MATLAB-Based Stochastic Modeling Distribution Analysis of Commercial Fishery Length-Frequency Samples Taken From the Black Sea (Bulgaria)

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

[1,2,3]Department of Automation, Technical University of Varna, Bulgaria
[4]Institute of Oceanology - Varna, Bulgaria
Corresponding Authors Email: m_alexandrova@tu-varna.bg

Abstract**—** *Fish stock assessment procedure is initially based on the assumption that the frequencies in length/weight-frequency samples used for analysis of the stock status follow approximately the normal distribution. Many of the statistical procedures are based on specific distributional assumptions. The assumption of normality is very common in most classical statistical tests. In case that analysis of data implies techniques that make normality or some other distributional assumptions it is essential that this assumption is confirmed. If distributional assumption is proved, more powerful parametric techniques can be applied and if it is not justified an application of non-parametric or robust techniques may be required.*

The present article aims to present MATLAB-based algorithm for commercial fisheries length-frequency samples distribution analysis of samples of Sprat and Anchovy caught in the Bulgarian waters in the Black sea. The statistical analysis uses engineering approaches in statistical data processing and the method used for analysis of sample frequencies distribution is chi-square normality or goodness of fit test. For the provision of this analysis, specific program is developed in MATLAB programming environment to support and confirm the assumption that length-frequency samples follow the normal distribution.

Keywords— MATLAB, fisheries, normal distribution, length-frequency samples, stock assessment, chi-square normality test.

I. INTRODUCTION

Fish stock assessment procedure is initially based on the assumption that the frequencies in length-frequency samples used for analysis follow approximately the normal distribution [9]. The normal (Gaussian) distribution is a very common continuous probability distribution and it is important in statistics because it is often used in natural and social sciences to represent real-valued random variables whose distributions are not known [1,4,5]. The normal distribution is also useful due to its relation to the central limit theorem. Under certain conditions, in its most general form, the normal distribution determines that averages (mean values or expectation) of samples of observations of random variables, independently drawn from a population with unknown distribution, converge in distribution approximate to normal. The sample becomes normally distributed when the number of observations is sufficiently large. Any real values expected to indicate the sum of many independent processes (such as measurement errors) often have distributions that are relatively normal. Moreover, many results and methods (such as propagation of uncertainty and least squares parameter fitting) can be calculated analytically in explicit form when the relevant variables are normally distributed [5,7]. There are in use many non-parametric and robust statistical techniques, which are not based on strong distributional assumptions, however those based on specific distributional assumptions are considered in general most powerful and respectively preferred [6].

The present article deals with distribution analysis of commercial fisheries samples (length-frequency samples) of species taken from commercial catch caught in the Bulgarian part of the Black sea – i.e. sprat as a targeted catch and anchovy as by-catch and is aiming to confirm and justify the assumption for normality of length-frequencies samples distribution. The samples are analyzed by using engineering approaches in stochastic modeling and [2,3,6] the calculations are held in MATLAB programming environment by using specific program developed for the provision of this analysis.

II. METHODS

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 7.5mm). 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 a specified 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.

The probability density function of the normal distribution is

$$f(x) = \frac{1}{s\sqrt{2\pi}} exp\left[-\frac{(x-M)^2}{2s^2}\right] \quad (1)$$

where: the expectation *M* (also median and mode) and the standard deviation sare distribution parameters, which characterize the center of the distribution and its scale and s^2 is the variance of *M:*

$$M = \int_{-\infty}^{\infty} xf(x)dx \quad (2)$$

$$s^2 = \int_{-\infty}^{\infty} (x-M)^2 f(x) \quad (3)$$

here: $-\infty < x > \infty$, $-\infty < M > \infty$, s> 0.

Significant and unbiased estimates of the expectation and variance when the sample is broken to *k*-intervals (where: $k \approx 1 + 3.22\log_{10}(n)$ and n is the number of observations or observed frequencies) are:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i^* n_i = \sum_{i=1}^{k} x_i^* P_i,\ P_i = \frac{n_i}{n} \quad (4)$$

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

where: x_i^* is the mid-point of the, *i-th*" interval, and n_i are the observed frequencies in a given interval.

Important parameters in distribution analysis are the skewness (a parameter, characterizing the asymmetry) of the distribution and the excess. They both are used to identify the deviation of a certain distribution from the normal distribution. Their estimates for a finite number of values of a random variable are:

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

$$m_4 = \sum_{i=1}^{k} (x_i^* - \bar{x})^4 P_i \quad (7)$$

For symmetric distribution the skewness (asymmetry) is zero. Depending on the sign of the asymmetry it could be negative (the distribution is left-skewed and it has a long tail in the negative direction of the number line) and positive (right-skewed). In both cases the mean is also shifted, following the asymmetry sign.

The asymmetry relation to variance is an important indicator, which allows comparative analysis of two distributions, having a different scale. The estimate of this indicator is obtained with:

$$\sqrt{b_1} = \frac{m_3}{m_2^{3/2}} \quad (8)$$

The excess kurtos is m_4 is a measure of whether the data is heavy-tailed or light-tailed in relation to the normal distribution. That is, data set with high kurtosis tend to have heavy tails, or outliers. Data set with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be an extreme case.

Its estimate is obtained with:

$$b_2 = \frac{m_4}{m_2^2} \quad (9)$$

For normal distribution the real precise value of proportion (9) is 3.

Normality tests are used to determine if a data set is well-modeled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed. More precisely, the tests are a form of model selection, and can be interpreted in several ways, depending on the aim of analysis and interpretations of probability and certain distribution parameters.

The experimental data is tested for normality using the chi-square normality test:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n_{i\_t})^2}{n_{i\_t}} \quad (10)$$

where: n_{i_t} are the theoretical frequencies, n_i are the observed frequencies [3].

The chi-square normality test (or chi-square goodness of fit test) is used to test if a sample of data came from a population with a specific distribution [6,8]. The chi-square test is defined for the hypothesis H_0 – the data (the frequency distribution of certain eventsor variables) observed in a sample is consistent with a specific theoretical distribution and H_a – the data do not follow the specified distribution. For the provision of this analysis the events considered as hypothesis must be independent and have a total probability which equals 1. The null hypothesis is retained if the calculated value for χ^2 is less than a certain critical value $\chi^2_T(v; p)$ under significance level α, (where *v are the degrees of freedom and p is the probability).*

2.1 MATLAB based calculation procedure

The MATLAB script is developed in accordance with the following calculation procedure

Step	Matlab function	Description
1	*L=load('L.mat')*	Form the input data massive
2	*n = length(L)*	Define the size of input data massive loaded
3	x_1*=min(L)*	Calculate the min value
4	x_2*= max(L)*	Calculate the max value
5	*K*= 1 +*3,22*log*$_{10}$*(n)*	Calculate the interval numbers K. Round K to the nearest integer valuek
6	*delta=(*x_2*-*x_1*)/k*	Define the intervals length.
7	*hist(L, k);grid*	Build the histogram
8	*ni=hist(L,k)*	Calculate intervals frequencies, (vector row)
9	*Pe=*n_i*/n*	Calculate the frequencies by intervals, (vector row)
10	*x(1)=*x_1*+delta/2;* *for i=2:k* *x(i)=x(i-1)+delta;end*	Calculate interval midpoints, (vector row)
11	*M=x*Pe'*	Calculate the expectation value M
12	*s2=((x-M).^2)*Pe';* *S2=S2*n/(n-1)*	Calculate the variance
13	*s=sqrt(S2)*	Calculate the standard deviation
14	*a=1/(sqrt(2**p_i*)*S);* *b=((x-M).^2)/(2*S2);* *f=a*delta*exp(-b)*	Calculate the theoretical distribution f by intervals
15	*ff=sum(f)*	Calculate the sum f
16	*ni_t=f*n*	Calculate the theoretical frequencies by intervals
17	n_{ic} *and* n_{i_tc}	Combine intervals (bins) with frequency count less than 5
18	*cap=((*n_{ic}*-*n_{i_tc}*).^2)./*n_{i_t}	Calculated the weighed residuals
19	χ^2 *=sum(cap)*	Calculate χ^2 value
20	*Pef=[Pe'f']; plot(x,[Pef]), grid*	Build the histograms of the empirical and theoretical distributions
21	$f(x) = \frac{1}{s\sqrt{2\pi}} exp\left[-\frac{(x-M)^2}{2s^2}\right]$	Build the Normal distribution with the stochastic models delivered

The chi-square test is found sensitive to the selection of bins number (intervals) and for the approximation to be valid the expected (calculated, theoretical) frequency should be at least 5. It is not well applied to small samples and if some of the counts in sample frequencies are less than 5 it is recommended that they are combined.

2.2 MATLAB program script

```
clc
clear
load('spratLW.mat')
n=length(L);
```

```
x1=min(L)
x2=max(L)
K=1+(3.2*log10(n))
k=round(K)
delta=(x2-x1)/k
disp('Give precise value for deltal deltal=:')
deltal = input('deltal=')
hist(L,k);grid
Ni=hist(L,k)
ni=Ni;
Pe=ni/n
Pes=sum(Pe)
x(1)=x1+(deltal/2)
for i=2:k
    x(i)=x(i-1)+deltal;
end
x(i)
M=x*Pe'
S2=((x-M).^2)*Pe'
s2=S2*n/(n-1)
S=sqrt(S2)
a=1/(sqrt(2*pi)*S)
b=((x-M).^2)/(2*S2)
ft=a*(deltal*exp(-b))
ff=sum(ft)
ni_t=ft*n
disp('Combine intervals with values of ni_t less than 5')
disp('number of intervals with frequency < 5 r=:')
r = input('r=')
for i=1:k-r
ni_tc(i)=input('ni_tc=:')
end
disp('Combine intervals with values of ni less than 5')
ni=Ni
for i=1:k-r
nic(i)=input('nic=:')
end
cap=((nic-ni_tc).^2)./ni_tc;
chi2=sum(cap)
disp('Compare chi2 value with critical value of Chi2 distribution table:')
Chi2cr=input('enter Chi2cr Table value Chi2cr=:')
if chi2<Chi2cr
disp('sample frequencies follow the normal distribution')
else
disp('sample frequencies do not follow the normal distribution')
end
```

III. RESULTS

3.1 Chi-square test results length-frequency sample distribution analysis of sprat

The interim and final test results are presented in Table 1.

TABLE 1
NORMALITY TEST RESULTS FOR LENGTH-FREQUENCY SAMPLE OF SPRAT

Interval numbers k=11, observations interval $[x_{min}:x_{max}]$=[6.30;11.00] (cm), dl=0.61							
*Expectation **M=9.2676**; Variance S^2=0.5945; Standard deviation S=0.7711;*							
Intervals (cm)	***Intervals midpoint x***	***Observed frequencies n_i***	***Theoretical Frequencies n_{it}***	***Corrected Observed Frequencies n_{ic}***	***Corrected Theoretical Frequencies n_{i_tc}***	***Empirical Probability $Pe=n_i/N$***	***Theoretical Probability f***
6.30-6.91	6.6050	***2***	***0.8123***			0.0020	0.0008
6.92-7.53	7.2150	***13***	***9.1255***	15	9.9378	0.0130	0.0091
7.54-8.15	7.8250	52	54.8276	52	54.8276	0.0520	0.0548
8.16-8.77	8.4350	160	176.1687	160	176.1687	0.1600	0.1762
8.78-9.39	9.0450	316	302.7235	316	302.7235	0.3160	0.3027
9.40-10.01	9.6550	273	278.1964	273	278.1964	0.2730	0.2782
10.02-10.63	10.2650	159	136.7240	159	136.7240	0.1590	0.1367
10.64-11.25	10.8750	19	35.9357	19	35.9357	0.0190	0.0359
11.26-11.87	11.4850	***3***	***5.0512***			0.0030	0.0051
11.88-12.49	12.0950	***2***	***0.3797***	6	5.4462	0.0020	0.0004
12.50-13.11	12.7050	***1***	***0.0153***			0.0010	0.0000
χ^2=16.55 <χ^2_T (5; 0.005)=16.75➔ the null hypothesis is retained and the sample data follows the normal distribution							

Expected (theoretical) frequencies in count less than 5 are combined to ensure validity of chi-square test results. They are marked in different color, presented in bold and denoted in Table1.

The histogram of the empirical probabilities P_e, theoretical probabilities f_t by bins (intervals) and the shape of the theoretical probability distribution are presented in Fig. 1.

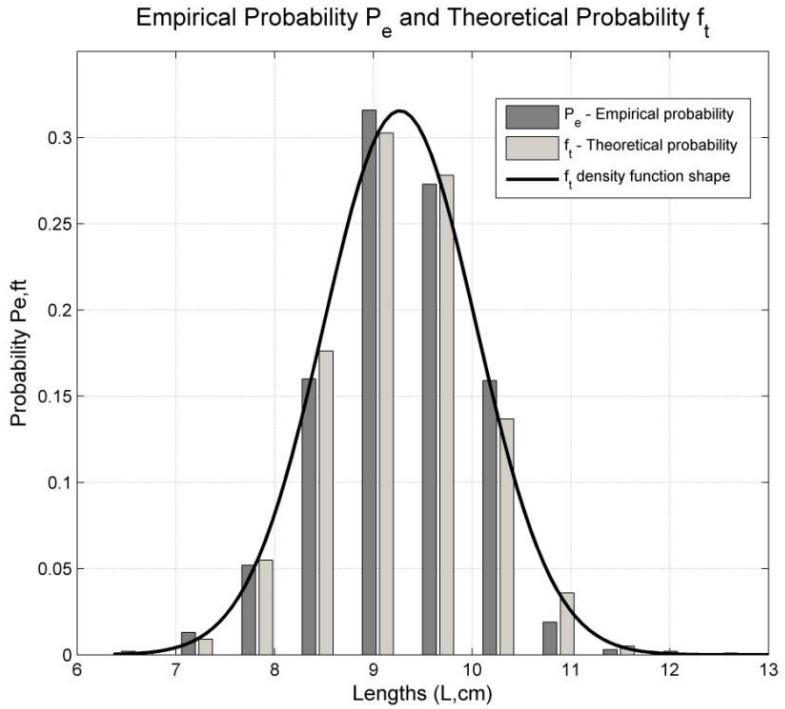

FIGURE 1. HISTOGRAM OF THE EMPIRICAL PROBABILITY P_e, THEORETICAL PROBABILITY f_t AND THE SHAPE OF THE THEORETICAL PROBABILITY DISTRIBUTION OF SPRAT LENGTH-FREQUENCY SAMPLE

3.2 Chi-square test result length-frequency sample distribution analysis of anchovy

The interim and final test results are presented in Table 2.

TABLE 2
NORMALITY TEST RESULTS FOR LENGTH-FREQUENCY SAMPLE OF ANCHOVY

Interval numbers k=9, observations interval [x_{min}:x_{max}]=[9.00;14.50] (cm), dl=0.61							
*Expectation **M=12.0686**; Variance S^2=1.0343; Standard deviation S=1.0170;*							
Intervals (cm)	***Intervals midpoint x***	***Observed frequencies n_i***	***Theoretical Frequencies n_{it}***	***Corrected Observed Frequencies n_{ic}***	***Corrected Theoretical Frequencies n_{i_tc}***	***Empirical Probability Pe=n_i/N***	***Theoretical Probability f***
9.00-9.61	9.3050	***1***	***1.3714***			0.0043	0.0060
9.62-10.23	9.9150	***10***	***5.8468***	11	7.2182	0.0435	0.0254
10.24-10.85	10.5250	17	17.3947	17	17.3947	0.0739	0.0756
10.86-11.47	11.1350	34	36.1134	34	36.1134	0.1478	0.1570
11.48-12.09	11.7450	44	52.3206	44	52.3206	0.1913	0.2275
12.10-12.71	12.3550	56	52.8966	56	52.8966	0.2435	0.2300
12.72-13.33	12.9650	40	37.3194	40	37.3194	0.1739	0.1623
13.34-13.95	13.5750	24	18.3735	24	18.3735	0.1043	0.0799
13.96-14.57	14.1850	4	6.3125	4	6.3125	0.0174	0.0274
χ^2=6.63<χ^2_T (5; 0.1)=9.236➔ the null hypothesis is retained and the sample data follows the normal distribution							

Expected (theoretical) frequencies in count less than 5 are combined to ensure validity of chi-square test results. They are marked in different color, presented in bold and denoted in Table 2.

The histogram of the empirical probabilities P_e, theoretical probabilities f_t by bins (intervals) and the shape of the theoretical probability distribution are presented in Fig. 2.

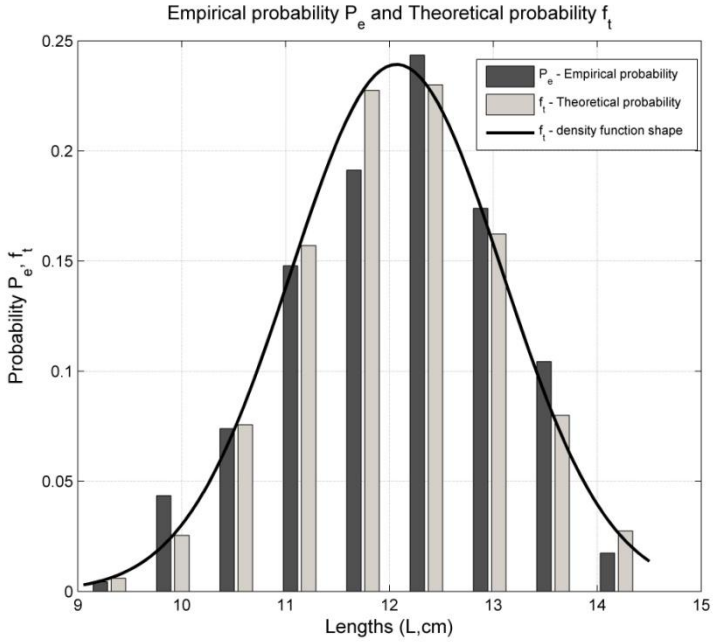

FIGURE 2. HISTOGRAM OF THE EMPIRICAL PROBABILITY P_e, THEORETICAL PROBABILITY f_t AND THE SHAPE OF THE THEORETICAL PROBABILITY DISTRIBUTION OF ANCHOVY LENGTH-FREQUENCY SAMPLE

IV. CONCLUSIONS

The experimental data distribution (length-frequency commercial fisheries samples of sprat n=1000 and anchovy (by-catch) n=230 stocks) is tested for normality with chi-square goodness of fit test and results delivered confirm the null hypothesis, stating that data is following the normal distribution.

The procedure developed implies engineering approaches in analysis of statistical data and can be successfully used to confirm assumptions for normally distributed sample data, which one justified, might be successfully used further to facilitate unbiased parametric estimates.

The program developed to support the analysis might be successfully used for testing any sample data and respectively confirm or reject the assumptions for normal distribution of sample data.

ACKNOWLEDGEMENTS

This paper is developed in the frames of the project НП6 "Research and Synthesis of Algorithms and Systems for Adaptive Observation, Filtration and Control", ДН997-НП/09.05.2017.

REFERENCES

[1] Cover, T. M. and Thomas, J.A., 2006. Elements of Information Theory 2nd Ed. Hoboken, New Jersey: John Wiley& Sons, Inc., 792p.

[2] Genov, D., 2000. Modeling and Optimization of industrial processes Manual Lab, Varna: Technical university-Varna, 192p.

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

[4] Lukacs, E., 1942. A Characterization of the Normal Distribution. The Annals of Mathematical Statistics. 13 (1): 91–93.

[5] Lyon, A., 2014. Why are Normal Distributions Normal?, The British Journal for the Philosophy of Science, 65 (3): 621-649.

[6] NIST/SEMATECH e-Handbook of Statistical Methods, 2003, last updated: 10/30/2013: http://www.itl.nist.gov/div898/handbook/03.11.2017.

[7] Papoulis, A., 2002. Probability, Random Variables and Stochastic Processes (4th Edition). New York: McGraw-Hill Higher Education p.168.

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

[9] Sparre P. and Venema S.C. 1998. Introduction to Tropical Fish Stock Assessment - Part 1: Manual. Rome: FAO Fish Tech. Pap., 306/1 (Rev.2), 407p.