Chapter 1

Monte Carlo Statistical Tests for Identity of Theoretical and Empirical Distributions of Experimental Data

Natalia D. Nikolova, Daniela Toneva-Zheynova, Krasimir Kolev and Kiril Tenekedjiev

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/53049

1. Introduction

Often experimental work requires analysis of many datasets derived in a similar way. For each dataset it is possible to find a specific theoretical distribution that describes best the sample. A basic assumption in this type of work is that if the mechanism (experiment) to generate the samples is the same, then the distribution type that describes the datasets will also be the same [1]. In that case, the difference between the sets will be captured not through changing the type of the distribution, but through changes in its parameters. There are some advantages in finding whether a type of theoretical distribution that fits several datasets exists. At first, it improves the fit because the assumptions concerning the mechanism underlying the experiment can be verified against several datasets. Secondly, it is possible to investigate how the variation of the input parameters influences the parameters of the theoretical distribution. In some experiments it might be proven that the differences in the input conditions lead to qualitative change of the fitted distributions (i.e. change of the type of the distribution). In other cases the variation of the input conditions 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 [2], 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 merging datasets and operating on larger amount of data, which is a prerequisite for higher precision of the statistical results. This task is similar to testing for stability in regression analysis [3].

Formulating three separate tasks, this chapter solves the problem of identifying an appropriate distribution type that fits several one-dimensional (1-D) datasets and testing the statistical significance of the observed differences in the empirical and in the fitted distributions for each pair of samples. The first task (Task 1) aims at identifying a type of 1-D theoretical distribution that fits best the samples in several datasets by altering its parameters. The second task (Task 2) is to test the statistical significance of the difference between two empirical distributions of a pair of 1-D datasets. The third task (Task 3) is to test the statistical significance of the difference between two fitted distributions of the same type over two arbitrary datasets.

Task 2 can be performed independently of the existence of a theoretical distribution fit valid for all samples. Therefore, comparing and eventually merging pairs of samples will always be possible. This task requires comparing two independent discontinuous (stair-case) empirical cumulative distribution functions (CDF). It is a standard problem and the approach here is based on a symmetric variant of the Kolmogorov-Smirnov test [4] called the Kuiper two-sample test, which essentially performs an estimate of the closeness of a pair of independent stair-case CDFs by finding the maximum positive and the maximum negative deviation between the two [5]. The distribution of the test statistics is known and the p value of the test can be readily estimated.

Tasks 1 and 3 introduce the novel elements of this chapter. Task 1 searches for a type of theoretical distribution (out of an enumerated list of distributions) which fits best multiple datasets by varying its specific parameter values. The performance of a distribution fit is assessed through four criteria, namely the Akaike Information Criterion (AIC) [6], the Bayesian Information Criterion (BIC) [7], 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 merged. In Task 1 the distribution of the Kuiper statistic cannot be calculated in a closed form, because the problem is to compare an empirical distribution with its own fit and the independence is violated. A distribution of the Kuiper statistic in Task 3 cannot be estimated in close form either, because here one has to compare two analytical distributions, but not two stair-case CDFs. For that reason the distributions of the Kuiper statistic in Tasks 1 and 3 are constructed via a Monte Carlo simulation procedures, which in Tasks 1 is based on Bootstrap [8].

The described approach is illustrated with practical applications for the characterization of the fibrin structure in natural and experimental thrombi evaluated with scanning electron microscopy (SEM).

2. Theoretical setup

The approach considers N 1-D 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>64 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 contain samples of the same random quantity, but under slightly different conditions.

http://dx.doi.org/10.5772/53049

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

2.1. Task 1 – Theoretical solution

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 two 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. This is the most frequent case when the sample values are positive and the lower external node will never be with a negative abscissa because $\left(x_1^i - \Delta_d^i\right) \geq 0$. If both negative and positive sample values are acceptable then $\Delta_d^i = \left(x_{16}^i - x_1^i\right)/30$ and $\Delta_u^i = \left(x_{n_i}^i - x_{n_i-15}^i\right)/30$. Of course if all the sample values have to be negative then $\Delta_d^i = \left(x_{16}^i - x_1^i\right)/30$ and $\Delta_u^i = \min\left(-x_{n_i}^i, \left(x_{n_i}^i - x_{n_i-15}^i\right)/30\right)$. In that rare case the upper external node will never be with positive abscissa because $\left(x_{n_i}^i + \Delta_u^i\right) \leq 0$.

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 $x_{k-1}^i < x_k^i = x_{k+1}^i = x_{k+2}^i = \ldots = x_{k+p}^i < x_{k+p+1}^i$, 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 $CDF_e^i\left(x_k^i\right) = (k + p/2 - 1/2)/n_i$. The described two-step procedure [2] 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: $std_e^i = \sqrt{\frac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(x_k^i - mean_e^i\right)^2}$;
- inter-quartile 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 [9], Sturges [10] and Freedman-Diaconis [11] suggestions: $b_i = \min\left\{b_i^{Sc},\, b_i^{St},\, b_i^{FD}\right\}$, where $b_i^{Sc} = fl\left(0.2865\left(x_{n_i}^i - x_1^i\right)\sqrt[3]{n_i} \,/\, std_e^i\right)$, $b_i^{St} = fl\left(1 + log_2(n_i)\right)$, and $b_i^{FD} = fl\left(0.5\left(x_{n_i}^i - x_1^i\right)\sqrt[3]{n_i} \,/\, iqr_e^i\right)$. In the last three formulae, $fl(y)$ stands for the greatest whole number less or equal to y. 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 as $PDF_e^i(x) = b_i^{-1} / \left(m_{u,k}^i - m_{d,k}^i\right)$. The described procedure [2] results in a histogram, where the relative error of the worst $PDF_e^i(.)$ estimate is minimal from all possible splitting of the samples into b_i bins. This is so because the PDF estimate of a bin is found as the probability that the random variable would have a value in that bin divided to the bin's width. This probability is estimated as the relative frequency to have a data point in that bin at the given data set. The closer to zero that frequency is the worse it has been estimated. That is why the worst PDF estimate is at the bin that contains the least number of data points. Since for the proposed distribution each bin contains equal number of data points, any other division to the same number of bins would result in having a bin with less data points. Hence, the relative error of its PDF estimate would be worse.

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, …, b_i.

If the samples are distributed with density $PDF_j(x, \vec{p}_j)$, then the likelihood of the dataset χ^i is $L_j^i(\vec{p}_j) = \prod_{k=1}^{n_i} PDF_j\left(x_k^i, \vec{p}_j\right)$. 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: $mean_j^i = \int_{-\infty}^{+\infty} x.PDF_j\left(x, \vec{p}_j^{\,i}\right)dx$
- median: $med_j^i = invCDF_j\left(0.5, \vec{p}_j^{\,i}\right)$
- mode: $mode_j^i = \arg\left\{\max_x\left[PDF_j(x, \vec{p}_j)\right]\right\}$
- standard deviation: $std_j^i = \sqrt{\int_{-\infty}^{+\infty}\left(x - mean_j^i\right)^2 PDF_j\left(x, \vec{p}_j^{\,i}\right)dx};$

http://dx.doi.org/10.5772/53049

- inter-quartile range: $iqr_j^i = invCDF_j\left(0.75, \vec{p}^{\,i}_j\right) - invCDF_j\left(0.25, \vec{p}^{\,i}_j\right)$.

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 $CDF_j\left(x, \vec{p}^{\,i}_j\right)$, which means that the sample χ^i is drawn from $CDF_j\left(x, \vec{p}^{\,i}_j\right)$. The alternative hypothesis H_1 is that $CDF_e^i(x)$ is different from $CDF_j\left(x, \vec{p}^{\,i}_j\right)$, which means that the fit is not good. The Kuiper statistic V_j^i [12] is a suitable measure for the goodness-of-fit of the theoretical cumulative distribution functions $CDF_j\left(x, \vec{p}^{\,i}_j\right)$ to the dataset χ^i:

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

The theoretical Kuiper's distribution is derived just for the case of two independent staircase distributions, but not for continuous distribution fitted to the data of another [5]. That is why the distribution of V from (1), if H_0 is true, should be estimated by a Monte Carlo procedure. The main idea is that if the dataset $\chi^i = \left(x_1^i, x_2^i, \ldots, x_{n_i}^i\right)$ is distributed in compliance with the 1-D theoretical distributions of type j, then its PDF would be very close to its estimate $PDF_j\left(x, \vec{p}^{\,i}_j\right)$, and so each synthetic dataset generated from $PDF_j\left(x, \vec{p}^{\,i}_j\right)$ would produce Kuiper statistics according to (1), which would be close to zero [1].

The algorithm of the proposed procedure is the following:

1. Construct the empirical cumulative distribution function $CDF_e^i(x)$ describing the data in χ^i.
2. Find the MLE of the parameters for the distributions of type j fitting χ^i as $\vec{p}^{\,i}_j = \arg\left\{\max_{\vec{p}_j}\left[\prod_{k=1}^{n_i} PDF_j\left(x_k^i, \vec{p}_j\right)\right]\right\}$.
3. Build the fitted cumulative distribution function $CDF_j\left(x, \vec{p}^{\,i}_j\right)$ describing χ^i.
4. Calculate the actual Kuiper statistic V_j^i according to (1).
5. Repeat for r=1,2,…, n^{MC} (in fact use n^{MC} simulation cycles):
   a. generate a synthetic dataset $\chi_r^{i,syn} = \left\{x_{1,r}^{i,syn}, x_{2,r}^{i,syn}, \ldots, x_{n_i,r}^{i,syn}\right\}$ from the fitted cumulative distribution function $CDF_j\left(x, \vec{p}^{\,i}_j\right)$. The dataset $\chi_r^{i,syn}$ contains n_i sorted samples $\left(x_{1,r}^{i,syn} \le x_{2,r}^{i,syn} \le \ldots \le x_{n_i,r}^{i,syn}\right)$;
   b. construct the synthetic empirical distribution function $CDF_{e,r}^{i,syn}(x)$ describing the data in $\chi_r^{i,syn}$;
   c. find the MLE of the parameters for the distributions of type j fitting $\chi_r^{i,syn}$ as

$$\vec{p}_{j,r}^{\,i,syn} = \arg\left\{\max_{\vec{p}_j}\left[\prod_{k=1}^{n_i} PDF_j\left(x_{k,r}^{i,syn}, \vec{p}_j\right)\right]\right\};$$

d. build the theoretical distribution function $CDF_{j,r}^{syn}\left(x, \vec{p}_{j,r}^{\,i,syn}\right)$ describing $\chi_r^{i,syn}$;

e. estimate the r^{th} instance of the synthetic Kuiper statistic as $V_{j,r}^{i,syn} = \max_x\left\{CDF_{e,r}^{i,syn}(x) - CDF_{j,r}^{syn}\left(x, \vec{p}_{j,r}^{\,i,syn}\right)\right\} + \max_x\left\{CDF_{j,r}^{syn}\left(x, \vec{p}_{j,r}^{\,i,syn}\right) - CDF_{e,r}^{i,syn}(x)\right\}$.

6. 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 from step 4:

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

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:

$$min P_{value,j}^{fit} = \min\left\{P_{value,j}^{fit,1}, P_{value,j}^{fit,2}, ..., P_{value,j}^{fit,N}\right\}, \ for \ \ j=1, 2, ...,M. \tag{3}$$

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

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

The second and the third auxiliary criteria are the AIC-Akaike Information Criterion [6] and the BIC-Bayesian Information Criterion [7], which corrects the negative log-likelihoods with the number of the assessed parameters:

http://dx.doi.org/10.5772/53049

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

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

for j=1,2,..,M. The best theoretical distribution type should have maximal values for $minP^{fit}_{value,\,j}$ and $meanP^{fit}_{value,\,j}$, whereas its values for AIC_j and BIC_j should be minimal. On top, the best theoretical distribution type should have $minP^{fit}_{value,\,j}$>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.

2.2. Task 2 – Theoretical solution

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 [13] are built from the two datasets χ^{i1} and χ^{i2}:

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

The "staircase" empirical $CDF^{\,i}_{sce}(.)$ is a discontinuous version of the already defined empirical $CDF^{\,i}_{e}(.)$. The Kuiper statistic $V^{i1,i2}$ [12] is a measure for the closeness of the two 'staircase' empirical cumulative distribution functions $CDF^{\,i1}_{sce}(.)$ and $CDF^{\,i2}_{sce}(.)$:

$$V^{i1,i2} = \max_x\left\{CDF^{\,i1}_{sce}(x) - CDF^{\,i2}_{sce}(x)\right\} + \max_x\left\{CDF^{\,i2}_{sce}(x) - CDF^{\,i1}_{sce}(x)\right\} \tag{8}$$

The distribution of the test statics $V^{i1,i2}$ is known and the *p*-value of the two tail statistical test with 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 [5] according to formulae (9) and (10).

The algorithm for the theoretical solution of Task 2 is straightforward:

1. Construct the "staircase" empirical cumulative distribution function describing the data in χ^{i1} as $CDF_{sce}^{i1}(x) = \sum_{\substack{k=1 \\ x_k^{i1} \le x}}^{n_{i1}} 1/n_{i1}$.

2. Construct the "staircase" empirical cumulative distribution function describing the data in χ^{i2} as $CDF_{sce}^{i2}(x) = \sum_{\substack{k=1 \\ x_k^{i2} \le x}}^{n_{i2}} 1/n_{i2}$.

3. Calculate the actual Kuiper statistic $V^{i1,i2}$ according to (8).

4. The *p*-value of the statistical test (the probability to reject a true null hypothesis H_0) is estimated as:

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

where

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

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

2.3. Task 3 – Theoretical solution

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 $V_j^{i1,i2}$:

http://dx.doi.org/10.5772/53049

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

As it has already been mentioned the theoretical Kuiper's distribution is derived just for the case of two independent staircase distributions, but not for the case of two independent continuous cumulative distribution functions. That is why the distribution of V from (11), if H_0 is true, should be estimated by a Monte Carlo procedure. The main idea is that 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 merged dataset χ^{i1+i2} formed by merging the samples of χ^{i1} and χ^{i2} [1].

The algorithm of the proposed procedure is the following:

1. Find the MLE of the parameters for the distributions of type j fitting χ^{i1} as $\vec{p}_j^{i1} = \arg\left\{ \max_{\vec{p}_j} \left[\prod_{k=1}^{n_{i1}} PDF_j\left(x_k^{i1}, \vec{p}_j\right) \right] \right\}$.

2. Build the fitted cumulative distribution function $CDF_j\left(x, \vec{p}_j^{i1}\right)$ describing χ^{i1}.

3. Find the MLE of the parameters for the distributions of type j fitting χ^{i2} as $\vec{p}_j^{i2} = \arg\left\{ \max_{\vec{p}_j} \left[\prod_{k=1}^{n_{i2}} PDF_j\left(x_k^{i2}, \vec{p}_j\right) \right] \right\}$.

4. Build the fitted cumulative distribution function $CDF_j\left(x, \vec{p}_j^{i2}\right)$ describing χ^{i2}.

5. Calculate the actual Kuiper statistic $V_j^{i1,i2}$ according to (11).

6. Merge the samples χ^{i1} and χ^{i2}, and form the merged data set χ^{i1+i2}.

7. Find the MLE of the parameters for the distributions of type j fitting χ^{i1+i2} as $\vec{p}_j^{i1+i2} = \arg\left\{ \max_{\vec{p}_j} \left[\prod_{k=1}^{n_{i1}} PDF_j\left(x_k^{i1}, \vec{p}_j\right) \prod_{k=1}^{n_{i2}} PDF_j\left(x_k^{i2}, \vec{p}_j\right) \right] \right\}$.

8. Fit the merged fitted cumulative distribution function $CDF_j\left(x, \vec{p}_j^{i1+i2}\right)$ to χ^{i1+i2}.

9. Repeat for r=1,2,…, n^{MC} (in fact use n^{MC} simulation cycles):

   a. a. generate a synthetic dataset $\chi_r^{i1,syn} = \left\{ x_{1,r}^{i1,syn}, x_{2,r}^{i1,syn}, \ldots, x_{n_{i1},r}^{i1,syn} \right\}$ from the fitted cumulative distribution function $CDF_j\left(x, \vec{p}_j^{i1+i2}\right)$;

   b. b. find the MLE of the parameters for the distributions of type j fitting $\chi_r^{i1,syn}$ as

   $$\vec{p}_{j,r}^{i1,syn} = \arg\left\{ \max_{\vec{p}_j} \left[\prod_{k=1}^{n_{i1}} PDF_j\left(x_{k,r}^{i1,syn}, \vec{p}_j\right) \right] \right\};$$

   c. c. build the theoretical distribution function $CDF_{j,r}^{syn}\left(x, \vec{p}_{j,r}^{i1,syn}\right)$ describing $\chi_r^{i1,syn}$;

d. d. generate a synthetic dataset $\chi_r^{i2,syn}=\left\{x_{1,r}^{i2,syn}, x_{2,r}^{i2,syn}, ..., x_{n_{i2},r}^{i2,syn}\right\}$ from the fitted cumulative distribution function $CDF_j\left(x, \vec{p}_j^{i1+i2}\right)$;

e. e. find the MLE of the parameters for the distributions of type j fitting $\chi_r^{i2,syn}$ as

$$\vec{p}_{j,r}^{i2,syn}=\arg\left\{\max_{\vec{p}_j}\left[\prod_{k=1}^{n_{i2}} PDF_j\left(x_{k,r}^{i2,syn}, \vec{p}_j\right)\right]\right\};$$

f. f. build the theoretical distribution function $CDF_{j,r}^{syn}\left(x, \vec{p}_{j,r}^{i2,syn}\right)$ describing $\chi_r^{i2,syn}$;

g. g. estimate the r^{th} instance of the synthetic Kuiper statistic as:

$$V_{j,r}^{i1,i2,syn}=\max_x\left\{CDF_{j,r}^{syn}\left(x, \vec{p}_{j,r}^{i1,syn}\right)-CDF_{j,r}^{syn}\left(x, \vec{p}_{j,r}^{i2,syn}\right)\right\}+$$
$$+\max_x\left\{CDF_{j,r}^{syn}\left(x, \vec{p}_{j,r}^{i2,syn}\right)-CDF_{j,r}^{syn}\left(x, \vec{p}_{j,r}^{i1,syn}\right)\right\}.$$

10. The p-value $P_{value,j}^{i1,i2}$ of the statistical test (the probability to reject a true hypothesis H_0 that the j^{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 $V_j^{i1,i2}$ from step 5:

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

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.

3. Software

A platform of program functions, written in MATLAB environment, is created to execute the statistical procedures from the previous section. At present 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 [14, 15]. The platform also permits the user to add optional types of distribution.

The platform contains several main program functions. The function *set_distribution* contains the information about the 11 distributions, particularly their names, and the links to the functions that operate with the selected distribution type. Also, the function permits the inclusion of new distribution type. In that case, the necessary information the user must provide as input

http://dx.doi.org/10.5772/53049

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.

	Beta distribution		**Lognormal distribution**
Parameters	$a>0, \beta>0$	Parameters	$\mu \in (-\infty;\ +\infty)$, $\sigma>0$,
Support	$x \in [0;\ 1]$	Support	$x \in [0;\ +\infty)$
PDF	$f(x;\ a,\ \beta)=\frac{x^{a-1}(1-x)^{\beta-1}}{B(a,\ \beta)}$, where $B(a,\ \beta)$ 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	$a, \beta\neq 0$	Parameters	$\sigma>0$
Support	$x \in (-\infty;\ +\infty)$	Support	$x \in [0;\ +\infty)$
PDF	$f(x\ ;\ a,\ \beta)=\frac{e^{\left[(a-x)/\beta\right]}-e^{(a-x)/\beta}}{\beta}$	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	$x>\mu-\sigma/\xi \quad (\xi>0), x<\mu-\sigma/\xi \quad (\xi<0),$ $x \in (-\infty;\ +\infty) \quad (\xi=0)$	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	$f(x;\ x_m,\ k)=\frac{k x_m^k}{x^{k+1}}$		

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

The program function *kutest2* 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 Task 2 (see section 2.2) (to check whether two different datasets are drawn from the same general population).

Another key function is *fitdata*. 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 Task 1 and 3 (see sections 2.1 and 2.2). To execute the Kuiper test the function calls *kutest*. Finally, the program function *plot_print_data* provides the on-screen results from the statistical analysis and plots figures containing the pair of distributions that are analyzed. The developed software is available free of charge upon request from the authors provided proper citation is done in subsequent publications.

4. Source of experimental data for analysis

The statistical procedures and the program platform introduced in this chapter 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 (thrombin) activated in the damaged tissue 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 3 ml 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 %), 1:1 mixture of 96 %(v/v) ethanol/acetone and pure acetone followed by critical point drying with CO2 in E3000 Critical Point Drying Apparatus (Quorum Technologies, Newhaven, UK). 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 presented in Fig. 1). Some main parameters of the 12 collected datasets are given in Table 2.

http://dx.doi.org/10.5772/53049

An automated procedure is elaborated in MATLAB 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 SEM images. The procedure takes the file name of the fibrin image (see Fig. 1) and the planned number of measurements as input. 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 setting of the scale. A prompt appears, asking the user to type the numerical value of the length of the scale in μm. Then the image appears on screen and a red line has to be moved and resized to fit the scale (Fig. 2a and 2b). The third step requires a red rectangle to be placed over the actual image of the fibrin for selection of the region of interest (Fig. 2c). With this, the preparations of the image are done, and the user can start taking the desired number of measurements for the distances between adjacent nodes (Fig. 2d).

Using this approach 12 datasets containing measurements of lengths between branching points of fibrin have been collected (Table 2) and the three statistical tasks described above are executed over these datasets.

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

Table 2. Distance between branching points of fibrin fibers. Sample size (*N*), mean ($mean_e$ in μm), median (med_e in μm), standard deviation (std_e), inter-quartile range (iqr_e, in μm) of the empirical distributions over the 12 datasets for different thrombin concentrations (in U/ml) and buffers are presented

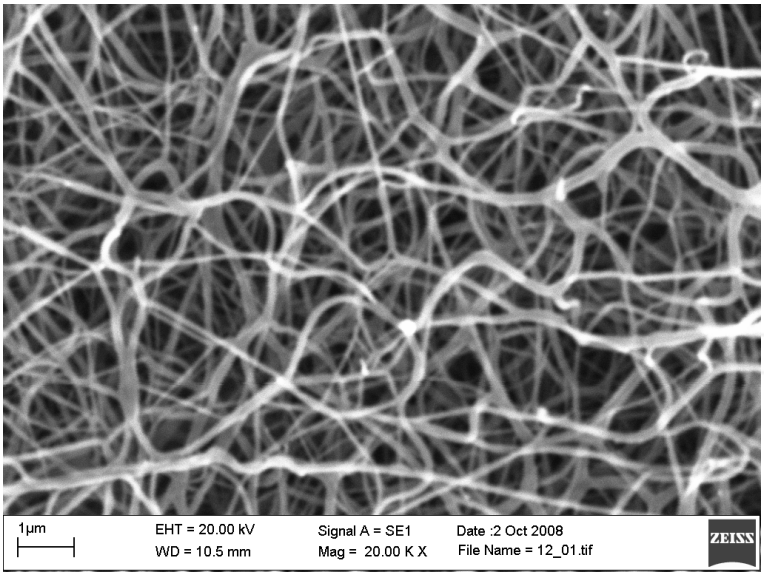


Figure 1. SEM image of fibrin used for morphometric analysis

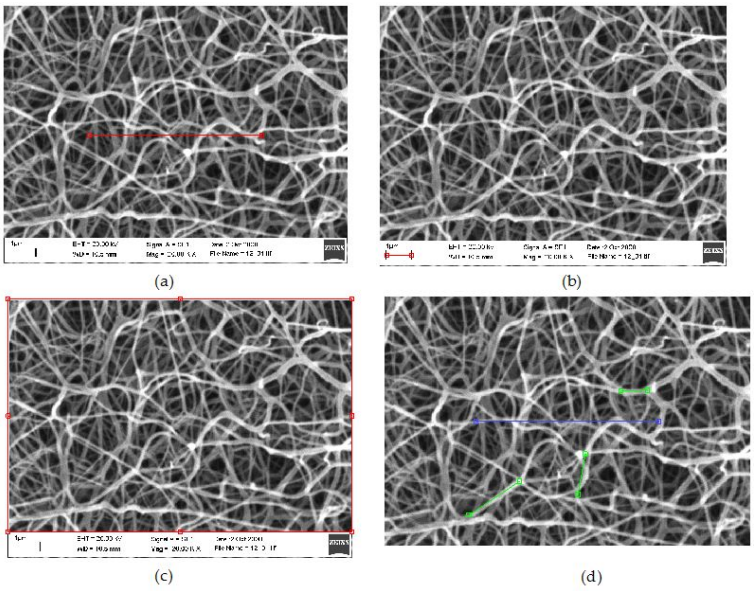


Figure 2. Steps of the automated procedure for measuring distances between branching points in fibrin. Panels *a* and *b*: scaling. Panel c: selection of region of interest. Panel d: taking a measurement

4.1. Task 1 – Finding a common distribution fit

A total of 11 types of distributions (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 $minP^{fit}_{value,\,j}$ and $meanP^{fit}_{value,\,j}$, along with 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 output of the former is NaN (not-a-number) since it does not apply to values of $x \notin [0; 1]$. The latter 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 $meanP^{fit}_{value,\,j}$). Figure 3 presents 4 of the 11 distribution fits to DS4. Similar graphical output is generated for all other datasets and for all distribution types.

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
$minP^{fit}_{value}$	**5.490e–1**	*0*	*0*	5.000e–3	1.020e–1	*0*
$meanP^{fit}_{value}$	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	
$minP^{fit}_{value}$	8.200e–2	*0*	*0*	*0*	*0*	
$meanP^{fit}_{value}$	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 - generalized extreme value, 6 – generalized Pareto; 7 – lognormal, 8 – normal, 9 – Rayleigh, 10 – uniform, 11 – Weibull

Table 3. Values of the criteria used to evaluate the goodness-of-fit of 11 types of distributions over the datasets with 1000 Monte Carlo simulation cycles. The table contains the maximal values for $minP^{fit}_{value,\,j}$ and $meanP^{fit}_{value,\,j}$, and the minimal values for AIC_j and BIC_j across the datasets for each distribution type. The bold and the italic values are the best one and the worst one achieved for a given criterion, respectively.

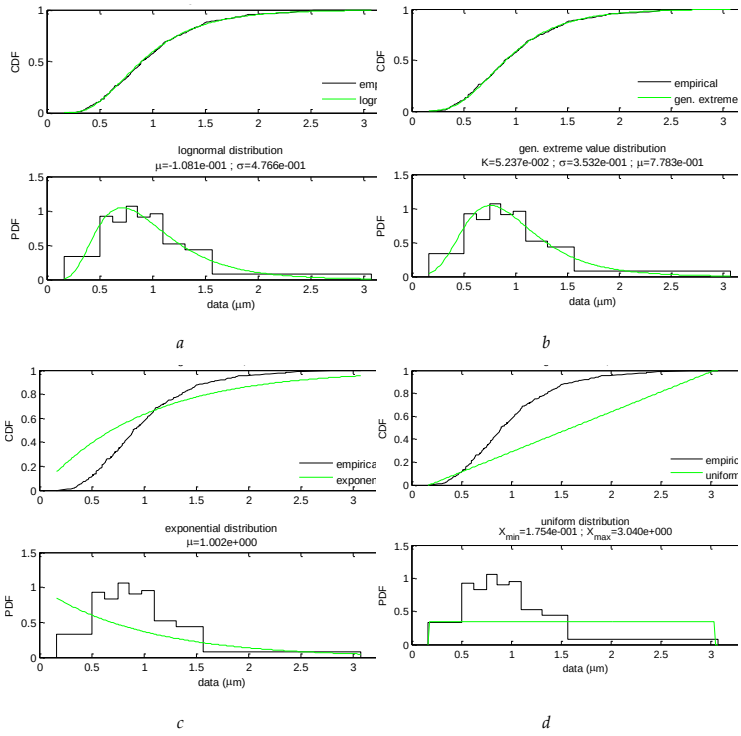


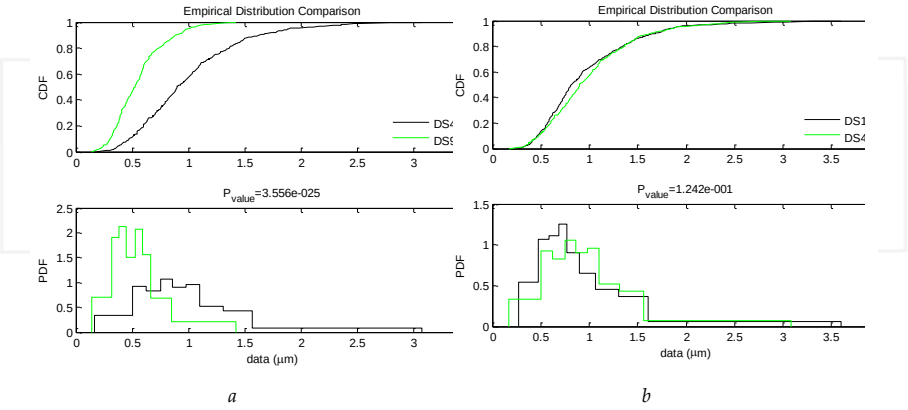
Figure 3. Graphical results from the fit of the lognormal (a), generalized extreme value (b), exponential (c), and uniform (d) distributions over DS4 (where μ, σ, X_{min}, X_{max}, k are the parameters of the theoretical distributions from Table 1)

4.2. Task 2 – Identity 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 merge the following datasets: 1) DS1, DS2, DS3, D4 and DS8; 2) DS7, DS10, and DS11; 3) DS5 and DS12. All other combinations (except DS5 and DS10) 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) $P_{value,e}^{4,9}$=2.024e–24<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 $mean_e^1$

http://dx.doi.org/10.5772/53049

=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) $P_{value,e}^{1,4}$=0.1957>0.05. Therefore the null hypothesis fails to be rejected, which is also confirmed by the graphical output.



a *b*

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

Datasets	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	DS10	DS11	DS12
DS1	1.00e+00	**3.81e-01**	**6.18e-01**	**1.96e-01**	5.80e-06	8.88e-125	3.46e-03	**5.21e-02**	4.57e-19	1.73e-04	1.89e-02	2.59e-10
DS2	**3.81e-01**	1.00e+00	**6.77e-01**	**6.11e-01**	1.94e-05	5.13e-44	2.13e-03	**2.92e-01**	1.71e-09	7.17e-04	5.34e-03	3.96e-08
DS3	**6.18e-01**	**6.77e-01**	1.00e+00	**2.01e-01**	1.46e-07	1.84e-101	6.94e-05	**1.47e-01**	1.79e-20	5.05e-06	1.55e-03	1.53e-12
DS4	**1.96e-01**	**6.11e-01**	**2.01e-01**	1.00e+00	5.47e-11	1.73e-123	5.14e-05	**8.57e-01**	2.02e-24	9.34e-08	3.50e-05	2.02e-17
DS5	5.80e-06	1.94e-05	1.46e-07	5.47e-11	1.00e+00	2.61e-100	9.67e-03	1.59e-11	6.68e-04	**2.32e-01**	1.65e-02	**1.52e-01**
DS6	8.88e-125	5.13e-44	1.84e-101	1.73e-123	2.61e-100	1.00e+00	7.45e-124	1.69e-125	3.14e-94	7.35e-125	9.98e-126	1.75e-124
DS7	3.46e-03	2.13e-03	6.94e-05	5.14e-05	9.67e-03	7.45e-124	1.00e+00	9.53e-05	7.13e-11	**1.64e-01**	**4.59e-01**	2.49e-05
DS8	**5.21e-02**	**2.92e-01**	**1.47e-01**	**8.57e-01**	1.59e-11	1.69e-125	9.53e-05	1.00e+00	1.04e-25	1.19e-08	6.36e-06	8.47e-19
DS9	4.57e-19	1.71e-09	1.79e-20	2.02e-24	6.68e-04	3.14e-94	7.13e-11	1.04e-25	1.00e+00	3.48e-06	6.05e-12	4.64e-03
DS10	1.73e-04	7.17e-04	5.05e-06	9.34e-08	**2.32e-01**	7.35e-125	**1.64e-01**	1.19e-08	3.48e-06	1.00e+00	**1.55e-01**	9.18e-03
DS11	1.89e-03	5.34e-03	1.55e-03	3.50e-05	1.65e-02	9.98e-126	**4.59e-01**	6.36e-06	6.05e-12	1.55e-01	1.00e+00	2.06e-04
DS12	2.59e-10	3.96e-08	1.53e-12	2.02e-17	**1.52e-01**	1.75e-124	2.49e-05	8.47e-19	4.64e-03	9.18e-03	2.06e-04	1.00e+00

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

4.3. Task 3 – Identity 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 bold 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, $mode_7^4$=0.7169, std_7^4=0.5077, $iqr_7^4 dy$=0.5870) and DS9 (with μ= –0.6694, σ=0.4181, $mean_7^9$=0.5587, med_7^9=0.5120, $mode_7^9$ =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, Fig. 5b presents the lognormal distribution fit over DS1 (with μ= –1477, σ=0.4843, $mean_7^1$=0.9701, med_7^1=0.8627, $mode_7^1$=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.

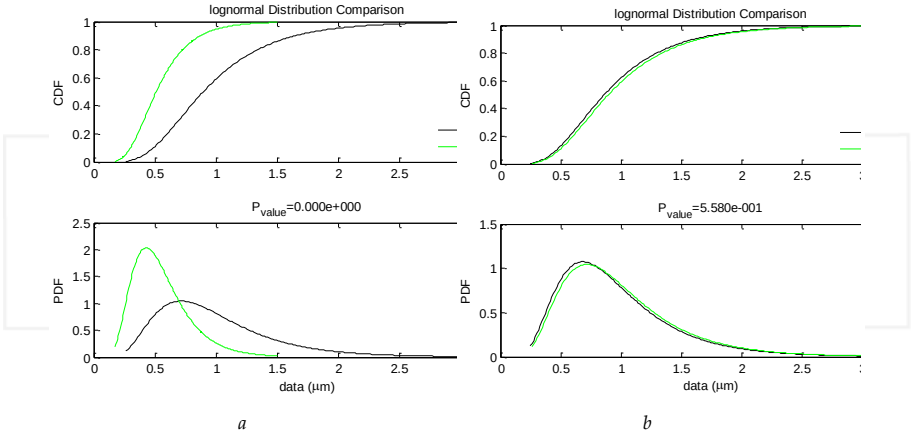


Figure 5. Comparison of the lognormal distribution fits over DS4 and DS9 (a) and over DS1 and DS4 (b)

http://dx.doi.org/10.5772/53049

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

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 bold values indicate the distribution fit pairs that may be assumed as identical.

The statistical procedures described above have been successfully applied for the solution of important medical problems [16; 17]. At first we could prove the role of mechanical forces in the organization of the final architecture of the fibrin network. Our *ex vivo* exploration of the ultrastructure of fibrin at different locations of surgically removed thrombi evidenced gross differences in the fiber diameter and pore area of the fibrin network resulting from shear forces acting in circulation (Fig. 6). *In vitro* fibrin structures were also generated and their equivalence with the *in vivo* fibrin architecture was proven using the distribution analysis described in this chapter (Fig. 7). Stretching changed the arrangement of the fibers (Fig. 7A) to a pattern similar to the one observed on the surface of thrombi (Fig. 6A); both the median fiber diameter and the pore area of the fibrins decreased 2-3-fold and the distribution of these morphometric parameters became more homogeneous (Fig. 7B). Thus, following this verification of the experimental model ultrastructure, the *in vitro* fibrin clots could be used for the convenient evaluation of these structures with respect to their chemical stability and resistance to enzymatic degradation [16].

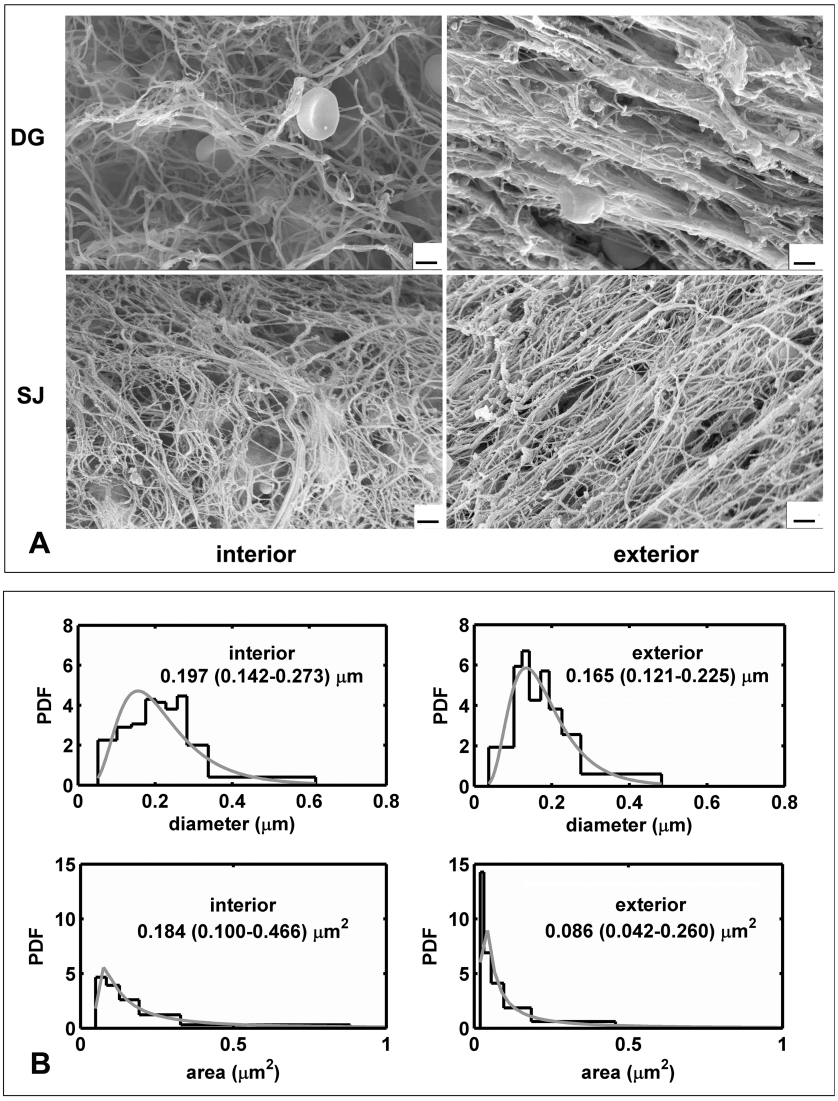


Figure 6. Fibrin structure on the surface and in the core of thrombi. A. Following thrombectomy thrombi were washed, fixed and dehydrated. SEM images were taken from the surface and transverse section of the same thrombus sample, scale bar = 2 μm. DG: a thrombus from popliteal artery, SJ: a thrombus from aorto-bifemoral by-pass Dacron graft. B. Fiber diameter (upper graphs) and fibrin pore area (lower graphs) were measured from the SEM images of the DG thrombus shown in A using the algorithms described in this chapter. The graphs present the probability density function (PDF) of the empirical distribution (black histogram) and the fitted theoretical distribution (grey curves). The numbers under the location of the observed fibrin structure show the median, as well as the bottom and the top quartile values (in brackets) of the fitted theoretical distributions (lognormal for fiber diameter and generalized extreme value for area). The figure is reproduced from Ref. [16].

http://dx.doi.org/10.5772/53049

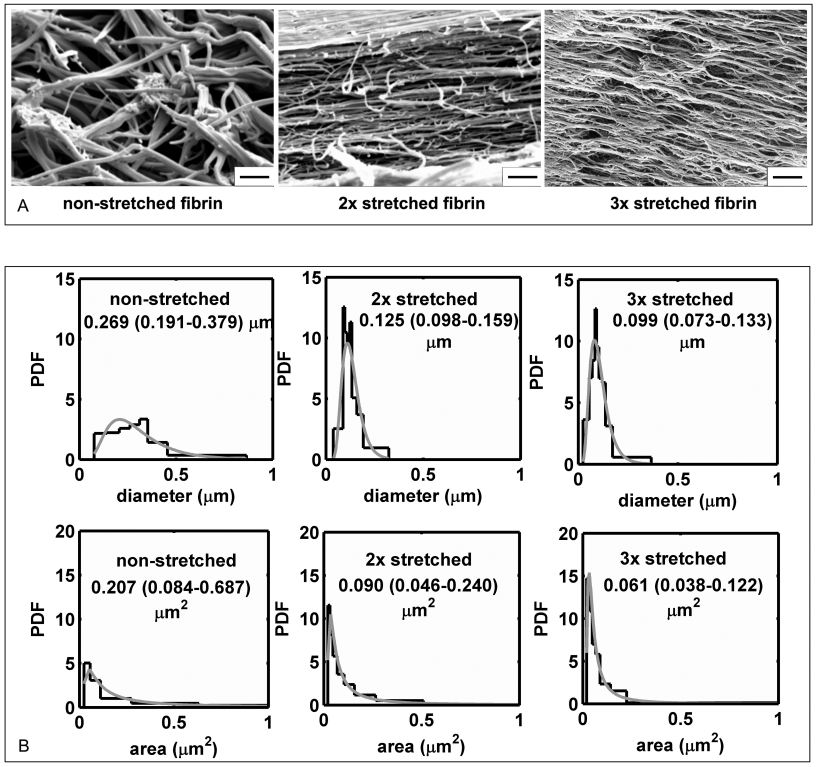


Figure 7. Changes in fibrin network structure caused by mechanical stretching. A. SEM images of fibrin clots fixed with glutaraldehyde before stretching or following 2-and 3-fold stretching as indicated, scale bar = 2 μm. B. Fiber diameter (upper graphs) and fibrin pore area (lower graphs) were measured from the SEM images illustrated in A using the algorithms described in this chapter. The graphs present the probability density function (PDF) of the empiric distribution (black histogram) and the fitted theoretical distribution (grey curves). The numbers under the fibrin type show the median, as well as the bottom and the top quartile values (in brackets) of the fitted theoretical distributions (lognormal for fiber diameter and generalized extreme value for area). The figure is reproduced from Ref. [16].

Application of the described distribution analysis allowed identification of the effect of red blood cells (RBCs) on the structure of fibrin [17]. The presence of RBCs at the time of fibrin formation causes a decrease in the fiber diameter (Fig. 8) based on a specific interaction between fibrinogen and a cell surface receptor. The specificity of this effect could be proven partially by the sensitivity of the changes in the distribution parameters to the presence of a drug (eptifibatide) that blocks the RBC receptor for fibrinogen (compare the median and interquartile range values for the experimental fibrins in the presence and absence of the drug illustrated in Fig. 8). It is noteworthy that the type of distribution was not changed by the drug, only its parameters were modified. This example underscores the applicability of the designed procedure for testing of statistical hypotheses in situations when subtle quantitative biological and pharmacological effects are at issue.

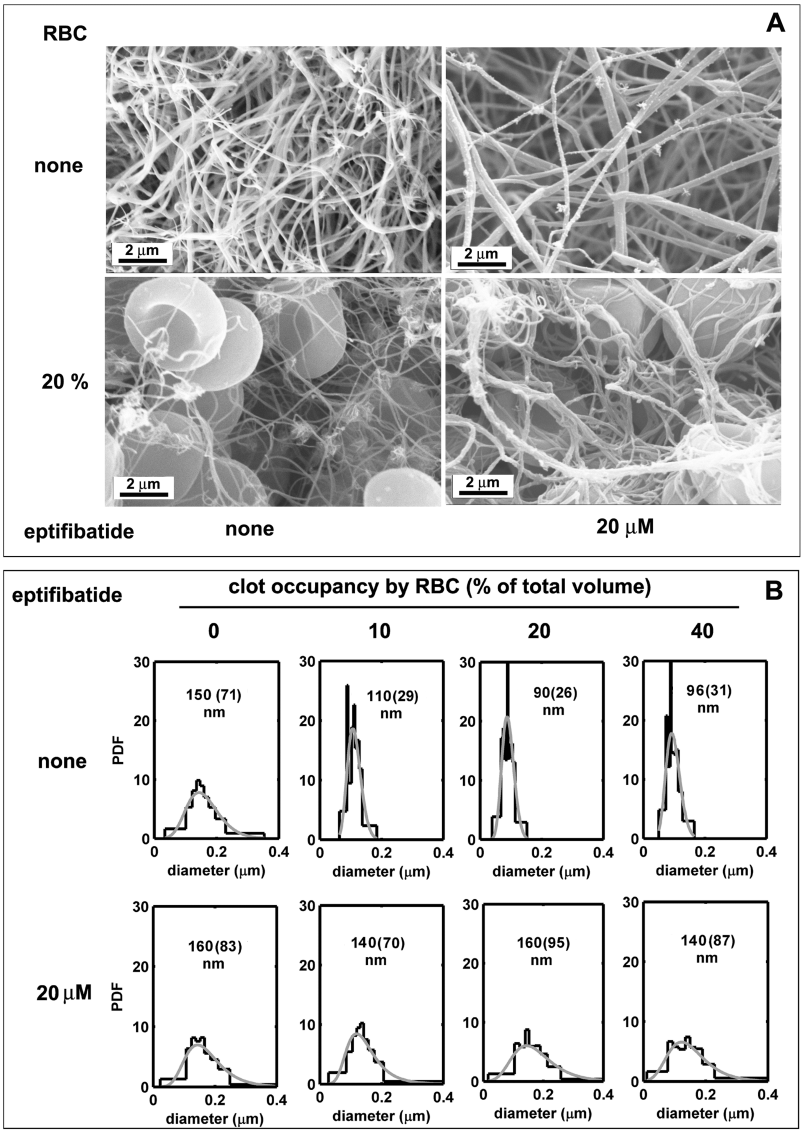


Figure 8. Changes. in the fibrin network structure caused by red blood cells and eptifibatide. The SEM images in Panel A illustrate the fibrin structure in clots of identical volume and fibrinogen content in the absence or presence of 20 % RBC. Panel B shows fiber diameter measured from the SEM images for a range of RBC-occupancy in the same clot model. Probability density functions (PDF) of the empirical distribution (black histogram) and the fitted lognormal theoretical distribution (grey curves) are presented with indication of the median and the interquartile range (in brackets) of the fitted theoretical distributions. In the presence of RBC the parameters of the fitted distributions of the eptifibatide-free and eptifibatide-treated fibers differ at $p<0.001$ level (for the RBC-free fibrins the eptifibatide-related difference is not significant, $p>0.05$). The figure is reproduced from Ref. [17].

5. Discussion and conclusions

This chapter addressed the problem of identifying 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 addressing the identity of empirical (stair-case) and of fitted distributions. In the case of empirical distributions, the failure to reject the null hypothesis proves that samples come from one and the same general population. In the case of fitted distributions, 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 that executes the necessary calculation and evaluation procedures.

Some parts of the three problems analyzed in this chapter may be solved using similar methods or software tools different from the MATLAB procedures described in section 3. Some software packages solve the task of choosing the best distribution type to fit the data [18, 19]. The appropriateness of the fit is defined by the goodness-of-fit metrics, which may be selected by the user. The Kolmogorov-Smirnov statistics is recommended for the case of samples with continuous variables, but strictly speaking the analytical Kolmogorov-Smirnov distribution should not be used to calculate the *p*-value in case any of the parameters has been calculated on the basis of the sample as explicitly stated in [19]. Its widespread application, however, is based on the fact that it is the most conservative, i.e. the probability to reject the null hypothesis is lower compared to the other goodness-of-fit criteria. Some available tools [20] also use analytical expressions to calculate the *p*-value of the Kolmogorov-Smirnov test in the case of a sample that is normally distributed, exponentially distributed or extreme-value distributed [21, 22]. Those formulae are applied in the *lillietest* MATLAB function from the Statistical toolbox, where Monte-Carlo simulation is conducted for the other distributions. It is recommended to use Monte-Carlo simulation even for the three aforementioned distributions in case any of the parameters has been derived from the sample. Some applications calculate a goodness-of-fit metrics of a single sample as a Kuiper statistics (e.g. in the awkwardly spelled *kupiertest* MATLAB function of [23]) and the *p*-value is calculated analytically. The main drawback of that program is that the user must guarantee that the parameters of the theoretical distribution have not been calculated from the sample. Other available applications offer single-sample Kuiper test (e.g. *v.test* function in [24]) or single- and two-sample Kuiper tests (e.g. *KuiperTest* function in [25]), which use Monte-Carlo simulation. The results of the functions *v.test* and *KuiperTest* are quite similar to those presented in this chapter, the main difference being our better approximation of the empirical distribution with a linear function, rather than with a histogram. Our approach to calculate *p*-values with Monte-Carlo simulation stems from the previously recognized fact that "…if one or more parameters have to be estimated, the standard tables for the Kuiper test are no longer valid …" [26]. Similar concepts have been proposed by others too [27].

An advantage of the method applied by us is that the Kuiper statistics is very sensitive to discrepancies at the tails of the distribution, unlike the Kolmogorov-Smirnov statistics, whereas at the same time it does not need to distribute the data into bins, as it is for the chi-square statistics. Another advantage is that the method is very suitable for circular probability distributions [23, 24], because it is invariant to the starting point where cyclic variations are observed in the sample. In addition it is easily generalized for multi-dimensional cases [25].

A limitation of our method is that it cannot be used for discrete variables [25], whereas the Kolmogorov-Smirnov test could be easily modified for the discrete case. The second drawback is that if the data are not *i.i.d*. (independent and identically distributed), then all Bootstrap and Monte-Carlo simulations give wrong results. In that case, the null hypothesis is rejected even if true, but this is an issue with all Monte-Carlo approaches. Some graphical and analytical possibilities to test the *i.i.d.* assumption are described in [19].

Further extension of the statistical procedures proposed in this chapter may focus on the inclusion of additional statistical tests evaluating the quality of the fits and the identity of the distributions. The simulation procedures in Task 3 may be modified to use Bootstrap, because this method relies on fewer assumptions about the underlying process and the associated measurement error [28]. Other theoretical distribution types could also be included in the program platform, especially those that can interpret different behaviour 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) to implement the described procedures.

Acknowledgements

This. research is funded by the Hungarian Scientific Research Fund OTKA 83023. The authors wish to thank Imre Varju from the Department of Medical Biochemistry, Semmelweis University, Budapest, Hungary for collecting the datasets with length measurements, and Laszlo Szabo from the Chemical Research Center, Hungarian Academy of Sciences, Budapest, Hungary for taking the SEM images.

Author details

Natalia D. Nikolova[1], Daniela Toneva-Zheynova[2], Krasimir Kolev[3] and Kiril Tenekedjiev[1]

*Address all correspondence to: Kolev.Krasimir@med.semmelweis-univ.hu

1 Department of Information Technologies, N. Vaptsarov Naval Academy, Varna, Bulgaria

2 Department of Environmental Management, Technical University – Varna, Varna, Bulgaria

3 Department of Medical Biochemistry, Semmelweis University, Budapest, Hungary

References

[1] Nikolova ND, Toneva D, Tenekedjieva A-M. Statistical Procedures for Finding Distribution Fits over Datasets with Applications in Biochemistry. Bioautomation 2009; 13(2) 27-44.

[2] Tenekedjiev K, Dimitrakiev D, Nikolova ND. Building Frequentist Distribution of Continuous Random Variables. Machine Mechanics 2002; 47 164-168,

[3] Gujarati DN. Basic Econometrics, Third Edition. USA: McGraw-Hill, pp. 15-318; 1995

[4] Knuth DE. The Art of Computer Programming, Vol. 2: Seminumerical Algorithms, 3rd ed. Reading, MA: Addison-Wesley, pp. 45-52; 1998.

[5] Press W, Flannery B, Teukolsky S, Vetterling W. Numerical Recipes in FORTRAN: The Art of Scientific Computing 2nd ed. England: Cambridge University Press, pp. 620-622; (1992).

[6] Burnham KP, Anderson DR. Model Selection and Inference: A Practical Information-Theoretic Approach. Springer, pp. 60-64; 2002.

[7] Schwarz G. Estimating the Dimension of a Model. Annals of Statistic 1974; 6 461-464.

[8] Politis D. Computer-intensive Methods in Statistical Analysis. IEEE Signal Processing Magazine 1998; 15(1) 39-55.

[9] Scott DW. On Optimal and Data-based Histograms, Biometrika 1979; 66 605-610.

[10] Sturges HA. The Choice of a Class Interval. J.Am.Stat.Assoc. 1926; 21 65-66.

[11] Freedman D, Diaconis P. On the Histogram as a Density Estimator: L_2 Theory, Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete 1981; 57 453–476.

[12] Kuiper NH. Tests Concerning Random Points on a Circle. Proceedings of the Koninklijke Nederlandse Akademie van Wetenshappen 1962; A(63) 38-47.

[13] The MathWorks. Statistical ToolboxTM 7.0 User's Guide. USA: the MathWorks Inc.; 2008.

[14] Finch SR. Extreme Value Constants. England: Cambridge University Press, pp. 363-367; 2003.

[15] Hanke JE, Reitsch AG. Understanding Business Statistics. USA: Irwin, pp. 165-198; 1991.

[16] Varjú I, Sótonyi P, Machovich R, Szabó L, Tenekedjiev T, Silva M, Longstaff C, Kolev K. Hindered Dissolution of Fibrin Formed under Mechanical Stress. J Thromb Haemost 2011; 9 979-986.

[17] Wohner N, Sótonyi P, Machovich R, Szabó L, Tenekedjiev K, Silva MMCG, Longstaff C, Kolev K. Lytic Resistance of Fibrin Containing Red Blood Cells. Arterioscl Thromb Vasc Biol 2011; 31 2306-2313.

[18] Palisade Corporation. Guide to Using @RISK – Risk Analysis and Simulation Add-in for Microsoft Excel, Version 4.5. USA: Palisade Corporation; (2004).

[19] Geer Mountain Software Corporation Inc. Stat::Fit - Version 2. USA: Geer Mountain Software Corporation Inc.; (2001).

[20] The MathWorks. Statistics Toolbox Software – User's Guide: Version 8.0. USA: The MathWorks; (2012).

[21] Lilliefors HW. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. J.Am.Stat.Assoc.: 1967; 62 399-402

[22] Lilliefors HW. On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. J.Am.Stat.Assoc: 1969; 64 387-389.

[23] Mossi D. Single Sample Kupier Goodness-Of-Fit Hypothesis Test. (2005) http://www.mathworks.com/matlabcentral/fileexchange/8717-kupiertest

[24] Venables WN, Smith DM, the R Core Team. An Introduction to R. USA: R Core Team; (2012).

[25] Weisstein EW. Kuiper Statistic. From MathWorld--A Wolfram Web Resource, http://mathworld.wolfram.com/KuiperStatistic.html, retrieved 2012 September

[26] Louter AS, Koerts J. On the Kuiper Test for Normality with Mean and Variance Unknown. Statistica Neerlandica 1970; 24 83–87.

[27] Paltani S. Searching for Periods in X-ray Observations using Kuiper's Test. Application to the ROSAT PSPC Archive. Astronomy and Astrophysics: 2004; 420 789-797.

[28] Efron B, Tibshirani RJ. An Introduction to the Bootstrap. USA: Chapman &Hall, pp. 45-59; 1993;