SHANNON'S ENTROPY USAGE AS STATISTIC IN ASSESSMENT OF DISTRIBUTION

Lorentz Jäntschi

Sorana D. Bolboacă

MINISTRY OF EDUCATION AND RESEARECH
TECHNICAL UNIVERSITY OF CLUJ-NAPOCA

UNIVERSITAS MEDICINAE ET PHARMACIAE NAPOCENSIS

Technical University of Cluj-Napoca, Romania
Iuliu Haţieganu University of Medicine and Pharmacy, Romania

Introduction

- General null hypothesis H_0: *data follow a specific distribution*
  - Kolmogorov-Smirnov (KS)
  - Anderson-Darling (AD)
  - Cramér-von-Mises (CM)
  - Kuiper V (KV)
  - Watson U^2 (WU)
  - Shannon's H1 – introduced here as statistic

Computing of statistics

$$AD = -n - \frac{1}{n}\sum_{i=0}^{n-1}(2\cdot i+1)\cdot \ln(f_i\cdot(1-f_i))$$

$$KS = \sqrt{n}\cdot \max_{0\le i<n}(f_i - \frac{i-1}{n}, \frac{i}{n} - f_i)$$

$$KV = \sqrt{n}\cdot(\max_{0\le i<n}(f_i - \frac{i-1}{n}) + \max_{0\le i<n}(\frac{i}{n} - f_i))$$

$$CM = \frac{1}{12n} + \sum_{i=0}^{n-1}(\frac{2i+1}{2n} - f_i)^2$$

$$WU = \frac{1}{12n} + \sum_{i=0}^{n-1}(\frac{2i+1}{2n} - f_i)^2 - n(\frac{1}{2} - \frac{1}{n}\sum_{i=0}^{n-1} f_i)^2$$

$$H1 = -\sum_{i=0}^{n-1} f_i\cdot\ln(f_i) + (1-f_i)\cdot\ln(1-f_i)$$

where

- n: sample size
- f_i: cumulative distribution function (of the distribution being tested) associated with the i^{th} (from 0 to n-1) observation sorted in ascending order

Monte-Carlo building of statistic-probability map

For $0 \le k \le 1000 \cdot K$

$f_i \leftarrow \mathrm{Random}_{\mathrm{Uniform}[0,1]}$, for $0 \le i < n$

$(f_i)_{0 \le i < n} \leftarrow \mathrm{Sort}_{\mathrm{ASC}}((f_i)_{0 \le i < n})$

$\mathrm{Observed}_k \leftarrow \mathrm{Formula}((f_i)_{0 \le i < n})$

The formula of each statistic enters here

EndFor

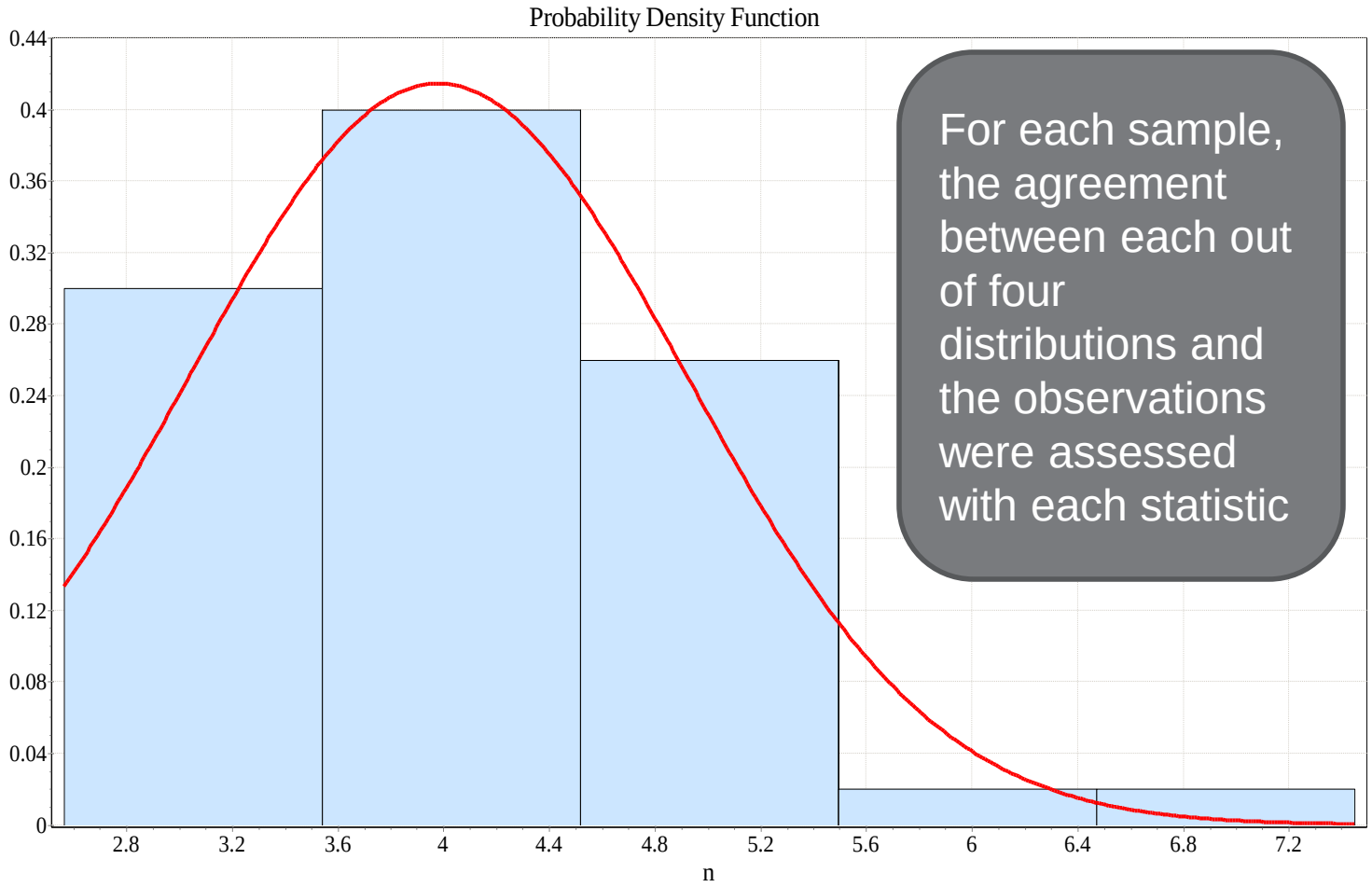
$(\mathrm{Observed}_k)_{0 \le k < K} \leftarrow \mathrm{Sort}_{\mathrm{ASC}}((\mathrm{Observed}_k)_{0 \le k < K})$

For $1 \le j \le 999$

$\mathrm{Statistic}_{j/1000} \leftarrow \mathrm{Mean}(\mathrm{Observed}_{1000 \cdot K \cdot j - 1}, \mathrm{Observed}_{1000 \cdot K \cdot j})$

EndFor

Material: 50 samples of properties and activities measurements of chemicals



For each sample, the agreement between each out of four distributions and the observations were assessed with each statistic

Experimental values taken from literature

Sample of samples sizes are log-normal distributed ranging from n = 13 to n = 1714

Common distributions were included into analysis. Two of them are two-parametrical (LN and N) and two are three-parametrical

Log-Normal	$$LN(x;\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{\left\lvert\frac{\ln x-\mu}{\sigma}\right\rvert^{2}}{2}\right)$$
Normal	$$N(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\left\lvert\frac{x-\mu}{\sigma}\right\rvert^{2}}{2}\right)$$
Gauss-Laplace	$$GL(x;\mu,\sigma,q)=\frac{p}{2\sigma}\frac{\Gamma^{1/2}(3/q)}{\Gamma^{3/2}(1/q)}\exp\left(-\frac{\left\lvert\frac{x-\mu}{\sigma}\right\rvert^{q}}{\left(\frac{\Gamma(1/q)}{\Gamma(3/q)}\right)^{q/2}}\right)$$
Fisher-Tippett	$$FT(x;\mu,\sigma,q)=\frac{1}{\sigma\left(1+\kappa\frac{x-\mu}{\sigma}\right)^{1+1/\kappa}}\exp\left(-\left(1+\kappa\frac{x-\mu}{\sigma}\right)^{-1/\kappa}\right)$$

Combining multiple tests

- **Pearson-Fisher Chi-Square**

COMBINING INDEPENDENT TESTS OF SIGNIFICANCE

Question 14. I have made several tests of significance on independent data. The tests I have made concern the difference between mean scores under two different treatments. Retaining the same order of differencing in the numerator of my t-test, I get t values of $-.68$, $+1.53$, $+2.21$, $+1.85$ with 6, 18, 22, and 25 degrees of freedom respectively. I would like to make a combined test of significance to test whether the difference between the means is positive. I do not want to try to form a single value of t from the original raw data for various experimental reasons. In *Statistical Methods for Research Workers* R. A. Fisher presents a method of combining tests of significance, but it seems not to apply to the present situation since one of my differences is in the wrong direction. Also I fail to see how Fisher's method takes account of the varying degrees of freedom in the four sets of data. Can you tell me an appropriate procedure for combining these data? If you have any other advice about combining independent tests of significance, I would be glad to get it.

Answer. If the test based on the product of probabilities from different trials is adopted, allowance can be made for the occasional occurrence of negative values of t by calculating the corresponding probability for a single tail. Negative values of t will then supply probabilities greater than .5, which do not greatly reduce the product and, since they introduce two additional degrees of freedom, will tend to lower its significance.

Thus with the illustrative values given in the problem, I obtain the following table:

n	t	P(single tail)	$\log_e 1/P$
6	$-.68$	.739	.3025
18	1.53	.0717	2.6353
22	2.21	.01932	3.9466
25	1.85	.03809	3.2698
			10.1542

Doubling the total of the natural logarithms we have chi-square equal to 20.31 for 8 degrees of freedom, corresponding to a probability rather less than 1%. If, as seems here to be the case, consistent deviations in either direction would be deemed relevant, we may say that the probability is less than 2% for the observation of deviations so large and so consistent in direction as those observed.

The differences in the degrees of freedom of the several trials, and also differences in their respective precisions, are, of course, reflected in the values of P to be multiplied together. The inclusion of the first trial lowers the significance of the test, and it may well be that this first trial was on so small a scale that it was likely *a priori* to do so. It would have been open to the experimenter to decide in advance not to conduct any trials with less than 15 degrees of freedom, or, equally legitimately, not to include such trials in his test. Such a decision must, however, be taken in advance and not after inspection of the resultt of the trials. Generally, therefore, if combination by the producs method is intended, it will be worth while to aim at a succession of trials of approximately equal sensitiveness. More usually, however, the test will be employed to give provisional guidance based on all the evidence to hand, and this will often be of very variable precision.

It may be noticed that the same body of data may give a significant combined test in whichever direction the effect is tested. This would, of course, indicate reality of departure from the null hypothesis, but in discrepant directions in different cases, so that the results of the several trials would be heterogeneous.

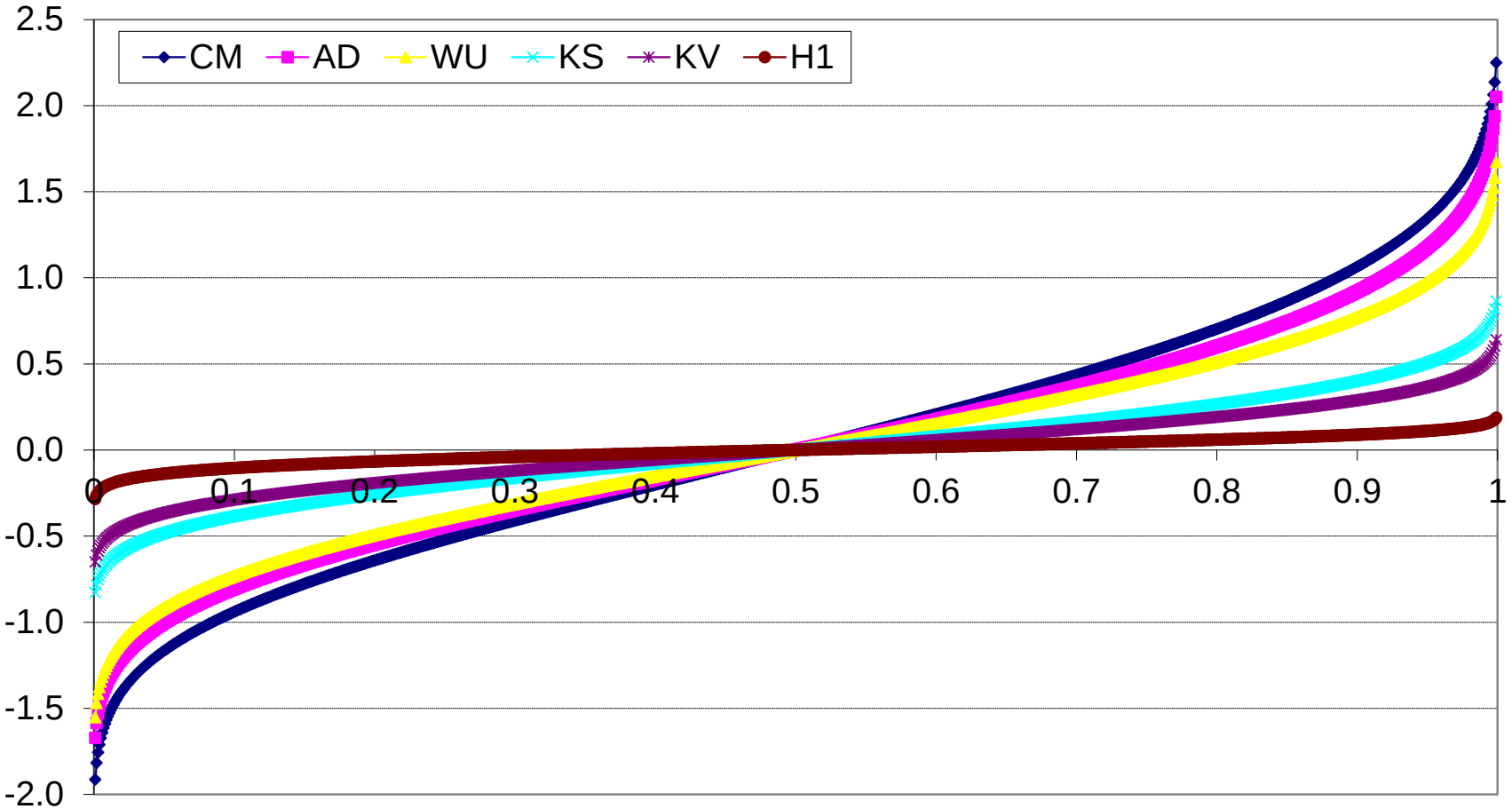
R. A. Fisher

Fisher RA, 1948. "Questions and answers #14". The American Statistician 2(5):30-31

RESULTS

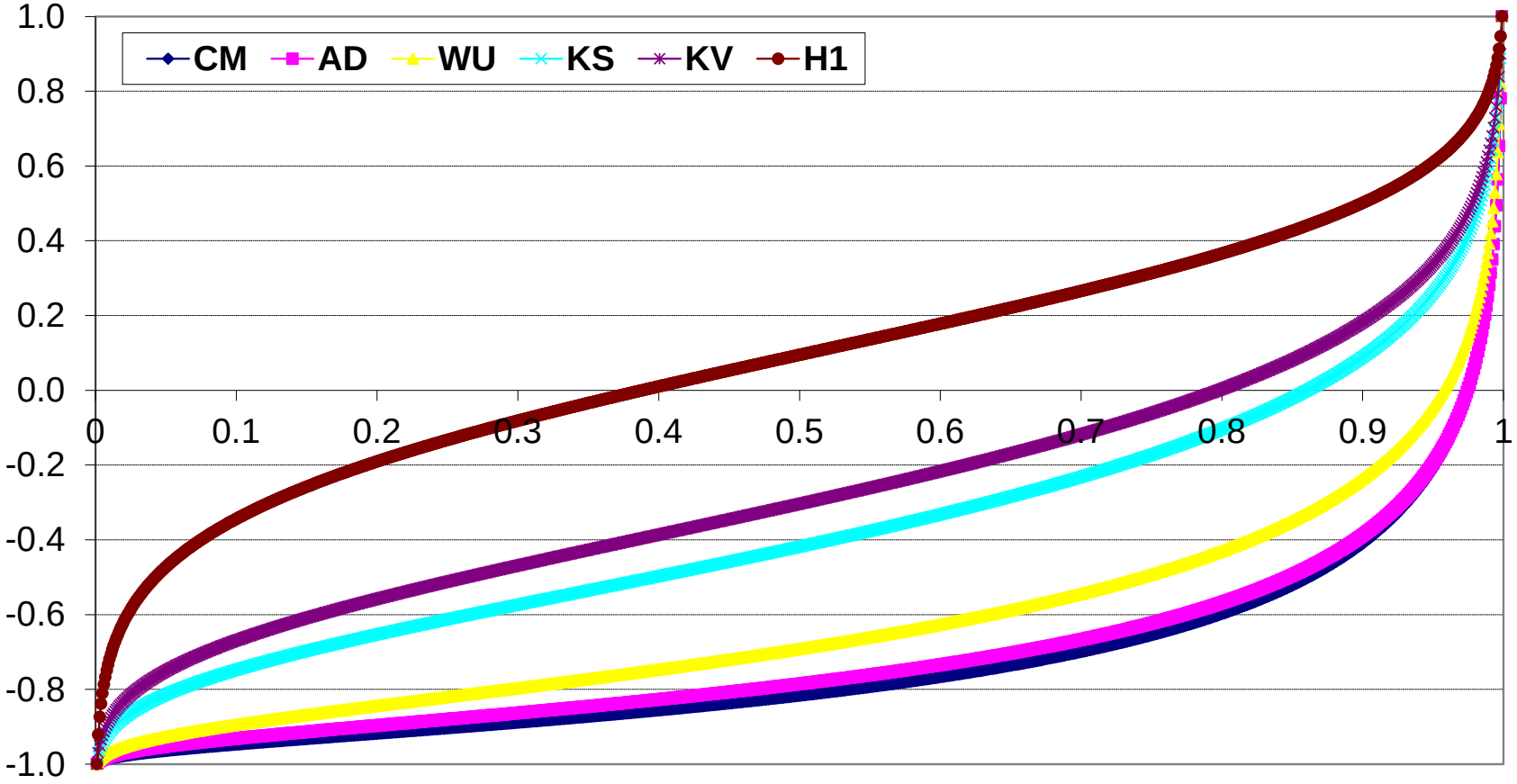
Plots of statistic-probability maps

Statistics in log scale for n = 25



$$\ln(S_p / S_{0.500}), \text{ for } p = 0.001, 0.002, \ldots, 0.999$$
$$S \in \{CM, AD, WU, KS, KV, H1\}$$

Statistics in relative scale for n = 25



$$\frac{2 \cdot S_p}{S_{0.001} + S_{0.999}} - 1, \text{ for } \begin{matrix} p = 0.001, 0.002, \ldots, 0.999 \\ S \in \{CM, AD, WU, KS, KV, H1\} \end{matrix}$$

RESULTS

Scenario 1: combining probabilities from AD, KS, CM, KV and WU

Scenario 2: combining probabilities from AD, KS, CM, KV, WU and H1

Individual rejections at 5% risk being in error for each statistic

Distribution	AD	KS	CM	KV	WU	H1
Gauss-Laplace	9	12	11	19	17	0
Fisher-Tippett	6	5	4	13	11	3
Lognormal	4	7	4	18	16	3
Normal	8	14	10	21	20	0

50 samples were analyzed

H_0 (data follows a certain distribution) were rejected at 5% risk being in error differently by each statistic

Combined rejections at 5% risk being in error for each scenario

Distribution	Scenario 1	Scenario 2
Gauss-Laplace	19	19
Fisher-Tippett	13	13
Lognormal	20	18
Normal	21	21

50 samples were analyzed

H_0 (data follows a certain distribution) were rejected at 5% risk being in error differently by each scenario of combining statistics

DISCUSSION

By taking the 5% risk being in error as the threshold of rejecting H_0 (data follows a certain distribution)

- The scenario (1) not including H1 have the tendency to reject the H_0 more often than any single statistic
- The scenario (2) including H1 have the tendency to reject the H_0 closely to the highest rejection rate of any single statistic (was obtained the same rejection rate as KV)

CONCLUSIONS

- Shannon's statistic seems to have the tendency to fail to reject H_0 more often than all another investigated statistics
- However, its use in a battery of statistics in testing the H_0, it changes the outcome not significantly (2 out of 73 less rejections of H_0) and making it more closely to the maximum rejection rate of a single statistic

References

1. Kolmogorov, A. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari* **1933**, *4*, 83-91.
2. Smirnov, N. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics* **1948**, *19*, 279-281.
3. Anderson, T. W.; Darling, D. A. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics* **1952**, *23*, 193-212.
4. Anderson, T.W.; Darling, D.A. A Test of Goodness-of-Fit. *Journal of the American Statistical Association* **1954**, *49*, 765-769.
5. Pearson, K. Contribution to the mathematical theory of evolution, II. Skew variation in homogenous material. *Philosophical Transactions of the Royal Society of London* **1895**, *91*, 343-414.
6. Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5* **1900**, *50(302)*, 157-175.
7. Cramér, H. On the composition of elementary errors. *Skand. Akt.* **1928**, *11*, 141-180.
8. von Mises, R.E. Wahrscheinlichkeit, Statistik und Wahrheit. Julius Springer: Vienna, Austria; 1928.
9. Shapiro, S.S.; Wilk, M.B. An analysis of variance test for normality (complete samples). *Biometrika* **1965**, *52(3-4)*, 591-611.
10. Jarque, C.M.; Bera, A.K. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters* **1980**, *6(3)*, 255-259.
11. Jarque, C.M.; Bera, A.K. Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. *Economics Letters* **1981**, *7(4)*, 313-318.
12. Jarque, C.M.; Bera, A.K. A test for normality of observations and regression residuals. *International Statistical Review* **1987**, *55(2)*, 163-172.
13. D'Agostino, R.B.; Belanger, A.; D'Agostino, R.B.Jr. A suggestion for using powerful and informative tests of normality. *The American Statistician* **1990**, *44(4)*, 316-321.
14. Lilliefors, H.W. On the Kolmogorov-Smirnov for normality with mean and variance unknown. *Journal of the American Statistical Association* **1967**, *62*, 399-402.
15. Shapiro, S.S.; Francia, R.S. An approximate analysis of variance test for normality. *Journal of the American Statistical Association* **1972**, *67*, 215-216.
16. Jäntschi, L.; Bolboacă, S.D. Distribution fitting 2. Pearson-Fisher, Kolmogorov-Smirnov, Anderson-Darling, Wilks-Shapiro, Kramer-von-Misses and Jarque-Bera statistics. *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Horticulture* **2009**, *66(2)*, 691-697.
17. Fisher, R.A, Questions and answers #14. *The American Statistician* **1948**, *2(5)*, 30-31.
18. Bolboacă, S.D.; Jäntschi, L.; Sestraş, A.F.; Sestraş, R.E.; Pamfil, D.C. Supplementary material of 'Pearson-Fisher chi-square statistic revisited'. *Information* **2011**, *2(3)*, 528-545.
19. Shannon, C.E. A Mathematical Theory of Communication. *Bell System Technical Journal* **1948**, *27(3)*, 379-423.
20. Bolboacă, S.D.; Jäntschi, L. Predictivity Approach for Quantitative Structure-Property Models. Application for Blood-Brain Barrier Permeation of Diverse Drug-Like Compounds. *International Journal of Molecular Science* **2011**, *12*, 4348-4364.
21. Bolboacă, S.D.; Jäntschi, L. *From molecular structure to molecular design through the Molecular Descriptors Family Methodology*, In: Castro, E.A. (Ed.), *QSPR-QSAR Studies on Desired Properties for Drug Design*. Research Signpost, Transworld Research Network, 2010, pp. 117-166.
22. Jäntschi, L.; Bolboacă, S.D., Diudea, M.V. Chromatographic Retention Times of Polychlorinated Biphenyls: from Structural Information to Property Characterization. *International Journal of Molecular Sciences* **2007**, *8(11)*, 1125-1157.