Thailand Statistician
July 2010; 8(2) : 143-155
http://statassoc.or.th
Contributed paper

The Modified Kolmogorov-Smirnov One-Sample Test Statistic

Jantra Sukkasem
Department of Mathematics, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla, 90112, Thailand.
E-mail: jantra.s@psu.ac.th

Received: 31 August 2009
Accepted: 1 June 2010

Abstract

The Kolmogorov-Smirnov (K-S) statistic is a well known nonparametric test statistic used to solve goodness-of fit problems. However, a disadvantage of the K-S is its low power when the true distribution differs from the hypothesized distribution in the tails. Here, a new methodology is proposed in which a modified K-S statistic is obtained. The simulation results indicate that the modified K-S statistic is preferable when the sample size is large.

Keywords : Goodness-of fit test, Kolmogorov-Smirnov (K-S) statistic, power.

1. Introduction

To test if a random sample comes from a hypothesized distribution, the Kolmogorov-Smirnov (K-S) one-sample test, introduced by Kolmogorov [1], may be used. The K-S statistic is based on the maximum absolute difference between the empirical distribution function (e.d.f) and specified cumulative distribution function (c.d.f). It is well known that the Kolmogorov-Smirnov (K-S) test exhibits poor sensitivity to deviations from the hypothesized distribution that occurs in the tails [2]. Modifications of the K-S statistic include Kuiper's test, which was proposed by Kuiper [3] and is more efficient than the K-S statistic (see also Abrahamson [4]). More recently, Dryver and

Sukkasem [5] applied the K-S test in the validation of risk models with a focus on credit scoring models.

The aim of this current study is to investigate a modified K-S statistic through a proposed methodology. The idea centres around partitioning a random sample by the optimal criteria and then obtaining a modified K-S statistic. Section 2 focuses on the goodness-of-fit test based on empirical distribution functions for one-sample situations. The considered statistic is a K-S statistic. Section 3 presents the underlying idea of the proposed methodology and the modified K-S statistic. Section 4 investigates the necessary condition on the modified K-S statistic. The performances of the modified K-S test and the original underlying K-S test are demonstrated and compared with each other via a simulation study in certain circumstances. Section 5 illustrates the application of the modified K-S statistic using a numerical example. Conclusions and discussion are described in Section 6.

2. The Kolmogorov-Smirnov (K-S) One-Sample Statistic for Goodness-of-Fit Test

Given a random sample of size n from an unknown continuous c.d.f F the problem of goodness of fit test is about testing if the random sample has arisen from the hypothesized c.d.f F^*. The testing hypothesis for a distribution test is, for $x \in (-\infty, \infty)$,

$$H_0 : F(x) = F^*(x) \text{ for all x} \qquad (1)$$

against the alternative

$$H_1 : F(x) \neq F^*(x) \text{ for some x.}$$

A common statistic used to test the hypothesis in (1) is the K-S one-sample statistic to decide if the unknown c.d.f F, is in fact, the hypothesized c.d.f F^*. The K-S one-sample statistic is defined as follows [1] :

$$D_n = \sup\left\{ | \hat{F}(x) - F^*(x) |; x \in (-\infty, \infty) \right\}. \qquad (2)$$

The notation $\hat{F}$ is used to represent the empirical distribution function (e.d.f) calculated from a random sample of size n corresponding to the c.d.f F.

The hypothesis in (1) is rejected if and only if the K-S statistics, D_n in (2), is large i.e.,

$$D_n > c \text{, for } c \geq 0, \tag{3}$$

where c is a constant value such that $P\left(D_n > c\right) = \alpha$.

Kolmogorov [1] and Smirnov [6] derived the probability distribution of the K-S one-sample statistic, D_n in (2), when $n \to \infty$ under the null hypothesis of equality. Below is the asymptotic null distribution of the K-S one-sample statistic. If F is any continuous distribution function, then for every $c \geq 0$,

$$\lim_{n\to\infty} P(\sqrt{n}D_n \leq c) = L(c), \tag{4}$$

where $L(c) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2c^2)$.

3. The Methodology

The idea focuses on the condition that the c.d.f's (F and F^*) have identical mean and variance. The methodology is to partition a random sample of size n, say, $\mathbf{X_1, X_2, X_3, \ldots, X_n}$, into 2 sub-samples using a partitioning rule based on the population median. The studied population is similarly partitioned into 2 sub-populations using the same rule. The two K-S statistics are then calculated by measuring the difference between each sub-sample and its corresponding sub-population. The hypothesis testing procedure using the calculated K-S values is as follows:

For $x \in (-\infty, \infty)$,

$$H_0 : F_1(x) = F_1^*(x) \text{ for all x versus } H_1 : F_1(x) \neq F_1^*(x) \text{ for some x}$$

and $$H_0 : F_2(x) = F_2^*(x) \text{ for all x versus } H_1 : F_1(x) \neq F_1^*(x) \text{ for some x,} \tag{5}$$

where F_1 and F_2 are unknown c.d.f's and estimated by using sub-samples X_1 and X_2 respectively, and F_1^* and F_2^* represent the hypothesized c.d.f's corresponding to the sub-populations.

Then the two K-S statistics, D_{n_1} and D_{n_2}, used to test the hypotheses in (5), are defined as follows

$$D_{n_1} = \text{Sup}\left\{|\hat{F}_1(x) - F_1^*(x)|; x \in (-\infty, \infty)\right\}$$

and $$D_{n_2} = \text{Sup}\left\{|\hat{F}_2(x) - F_2^*(x)|; x \in (-\infty, \infty)\right\}. \quad (6)$$

The notations $\hat{F}_1$ and $\hat{F}_2$ are the e.d.f's calculated from the sub-samples X_1 and X_2 and are the estimators of unknown c.d.f's, F_1 and F_2 , respectively.

The first K-S statistic, D_{n_1}, is used to test the first (null) hypothesis in (5) to decide if the c.d.f F_1 is, in fact, the hypothesized c.d.f F_1^*. The second K-S statistic, D_{n_2}, is used to test the second (null) hypothesis in (5) to decide if the c.d.f F_2 is, in fact, the hypothesized c.d.f F_2^*.

The hypothesis in (1) is rejected if and only if at least one of the hypotheses in (5) is rejected. Thus, the testing procedure is that the hypothesis (1) is rejected if and only if at least one of two K-S statistics, D_{n_1} and D_{n_2}, is large i.e.,

$$D_{n_1} > c_1 \text{ or } D_{n_2} > c_2 \text{ for } c_1 \geq 0,\ c_2 \geq 0. \quad (7)$$

The probability of not rejecting both null hypotheses in (5) when they are true is:

$$P(C_1 \cap C_2) = \prod_{i=1}^{2} P(C_i) = (1-\alpha')^2, \quad (8)$$

where C_1 and C_2 are the events that $D_{n_1} \leq c_1$ and $D_{n_2} \leq c_2$ respectively, α' is the adjusted alpha for two comparisons investigated so that

$$1-(1-\alpha')^2 \leq \alpha. \quad (9)$$

Note here that the above procedure is based on partitioning a random sample into two sub-samples. The median of population is the partitioning rule so that the two K-S statistics, D_{n_1} and D_{n_2}, are independent or at least their correlation coefficient is close to zero.

4. Investigation of the correlation between the two statistics, D_{n_1} and D_{n_2}

This section investigates the performance of the modified K-S test compared to the original K-S test using simulation on some selected symmetric distributions, namely Uniform, Student's t, and Gaussian. The simulations were implemented in S–PLUS 6.2 for Windows with 10,000 replications of each Monte Carlo experiment. The actual distributions F and the sample sizes n will be randomly selected in each replication. The population median rule of partitioning a dataset for the modified K-S test and the significance level (α) 0.05 will be fixed. With the same mean and variance, the hypothesized distribution F^* and the actual distribution F are simulated to study the performance of the modified K-S test and that of the original K-S test.

The following case (i) is simulated under the null hypothesis to study the empirical significance level of the modified K-S test and that of the original K-S test.

Case (i) : A population with the uniform (0,1) distribution F^* versus a random sample taken under the actual uniform (0,1) distribution F.

For this case, the empirical significance level of the tests will be explored under the null hypothesis of equality $H_0 : F = F^*$. A random sample X is taken under the actual uniform (0,1) distribution F and a uniform (0,1) population distribution F^* is the hypothesized distribution.

The random samples X of the various sizes n equal to 20, 30, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1,000 are generated under the actual distribution F, the uniform distribution on the unit interval (0,1). For each n, the simulations were implemented with 10,000 replications. In each replication, the population median (0.5) is used to be a partitioning rule for the modified K-S test. The 10,000 K-S tests and the 10,000 modified K-S tests are then performed at the significance level (α) 0.05 and the adjusted alpha (α') 0.02532 respectively. The empirical significance levels for both statistics at each n are calculated by dividing their numbers of rejection the null hypothesis of equality by 10,000. The simulation result is presented in Figure 1.

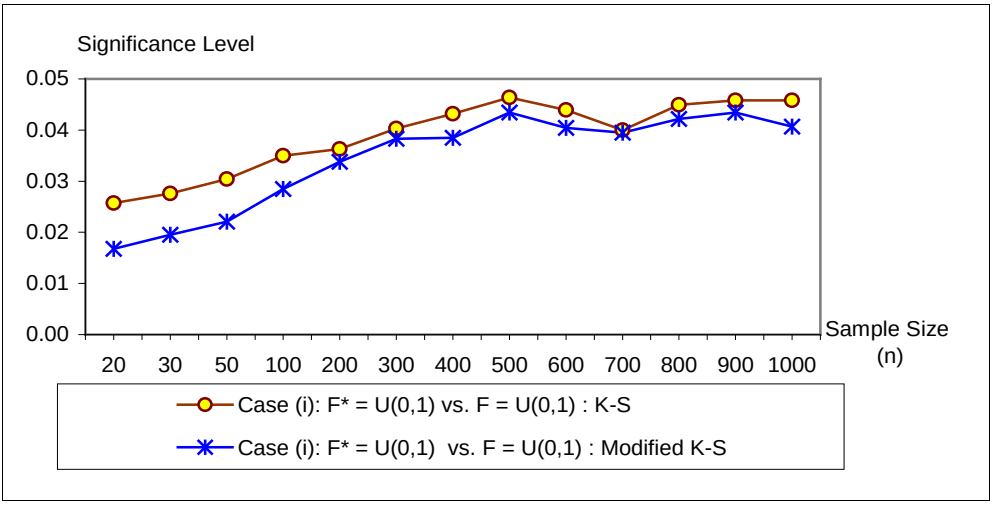

Figure 1. Comparison between the empirical significance level of the modified K-S test based on the median partitioning rule and that of the original K-S test under the null hypothesis of equality between the hypothesized uniform (0,1) distribution and the actual uniform (0,1) distribution of a random sample using 10,000 iterations.

The result shows that the empirical significance levels of the modified K-S test and the original K-S test are similar and close to the significance level (0.05) when the sample size is large for the case (i).

The following three cases (ii)-(iv) are cases of study under the alternative hypothesis of inequality, $H_1 : F \neq F^*$. They are simulated to study the power of the original K-S test and that of the modified K-S test.

Case (ii) : An artificial population taken under the hypothesized uniform (0,1) distribution F^* versus a random sample taken under the actual normal (0.5,1/12) distribution F.

Case (iii) : An artificial population taken under the hypothesized Student's t distribution F^* with degree of freedom 10 versus a random sample taken under the actual uniform (-1.9326684 ,1.9326684) distribution F.

Case (iv) : An artificial population taken under the hypothesized Student's t distribution F^* with degree of freedom 10 versus a random sample taken under the actual normal (0 ,1.115827^2) distribution F.

Here, the artificial population is the data of size 10,000 simulated under the hypothesized distribution F^*. In each of the cases (ii)-(iv), the artificial population is generated to be partitioned into 2 sub-populations by using the population median for the modified K-S test. The random samples of various size n equal to 20, 30, 50, 100, 200, 300, 400, 500, 1,000, 2,000 are generated under the actual distribution F. For each n, the simulations were implemented with 10,000 replications. In each replication, the population median is used to be a partitioning rule for the modified K-S test. The 10,000 original K-S tests and the 10,000 modified K-S tests are then performed at the significance level (α) 0.05 and the adjusted level (α') 0.02532 respectively. The empirical powers of the test under the alternative hypothesis for both statistics at each n are calculated by dividing their numbers of rejection of the null hypothesis of equality by 10,000. The simulation result is presented in Figure 2.

The simulation result indicated that when the sample size n is large, the power of the modified K-S test is superior to that of the original K-S test for any distribution (ii)-(iv).

It is interesting to note that the power of the test in case (iv) is smaller than in the other cases. This is because Student's t distribution and the Gaussian distribution have similar shapes with the same mean and variance. Consequently, the null hypothesis of equality rarely ever fails to be rejected. However, the power of these two tests increases as the sample size n increases.

In conclusion, for testing if a sample comes from the hypothesized distribution F^*, the empirical significance level and the power of the modified K-S test are approximated under the null hypothesis of equality and under fixed alternatives, respectively. The empirical significance levels of the modified K-S test and the original K-S test are similar and close to the significance levels when the sample size n is large. When the sample size is large, the power of the modified K-S test is superior to that of the original K-S test under the median portioning rule.

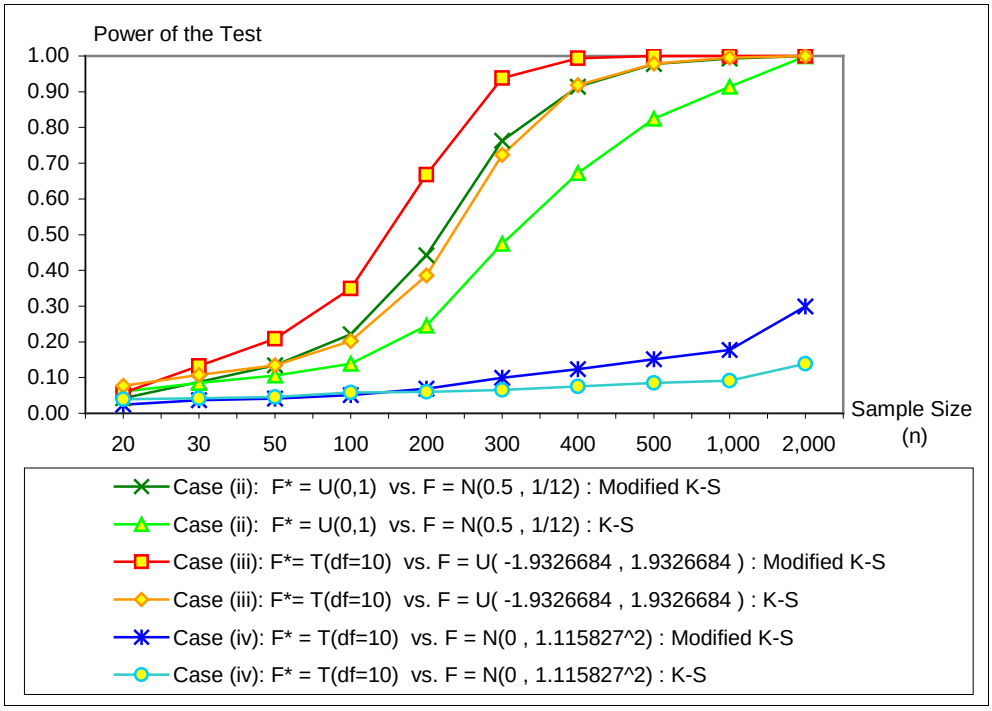

Figure 2. Comparison between the power of the modified K-S test based on the median partitioning rule and that of the original K-S test under the alternative hypothesis of inequality between the hypothesized distribution F^* and the actual distribution F using 10,000 iterations.

5. Numerical Example

In this section we illustrate how to apply the modified K-S statistic. For the simplest case, an example of the case (ii) in section 4 is considered. The hypothesized uniform (0,1) distribution F^* is considered and a random sample X of size n=10 is generated from the actual normal (0.5, 1/12) distribution F as following: 0.0874994, 0.2827867, 0.4177127, 0.5016324, 0.5177516, 0.5226764, 0.5510966, 0.5542756, 0.5845493, 1.0241984. The histogram of the random sample X is shown in Figure 3.

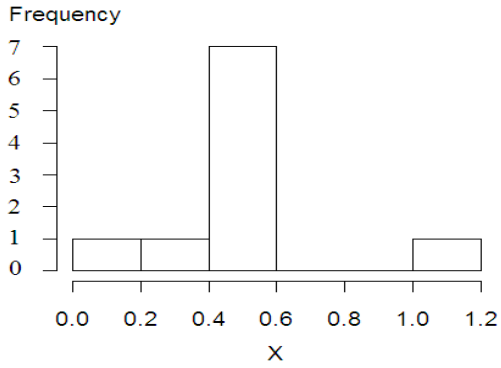

Figure 3. Histogram of the random sample X of size n=10 taken under the actual normal (0.5,1/12) distribution F.

Here it is known that the random sample X has not been arisen from the uniform (0,1) distribution, but the normal (0.5,1/12) distribution. Notice that these two distributions have the same mean 0.5 and variance 1/12. The desire is to test whether or not this sample comes from the uniform (0,1) distribution at the significance level (α) 0.05.

The original K-S test and the modified K-S test are used. The performance of the original K-S test and that of the modified K-S test are illustrated as follows.

(1) Performance of the original K-S Test. Let F^* denotes the uniform (hypothesized) c.d.f and F denotes the normal c.d.f of the random sample X with the corresponding e.d.f $\hat{F}$. The two-sided hypothesis is, for $x \in (-\infty, \infty)$,

$$H_0 : F(x) = F^*(x) \text{ for all x versus } H_1 : F(x) \neq F^*(x) \text{ for some x.}$$

Table 1. Maximum distance between the e.d.f $\hat{F}$ calculated from the random sample X of size n=10 and the Uniform (0,1) c.d.f F^*.

Observation	X~N(0.5,0.288675^2)	E.D.F. $\hat{F}(x)$	Uniform(0,1) C.D.F. $F^*(x)$	$\lvert\hat{F}(x)-F^*(x)\rvert$
1	0.0874994	0.1	0.0874994	0.0125006
2	0.2827867	0.2	0.2827867	0.0827867
3	0.4177127	0.3	0.4177127	0.1177127
4	0.5016324	0.4	0.5016324	0.1016324
5	0.5177516	0.5	0.5177516	0.0177516
6	0.5226764	0.6	0.5226764	0.0773236
7	0.5510966	0.7	0.5510966	0.1489034
8	0.5542756	0.8	0.5542756	0.2457244
9	0.5845493	0.9	0.5845493	0.3154507
10	1.0241984	1.0	1.0000000	0.0000000
			The K-S Statistic: D_n=	0.3154507
			p-value=	0.2726431

For this case, the K-S statistic is 0.3154507 and the p-value is 0.2726431. Therefore, the null hypothesis is failed to reject at the significance level 0.05.

(2) Performance of the modified K-S Test. Consider how to apply the modified K-S statistic to test if the random sample X comes from the uniform (0,1) distribution F^*.

By using the median (m=0.5) of uniform (0,1) population, the random sample X is partitioned into two uncorrelated sub-samples X_1 of size n_1=3 and X_2 of size n_2=7 as following:

X_1= (0.0874994, 0.2827867, 0.4177127) and
X_2= (0.5016324, 0.5177516, 0.5226764, 0.5510966, 0.5542756, 0.5845493, 1.0241984)

The first sub-sample X_1 is used to calculate the first K-S statistic (D_{n_1}). The following Table 2 shows the maximum distance between the e.d.f $\hat{F}_1$, and the Uniform (0,0.5) c.d.f F_1^*.

Table 2. Maximum distance between the e.d.f $\hat{F}_1$ calculated from the first sub-sample X_1 of size n_1= 3 and the Uniform (0,0.5) c.d.f F_1^* under the null hypothesis $H_0 : F_1(x) = F_1^*(x)$.

Observation	The first sample X_1	E.D.F. $\hat{F}_1(x)$	Uniform (0,0.5) C.D.F. $F_1^*(x)$	$\lvert\hat{F}_1(x) - F_1^*(x)\rvert$
1	0.0874994	0.3333333	0.1749988	0.1583345
2	0.2827867	0.6666667	0.5655734	0.1010933
3	0.4177127	1.0000000	0.8354254	0.1645746
			The first K-S Statistic: D_{n1}=	0.1645746
			p-value=	0.9999979

For this case, the first K-S statistic (D_{n_1}) is 0.1645746 and the p-value is 0.9999979. The adjusted significance level (α') for this case is 0.02532 so that $(1-\alpha')^2$ is close to 0.95 as required. Therefore, the null hypothesis $H_0 : F_1(x) = F_1^*(x)$ is failed to reject at the adjusted significance level 0.02532.

Similarly, the observation in the second sub-sample X_2 is used to calculate the second K-S statistic (D_{n_2}). The following Table 3 shows the maximum distance between the e.d.f $\hat{F}_2$ calculated from the second sub-sample X_2 and the c.d.f F_2^*.

Table 3. Maximum distance between the e.d.f $\hat{F}_2$ calculated from the second sub-sample X_2 of size n_2= 7 and the Uniform (0.5,1) c.d.f F_2^* under the null hypothesis $H_0 : F_2(x) = F_2^*(x)$.

Observation	The second sample X_2	E.D.F. $\hat{F}_2(x)$	Uniform (0.5,1) C.D.F. $F_2^*(x)$	$\lvert\hat{F}_2(x) - F_2^*(x)\rvert$
1	0.5016324	0.1428571	0.0032648	0.1395923
2	0.5177516	0.2857143	0.0355032	0.2502111
3	0.5226764	0.4285714	0.0453528	0.3832186
4	0.5510966	0.5714286	0.1021932	0.4692354
5	0.5542756	0.7142857	0.1085512	0.6057345
6	0.5845493	0.8571429	0.1690986	0.6880443
7	1.0241984	1.0000000	1.0000000	0.0000000
			The second K-S Statistic: D_{n2}=	0.6880443
			p-value=	0.0026465

For this case, the second K-S statistic (D_{n_2}) is 0.6880443 and the p-value is 0.0026465. Therefore, the null hypothesis $H_0 : F_2(x) = F_2^*(x)$ is rejected at the significance level (α') 0.02532.

Since one of the two K-S statistics is large, it implies that the difference is significant at the significance level (α) 0.05. That is the random sample X does not come from the uniform (0,1) distribution.

Hence, for this example, the performance of the original K-S test and the modified K-S test are different. The K-S test cannot be used to conclude that the data is not the normal (0.5,1/12) distribution, but the modified K-S test can.

6. Conclusion and Discussion

The proposed methodology is based on partitioning a random sample by using the population median. The modified K-S statistic is then obtained for testing whether or not a random sample comes from the specific distribution F^*. In order to use the modified K-S statistic, the two hypotheses will be tested namely, (1) $H_0 : F_1(x) = F_1^*(x)$ (2) $H_0 : F_2(x) = F_2^*(x)$, instead of testing the hypothesis of equality $H_0 : F = F^*$. With independence of these two tests, an adjusted significance level (α') will be used. The simulation result indicated that (1) the correlation between two K-S statistics (D_{n_1} and D_{n_2}) obtained from the partitioning rule is slightly negative for small sample size n; the correlation approaches zero as the sample size n becomes larger. (2) The result shows that the empirical significance levels of the modified K-S test and the original K-S test are similar and close to the significance level when the sample size is large. (3) The power of the modified K-S test is superior to that of the original K-S test when the sample size n is large. In summary, the modified K-S test is preferable when the sample size n is large in certain circumstances.

The study is only on symmetric distributions, specifically Uniform, Student's t, and Gaussian. The target is to compare the difference between two c.d.f's with the same mean and variance when the true c.d.f differs from the hypothesized c.d.f in the tails.

The idea of the proposed methodology is based on a partition of a random sample into 2 sub-samples by using the median population so that the two K-S statistics (D_{n_1} and D_{n_2}) will be independent or at least asymptotically the correlation will be

closed to zero. The correlation between the two K-S statistics (D_{n_1} and D_{n_2}) can be seen to go to zero in the simulation result (Figure 1).

The application of the modified K-S statistic is more complicated than that of the original K-S statistic. This is because the two hypotheses of equality depending on the two c.d.f's are used instead of the single hypothesis of equality.

It is interesting to apply the same idea in the context of partitioning a random sample to other goodness of fit test statistics, such as the Kuiper statistic, the Gini coefficient for both one and two-sample problems.

Acknowledgments

I would like to acknowledge Dr. Arthur L. Dryver and Dr. Edward McNeil for help and guidance and Dr. Naratip Jansakul for all support.

References

[1] Kolmogorov, A.N. Sulla Determinazione Empirica Di una Legge Di Distribuzione, Giornale Dell' Istituto Italiano Degli Attuari, 1933; 4: 83-91.

[2] Mason, M.D. and Schuenemeyer, H.J. A Modified Kolmogorov-Smirnov Test Sensitive to Tail Alternatives, The Annals of Statistics, 1983; 11: 933-946.

[3] Kuiper, N.H. Tests Concerning Random Points on a Circle, Proc. Konink. Ned. Akad. Van Wettenschaffen, A., 1960; 63: 38-47.

[4] Abrahamson, I.G. Exact Bahadur Efficiencies for the Kolmogorov-Smirnov and Kuiper One- and Two-Sample Statistics, The Annals of Mathematical Statistics, 1967; 38: 1475-1490.

[5] Dryver, A.L, and Sukkasem, J. Validating Risk Models with a Focus On Credit Scoring Models, Journal of Statistical Computation and Simulation, 2009; 79: 181-193.

[6] Smirnov, N.V. Estimate of Deviation Between Empirical Distribution Functions in Two Independent Samples. (Russian) Bulletin Moscow University**.** 1939; 2 (2): 3-16.