Constructing an Algorithm for Selecting the Number of Histogram Bins in Statistical Hypothesis Testing for Normal Distribution of Sample Data

Ivelina Zlateva[1], Nikola Nikolov[2], Mariela Alexandrova[3], Violin Raykov[4]

[1,2,3]Department of Automation, Technical University of Varna, Bulgaria
[4]Institute of Oceanology, Technical University of Varna, Bulgaria

Abstract— *Practice, on the whole, makes extensive use of the vast range of assumptions and conjectures in regards to the type of frequency distribution in statistical samples, the deviations from which would significantly affect the qualities of the model and the estimation accuracy of its parameters. Regrettably, a reliable and clearly defined criterion as to their permissibility is completely absent.*

For instance the fish stock assessment procedure is initially based on assumption that the frequencies in the length-frequency samples used for estimation of growth parameters of fish and analysis of the stock status are normally distributed or follow approximately the normal distribution [15,17].

The purpose of the present study is to construct an algorithm for identification of the statistical distribution of a random variable focusing on the proper selection of the number of histogram bins and further assessment of its impact on the stochastic models delivered. To that effect, appropriate simulation studies have been carried out to compensate for the lack of any concrete evidence related to the potential impact of the number of bins in the histogram and the overall data accuracy on the results of the application of the statistical criterion for the verification of the law of distribution. Applied has been the direct statistical method for determining the law of the distribution - chi-square criteria along with some indirect methods. Provided for the simulation studies were machine-generated data sets and the relevant simulations were held in MATLAB programming environment.

Keywords— histogram bins, length-frequency samples, normal distribution, stochastic modeling, stock assessment.

I. INTRODUCTION

Exploring the law of random variable distribution is the first fundamental step in a researcher's journey into the possibility for obtaining specific targeted information about the object of their study. Analyzing experimental data and displaying it graphically in a histogram offers the scientist a better insight into the intricate pattern of statistical regularity, which, in turn, will help them draw the relevant inferences about the events and processes under study. Undoubtedly, the information thus obtained is often insufficient and requires further refinement through the use of more scientifically-based methods of knowledge acquisition and attainment of improved objectivity and decision quality.

Indeed, thorough awareness of the distribution law, along with its underlying parameters, opens up the possibility for the parameters of the object under exploration to be modeled with sufficient accuracy, and to be validated as unbiased estimates of the general population (herein, class biological objects) with sufficient accuracy and thus, provides an effective means of solving various prediction problems.

In probability theory and mathematical statistics, the normal distribution, or the Gaussian distribution is continuous and gives a good approximate description of the samples, with the data values being tightly grouped round the mean, and distributed symmetrically to form a bell-shaped density curve.

It is widely applicable for mathematical descriptions of real-world phenomena and processes as well. This is ascribed to the validity of the central limit theorem that the sum of a large number of independent random variables with arbitrary laws of distribution is considered as such with a normal distribution of the variables.

The density function of the normal distribution has the form:

$$f(x; m, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right] \quad (1)$$

where, the mathematical expectation m and the standard deviation σ are the distribution parameters characterizing: the distribution center and its scale, and σ^2 the variance around the mean m:

$$m = \int_{-\infty}^{\infty} x.f(x)dx \tag{2}$$

$$\sigma^2 = \int_{-\infty}^{\infty}(x-m)^2.f(x) \tag{3}$$

Here$-\infty < x > \infty, -\infty < m > \infty$, $\sigma > 0$.

.....The unbiased and significant estimates of the mathematical expectation and variance, upon the splitting of sample into "k" bin intervals are:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i^* n_i = \sum_{i=1}^{k} x_i^*.P_i, \text{ where: } P_i = \frac{n_i}{n} \tag{4}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i.(x_i^* - \bar{x})^2 = \frac{n}{n-1}\sum_{i=1}^{k}(x_i^* - \bar{x})^2.P_i \tag{5}$$

Here x_i^* is the middle of "$i_{\text{-}th}$" interval, and n_i is the number of data occurrences (observed frequencies) within that interval.

Asymmetry (skewness) and excess play an important role in the normal distribution. Their values characterize the deviation of a particular distribution from the norm. Their estimates of the finite number of values of random variables are:

$$m_3 = \sum_{i=1}^{k}(x_i^* - \bar{x})^3/n \tag{6}$$

$$m_4 = \sum_{i=1}^{k}(x_i^* - \bar{x})^4/n - 3 \tag{7}$$

The asymmetry for a symmetric distribution is zero. Conditional upon the sign of asymmetry, the distribution can be left-skewed with (negative) asymmetry with the tail to the left of the centre of the grouped frequency distribution, or right-skewed, i. e. with positive asymmetry.

II. DETERMINING THE LAW OF RANDOM VARIABLE DISTRIBUTION ON THE BASIS OF STATISTICAL ANALYSIS OF EXPERIMENTAL DATA

The most common type of problem is when they obtained experimental data is used to determine the statistical parameters, such as point estimates of the distribution, with the law of distribution not having been completely settled.

Formulated, on this basis, is the following general algorithm with the subsequent estimates of the specific distribution parameters (all the observations have been made by assumed normal distribution of the data in the statistical inferences of experimental sampling):

Step 1. The sample is split into k bin intervals;

Step 2. The indirect criteria are determined–moments of the distribution, the mode, the median, the mathematical expectation and the number of occurrences within the intervals: $M \pm \sigma - 68{,}25\%$; $M \pm 2\sigma - 95.45\%$;$M \pm 3\sigma - 99{,}75\%$. The proper selection of the model for describing empirical data is not determined unilaterally through the above-defined indicators and does not guarantee its adequacy. Statistical methods are required to assess the adequacy of the selected model.

Step 3. Formulating the null hypothesis at level of significance $\alpha$and probability p allows for the empirical distribution to approximate the selected theoretical distribution. Applying statistical methods to verify the consistency of the empirical distribution with a given theoretical distribution–whereupon, the research practice encourages the extensive use of normality tests: the Pearson's criterion χ^2 (chi-square), Kolmogorov-Smirnov test, Fisher exact test and *etc.*

MATLAB with its highly advanced problem-solving tools and capabilities is, indisputably, the most productive programming environment. In cases where certain physical, even biological, processes require further elucidation, additional simulation methods are there to bring more accuracy and efficiency.

Addressed, in light of the discussion so far, are the following issues and tasks related to the analysis and determination of the law of distribution of statistical and experimental samples (growth-frequency, weight characteristics or experimental samples for determining the indices of abundance /biomass of renewable marine living resources):

1) The study of the above described biological objects (BO) should meet the requirement for consistency of data in the experimental and statistical samples with the normal distribution;

2) Determining the law of distribution of random variables, which may refer to the growth (length or weight) characteristics of the surveyed BO, creates two problems associated with data collection and processing: the selection of data splitting bin intervals for the construction of the histogram and the impact of the data accuracy on the derivation of theoretical distribution. The right selection of the number of bins should be considered carefully when applying this criterion since it directly affects the respective number of the degrees of freedom. The larger the number, the more reliable the criterion is in recognizing the correct distribution to the given data.

3) The insufficient information as regards the issues stated above calls for their further elucidation and detailed exploration through the use of simulation modelling.

4) Deriving universal algorithm for research studies into the law of distribution of a random variable and its successful adaptation and application to the actual experimental and statistical data obtained from technical, natural and human systems.

III. SIMULATION STUDY OF THE EFFECT OF THE ACCURACY OF THE EXPERIMENTAL DATA AND THE NUMBER OF HISTOGRAM BIN INTERVALS ON THE PATTERN OF STATISTICAL DISTRIBUTION

The density function of the empirical distribution provides a complete description of the particular characteristic features of a given random variable. The algorithm, adopted to help determine the law of the empirical distribution according to the experimental data, is as follows:

1) Collection of information about the examined random variable through appropriate observations or experiments. As a result, the statistical sequence $x_1, x_2, x_3 \dots x_n$ is registered and x_{min} and x_{max} are there upon computed;

2) The range of x_{min} - x_{max} is split into *k* intervals;

3) Identified are the observed frequencies within every interval n_i;

4) Determined are the probabilities (empirical) for every interval: $p_i = \frac{n_i}{n}$;

5) A histogram is under construction, which is a corresponding approximation to the density function of the random variable distribution;

6) The resultant histogram is compared to the theoretical distributions and the most appropriate pattern is selected for the purposes of obtaining data approximation with sufficient accuracy;

7) The distribution parameters have been established;

8) The selected law is put to the test to validate the extent to which its underlying parameters are consistent with the relevant experimental data.

The algorithm, thus outlined, makes no reference as to the selected number of intervals k to split the volume of the sample. As specified in [7,8,10,13], the number of the groups k varies from 5 to 20, being contingent upon the amount of data, and when conducting hypothesis testing as to the criterion χ^2, the theoretical frequency n_{i_t} in each group (interval) should be > 5. Serious disruption or failure to comply with this condition is likely to result in intervals being integrated or joined together.

In line with [3], what is normally selected is $k = 7 - 20$ intervals. The width (size) of the group intervals shapes or presupposes the type of the respective histogram. With smaller intervals there are few possibilities of occurrences in them and the subsequent histogram is then a poor "transmitter" of the distinctive characteristic features of the surveyed distribution. Similarly, the larger the intervals, the more distorted is the conjecture of the specific properties of examined distribution.

As stated in [4] a normal approximation which will be sufficiently accurate in practice implies that all the frequencies observed in all of the intervals are $n.p_i \geq 10$, and if not, recommended is consolidation of neighboring groups, so that the condition is fulfilled. With a large number of observations within the range of 200-300 or more, $k = 10 - 20$. Frequently, with an assumed normal distribution, $k = 12$. With a larger number of intervals, the pattern of distribution gets distorted and results in random zigzag motion following the specific changes in the frequency. With a smaller number of intervals, the characteristic features of the distribution are also modified.

In keeping with [6,8], the number of the intervals can be specified in terms of the semi-empirical formula $k \approx 1 + 3.22 * \log_{10}(n)$.

Equally applicable as well are other formulas for calculating the number of intervals:

- $k = \left\lceil \frac{x_{max} - x_{min}}{h} \right\rceil$ such as that of Venables and Ripley, where: h - is a pre-selected value for the width of the interval;
- $k = \sqrt{n}$, *where*: n is the number of observations (the size of the sample), integrated in the statistical analysis functions in Excel (univariate:histograms) and other specialized programs;
- Sturgis' formula: $k = \lceil log_2 n \rceil + 1$ – not recommended for $n < 30$, as the number of intervals will be too small to realistically reflect the actual pattern of distribution and is considered therefore inappropriate for distributions other than normal [9,16];
- The Rice rule [11]: $k = \left\lceil 2n^{\frac{1}{3}} \right\rceil$;
- Doane's formula – modification of the Sturgis' formula proposed by Doane to improve the results in the study of data that do not follow a normal distribution [5]:

  $k = 1 + log_2(n) + log_2(1 + \frac{|g_1|}{\sigma_{g_1}})$, where: $g_1$the calculated value for the third moment of distribution m_3 – or the asymmetry a:$\sigma_{g_1} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}$;
- Scott's normal reference rule for determining the optimal bin interval width: $h = \frac{3.5\hat{\sigma}}{n^{\frac{1}{3}}}$, where: $\hat{\sigma}$is the standard deviation of the sample. The Scott's rule is optimal in reference to random normally distributed samples in the sense of minimizing mean integrated square error of the distribution density estimates (empirical and theoretical) [12];
- Friedman-Diaconis' rule to determine the width of the interval: $h = 2\frac{IQR(x)}{n^{1/3}}$, where: $IQR(x)$is the interquartile range (or the distance between the third and the first quartile of the distribution) [6].

The development of computer technologies creates ample opportunities for the research studies to be undertaken in simulated environment which, in turn, allows for further complementation and concretization and, to a certain extent, customization of the recommendations in relation to the analysis and validation of the assumptions that the data follows a normal distribution in the relevant statistical (experimental) samples.

IV. IDENTIFICATION OF THE LAW OF DISTRIBUTION

The law of distribution plays a crucial role in the quality of the estimates for the process and object parameters. There is no data in the literature as to the impact of the random errors accompanying distributed data sets that help assess the quality of the applied criteria for testing normal distribution hypotheses. It is also evident from the previous paragraph that there is a lack of uniformity in the process of selection of the sample splitting bin intervals in the construction of the empirical distribution. This necessitates additional targeted research in this direction.

The present and the following paragraphs deal with the issue of the most effective method for verifying the consistency of the experimental data with a selected theoretical distribution (normal distribution) in the presence of poor data quality (added error of measurement), as well as the effect of the selected number of bin intervals.

The research will focus on the law of normal distribution, since the analysis of the growth-frequency samples and the subsequent estimation of the BO's growth parameters rely on its assumed validity.

According to literature data, there exist two types of criteria for testing the validity of the normal distribution: direct and indirect.

The indirect methods for determining the parameters of the stochastic distribution are reduced to the analysis of: mode, median, mathematical expectation (mean) – and when the data follows a normal distribution, the values for the three mathematical characteristics of the distribution are the same, deviations of asymmetry and excess from the standard type of distribution, range of distribution and the number of occurrences within the intervals:$M \pm \sigma - 68{,}25\%$; $M \pm 2\sigma - 95.45\%$;$M \pm 3\sigma - 99{,}75\%$.

Among the direct methods for determining the law of distribution (criteria), a well–established criterion with practically proven performance characteristics is the χ^2 (chi-square) criterion, which is highly efficient for samples with volume values $n \geq 100$ [3,4,7,13,14,19].

If the observation frequency that has been determined experimentally by data analysis techniques does not differ significantly from the frequency predicted by the selected theoretical law, then, it may be selected as a mathematical model describing the distribution of the random variable.

A parameter that facilitates the estimation of difference between the observed and expected (theoretical) frequencies (or the deviation of the observed distribution from the theoretical one) is the variable χ^2[3,4,7,13,14,19]. In mathematics, it is generally known as chi-square criterion for testing statistical hypotheses. By definition, $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n_{i\_t})^2}{n_{i\_t}}$, where: n_{i_t}– is the theoretical number of occurrences, in accordance with the selected theoretical law of assumed distribution. For practical implementation of the criterion, raised is a null hypothesis H_0, (at level of significance α, probability p and degrees of freedom $v = n - l - 1$, where l is the number of parameters of the law of distribution) which is to certify that the difference between the empirical and theoretical distributions, with the pertinent parameters, is insignificant. If the hypothesis testing validates that the calculated value χ^2 is less than the critical tabular value χ^2_{cr}, then, H_0 shall be accepted.

The proposed research applies the method of simulation and explores lenght-frequency samples obtained from scientific experiments. Conducted in MATLAB programming environment have been simulation studies for 2 biological objects (BO): BO_1 (length of sprat) and similarly for BO_2 (anchovy), with $k = var$ and different accuracy of the data, reflecting the inaccurate measurements of the growth parameters or the variability in the natural environment of the surveyed BOs. Presumably, these effects exert a profound influence on the accuracy of the results when determining the law of BOs parameters distribution. The parameters of the simulation models are close to those of the experimental measurements of randomly selected samples from commercial catches and are:

a) BO_1 – for length: $M = 9.3\ cm; S = 0.8\ cm$; sample size: $n = 1000$ individuals;

b) BO_2 - for length; $M = 12.3\ cm; S = 1.15\ cm$; sample size: $n = 230$ individuals.

The values for L_{sim}, simulated according to the models described, have been contaminated with normally distributed noise, characterized by zero mathematical expectation $M = 0$ and standard deviationσ, generated in the MATLAB programming environment using the function R = normrnd(mu,sigma), which generates random numbers with a normal distribution with a mathematical expectation mu and a standard deviation sigma (which may be vectors, matrices, or multi-dimensional arrays) [18].

- Determined is the standard deviation of the input data: $L(t)$: $S_L = Std(L_{sim})$;
- Calculated is the noise-to-signal ratio: $S_{relL} = \frac{S_e}{S_L} 100(\%)$.

MATLAB program has been developed with the purpose of facilitating the implementation of an algorithm for stochastic analysis and modeling the data distribution in statistical samples with the appropriate graphic representation of the results. In addition, provided is information as to prompt decision-making when the qualities of the model are being verified, by applying the chi-square criteria to validate the consistency of the model with the theoretical distribution. Proposed also is information about the indirect criteria for assessing the consistency of the empirical and theoretical distribution. Since the required arrays are of large dimensions, applied, in statistical processing, is bin interval splitting of the algorithm input data.

The program was developed under the presumed pursuit for a normal distribution and with the explicit aim of eliminating the routine calculations. Created, thus, is an opportunity to explore different numbers of bins and their effect on the stochastic model of empirical distribution.

The results of the program implementation are: the parameters of the distribution and the statistical verification of the law of distribution according to the chi-square criteria. Attained is also information about the indirect indicators of the normal distribution.

V. RESULTS OF SIMULATION STUDIES INTO THE LAW OF FREQUENCY DISTRIBUTION IN THE SIMULATED SAMPLES

5.1 Simulation results of BO_1 - length

The research was conducted under the following initial conditions:

1) $L_{BO_1} = L_{sim}$; characterized by: $M = 9.3086$ cm, $S_{L_{sim}} = 0.8261$;

2) $L_{BO_{1e1}} = L_{sim} + e_1$, the standard deviation of the error being: $S_{e_1} = 0.0457$, the noise-to-signal ratio:

$S_{relL} = \left(\frac{S_{e1}}{S_{L_{sim}}}\right) * 100 = 5.5283\%$ or an added measurement error ≈ 5,5%;

3) $L_{BO_{1e2}} = L_{sim} + e_2$, *the standard deviation of the error being* $S_{e_2} = 0.0997$, *the noise-to-signal ratio:*

$S_{relL} = \left(\frac{S_{e2}}{S_{L_{sim}}}\right) * 100 = 12.0637\%$ or an added measurement error ≈ 12%;

4) Pearson's chi-square test has been applied to a level of significance α.

The results of the simulation studies and graphic interpretation are given in *Appendix1*, tables 1-1to 1-3.

Varying with the number of splitting bin intervals $k(5-20)$ of the array data L, calculated, through the use of the program, are the values of the chi-square criteria χ^2 and the indirect methods of validating the law of distribution: the number of occurrences in the intervals M$\pm S$; M$\pm 2S$; M$\pm 3S$, the asymmetry, the excess of distribution and the number of intervals for which $n_{i\_t} < 5$.

Table 1-1 (*Appendix 1*) presents the results of the simulation study for $L_{BO_1} = L_{sim}$. Parameters of distribution are: $Me = 9.4326$; $Mo = 9.3067$; $M = 9.3086$, sufficiently close values, which may serve as a basis for the adoption of the normal distribution.

For all the values of k from 5 to 20, the chi-square criterion and the indirect criteria recognize the normal distribution as valid.

Table 1-2 (*Appendix 1*) shows the results of the simulation study for $L_{BO_{1e1}} = L_{sim} + e_1$.Parameters of distribution are: $Me = 9.4710$; $Mo = 9.3098$; $M = 9.3118$, sufficiently close values, which may serve as a basis for formally acknowledging the status of the normal distribution.

For all the values of k from 5 to 20, the chi-square criterion and the indirect criteria recognize the normal distribution as valid.

Table 1-3(*Appendix 1*) displays the results of the simulation study for $L_{BO_{1e2}} = L_{sim} + e_2$. Parameters of distribution are: $Me = 9.3623$; $Mo = 9.3036$; $M = 9.3026$, sufficiently close values, which may serve as a basis for formally acknowledging the status of the normal distribution.

For all the values of k from 5 to 20, the chi-square criterion and the indirect criteria recognize the normal distribution as valid.

5.2 Simulation results of BO_2 - length

The research was conducted under the following initial conditions:

1) $L_{BO_1} = L_{sim}$; characterized by: $M = 12.3323$см, $S_{L_{sim}} = 1.1035$;

2) $L_{BO_{1e1}} = L_{sim} + e_1$, the standard deviation of the error being: $S_{e_1} = 0.0596$, the noise-to-signal ratio:

$S_{relL} = \left(\frac{S_{e1}}{S_{L_{sim}}}\right) * 100 = 5.3988\%$ or an added measurement error ≈ 5.5%;

3) $L_{BO_{1e2}} = L_{sim} + e_2$, the standard deviation of the error being: $S_{e_2} = 0.1253$, the noise-to-signal ratio:

$S_{relL} = \left(\frac{S_{e2}}{S_{L_{sim}}}\right) * 100 = 11.3523\%$ or an added measurement erro r ≈ 11%;

4) Pearson's chi-square test has been applied to a level of significance α.

The results of the simulation studies and graphic interpretation are given in *Appendix 1*, tables 1-4 to 1-6.

Varying with the number of splitting bin intervals $k(5-20)$ of the array data L,calculated, through the use of the program, are the values of the chi-square criteria χ^2 and the indirect methods of validating the law of distribution: the number of occurrences in the intervals M$\pm S$; M$\pm 2S$; M$\pm 3S$, the asymmetry, the excess of distribution and the number of intervals for which $n_{i\_t} < 5$.

Table 1-4 (*Appendix 1*) presents the results of the simulation study for $L_{BO_1} = L_{sim}$. Parameters of distribution are: $Me = 12.6105$; $Mo = 12.3313$; $M = 12.3323$, sufficiently close values, which may serve as a basis for the adoption (assumption) of the normal distribution.

For all the values of k from 5 to 20, the chi-square criterion and the indirect criteria recognize the normal distribution as valid.

Table 1-5 (*Appendix 1*) displays the results of the simulation study for $L_{BO_{1e1}} = L_{sim} + e_1$.Parameters of distribution are: $Me = 12.5559$; $Mo = 12.3658$; $M = 12.3655$, sufficiently close values, which may serve as a basis for formally acknowledging the status of the normal distribution.

For all the values of$k$from 5 to 20, the chi-square criterion and the indirect criteria recognize the normal distribution as valid.

Table 1-6 (*Appendix 1*) displays the results of the simulation study for $L_{BO_{1e2}} = L_{sim} + e_2$. Parameters of distribution are: $Me = 12.5165$; $Mo = 12.3371$; $M = 12.3397$, sufficiently close values, which may serve as a basis for formally acknowledging the status of the normal distribution.

For all the values of k from 5 to 20, the chi-square criterion and the indirect criteria recognize the normal distribution as valid.

VI. RESEARCH INTO THE LAWS OF BO PARAMETER DISTRIBUTION OF REAL DATA

An experimental approach was adopted for collection of statistical data (total body length measurements of sprat and anchovy) to support the stochastic modeling process and distribution analysis. The samples are taken from commercial catches (stationary pound nets – with mesh size 7.5 mm). The fish was caught on 1st of May 2017, near Varna, Bulgaria - "Trakata" area. The catch composition was presented by two species– Sprat (Sprattus Sprattus) as a targeted catch and anchovy (Engraulis Encrasicolus) as a by-catch. The samples processed for further analysis are: n=1000 individuals of sprat and $n = 230$ individuals of anchovy. The body length measurements of the samples have been recorded and processed to form the input massive for calculations done by the specified in paragraph 4 script developed in MATLAB programming environment. The null hypothesis is formed under the above-described conditions, stating that sample data follows the normal distribution. Respectively an alternative hypothesis is that the sample data do not follow the normal distribution.

6.1 Stochastic model of length frequencies in the sample of BO_1

The results of the research study are registered in table 1-7 of *Appendix 1*.

The minimum and maximum values for this particular BO are: $x_{min} = 6.3$cm, $x_{max} = 13$cm; $n = 1000$; $M = 9.3086$ cm; $S = 0.8261$ cm.

Application of Pearson's chi-square test produces positive results when $k = 6, 8, 11, 13$, i.e. the distribution of the length frequencies in the sample of BO_1 does not contradict *Ho* for normal distribution.

The indirect criteria as well as the close values: $Me = 9.6500$cm; $Mo = 9.3217$; $M = 9.3217$; also point to this conclusion.

The model has the following form:

$$f(x; m,\sigma) = \frac{1}{0.8261\sqrt{2\pi}} \exp[-\frac{(L_i - 9.3086)^2}{2 * 0.6824}]$$

Fig. 1 introduces the empirical and theoretical probabilities in intervals, as well as the predicted values of the approximating model when $k = 6,8,11,13$.

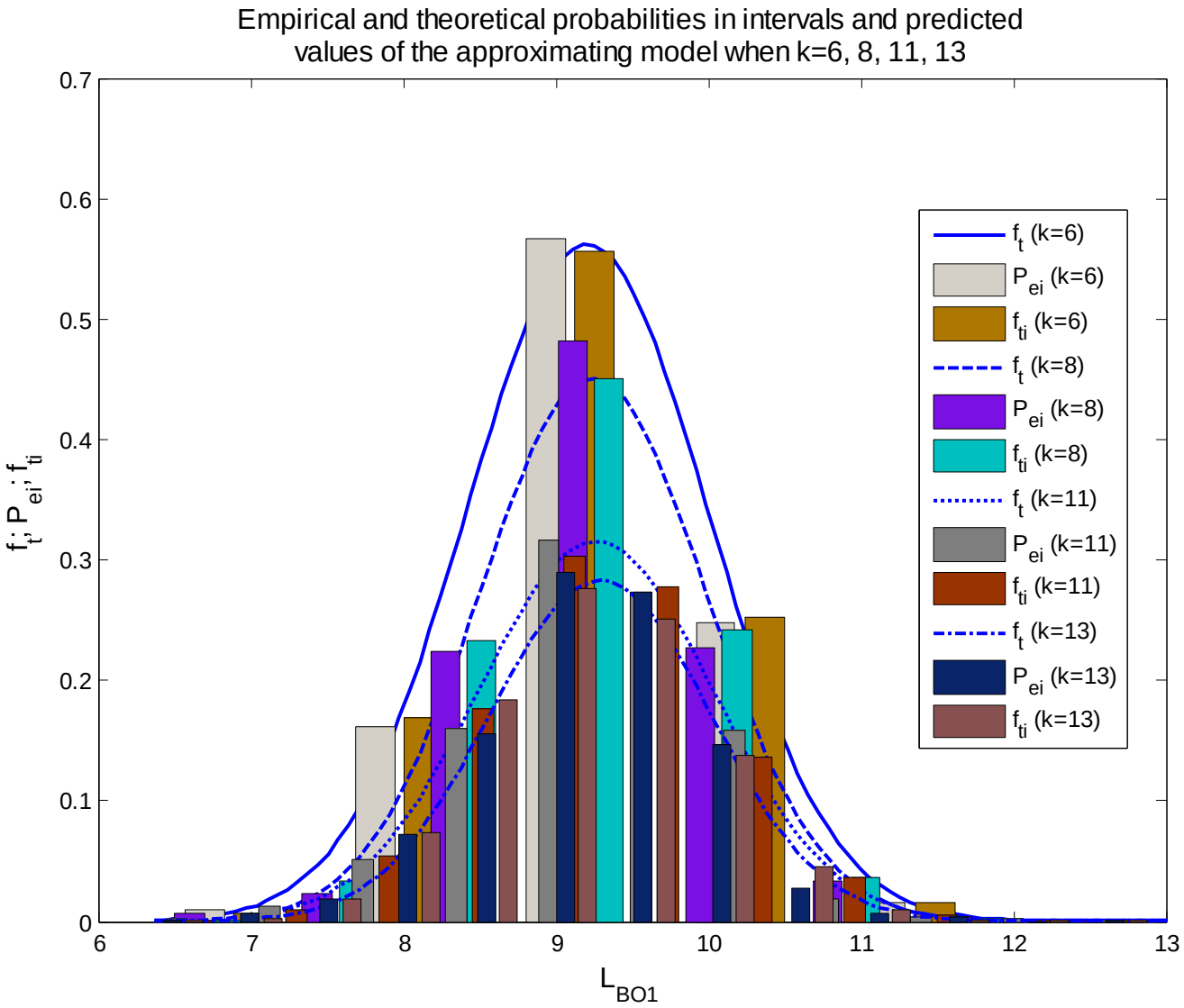

FIGURE 1. Empirical and theoretical probabilities in intervals and predicted values of the approximating model when k =6,8,11,13 (length BO1)

6.2 Stochastic model of linear dimensions (length) of BO_2

The results of the research study are registered in table 2-15 of *Appendix 2*.

The minimum and maximum values for the sample are: $x_{min} = 9.00$ cm, $x_{max} = 14.50$ cm.$n = 230$, $M = 12.0226$ cm; $S = 1.0197$cm.

Application of Pearson's chi-square test produces positive results when $k = 6,8,9,10,12,16$, i.e. the distribution of the linear dimensions of BO_2 does not contradict the raised *Ho* for normal distribution of sample data.

The indirect criteria as well as the close values: $Me = 11.75$cm; $Mo = 12.0250$; $M = 12.0226$; also point to this conclusion.

The model has the following form:

$$f(x; m,\sigma) = \frac{1}{1.0197\sqrt{2\pi}} \exp[-\frac{(L_i - 12.0226)^2}{2 * 1.0398}]$$

Fig. 2 introduces the empirical and theoretical probabilities by intervals, as well as the predicted values of the approximating model when $k = 6,8,9,10,12,16$.

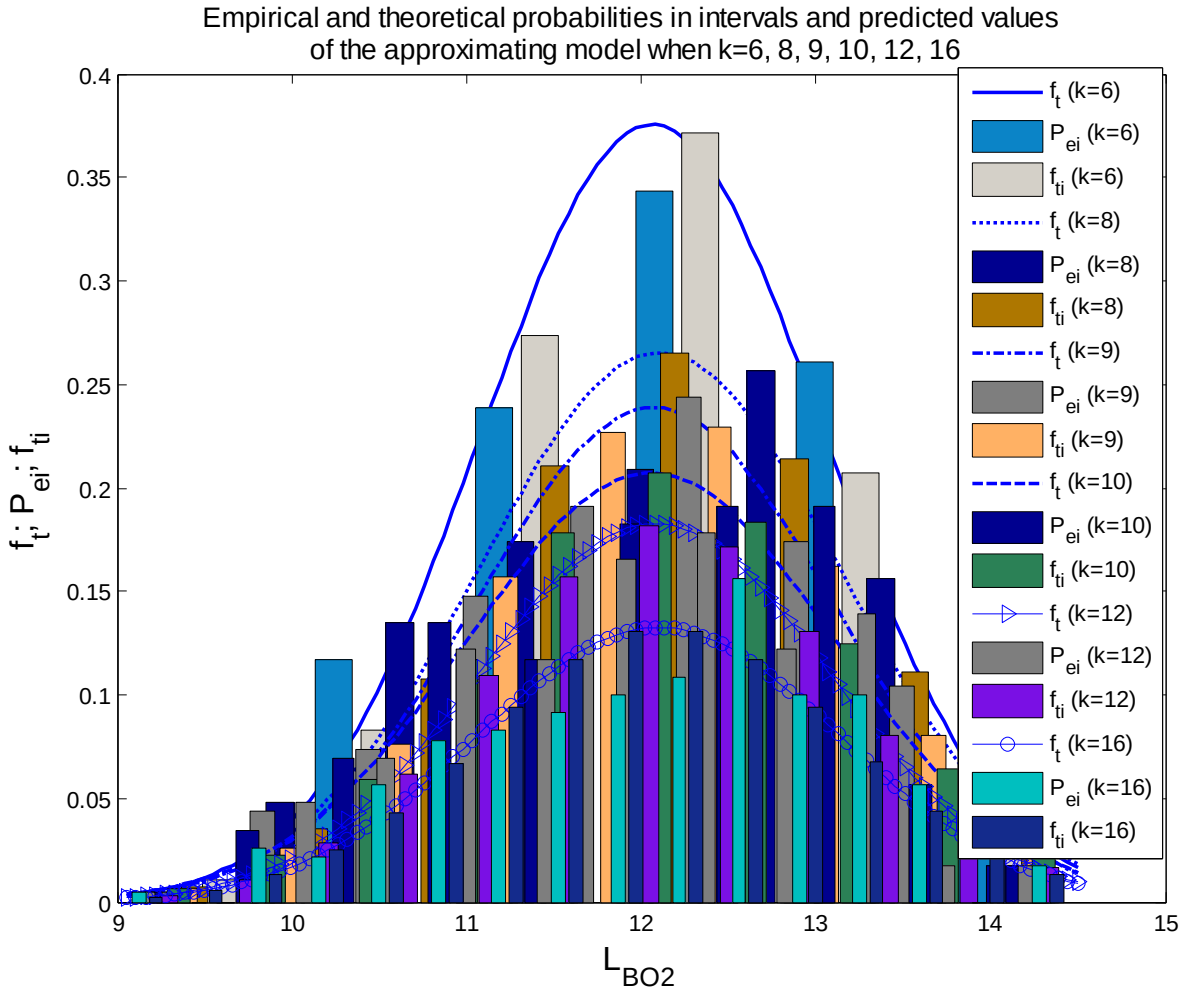

FIGURE 2. Empirical and theoretical probabilities by intervals and predicted values of the approximating model when k =6,8,9,10,12,16 (Length BO_2)

VII. CONCLUSION

Through the adoption of an experimental and statistical approach, a passive experiment was carried out to collect relevant information about the growth parameters of BO in the Bulgarian Black Sea coast in the area of "Trakata" in the vicinity of the town of Varna. The aim is to determine the law of statistical distribution of the length of the two types of BO. The lack of specific information on the impact of the accuracy of the data used on the results of the application of statistical criterion for validating the law of distribution has necessitated the completion of additional simulation studies. Accordingly, conducted have been further studies to clarify the number of the splitting sample bin intervals with the formation of empirical distribution, which directly affects the selection of the theoretical law of distribution. A direct approach is used to determine the law of distribution through chi- square criteria in combination with indirect methods. Employed in the simulations were computer-generated data with a normal distribution in MATLAB environment function randn.

The following primary conclusions have been reached:

1) In the study of the distribution law, combining the direct method (chi-square), the recommendations$n_i . p_i < 5$and indirect methods improves the quality of the end solution. The considerable computational work while fusing them together does not pose a problem with the present-day state-of-the-art computer technology.

2) The number of bin intervals k, to which the data necessary for the construction of the histogram is split has a profound effect upon the results obtained in the process of determining the law of the random variable distribution. The selection of only one specific value of k is found to be quite insufficient to bring about a reasonable conclusion. The use of computer equipment with appropriate software provides the opportunity for multiple values to be included in the study towards a more informed decision.

3) The use of k from 5 to 13 is considered sufficient enough to reveal the stochastic regularity. With significant data noise-contamination the smaller values of k produce reliable results, although the degrees of freedom are on decrease. With substantial data noise-contamination, the smaller values of k (5,6), the chi- square is able to detect the normal distribution in spite of the curtailed degrees of freedom.

4) With both uncontaminated and contaminated data, the increase of k, is likely to result in intervals of $n.pi < 5$. This indicator increases with increased number of contamination intervals. The χ^2 criterion recognizes the normal distribution easily, when there are intervals with $n_i.p_i < 5$, and in both cases, following their integration.

5) The proposed recommended values for the number of sample splitting intervals is $k = 5 - 13$, with n > 200. Modern computer technology makes it possible for the distribution of data to be explored with multiple intervals, rather than only one selected value for k, subsequent to the process of decision-making. The availability of information about the level of data contamination is of utmost convenience.

6) The distribution of BO_1 and BO_2 lengths is subject to the law of normal distribution. The accuracy of the experimental data, of 0.1 cm with which they have been obtained is seen as sufficient.

7) The obtained models of the laws of probability distribution with the underlying parameters are viewed as adequate can be used for solving research and practical tasks as well.

REFERENCES

[1] Genov, D., 2000. Modeling and Optimization of industrial processes Manual Lab, Varna: Technical university-Varna, 192p, ISBN-954-20-0263-7.

[2] Krug, G.K. 1973. Planning an experimental Research. Theoretical basis. MEI, M.

[3] Novikov, N.M. 2000. Methodological guidelines for Mathematical statistics for students, (Part II), Voronezh.

[4] Smirnov, N.V. Dunin-Barkovskii, I.V. 1959. Short course of Mathematical Statistics for technical applications, Fizmatgiz.

[5] Doane, D. P. 1976. Aesthetic frequency classification. American Statistician, vol. 30: pp. 181–183.

[6] Freedman, D., Diaconis, P. 1981. "On the histogram as a density estimator: L2theory". Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete. 57 (4): pp. 453–476. doi:10.1007/BF01025868.

[7] Hahn, G. J. and Shapiro, S.S., 1967. Statistical Modeling in engineering, New York, London, Sydney: John Wiley and Sons Inc., 376p.

[8] He, K., Meeden, G. 1997, Selecting the number of bins in a Histogram: A decision Theoretic approach, Journal of statistical Planning and Inference, Vol. 61, Issue 1, pp. 49-59.

[9] Hyndman, R. (1995). The problem with Sturges' rule for constructing histograms. Retrieved December 13, 2017 from: https://robjhyndman.com/papers/sturges.pdf

[10] Knuth, K. 2013. Optimal Data Binning for Histograms, arXiv;physics/0605197v2 [physics.data-an]

[11] Lane, D. M. Project Leader, Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/).:, Rice University (chapter 2 "Graphing Distributions", section "Histograms").

[12] Scott, David W. 1979. "On optimal and data-based histograms". Biometrika. 66 (3): pp.605–610. doi:10.1093/biomet/66.3.605.

[13] Shannon, R.E. 1975. SYSTEMS SIMULATION: The Art and Science, Prentice-Hall, Inc., New Jersey.

[14] Snedecor G.W and Cochran W.G, 1989, Statistical Methods, 8th Edition, Iowa State University Press, 491p.

[15] Sparre P., Venema S.C. 1998. Introduction to tropical fish stock assessment. Part 1. Manual. FAO Fish Tech. Pap., 306/1 (Rev.2), 407p.

[16] Sturges, H. A. (1926). "The choice of a class interval". Journal of the American Statistical Association: 65–66. doi:10.1080/01621459.1926.10502161. JSTOR 2965501.

[17] Food and Agriculture organization of the United Nations - http://www.fao.org/home/en/.

[18] MathWorks: Matlab-online: https://www.mathworks.com/products/matlab-online.html.

[19] NIST/SEMATECH e-Handbook of Statistical Methods web-site: http://www.itl.nist.gov/div898/handbook/.

APPENDIX 1

SIMULATION: LENGTH – BO_1 UNCONTAMINATED DATA (L_{sim})

Table 1-1

k	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	Direct method for determination of the law of the sample frequencies distribution, Chi-square normality test															
χ^2	*0.9347*	*0.8788*	*2.2176*	*2.1416*	*2.8800*	*4.0621*	*5.0216*	*1.3480*	*4.3055*	*6.8472*	*4.3793*	*12.3688*	*5.6879*	*9.3223*	*12.9344*	*12.8948*
χ^2_T	*1.39*	*2.37*	*3.36*	*4.35*	*5.35*	*6.35*	*6.35*	*2.18*	*8.34*	*9.34*	*4.57*	*13.70*	*11.34*	*11.34*	*15.98*	*13.34*
α	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.975*	*0.50*	*0.50*	*0.95*	*0.25*	*0.50*	*0.50*	*0.25*	*0.50*
v	*2*	*3*	*4*	*5*	*6*	*7*	*7*	*8*	*9*	*10*	*11*	*11*	*12*	*12*	*13*	*14*
	Indirect methods															
$M \pm S$ %	71.30	70.40	69.50	69.50	69.40	69.30	69.40	69.20	68.90	69.10	68.70	68.80	69.10	69	68.90	69.10
$M \pm 2.S$%	96.50	96.20	96	96	95.80	95.80	95.80	95.80	95.60	95.70	95.50	95.60	95.70	95.60	95.60	95.60
$M \pm 3.S$ %	100	100	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90
b_1	0.0023	$6.6*10^{-4}$	0.0054	0.0029	0.0043	0.0024	0.0091	0.0036	0.0030	0.0023	0.0036	0.0035	0.0028	0.0019	0.0024	0.0024
b_2	2.7163	2.7384	2.8097	2.7427	2.8505	2.8407	2.8622	2.8170	2.8082	2.7687	2.7867	2.8383	2.7901	2.8268	2.8517	2.8241
Ase	0.0480	0.0259	0.0737	0.0541	0.0659	0.0491	0.0954	0.0599	0.0544	0.0480	0.0604	0.0594	0.0531	0.0439	0.0492	0.0491
Exe	-0.2837	-0.2616	-0.1903	-0.2573	-0.1495	-0.1593	-0.1378	-0.1830	-0.1918	-0.2313	-0.2133	-0.1617	-0.2099	-0.1732	-0.1483	-0.1759
Intervals with $n.pi < 5$ *to the left*	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
Intervals with $n.pi < 5$ *to the right*	0	0	0	0	0	0	1	1	1	1	1	1	1	2	2	2
S	0.8261	0.8261	0.8261	0.8261	0.8261	0.8261	0.8261	0.8261	0.8261	0.8261	0.8261	0.8261	0.8261	0.8261	0.8261	0.8261
Mo	9.3067	9.3067	9.3067	9.3067	9.3067	9.3067	9.3067	9.3067	9.3067	9.3067	9.3067	9.3067	9.3067	9.3067	9.3067	9.3067
Me	9.4326	9.4326	9.4326	9.4326	9.4326	9.4326	9.4326	9.4326	9.4326	9.4326	9.4326	9.4326	9.4326	9.4326	9.4326	9.4326
M	9.3086	9.3086	9.3086	9.3086	9.3086	9.3086	9.3086	9.3086	9.3086	9.3086	9.3086	9.3086	9.3086	9.3086	9.3086	9.3086
H_0 *(accepted) Yes/No*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Simulation model parameters: simulated sample size $n = 1000$; $M = 9.30\ cm$; $S = 0.8\ cm$; $x_{min} = 7.0114\ cm$; $x_{max} = 11.8538\ cm$

SIMULATION: LENGTH – BO_1 CONTAMINATED DATA ($L_1 = L_{sim} + e_1$)

Table 1-2

k	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	Direct method for determination of the law of the sample frequencies distribution, Chi-square normality test															
χ^2	*1.2588*	*1.7520*	*2.6008*	*2.7670*	*2.6152*	*5.0463*	*4.3152*	*2.6806*	*6.3923*	*6.1589*	*7.2743*	*8.9571*	*8.8161*	*6.7539*	*10.6773*	*9.4611*
χ^2_T	*1.39*	*2.37*	*3.36*	*4.35*	*5.35*	*5.35*	*6.35*	*2.73*	*8.34*	*9.34*	*10.34*	*11.34*	*11.34*	*11.34*	*12.34*	*13.34*
α	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.95*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*
v	*2*	*3*	*4*	*5*	*6*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*12*	*12*	*13*	*14*
	Indirect methods															
$M \pm S$ %	71.20	70.20	70	69.80	69.80	69.90	69.60	69.70	69.50	69.70	69.40	69.50	69.70	69.40	69.40	69.70
$M \pm 2.S$%	96.50	96.30	95.80	95.70	95.70	95.80	95.60	95.70	95.30	95.60	95.30	95.40	95.60	95.40	95.30	95.40
$M \pm 3.S$ %	100	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90
b_1	0.0050	$6.1*10^{-4}$	0.0091	0.0077	0.0054	0.0073	0.0057	0.0011	0.0060	0.0045	0.0043	0.0043	0.0027	0.0027	0.0058	0.0070
b_2	2.7484	2.8088	2.8543	2.7835	2.8341	2.8815	2.8689	2.8321	2.8114	2.8139	2.8532	2.8522	2.8029	2.8483	2.8277	2.8746
Ase	0.0707	0.0248	0.0955	0.0877	0.0737	0.0852	0.0757	0.0338	0.0775	0.0671	0.0658	0.0653	0.0521	0.0523	0.0760	0.0838
Exe	-0.2516	-0.1912	-0.1457	-0.2165	-0.1659	-0.1185	-0.1311	-0.1679	-0.1886	-0.1861	-0.1468	-0.1478	-0.1971	-0.1517	-0.1723	-0.1254
Intervals with $n.pi < 5$ *to the left*	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
Intervals with $n.pi < 5$ *to the right*	0	0	0	0	0	1	1	1	1	1	1	1	2	2	2	2
S	0.8283	0.8283	0.8283	0.8283	0.8283	0.8283	0.8283	0.8283	0.8283	0.8283	0.8283	0.8283	0.8283	0.8283	0.8283	0.8283
Mo	9.3098	9.3098	9.3098	9.3098	9.3098	9.3098	9.3098	9.3098	9.3098	9.3098	9.3098	9.3098	9.3098	9.3098	9.3098	9.3098
Me	9.4710	9.4710	9.4710	9.4710	9.4710	9.4710	9.4710	9.4710	9.4710	9.4710	9.4710	9.4710	9.4710	9.4710	9.4710	9.4710
M	9.3118	9.3118	9.3118	9.3118	9.3118	9.3118	9.3118	9.3118	9.3118	9.3118	9.3118	9.3118	9.3118	9.3118	9.3118	9.3118
H_0 *(accepted) Yes/No*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Simulation model – contaminated data (measurement error or environmental variability impact on length of BO) with parameters: $S_L = 0.8261$; $L_1 = L_{sim} + e_1$; $S_{e1} = 0.0457$; signal-to-noise ratio: $S_{RelL} = \left(\frac{S_{e1}}{S_L}\right) * 100 = 5.5283\%$ (or added measurement error $\approx 5.5\%$) - $\boldsymbol{x_{min} = 7.0297\ cm;}$ $\boldsymbol{x_{max} = 11.9122\ cm}$

SIMULATION: LENGTH – BO_1 CONTAMINATED DATA ($L_2 = L_{sim} + e_2$)

Table 1-3

k	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	Direct method for determination of the law of the sample frequencies distribution, Chi-square normality test															
χ^2	*1.1681*	*0.8987*	*3.1955*	*1.3652*	*5.4989*	*7.4669*	*5.8386*	*6.4355*	*5.5585*	*9.8238*	*4.5142*	*10.9679*	*8.4898*	*15.4927*	*8.3760*	*18.0608*
χ^2_τ	*1.39*	*2.37*	*3.36*	*4.35*	*7.84*	*9.04*	*7.34*	*7.34*	*8.34*	*12.55*	*9.34*	*13.70*	*11.34*	*15.98*	*13.34*	*21.06*
α	*0.50*	*0.50*	*0.50*	*0.50*	*0.25*	*0.25*	*0.50*	*0.50*	*0.50*	*0.25*	*0.50*	*0.25*	*0.50*	*0.25*	*0.50*	*0.10*
v	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*8*	*9*	*10*	*10*	*11*	*12*	*13*	*14*	*14*
	Indirect methods															
$M \pm S$ %	71.20	69.90	68.90	70.10	68.70	69.20	68.90	68.70	68.70	68.80	68.70	69	68.70	68.60	68.90	68.70
$M \pm 2.S$%	96.60	95.80	95.80	96	95.40	95.70	95.70	95.50	95.30	95.50	95.30	95.60	95.50	95.30	95.60	95.40
$M \pm 3.S$ %	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
b_1	0.0029	$1.4*10^{-6}$	0.0013	$1.5*10^{-4}$	0.0036	0.0015	0.0022	$4.4*.10^{-4}$	$5.8*10^{-4}$	$8.02*10^{-4}$	0.0027	0.0018	0.0025	0.0032	0.0024	0.0022
b_2	2.7358	2.8116	2.8114	2.7131	2.8612	2.8725	2.8369	2.8527	2.8256	2.8453	2.8624	2.8263	2.8647	2.8925	2.8808	2.8661
Ase	0.0536	0.0012	0.0354	-0.0124	0.0601	0.0385	0.0469	0.0212	0.0242	0.0090	0.0516	0.0419	0.0496	0.0569	0.0492	0.0465
Exe	-0.2642	-0.1884	-0.1886	-0.2869	-0.1388	-0.1275	-0.1631	-0.1473	-0.1744	-0.1547	-0.1376	-0.1737	-0.1353	-0.1075	-0.1192	-0.1339
Intervals with $n.pi < 5$ *to the left*	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
Intervals with $n.pi < 5$ *to the right*	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	2
S	0.8351	0.8230	0.8230	0.8230	0.8230	0.8230	0.8230	0.8230	0.8230	0.8230	0.8230	0.8230	0.8230	0.8230	0.8230	0.8230
Mo	9.3036	9.3374	9.3374	9.3374	9.3374	9.3374	9.3374	9.3374	9.3374	9.3374	9.3374	9.3374	9.3374	9.3374	9.3374	9.3374
Me	9.3623	9.3893	9.3893	9.3893	9.3893	9.3893	9.3893	9.3893	9.3893	9.3893	9.3893	9.3893	9.3893	9.3893	9.3893	9.3893
M	9.3026	9.3177	9.3177	9.3177	9.3177	9.3177	9.3177	9.3177	9.3177	9.3177	9.3177	9.3177	9.3177	9.3177	9.3177	9.3177
H_0 *(accepted) Yes/No*	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Simulation model – contaminated data (measurement error or environmental variability impact on length of BO) with parameters: $S_L = 0.8261$; $L_2 = L_{sim} + e_2$; $S_{e2} = 0.0997$; signal-to-noise ratio: $S_{RelL} = \left(\frac{S_{e2}}{S_L}\right) * 100 = 12.0637\%$ (or added measurement error $\approx 12\%$ - $\boldsymbol{x_{min} = 6.9179\ cm;}$ $\boldsymbol{x_{max} = 11.8066\ cm}$

SIMULATION: LENGTH – BO_2 UNCONTAMINATED DATA (L_{sim})

Table 1-4

k	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	Direct method for determination of the law of the sample frequencies distribution, Chi-square normality test															
χ^2	*6.8265*	*4.5569*	*9.0120*	*2.1416*	*8.4972*	*9.7735*	*6.1660*	*13.3708*	*11.7931*	*11.0899*	*20.3029*	*8.9793*	*14.3961*	*14.8640*	*11.1304*	*15.5750*
χ^2_T	*7.38*	*5.99*	*9.49*	*3.36*	*9.24*	*11.07*	*6.63*	*14.45*	*12.02*	*13.36*	*26.12*	*10.22*	*14.68*	*16.92*	*12.55*	*17.27*
α	*0.025*	*0.05*	*0.05*	*0.50*	*0.10*	*0.05*	*0.25*	*0.025*	*0.10*	*0.10*	*0.001*	*0.25*	*0.10*	*0.05*	*0.25*	*0.10*
v	*2*	*2*	*4*	*4*	*5*	*5*	*5*	*6*	*7*	*8*	*8*	*8*	*9*	*9*	*10*	*11*
	Indirect methods															
$M \pm S$ %	66.09	66.52	66.08	66.08	66.08	65.22	65.22	65.22	65.22	66.08	65.22	65.22	65.22	66.08	65.22	65.22
$M \pm 2.S$%	96.09	96.08	95.65	95.65	96.08	95.65	95.65	95.65	95.65	95.65	95.65	95.65	95.65	95.65	95.65	95.65
$M \pm 3.S$ %	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
b_1	0.0167	0.0023	0.0054	$3.2*10^{-7}$	0.0021	0.0020	0.0012	$2.85*10^{-4}$	0.0021	$9.82*10^{-4}$	$6.68*10^{-4}$	$6.5*10^{-4}$	0.0016	0.0010	0.0048	0.0013
b_2	2.3500	2.4599	2.8097	2.6900	2.5083	2.5967	2.6608	2.5810	2.7190	2.6960	2.6725	2.7694	2.6916	2.6824	2.7233	2.6927
Ase	-0.1292	0.0477	0.0737	$-5.6*10^{-4}$	-0.0461	-0.0448	-0.0352	-0.0169	-0.0460	-0.0313	-0.0258	-0.0254	-0.0406	-0.0320	-0.0693	-0.0358
Exe	-0.6500	-0.5401	-0.1903	-0.3100	-0.4917	-0.4033	-0.3392	-0.4190	-0.2810	-0.3040	-0.3275	-0.2306	-0.3084	-0.3176	-0.2767	-0.3073
Intervals with $n.pi < 5$ *to the left*	0	0	0	0	0	1	1	1	1	1	1	2	2	2	2	2
Intervals with $n.pi < 5$ *to the right*	0	1	0	1	1	1	1	2	2	2	3	3	3	4	4	4
S	1.1035	1.1035	1.1035	1.1035	1.1035	1.1035	1.1035	1.1035	1.1035	1.1035	1.1035	1.1035	1.1035	1.1035	1.1035	1.1035
Mo	12.3313	12.3313	12.3313	12.3313	12.3313	12.3313	12.3313	12.3313	12.3313	12.3313	12.3313	12.3313	12.3313	12.3313	12.3313	12.3313
Me	12.6105	12.6105	12.6105	12.6105	12.6105	12.6105	12.6105	12.6105	12.6105	12.6105	12.6105	12.6105	12.6105	12.6105	12.6105	12.6105
M	12.3323	12.3323	12.3323	12.3323	12.3323	12.3323	12.3323	12.3323	12.3323	12.3323	12.3323	12.3323	12.3323	12.3323	12.3323	12.3323
H_0 *(accepted) Yes/No*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Simulation model parameters: simulated sample size $n = 230$; $M = 12.23$ *cm;* $S = 1.058$ *cm;* $x_{min} = 9.7018$ *cm;* $x_{max} = 15.5193$ *cm*

SIMULATION: LENGTH – BO_2 CONTAMINATED DATA ($L_1 = L_{sim} + e_1$)

Table 1-5

k	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	Direct method for determination of the law of the sample frequencies distribution, Chi-square normality test															
χ^2	*6.5619*	*5.1582*	*2.2685*	*8.1942*	*9.1862*	*6.6112*	*8.6436*	*15.1364*	*8.7381*	*6.0770*	*22.8924*	*9.7998*	*13.9518*	*16.0871*	*20.4804*	*18.5003*
χ^2_T	*7.38*	*6.25*	*4.11*	*9.49*	*9.24*	*6.63*	*9.24*	*16.81*	*9.04*	*7.34*	*26.12*	*10.22*	*14.68*	*18.31*	*23.21*	*19.68*
α	*0.025*	*0.10*	*0.25*	*0.05*	*0.10*	*0.25*	*0.10*	*0.01*	*0.25*	*0.50*	*0.001*	*0.25*	*0.10*	*0.05*	*0.01*	*0.05*
v	*2*	*3*	*3*	*4*	*5*	*5*	*5*	*6*	*7*	*8*	*8*	*8*	*9*	*10*	*10*	*11*
	Indirect methods															
$M \pm S$ %	**66.52**	**66.09**	**65.65**	**65.65**	**66.09**	**66.09**	**64.78**	**65.22**	**65.22**	**65.65**	**65.65**	**64.78**	**65.65**	**65.65**	**65.22**	**65.65**
$M \pm 2.S$%	96.09	96.09	95.65	95.65	96.09	96.09	95.65	95.65	95.65	95.65	95.65	95.65	96.09	95.65	95.65	95.65
$M \pm 3.S$ %	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
b_1	0.0216	0.0030	$5.4*10^{-4}$	0.0054	$5*10^{-5}$	0.0030	$8.9*10^{-5}$	0.0040	0.0054	0.0024	0.0048	0.0034	$5.6*10^{-4}$	0.0038	$3.4*10^{-4}$	0.0033
b_2	2.3852	2.4113	2.5903	2.6456	2.6514	2.6216	2.5900	2.5604	2.6974	2.6169	2.6359	2.7285	2.6737	2.7353	2.7286	2.7168
Ase	-0.1470	-0.0549	-0.0232	-0.0736	0.0071	-0.0545	-0.0094	-0.0632	-0.0737	-0.0492	-0.0694	-0.0585	-0.0236	-0.0619	-0.0184	-0.0573
Exe	-0.6148	-0.5887	-0.4097	-0.3544	-0.3486	-0.3784	-0.4100	-0.4396	-0.3026	-0.3831	-0.3641	-0.2715	-0.3263	-0.2647	-0.2714	-0.2832
Intervals with $n.pi < 5$ *to the left*	0	0	0	0	0	1	1	1	1	1	1	2	2	2	2	2
Intervals with $n.pi < 5$ *to the right*	0	0	1	1	1	1	2	2	2	2	3	3	3	3	4	4
S	1.1086	1.1086	1.1086	1.1086	1.1086	1.1086	1.1086	1.1086	1.1086	1.1086	1.1086	1.1086	1.1086	1.1086	1.1086	1.1086
Mo	12.3658	12.3658	12.3658	12.3658	12.3658	12.3658	12.3658	12.3658	12.3658	12.3658	12.3658	12.3658	12.3658	12.3658	12.3658	12.3658
Me	12.5559	12.5559	12.5559	12.5559	12.5559	12.5559	12.5559	12.5559	12.5559	12.5559	12.5559	12.5559	12.5559	12.5559	12.5559	12.5559
M	12.3655	12.3655	12.3655	12.3655	12.3655	12.3655	12.3655	12.3655	12.3655	12.3655	12.3655	12.3655	12.3655	12.3655	12.3655	12.3655
H_0 *(accepted) Yes/No*	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Simulation model – contaminated data (measurement error or environmental variability impact on length of BO) with parameters: $S_L = 0.8261$; $L_2 = L_{sim} + e_2$; $S_{e2} = 0.0997$; signal-to-noise ratio: $S_{RelL} = \left(\frac{S_{e1}}{S_L}\right) * 100 =$ 5.3988% (or added measurement error ≈ 5.5% - $\boldsymbol{x_{min} = 9.6752}$ cm; $\boldsymbol{x_{max} = 15.4366}$ cm

SIMULATION: LENGTH – BO_2 CONTAMINATED DATA ($L_2 = L_{sim} + e_2$)

Table 1-6

k	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
	Direct method for determination of the law of the sample frequencies distribution, Chi-square normality test															
χ^2	*3.0682*	*5.1004*	*3.0702*	*8.0021*	*7.4738*	*3.9229*	*9.`7750*	*13.6714*	*12.5797*	*10.5736*	*15.0008*	*15.8932*	*22.5019*	*16.5125*	*19.3644*	*22.4348*
χ^2_τ	*4.61*	*6.25*	*4.11*	*9.49*	*7.78*	*4.35*	*11.07*	*14.45*	*14.07*	*13.36*	*16.01*	*17.53*	*27.88*	*18.31*	*20.48*	*23.21*
α	*0.10*	*0.10*	*0.25*	*0.05*	*0.10*	*0.50*	*0.05*	*0.025*	*0.05*	*0.10*	*0.025*	*0.025*	*0.001*	*0.05*	*0.025*	*0.01*
v	*2*	*3*	*3*	*4*	*4*	*5*	*5*	*6*	*7*	*8*	*7*	*8*	*9*	*10*	*10*	*10*
	Indirect methods															
$M \pm S$ %	**66.09**	**67.82**	**66.52**	**66.09**	**66.09**	**65.22**	**66.09**	**66.09**	**66.09**	**66.09**	**64.78**	**66.09**	**66.09**	**66.09**	**66.09**	**66.09**
$M \pm 2.S$%	96.09	96.09	96.09	96.09	96.09	96.09	96.09	96.09	96.09	96.09	96.09	96.09	96.09	96.09	96.09	96.09
$M \pm 3.S$ %	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
b_1	0.0170	0.0210	0.0034	0.0056	$2.4*10^{-5}$	0.0063	$7*10^{-4}$	0.0060	0.0066	0.0065	0.0040	0.0041	0.0054	0.0083	0.0049	0.0091
b_2	2.5735	2.4786	2.6969	2.5375	2.5699	2.7263	2.6798	2.6310	2.6878	2.6671	2.7460	2.6965	2.6728	2.7124	2.7031	2.7579
Ase	-0.1303	-0.1448	-0.0584	-0.0746	-0.0049	-0.0797	-0.0265	-0.0776	-0.0814	-0.0804	-0.0630	-0.0642	-0.0732	-0.0910	-0.0700	-0.0953
Exe	-0.4265	-0.5214	-0.3031	-0.4625	-0.4301	-0.2737	-0.3202	-0.3690	-0.3122	-0.3329	-0.2540	-0.3035	-0.3272	-0.2876	-0.2969	-0.2421
Intervals with $n.pi < 5$ *to the left*	0	0	0	0	1	1	1	1	1	1	2	2	2	2	2	3
Intervals with $n.pi < 5$ *to the right*	0	0	1	1	1	1	2	2	2	2	3	3	3	3	4	4
S	1.1221	1.1221	1.1221	1.1221	1.1221	1.1221	1.1221	1.1221	1.1221	1.1221	1.1221	1.1221	1.1221	1.1221	1.1221	1.1221
Mo	12.3371	12.3371	12.3371	12.3371	12.3371	12.3371	12.3371	12.3371	12.3371	12.3371	12.3371	12.3371	12.3371	12.3371	12.3371	12.3371
Me	12.5165	12.5165	12.5165	12.5165	12.5165	12.5165	12.5165	12.5165	12.5165	12.5165	12.5165	12.5165	12.5165	12.5165	12.5165	12.5165
M	12.3397	12.3397	12.3397	12.3397	12.3397	12.3397	12.3397	12.3397	12.3397	12.3397	12.3397	12.3397	12.3397	12.3397	12.3397	12.3397
H_0 *(accepted) Yes/No*	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Simulation model – contaminated data (measurement error or environmental variability impact on length of BO) with parameters: $S_L = 1.058$; $L_2 = L_{sim} + e_2$; $S_{e2} = 0.1253$; signal-to-noise ratio:: $S_{RelL} = \left(\frac{S_{e2}}{S_L}\right) * 100 = 11.3523\%$ (or added measurement error $\approx 11\%$ $\boldsymbol{x_{min} = 9.5272}$ cm; $\boldsymbol{x_{max} = 15.5057}$ cm

STOCHASTIC MODEL LENGTH – BO_1 (REAL DATA)

Table 1-7

k	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	Direct method for determination of the law of the sample frequencies distribution, Chi-square normality test															
χ^2	*28.0442*	*1.1463*	*19.2485*	*4.2035*	*18.7612*	*58.4285*	*16.5549*	*23.3963*	*17.1820*	*26.9349*	*69.9811*	*27.1276*	*30.3511*	*45.6662*	*50.8067*	*142.85*
χ^2_T	-	*1.39*	-	*4.61*	-	-	*20.52*	-	*20.52*	-	-	-	-	-	-	-
α	-	*0.50*	-	*0.10*	-	-	*0.001*	-	*0.001*	-	-	-	-	-	-	-
v	*1*	*2*	*1*	*2*	*3*	*4*	*5*	*4*	*5*	*5*	*6*	*7*	*8*	*7*	*8*	*9*
	Indirect methods															
$M \pm S$ %	73.60	70.20	70.20	64.60	70.20	64.60	70.20	64.60	73.60	64.60	73.60	64.60	64.60	64.60	64.60	64.60
$M \pm 2.S$%	96.60	96.70	96.70	96.30	96	95.70	96	95.40	95.70	96	95.70	95.40	95.40	96	96.30	95.70
$M \pm 3.S$ %	99.20	99.50	99.40	99.10	99.40	98.90	99.40	99.10	98.90	99.10	99.10	98.90	99.10	99.10	99.10	99.10
b_1	0.1560	$3.6*10^{-4}$	0.0371	$3.1*10^{-5}$	$6.1*10^{-6}$	0.0164	0.0046	$5*10^{-5}$	0.0184	0.0093	0.0304	0.0050	0.0080	4.1869	0.0075	0.0019
b_2	4.0433	3.3042	3.6255	3.9483	4.0084	4.2176	3.5701	3.8289	3.8704	4.1181	3.8935	4.3849	3.7461	$3.8*10^{-4}$	3.9928	3.9440
Ase	-0.3950	0.0191	-0.1925	0.0056	0.0025	-0.1282	-0.0678	0.0070	-0.1356	-0.0965	-0.1743	0.0707	-0.0893	-0.0196	-0.0866	-0.0430
Exe	1.0433	0.3042	0.6255	0.9483	1.0084	1.2176	0.5701	0.8289	0.8704	1.1181	0.8935	1.3849	0.7461	1.1869	0.9928	0.9440
Intervals with $n.pi < 5$ *to the left*	0	0	1	1	1	1	1	2	2	2	2	2	2	3	3	3
Intervals with $n.pi < 5$ *to the right*	1	1	2	2	2	2	2	3	3	4	4	4	4	5	5	5
S	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158	0.7158
Mo	9.3150	9.3150	9.3150	9.3150	9.3150	9.3150	9.3150	9.3150	9.3150	9.3150	9.3150	9.3150	9.3150	9.3150	9.3150	9.3150
Me	9.6500	9.6500	9.6500	9.6500	9.6500	9.6500	9.6500	9.6500	9.6500	9.6500	9.6500	9.6500	9.6500	9.6500	9.6500	9.6500
M	9.3217	9.3217	9.3217	9.3217	9.3217	9.3217	9.3217	9.3217	9.3217	9.3217	9.3217	9.3217	9.3217	9.3217	9.3217	9.3217
H_0 *(accepted) Yes/No*	No	Yes	No	Yes	No	No	Yes	No	Yes	No	No	No	No	No	No	No

Distribution parameters: sample size $n = 1000$; $x_{min} = 6.3000\ cm$; $x_{max} = 13\ cm$

STOCHASTIC MODEL LENGTH – BO_2 (REAL DATA)

Table 1-8

k	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	Direct method for determination of the law of the sample frequencies distribution, Chi-square normality test															
χ^2	*11.2119*	*7.3917*	*21.0941*	*14.8384*	*6.2802*	*16.4486*	*21.1645*	*15.6186*	*29.2748*	*24.7673*	*29.3221*	*15.7895*	*37.7625*	*44.2246*	*41.2959*	*32.3429*
χ^2_T	-	*9.21*	-	*18.47*	*6.63*	*16.81*	-	*16.81*	-	-	-	*16.92*	-	-	-	-
α	-	*0.01*	-	*0.001*	*0.25*	*0.01*	-	*0.01*	-	-	-	*0.05*	-	-	-	-
v	*2*	*2*	*3*	*4*	*5*	*6*	*5*	*6*	*7*	*7*	*8*	*9*	*8*	*9*	*10*	*10*
	Indirect methods															
$M \pm S$ %	68.70	61.30	63.91	63.91	63.91	68.70	66.09	61.30	63.91	63.91	63.91	63.91	63.91	63.91	63.91	63.91
$M \pm 2.S$%	98.70	63.91	95.21	95.22	95.21	97.83	97.83	95.21	95.22	95.22	95.22	95.22	95.22	95.22	95.22	95.22
$M \pm 3.S$ %	100	99.57	99.57	99.57	99.57	100	100	99.57	99.57	99.57	100	100	99.57	99.57	99.57	99.57
b_1	0.0861	0.0353	0.0903	0.0798	0.0660	0.0874	0.0612	0.0284	0.0940	0.0731	0.0562	0.0751	0.0656	0.0663	0.0924	0.0767
b_2	2.4004	2.3413	2.2084	2.2579	2.5236	2.3481	2.3758	2.4063	2.3606	2.4817	2.4404	2.4540	2.4123	2.4208	2.3713	2.4183
Ase	-0.2935	-0.1879	-0.3005	-0.2825	-0.2569	-0.2957	-0.2475	-0.1685	-0.3066	-0.2704	-0.2372	-0.2740	-0.2561	-0.2576	-0.3040	-0.2769
Exe	-0.5996	-0.6587	-0.7916	-0.7421	-0.4764	-0.6519	-0.6242	-0.5937	-0.6394	-0.5183	-0.5596	-0.5460	-0.5877	-0.5792	-0.6287	-0.5817
Intervals with $n.pi < 5$ *to the left*	0	1	1	1	1	1	2	2	2	3	3	3	4	4	4	5
Intervals with $n.pi < 5$ *to the right*	0	0	0	0	0	0	1	1	1	1	1	1	2	2	2	2
S	1.0197	1.0197	1.0197	1.0197	1.0197	1.0197	1.0197	1.0197	1.0197	1.0197	1.0197	1.0197	1.0197	1.0197	1.0197	1.0197
Mo	12.0250	12.0250	12.0250	12.0250	12.0250	12.0250	12.0250	12.0250	12.0250	12.0250	12.0250	12.0250	12.0250	12.0250	12.0250	12.0250
Me	11.7500	11.7500	11.7500	11.7500	11.7500	11.7500	11.7500	11.7500	11.7500	11.7500	11.7500	11.7500	11.7500	11.7500	11.7500	11.7500
M	12.0226	12.0226	12.0226	12.0226	12.0226	12.0226	12.0226	12.0226	12.0226	12.0226	12.0226	12.0226	12.0226	12.0226	12.0226	12.0226
H_0 *(accepted) Yes/No*	No	Yes	No	Yes	Yes	Yes	No	Yes	No	No	No	Yes	No	No	No	No

Distribution parameters: sample size $n = 230$; $x_{min} = 9.00\ cm$; $x_{max} = 14.50\ cm$