average of the p-values of the theoretical distribution fits to the N datasets:

$$meanP_{value,j}^{fit} = \frac{1}{N}\sum_{j=1}^{N} P_{value,j}^{fit,i} \text{ , for } j=1, 2, .., M. \tag{4}$$

The second and the third auxiliary criteria are the AIC-Akaike Information Criterion [1] and the BIC-Bayesian Information Criterion [5], which corrects the negative log-likelihoods with the number of the assessed parameters:

$$\begin{aligned} AIC_j &= -2\sum_{i=1}^{N} log\left(L_j^i\left(\vec{p}_j^{\,i}\right)\right) + 2\,log\left(N.n_j^p\right) = \\ &= -2\sum_{i=1}^{N}\sum_{j=1}^{M} log\, PDF_j\left(x_k^i, \vec{p}_j^{\,i}\right) + 2\,log\left(N.n_j^p\right) \end{aligned}, \tag{5}$$

$$\begin{aligned} BIC_j &= -2\sum_{i=1}^{N} log\left(L_j^i\left(\vec{p}_j^{\,i}\right)\right) + 2\,log\left(N.n_j^p\right).log\left(\sum_{i=1}^{M} n_i\right) = \\ &= -2\sum_{i=1}^{N}\sum_{j=1}^{M} log\, PDF_j\left(x_k^i, \vec{p}_j^{\,i}\right) + 2\,log\left(N.n_j^p\right).log\left(\sum_{i=1}^{M} n_i\right) \end{aligned}, \tag{6}$$

for j = 1, 2, .., M. The best theoretical distribution type should have maximal values for $minP_{value,j}^{fit}$ and $meanP_{value,j}^{fit}$, whereas its values for AIC_j and BIC_j should be minimal. On top, the best theoretical distribution type should have $minP_{value,j}^{fit} > 0.05$, otherwise no theoretical distribution from the initial M types fits properly to the N datasets.

That solves the problem for selecting the best theoretical distribution type for fitting the samples in the N datasets.

Second problem

The second problem is the estimation of the statistical significance of the difference between two datasets. It is equivalent to calculating the p-value of a statistical hypothesis test, where the null hypothesis H_0 is that the samples of χ^{i1} and χ^{i2} are drawn from the same underlying continuous population, and the alternative hypothesis H_1 is that the samples of χ^{i1} and χ^{i2} are drawn from different underlying continuous populations. The two-sample asymptotic Kuiper test is designed exactly for that problem, because χ^{i1} and χ^{i2} are independently drawn datasets. That is why "staircase" empirical cumulative distribution functions [9] are build from the two datasets χ^{i1} and χ^{i2}:

$$CDF_{sce}^{i}\left(x\right) = \sum_{\substack{k=1 \\ x_k \le x}}^{n_i} 1/n_i \text{ , for } i \in \{i1, i2\}. \tag{7}$$

The "staircase" empirical cumulative distribution function $CDF_{sce}^{i}(.)$ is a discontinuous version of the already defined empirical cumulative distribution function $CDF_{e}^{i}(.)$. The Kuiper statistic $V^{i1,i2}$ [3] is a measure for the closeness of the two 'staircase' empirical cumulative distribution functions $CDF_{sce}^{i1}(.)$ and $CDF_{sce}^{i2}(.)$:

$$V^{i1,i2} = \max_x \left\{ CDF_{sce}^{i1}(x) - CDF_{sce}^{i2}(x) \right\} + \max_x \left\{ CDF_{sce}^{i2}(x) - CDF_{sce}^{i1}(x) \right\}. \quad (8)$$

The p-value $P_{value,e}^{i1,i2}$ of the statistical test (the probability to reject a true null hypothesis H_0, that the samples in χ^{i1} and in χ^{i2} result in the same 'staircase' empirical cumulative distribution functions) is estimated as a series [4]:

$$P_{value,e}^{i1,i2} = 2 \sum_{j=1}^{+\infty} \left(4j^2\lambda - 1 \right) e^{-2j^2\lambda^2}, \quad (9)$$

where

$$\lambda = \frac{1}{V^{i1,i2}} \left(\sqrt{\frac{n_{i1} n_{i2}}{n_{i1} + n_{i2}}} + 0.155 + 0.24 \sqrt{\frac{n_{i1} + n_{i2}}{n_{i1} n_{i2}}} \right). \quad (10)$$

If $P_{value,e}^{i1,i2} < 0.05$ the hypothesis H_0 is rejected.

Third problem

The last problem is to test the statistical significance of the difference between two fitted distributions of the same type. This type most often would be the best type of theoretical distribution, which was identified in the first problem, but the test is valid for any type. The problem is equivalent to calculating the p-value of statistical hypothesis test where the null hypothesis H_0 is that the theoretical distribution $CDF_j\left(x, \vec{p}_j^{i1}\right)$ and $CDF_j\left(x, \vec{p}_j^{i2}\right)$ fitted to the datasets χ^{i1} and χ^{i2} are identical, and the alternative hypothesis H_1 is that $CDF_j\left(x, \vec{p}_j^{i1}\right)$ and $CDF_j\left(x, \vec{p}_j^{i2}\right)$ are not identical.

The test statistic again is the Kuiper one $V_j^{i1,i2}$:

$$V_j^{i1,i2} = \max_x \left\{ CDF_j\left(x, \vec{p}_j^{i1}\right) - CDF_j\left(x, \vec{p}_j^{i2}\right) \right\} + \max_x \left\{ CDF_j\left(x, \vec{p}_j^{i2}\right) - CDF_j\left(x, \vec{p}_j^{i1}\right) \right\}. \quad (11)$$

The distribution of V, if H_0 is true, can be estimated by a Monte Carlo procedure, because the original Kuiper's distribution refers to two independent "staircase" empirical cumulative distribution functions, but not to the case of two independent theoretical cumulative distribution functions. If H_0 is true, then $CDF_j\left(x, \vec{p}_j^{i1}\right)$ and $CDF_j\left(x, \vec{p}_j^{i2}\right)$ should be identical to the merged distribution $CDF_j\left(x, \vec{p}_j^{i1+i2}\right)$, fitted to the artificial dataset χ^{i1+i2} formed by merging the samples of χ^{i1} and χ^{i2}. In n^{MC} simulation cycles, 2 datasets containing respectively n_{i1} samples and n_{i2} samples are drawn from the merged distribution $CDF_j\left(x, \vec{p}_j^{i1+i2}\right)$, and n^{MC} pairs of synthetic datasets $\chi_r^{i1,syn} = \left\{ x_{1,r}^{i1,syn}, x_{2,r}^{i1,syn}, ..., x_{n_{i1},r}^{i1,syn} \right\}$ and $\chi_r^{i2,syn} = \left\{ x_{1,r}^{i2,syn}, x_{2,r}^{i2,syn}, ..., x_{n_{i2},r}^{i2,syn} \right\}$, for $r = 1, 2, \ldots, n^{MC}$ are formed. It is possible to do the following for each synthetic dataset pair $\chi_r^{i1,syn}$ and $\chi_r^{i2,syn}$ using the described algorithms for:

1. Estimate the synthetic maximum likelihood estimates $\vec{p}_{j,r}^{i1,syn}$ and $\vec{p}_{j,r}^{i2,syn}$;

2. Fit the synthetic theoretical distribution functions $CDF_{j,r}^{syn}\left(x,\vec{p}_{j,r}^{i1,syn}\right)$ and $CDF_{j,r}^{syn}\left(x,\vec{p}_{j,r}^{i2,syn}\right)$ to $\chi_r^{i1,syn}$ and to $\chi_r^{i2,syn}$;

3. Estimate the synthetic Kuiper statistic $V_{j,r}^{i1,i2,syn}$.

The p-value $P_{value,j}^{i1,i2}$ of the statistical test (the probability to reject a true hypothesis H_0 that the $j^{\text{-th}}$ type theoretical distribution function $CDF_j\left(x,\vec{p}_j^{i1}\right)$ and $CDF_j\left(x,\vec{p}_j^{i2}\right)$ are identical) is estimated as the frequency of generating synthetic Kuiper statistic greater than the actual Kuiper statistic $V_j^{i1,i2}$:

$$P_{value,j}^{i1,i2}=\frac{1}{n^{mc}}\sum_{\substack{r=1\\ V_j^{i1,i2}<V_{j,r}^{i1,i2,syn}}}^{n^{mc}}1. \tag{12}$$

Formula (12), similar to (2), is the sum of the indicator function of the crisp set, defined as all synthetic dataset pairs with a Kuiper statistic greater than $V_j^{i1,i2}$.

If $P_{value,j}^{i1,i2}<0.05$ the hypothesis H_0 is rejected.

Software

A platform of program functions, written in MATLAB R2009a environment, is created to execute the statistical procedures from the previous section. At the present state of development, the platform allows users to test the fit of 11 types of distributions on the datasets. A description of the parameters and PDF of the embodied distribution types is given in Table 1 [12, 14]. The platform also permits the user to add additional types of distribution.

The platform contains several main functions. The function *set_distribution.m* contains the information about the 11 distributions, particularly their names, and the links to the functions that operate with the selected type distribution. Also, the function permits the inclusion of a new distribution type. In that case, the necessary information the user must provide as input is the procedures to find the CDF, PDF, the maximum likelihood measure, the negative log-likelihood, the mean and variance and the methods of generating random arrays from the given distribution type. The function also determines the screen output for each type of distribution.

The program function *kutest2.m* performs a two-sample Kuiper test to determine if the independent random datasets are drawn from the same underlying continuous population, i.e. it solves the second statistical problem, outlined in the Methods section (to check whether two different datasets are drawn from the same general population).

Another key function is *fitdata.m*. It constructs the fit of each theoretical distribution over each dataset, evaluates the quality of the fits, and gives their parameters. It also checks whether two distributions of one type fitted to two different arbitrary datasets are identical. In other words, this function is associated with the statistical problems involving the Monte Carlo procedures. To execute the Kuiper test the function calls *kutest2.m*. Finally, the function *plot_print_data.m* provides the on-screen results from the statistical analysis and plots figures containing the pair of distributions that are analyzed.

Materials

The statistical procedures and the program platform developed in this paper are implemented in an example, focusing on the morphometric evaluation of the effects of thrombin concentration on fibrin structure. Fibrin is a biopolymer formed from the blood-borne fibrinogen by an enzyme activated in the damaged tissue (thrombin) at sites of blood vessel wall injury to prevent bleeding. Following regeneration of the integrity of the blood vessel wall, the fibrin gel is dissolved to restore normal blood flow, but the efficiency of the dissolution strongly depends on the structure of the fibrin clots. The purpose of the evaluation is to establish any differences in the density of the branching points of the fibrin network related to the activity of the clotting enzyme (thrombin), the concentration of which is expected to vary in a broad range under physiological conditions.

For the purpose of the experiment, fibrin is prepared on glass slides in total volume of 100 µl by clotting 2 mg/ml fibrinogen (dissolved in different buffers) by varying concentrations of thrombin for 1 h at 37 °C in moisture chamber. The thrombin concentrations in the experiments vary in the range 0.3 – 10 U/ml, whereas the two buffers used are: 1) buffer1 – 25 mM Na-phosphate pH 7.4 buffer containing 75 mM NaCl; 2) buffer2 - 10 mM N-(2-Hydroxyethyl)piperazine-N'-(2-ethanesulfonic acid) (abbreviated as HEPES) pH 7.4 buffer containing 150 mM NaCl. At the end of the clotting time the fibrins are washed in 3ml 100 mM Na-cacodilate pH 7.2 buffer and fixated with 1% glutaraldehyde in the same buffer for 10 min. Thereafter the fibrins are dried in a series of ethanol dilutions (20 – 96 %), acetone and finally hexamethyldisilazane. The dry samples are examined in Zeiss Evo40 scanning electron microscope (Carl Zeiss, Jena, Germany) and images are taken at an indicated magnification. A total of 12 dry samples of fibrins are elaborated in this fashion, each having a given combination of thrombin concentration and buffer. Electron microscope images are taken for each dry sample (one of the analyzed dry samples of fibrins is given in Fig. 1). Some main parameters of the 12 collected datasets are given in Table 2.

An automated procedure is elaborated in MATLAB R2009a environment (embodied into the program function *find_distance.m*) to measure lengths of fibrin strands (i.e. sections between two branching points in the fibrin network) from the electron images. The procedure takes as input the file name of the fibrin image (see Fig. 1) and the planned number of measurements. Each file contains the fibrin image with legend at the bottom part, which gives the scale, the time the image was taken, etc.

The first step requires settling of the scale. A prompt appears, asking the user to type the numerical value of the length of the scale in microns. Then on screen appear the image and a red line, which has to be moved and resized to fit the scale (Fig. 2a and 2b). A double click signifies the end of scaling. The third step requires a red rectangle to be placed over the actual image of the fibrin (in other words, the legend and caption are excluded from the rectangle) (Fig. 2c). With this, the preparations of the image are done, and the user can start taking the desired number of measurements for the lengths of fibrins between adjacent nodes (Fig. 2d). After that, on screen appear the length of each selected fibrin part, and the numerical characteristics of this dataset.

Table 1. Parameters, support and formula for the PDF of the eleven types of theoretical distributions embodied into the MATLAB platform

	Beta distribution		**Lognormal distribution**
Parameters	$\alpha>0,\ \beta>0$	Parameters	$\mu\in(-\infty;\ +\infty)\ ,\ \sigma>0,$
Support	$x\in[0;\ 1]$	Support	$x\in[0;\ +\infty)$
PDF	$f(x;\ \alpha,\beta)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathbf{B}(\alpha,\beta)}$, where $\mathbf{B}(\alpha,\beta)$ is a beta function	PDF	$f(x;\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}}$

	Exponential distribution		**Normal distribution**
Parameters	$\lambda>0$	Parameters	$\mu,\ \sigma>0$
Support	$x\in[0;\ +\infty)$	Support	$x\in(-\infty;\ +\infty)$
PDF	$f(x;\ \lambda)=\begin{cases}\lambda e^{-\lambda x} & \text{for } x\geq 0\\ 0 & \text{for } x<0\end{cases}$	PDF	$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

	Extreme value distribution		**Rayleigh distribution**
Parameters	$\alpha,\ \beta\neq 0$	Parameters	$\sigma>0$
Support	$x\in(-\infty;\ +\infty)$	Support	$x\in[0;\ +\infty)$
PDF	$f(x;\ \alpha,\beta)=\frac{e^{\left[(\alpha-x)/\beta\right]-e^{(\alpha-x)/\beta}}}{\beta}$	PDF	$f(x;\sigma)=\frac{1}{\sigma^2}\times\left[x\,exp\left(\frac{-x^2}{2\sigma^2}\right)\right]$

	Gamma distribution		**Uniform distribution**
Parameters	$k>0,\ \theta>0$	Parameters	$a,b\in(-\infty;\ +\infty)$
Support	$x\in[0;\ +\infty)$	Support	$a\leq x\leq b$
PDF	$f(x;\ k,\ \theta)=x^{k-1}\frac{e^{-x/\theta}}{\theta^k\Gamma(k)}$, where $\Gamma(k)$ is a gamma function	PDF	$f(x;a,b)=\begin{cases}\frac{1}{b-a} & \text{for } a\leq x\leq b\\ 0 & \text{for } x<a \text{ or } x>b\end{cases}$

	Generalized extreme value distribution		**Weibull distribution**
Parameters	$\mu\in(-\infty;\ +\infty)$, $\sigma\in(0;\ +\infty)$, $\xi\in(-\infty;\ +\infty)$	Parameters	$\lambda>0, k>0$
Support	$x>\mu-\sigma/\xi\ \ (\xi>0)$, $x<\mu-\sigma/\xi\ \ (\xi<0)$, $x\in(-\infty;\ +\infty)\ \ (\xi=0)$	Support	$x\in[0;\ +\infty)$
PDF	$\frac{1}{\sigma}(1+\xi z)^{-1/\xi-1}e^{-(1+\xi z)-1/\xi}$ where $z=\frac{x-\mu}{\sigma}$	PDF	$f(x;\lambda,k)=$ $=\begin{cases}\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k} & \text{for } x\geq 0\\ 0 & \text{for } x<0\end{cases}$

	Generalized Pareto distribution
Parameters	$x_m>0,\ k>0$
Support	$x\in[x_m;\ +\infty)$
PDF	$f(x;x_m,k)=\frac{kx_m^k}{x^{k+1}}$

Table 2. Sample size (N), mean ($mean_e$ in microns), median (med_e in microns), standard deviation (std_e), interquartile range (iqr_e, in microns) of the empirical distributions over the 12 datasets (t given thrombin concentration (in U/ml) and buffer), containing measurements of lengths between branching points of fibrin fibers

Datasets	N	$mean_e$	med_e	std_e	iqr_e	Thrombin concentration	Buffer
DS1	274	0.9736	0.8121	0.5179	0.6160	1.0	buffer1
DS2	68	1.023	0.9374	0.5708	0.7615	10.0	buffer1
DS3	200	1.048	0.8748	0.6590	0.6469	4.0	buffer1
DS4	276	1.002	0.9003	0.4785	0.5970	0.5	buffer1
DS5	212	0.6848	0.6368	0.3155	0.4030	1.0	buffer2
DS6	300	0.1220	0.1265	0.04399	0.05560	1.2	buffer2
DS7	285	0.7802	0.7379	0.3253	0.4301	2.5	buffer2
DS8	277	0.9870	0.9326	0.4399	0.5702	0.6	buffer2
DS9	200	0.5575	0.5284	0.2328	0.2830	0.3	buffer1
DS10	301	0.7568	0.6555	0.3805	0.4491	0.6	buffer1
DS11	301	0.7875	0.7560	0.3425	0.4776	1.2	buffer1
DS12	307	0.65000	0.5962	0.2590	0.3250	2.5	buffer1

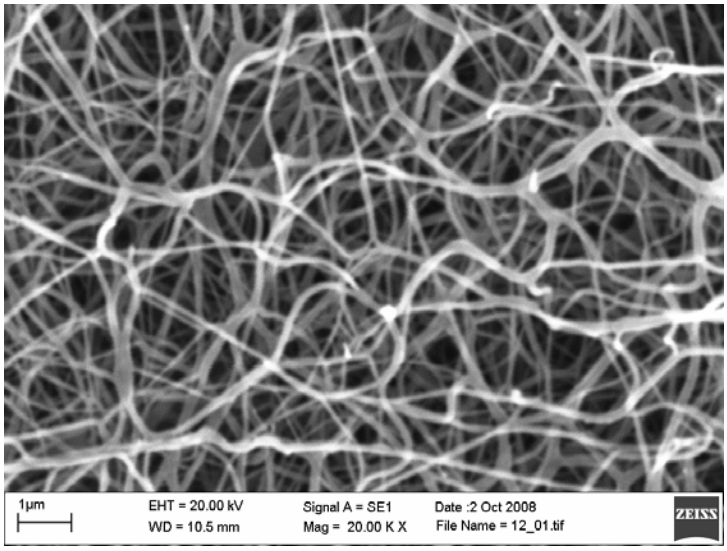
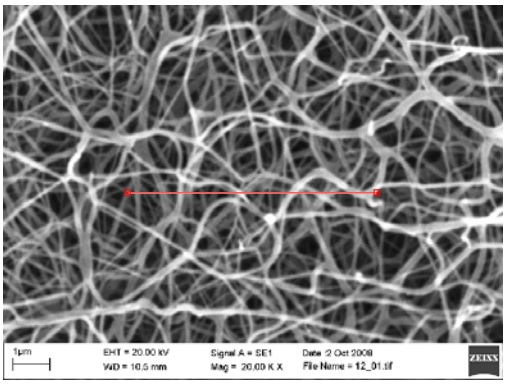
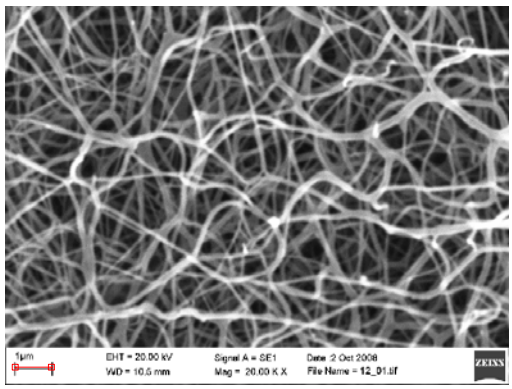


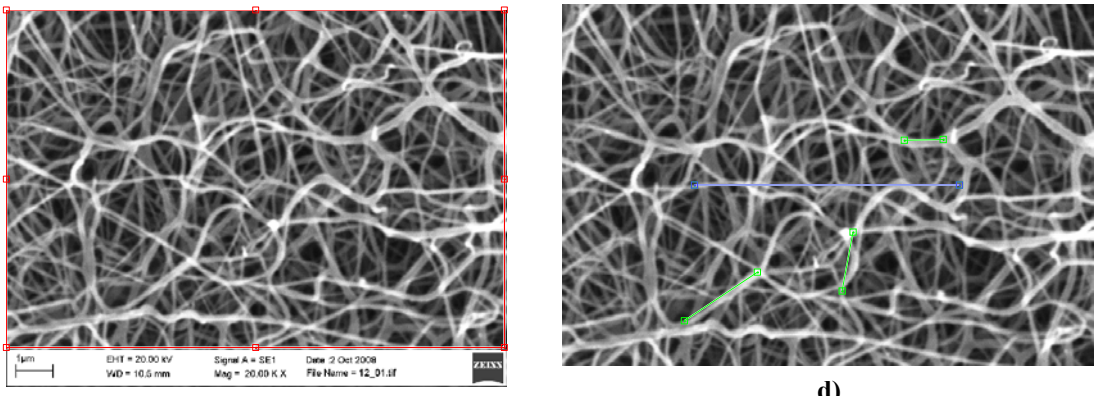
Fig. 1 Image of a dry sample of fibrin



a)



b)



c) d)

Fig. 2 Steps of the automated procedure for measuring lengths of fibrins from the dry sample images. Fig. 2a) and 2b) show scaling. Fig. 2c) shows the image selection. Fig. 2d) shows a phase of measurement taking.

Results

The three statistical tasks are applied over the 12 datasets containing measurements of lengths between branching points of fibrin fibers (see Table 2).

Task 1 – Finding a distribution fit

A total of 11 types of distributions (see table 1) are tested over the datasets, and the criteria (3)-(6) are evaluated. The Kuiper statistic's distribution is constructed with 1000 Monte Carlo simulation cycles. Table 3 presents the results regarding the distribution fits, where only the maximal values for $minP_{value,j}^{fit}$ and $meanP_{value,j}^{fit}$, and the minimal values for AIC_j and BIC_j across the datasets are given. The results allow ruling out the beta and the uniform distributions. The first outputs NaN since it does not apply to values of $x \notin [0; 1]$. The later has the lowest values of (3) and (4), and the highest of (5) and (6), i.e. it is the worst fit. The types of distributions worth using are mostly the lognormal distribution (having the lowest *AIC* and *BIC*), and the generalized extreme value (having the highest possible $meanP_{value,j}^{fit}$). Fig. 3 presents 4 of the 11 distribution fits to DS4. Similar graphical output is generated for all other datasets and for all distribution types.

Task 2 – Equality of empirical distributions

Table 4 contains the p-value calculated according to (9) for all pairs of distributions. The bolded values indicate the pairs, where the null hypothesis fails to be rejected and it is possible to assume that those datasets are drawn from the same general population. The results show that it is possible to unite the following datasets: 1) DS1, DS2, DS3, and DS4; 2) DS2, DS3, DS4, and DS8; 3) DS7, DS10, and DS11; 4) DS5 and DS10; 5) DS5 and DS12. All other combinations are not allowed and may give misleading results in a further statistical analysis, since the samples are not drawn from the same general population. Figure 4a presents the stair-case distributions over DS4 (with $mean_e^4 = 1.002$, $med_e^4 = 0.9003$, $std_e^4 = 0.4785$, $iqr_e^4 = 0.5970$) and DS9 (with $mean_e^9 = 0.5575$, $med_e^9 = 0.5284$, $std_e^9 = 0.2328$, $iqr_e^9 = 0.2830$). The Kuiper statistic for identity of the empirical distributions, calculated according to (8), is $V^{4,9} = 0.5005$, whereas according to (9), $P_{value,e}^{4,9} = 3.556\text{e}{-25} < 0.05$. Therefore the null hypothesis is rejected, which is also evident from the graphical

output. In the same fashion, figure 4b presents the stair-case distributions over DS1 (with $mean_e^1$ = 0.9736, med_e^1 = 0.8121, std_e^1 = 0.5179, iqr_e^1 = 0.6160) and DS4. The Kuiper statistic for identity of the empirical distributions, calculated according to (8), is $V^{1,4}$ =0.1242, whereas according to (9), $P_{value,e}^{4,9}$ = 0.1242 > 0.05. Therefore the null hypothesis fails to be rejected, which is also confirmed by the graphical output.

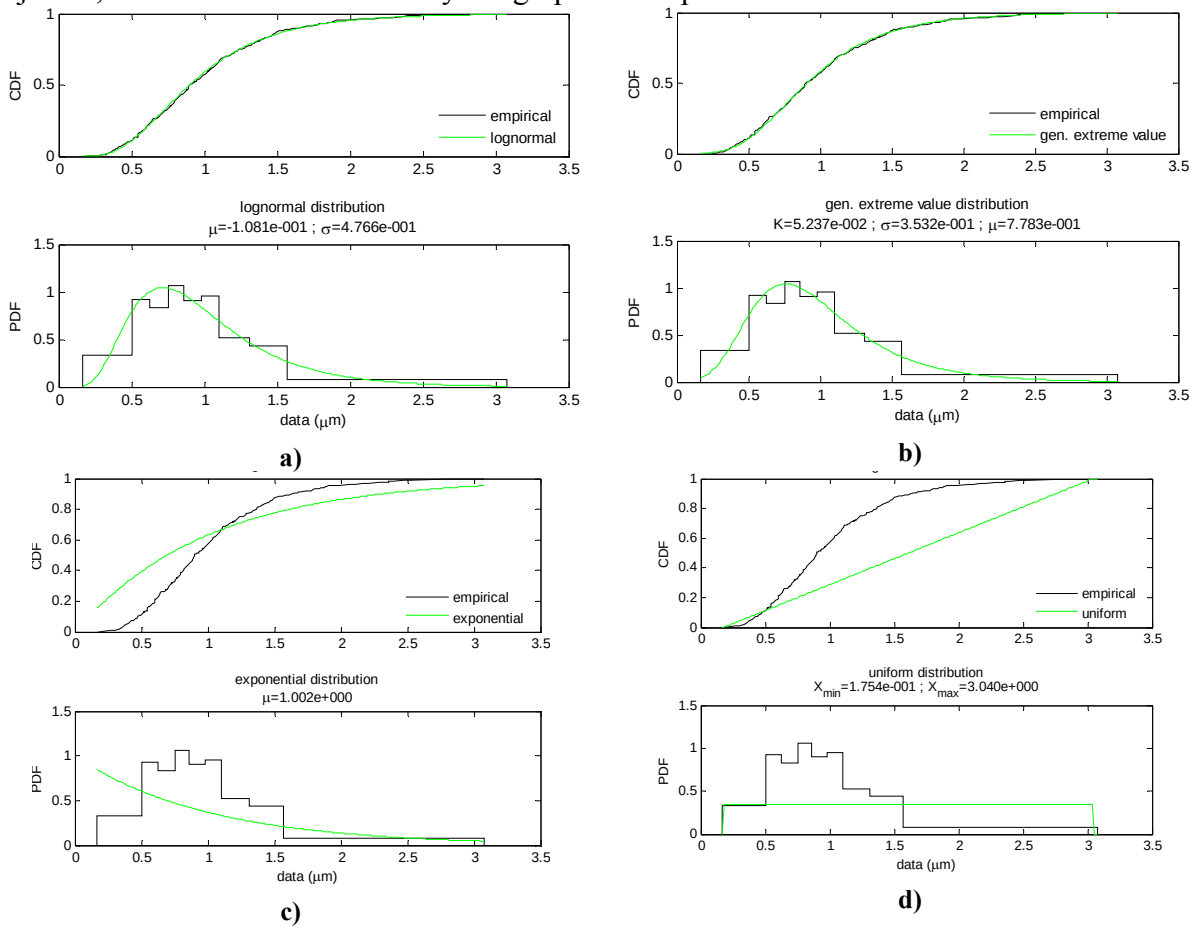


Fig. 3 Graphical results from the fit of the lognormal (a), generalized extreme value (b), exponential (c), and uniform (d) distributions over DS4

Task 3 – Equality of fitted distributions

As concluded in task 1, the lognormal distribution provides possibly the best fit to the 12 datasets. Table 5 contains the p-values calculated according to (12) for the lognormal distribution fitted to the datasets with 1000 Monte Carlo simulation cycles. The bolded values indicate the pairs, where the null hypothesis fails to be rejected and it is possible to assume that the distribution fits are identical. The results show that the lognormal fits to the following datasets are identical: 1) DS1, DS2, DS3, and DS4; 2) DS1, DS4, and DS8; 3) DS7, DS10, and DS11; 4) DS5 and DS10; 5) DS5 and DS12. These results correlate with the identity of the empirical distribution. Figure 5a presents the fitted lognormal distribution over DS4 (with μ = –0.1081, σ =0.4766, $mean_7^4$ = 1.005, med_7^4 = 0.8975, $mod\, e_7^4$ = 0.7169, std_7^4 = 0.5077, iqr_7^4 = 0.5870) and DS9 (with μ = –0.6694, σ = 0.4181, $mean_7^9$ = 0.5587, med_7^9 = 0.5120, $mod\, e_7^9$ = 0.4322, std_7^9 = 0.2442, iqr_7^9 = 0.2926). The Kuiper statistic for identity of the fits, calculated according to (11), is $V_7^{4,9}$ = 0.4671, whereas according to (12), $P_{value,7}^{4,9}$ = 0<0.05. Therefore the null hypothesis is rejected, which is also evident from the graphical output. In

the same fashion, figure 5b presents the lognormal distribution fit over DS1 (with $\mu = -1477$, $\sigma = 0.4843$, $mean_7^1 = 0.9701$, $med_7^1 = 0.8627$, $mod\, e_7^1 = 0.6758$, $std_7^1 = 0.4988$, $iqr_7^1 = 0.5737$) and DS4. The Kuiper statistic for identity of the fits, calculated according to (11), is $V_7^{1,4} = 0.03288$, whereas according to (12), $P_{value,7}^{1,4} = 0.5580 > 0.05$. Therefore the null hypothesis fails to be rejected, which is also evident from the graphical output.

Table 3. Results from the fit of 11 types of distributions over the datasets with 1000 Monte Carlo simulation cycles. The table contains the maximal values for $minP_{value,j}^{fit}$ and $meanP_{value,j}^{fit}$, and the minimal values for AIC_j and BIC_j across the datasets for each distribution type. The bolded and the italic values are respectively the best and the worst achieved for a given criterion.

Distribution type	1	2	3	4	5	6
AIC	NaN	3.705e+3	3.035e+3	8.078e+2	7.887e+2	1.633e+3
BIC	NaN	3.873e+3	3.371e+3	1.144e+3	1.293e+3	2.137e+3
$minP_{value}^{fit}$	**5.490e–1**	*0*	*0*	5.000e–3	1.020e–1	*0*
$meanP_{value}^{fit}$	NaN	*0*	*0*	5.914e–1	**6.978e–1**	7.500e–4

Distribution type	7	8	9	10	11
AIC	**7.847e+2**	1.444e+3	1.288e+3	*3.755e+3*	1.080e+3
BIC	**1.121e+3**	1.781e+3	1.457e+3	*4.092e+3*	1.416e+3
$minP_{value}^{fit}$	8.200e–2	*0*	*0*	*0*	*0*
$meanP_{value}^{fit}$	5.756e–1	2.592e–2	8.083e–2	*0*	1.118e–1

Legend: The numbers of the distribution types stand for the following: 1 – beta, 2– exponential, 3– extreme value, 4 – gamma, 5 – gen. extreme value, 6 – generalized Pareto; 7 – lognormal, 8 – normal, 9 – Rayleigh, 10 – uniform, 11 – Weibull

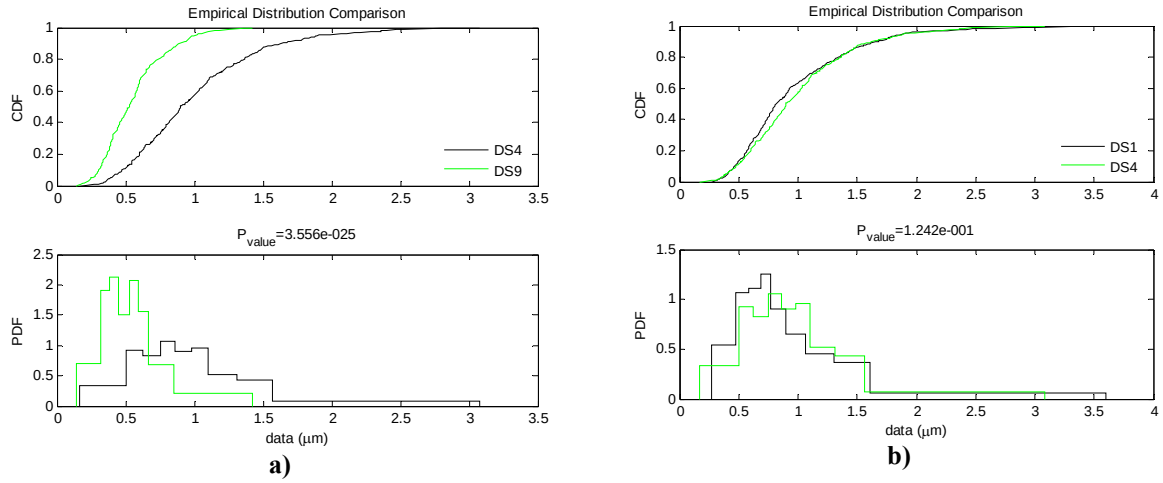


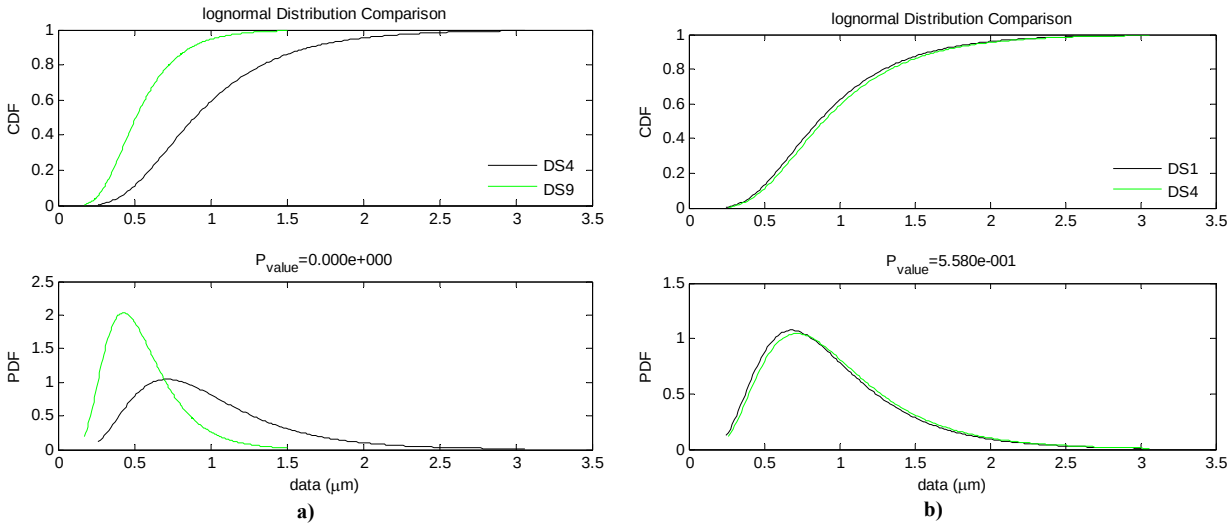
Fig. 4 Comparison of the stair-case empirical distributions over DS4 and DS9 (a), and over DS1 and DS4 (b).

Table 4. P-values of the statistical test for equality of stair-case distributions on pairs of datasets. The values on the main diagonal are shaded. The bolded values are those that exceed 0.05, i.e. indicate the pairs of datasets whose stair-case distributions are identical.

Datasets	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	DS10	DS11	DS12
DS1	1.00	**2.75e–1**	**5.21e–1**	**1.24e–1**	1.92e–6	7.20e–126	1.51e–3	2.79e–2	8.99e–20	6.48e–5	9.28e–3	6.72e–11
DS2	**2.75e–1**	1.00	**5.96e–1**	**5.13e–1**	6.69e–6	6.85e–45	9.03e–4	**1.99e–1**	4.62e–10	2.87e–4	2.39e–3	1.14e–8
DS3	**5.21e–1**	**5.96e–1**	1.00	**1.28e–1**	4.37e–8	1.65e–102	2.51e–5	**8.93e–2**	3.40e–21	1.66e–6	6.44e–4	3.65e–13
DS4	**1.24e–1**	**5.13e–1**	**1.28e–1**	1.00	1.38e–11	1.41e–124	1.83e–5	**8.94e–1**	3.56e–25	2.75e–8	1.23e–5	4.14e–18
DS5	1.92e–6	6.69e–6	4.37e–8	1.38e–11	1.00	2.35e–101	4.50e–3	3.93e–12	2.66e–4	**1.51e–1**	7.99e–3	**9.29e–2**
DS6	7.20e–126	6.85e–45	1.65e–102	1.41e–124	2.35e–101	1.00	6.06e–125	1.37e–126	2.92e–95	5.96e–126	8.07e–127	1.42e–125
DS7	1.51e–3	9.03e–4	2.51e–5	1.83e–5	4.50e–3	6.06e–125	1.00	3.48e–5	1.81e–11	**1.02e–1**	**3.49e–1**	8.65e–6
DS8	2.79e–2	**1.99e–1**	**8.93e–2**	**8.94e–1**	3.93e–12	1.37e–126	3.48e–5	1.00	1.78e–26	3.34e–9	2.11e–6	1.68e–19
DS9	8.99e–20	4.62e–10	3.40e–21	3.56e–25	2.66e–4	2.92e–95	1.81e–11	1.78e–26	1.00e	1.13e–6	1.47e–12	2.06e–3
DS10	6.48e–5	2.87e–4	1.66e–6	2.75e–8	**1.51e–1**	5.96e–126	**1.02e–1**	3.34e–9	1.13e–6	1.00	**9.51e–2**	4.26e–3
DS11	9.28e–3	2.39e–3	6.44e–4	1.23e–5	7.99e–3	8.07e–127	**3.49e–1**	2.11e–6	1.47e–12	**9.51e–2**	1.00	7.76e–5
DS12	6.72e–11	1.14e–8	3.65e–13	4.14e–18	**9.29e–2**	1.42e–125	8.65e–6	1.68e–19	2.06e–3	4.26e–3	7.76e–5	1.00

Table 5. P-values of the statistical test that the lognormal fitted distributions over two datasets are identical. The values on the main diagonal are shaded. The bolded values indicate the distribution fit pairs that may be assumed identical.

Datasets	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	DS10	DS11	DS12
DS1	1.00	**1.39e–1**	**1.90e–1**	**5.58e–1**	0.00	0.00	0.00	**3.49e–1**	0.00	0.00	0.00	0.00
DS2	**1.39e–1**	1.00	**6.37e–1**	**1.05e–1**	0.00	0.00	0.00	3.40e–2	0.00	0.00	1.00e–3	0.00
DS3	**1.90e–1**	**6.37e–1**	1.00	**2.01e–1**	0.00	0.00	0.00	3.20e–2	0.00	0.00	0.00	0.00
DS4	**5.58e–1**	**1.05e–1**	**2.01e–1**	1.00	0.00	0.00	0.00	**6.65e–1**	0.00	0.00	0.00	0.00
DS5	0.00	0.00	0.00	0.00	1.00	0.00	1.00e–3	0.00	0.00	**5.70e–2**	1.00e–3	**5.10e–2**
DS6	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
DS7	0.00	0.00	0.00	0.00	1.00e–3	0.00	1.00	0.00	0.00	**8.70e–2**	**7.90e–1**	0.00
DS8	**3.49e–1**	3.40e–2	3.20e–2	**6.65e–1**	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
DS9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
DS10	0.00	0.00	0.00	0.00	**5.70e–2**	0.00	**8.70e–2**	0.00	0.00	1.00	**1.86e–1**	0.00
DS11	0.00	1.00e–3	0.00	0.00	1.00e–3	0.00	**7.90e–1**	0.00	0.00	**1.86e–1**	1.00	0.00
DS12	0.00	0.00	0.00	0.00	**5.10e–2**	0.00	0.00	0.00	0.00	0.00	0.00	1.00



a) b)

Fig. 5 Comparison of the lognormal distribution fits over DS4 and DS9 (a) and over DS1 and DS4 (b).

Conclusions

The paper addressed the problem of finding a single type of theoretical distribution that fits to different datasets by altering its parameters. The identification of such type of distribution is a prerequisite for comparing the results, performing interpolation and extrapolation over the data, and studying the dependence between the input parameters (e.g. initial conditions of an experiment) and the distribution parameters. Additionally, the procedures included hypothesis tests over the equality of empirical (stair-case) and of fitted distributions. In the first case, the failure to reject the null hypothesis proves the samples come from one and the same general population. In the second case, the failure to reject the null hypothesis proves that although parameters are random (as the fits are also based on random data), the differences are not statistically significant. The implementation of the procedures is facilitated by the creation of a platform in MATLAB R2009a that executes the necessary calculation and evaluation procedures. The program functions of that platform are available free of charge upon request from the authors.

A current biochemical problem served as a demonstration of the procedures. The influence of thrombin concentration over the density of the fibrin network is tested. Datasets under different thrombin concentration and buffer are analyzed, each containing measures of the length of fibrin fibers between branching points. The measurements are taken from electron microscope images of dry fibrin samples using an automated procedure in MATLAB (its corresponding program function is also available free of charge upon request from the authors). The results proved that the most appropriate fit to the datasets is achieved by the lognormal distribution, whereas the worst fit is that of the uniform distribution. The comparison of the empirical and of the fitted lognormal distributions gives approximately the same results regarding the possibilities to unite samples and conduct further analysis over larger datasets.

Further extension of the statistical procedures developed in this paper may focus on the inclusion of additional statistical tests evaluating the quality of the fits and the equality of the distributions. The required simulation procedures may be realized also via Bootstrap, as this method relies on less assumptions about the underlying process and the associated measurement error [13]. Other theoretical distribution types should also be included in the program platform, especially those that can interpret different behavior of the data around the

mean and at the tails. Finally, further research could focus on new areas (e.g. economics, finance, management, other natural sciences, etc.) to implement the described procedures.

Acknowledgement

This research is partially funded by the INPORT project, No. DVU01/0031 (Bulgarian National Science Fund program for enhancement of scientific research). The authors wish to thank Imre Varju from the Department of Medical Biochemistry, Semmelweis University, Budapest, Hungary (1094 Budapest, Tűzoltó u. 37-47) for collecting the datasets with length measurements, and Laszlo Szabo from the Chemical Research Center, Hungarian Academy of Sciences, Budapest, Hungary (1025 Budapest, Pusztaszeri út 59-67) for taking the electron microscope images of the dry fibrin samples.

References:

1. Burnham K. P., D. R. Anderson (2002). Model Selection and Inference: A Practical Information-Theoretic Approach, Springer, 60-64.
2. Freedman, D., P. Diaconis (1981). On the Histogram as a Density Estimator: L_2 Theory, Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete, 57, 453–476.
3. Kuiper N. H. (1962). Tests Concerning Random Points on a Circle, Proceedings of the Koninklijke Nederlandse Akademie van Wetenshappen, ser. A, 63, 38-47.
4. Press W., B. Flannery, S. Teukolsky, W. Vetterling (1992). Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd ed., Cambridge University Press, Cambridge, England, 620-622.
5. Schwarz G. (1974). Estimating the Dimension of a Model, Annals of Statistic, 6, 461-464.
6. Scott D. W. (1979). On Optimal and Data-based Histograms, Biometrika, 66, 605-610.
7. Sturges H. A. (1926). The Choice of a Class Interval, J. Am. Stat. Assoc., 21, 65-66.
8. Tenekedjiev K., D. Dimitrakiev, N. D. Nikolova (2002). Building Frequentist Distribution of Continuous Random Variables, Machine Mechanics, 47, 164-168,
9. The MathWorks (2008). Statistics Toolbox™ 7.0 User's Guide.
10. Gujarati D. N. (1995). Basic Econometrics, Third Edition, McGraw-Hill, 15-318.
11. Politis D. (1998). Computer-intensive Methods in Statistical Analysis, IEEE Signal Processing Magazine, 15, 1, 39-55.
12. Finch S. R. (2003). Extreme Value Constants, Cambridge University Press, Cambridge, England, 363-367.
13. Efron B., R. J. Tibshirani (1993). An Introduction to the Bootstrap, Chapman &Hall, 45-59.
14. Hanke J. E., A. G. Reitsch (1991). Understanding Business Statistics, Irwin, 165-198.
15. Knuth D. E. (1998). The Art of Computer Programming, Vol. 2: Seminumerical Algorithms, 3rd ed. Reading, MA: Addison-Wesley, 45-52.

Assist. Prof. Natalia Nikolova, Ph.D.
E-mail: natalia@dilogos.com

Natalia Nikolova received her PhD in 2007 in the field of quantitative fuzzy-rational decision analysis. Since 2004 she works at the Dept. Of Economics and Management of Technical University – Varna, Bulgaria. Presently, she is an Assistant Professor. She has over 70 publications, of which journal and conference papers, as well as participation in three books. Main scientific interests: quantitative decision analysis, fuzzy-rational decision analysis, risk analysis, statistics, simulation modeling, econometrics.

Assist. Prof. Daniela Toneva, Ph.D.
E-mail: d_toneva@abv.bg

Daniela Toneva received her PhD in the field of economy in 2002 from Russian Academy of Science. Since 2006 she works as an Assistant Professor in dep. "Ecology and Environmental Protection" at Technical University – Varna. She has over 50 publications and 5 books. Main scientific interests: organization and management of manufacture; ecological monitoring; sustainable development.

Ana-Maria Tenekedjieva
E-mail: ana-maria.tenekedjieva@duke.edu

Ana-Maria Tenekedjieva graduated as a valedictorian from the High School of Mathematics – Varna in 2008. Since then, she is a majoring in mathematics at Duke University, North Carolina, USA. She is a recipient of the Angier B. Duke Scholarship, as well as Julia Dale Prize for excellence in mathematics. Ms Tenekedjieva is pursuing independent study in the laboratory of Michael Platt since January, 2009. She has 6 publications in Bulgaria. Main research interests: mathematics, neuroscience, statistics, quantitative decision analysis.