Proceedings of the Jangjeon Mathematical Society
20 (2017), No. 2. pp. 269 - 311

www.jangjeon.or.kr
http://dx.doi.org/10.23001/pjms2017.20.2.269

CALCULATON OF THE KOLMOGOROV-SMIRNOV AND KUIPER STATISTICS OVER FUZZY SAMPLES

Nikolova, N.[*#&], Ivanova, S.[*], Chin, C.[#], Tenekedjiev, K.[*#]

*Nikola Vaptsarov Naval Academy, Varna, Bulgaria

#Australian Maritime College (University of Tasmania), Australia

& Corresponding author, 1 Maritime Way, Launceston, TAS, Australia

e-mail: Natalia.Nikolova@utas.edu.au

Abstract

Fuzzy samples contain measurements that are only partially associated with their underlying population. This paper offers numerical indices for the difference of population distributions approximated over fuzzy samples. Formulae for the Kolmogorov-Smirnov (KS) and Kuiper (Ku) statistics in the case of fuzzy empirical cumulative distribution functions are given. It proves that the suprema in these criteria' standard formulae convert to maxima in the analyzed case, which substantially facilitates calculations. As a by-product, the paper also proves formulae for KS and Ku in the case of rigid samples (that are often used but never properly formalized). If Bootstrap and Monte Carlo simulations are employed to construct the distribution of KS and Ku and find the pvalue of the tests, then the quick and reliable calculation of the test statistics in each pseudo reality are of great importance. The derived formula for KS and Ku improve the quality of the simulation itself.

62E86 of the 2010 Mathematics Subject Classification

Keywords and phrases. Fuzzy samples; hypothesis testing; Kuiper; Kolmogorov-Smirnov, theorem

1. Introduction

In various cases of data collection one may identify situations, where measurements in a data sample are only partially associated with their underlying population. The presence of such data imposes challenges to any statistical procedure of comparison of distributions or numerical characteristics of variables. The work [Nikolova et al., 2015] presents procedures to test the identity of distributions on the basis of one key so called fuzzy samples. In the proposed definition, the authors use two 1D (one-dimensional) samples of a continuous parameter that contain respectively n_1 and n_2 observations. The fuzziness of the samples is based on the assumption that the observations z_k^1 and z_k^2 of each respective sample belong to two populations – Population 1 and Population 2 – with a given degree of membership respectively μ_k^1 and μ_k^2. In that way, the two fuzzy samples are presented as:

$$Z^1 = \left\{\left(z_1^1, \mu_1^1\right), \left(z_2^1, \mu_2^1\right), ..., \left(z_{n_1}^1, \mu_{n_1}^1\right)\right\} \tag{1.1}$$

$$Z^2 = \left\{\left(z_1^2, \mu_1^2\right), \left(z_2^2, \mu_2^2\right), ..., \left(z_{n_2}^2, \mu_{n_2}^2\right)\right\} \tag{1.2}$$

In the presence of fuzzy samples, one statistical procedure to perform is to define whether the two samples were drawn from two populations with the same characteristics.

Some examples of where fuzzy samples apply are given in [Viertl, 2011]. They show that the degree of membership may have different interpretations depending on the case analyzed. For example, in comparing subpopulations from two different populations, the membership of an observation from that subpopulation is calculated from a classificator. The resulting calculation has a degree of certainty that can be interpreted as a degree of membership to the subpopulation. An area, where problems with fuzzy samples arise is also medical research, where a given parameter may be measured in multiple spatial points for the same object (e.g. measurements in in different sections of an object, or measurement of the same section in different moment of time). Then, each measurement may be given a degree of membership, e.g. equal proportional weight so that to provide equal importance to each of the measured objects.

Let us denote with CDF_1 and CDF_2 any sample distribution function constructed on the data points in samples (1.1) and (1.2). In the presence of a rigid data sample, a common procedure to construct a distribution of the underlying random variable is to approximate the cumulative distribution function (CDF) by the so-called empirical distribution function (ECDF). The only assumption of this procedure is that the data points in the sample are independent and identically distributed (i.i.d.). ECDF assumes that the r.v. is discrete with possible values coinciding with the observations in the rigid

sample. The assigned probability to each value is its relative frequency in the sample. If fuzzy samples are present, then the idea and form of ECDF can be generalized to the fuzzy empirical distribution function (FECDF):

$$FECDF_1(z) = \sum_{\substack{k=1 \\ z_k^1 \le z}}^{n_1} \mu_k^1 \Big/ \sum_{k=1}^{n_1} \mu_k^1, \quad \text{for } z \in (-\infty; +\infty) \tag{1.3}$$

$$FECDF_2(z) = \sum_{\substack{k=1 \\ z_k^2 \le z}}^{n_2} \mu_k^2 \Big/ \sum_{k=1}^{n_2} \mu_k^2, \quad \text{for } z \in (-\infty; +\infty) \tag{1.4}$$

Similar to the ECDF approximations, the sample approximations (1.3) and (1.4) use no assumptions for the type of the approximated CDFs except for the standard i.i.d. assumption.

Comparing two population distributions (for equality of distributions as a whole, or for equality of any of their numerical characteristics) usually brings down to the use of a given test statistic *S*. When testing the equality of two population distributions, *S* is an estimator of the difference between two sample approximations of CDF under the assumption that their underlying populations had equal continuous distributions. *S* is a random variable that tends to increase when the difference between the sample distributions CDF_1 and CDF_2 increases. A statistical test is adopted with null hypothesis (H_0): "The continuous distributions of the two populations are equal" and alternative hypothesis (H_a): "The continuous distributions of the two populations are different". The p_{value} of the test may be calculated if the conditional distribution of the test statistic under a true null hypothesis was available.

There are three major groups of statistics that apply to the test of identity of two continuous distributions. The work [Chernobai, Rachev, Fabozzi, 2014] discusses the so-called quadratic class, along with some of its most typical representatives, such as the quadratic Anderson-Darling statistic and the Cramér–von Mises statistic. Another set of measures, with its very wide spread metrics, namely the Mann-Whitney *U* statistic and the Wilcoxon *T* statistic [Groebner et al., 2011] form the so-called rank class. By far the wide spread class is the supremum class [Chernobai, Rachev, Fabozzi, 2014]. A popular test statistic that belongs to this class is the Kolmogorov – Smirnov (*KS*) test statistic [Böhm, Hornik, 2010]. It is calculated as the supremum of the absolute value of the difference between the two available sample CDFs:

$$KS = \sup_z \left(\left| CDF_1(z) - CDF_2(z) \right| \right) \tag{1.5}$$

Another test statistic that exceeds the qualities of the *KS* is the Kuiper statistic (*Ku*) [Lemeshko, Gorbunova, 2013a]. It is the sum of the supremum of the positive difference and the supremum of the negative difference between two approximations of CDF on the available samples:

$$Ku = \sup_z \left(CDF_1(z) - CDF_2(z) \right) + \sup_z \left(CDF_2(z) - CDF_1(z) \right) \tag{1.6}$$

The *Ku* is an improved modification of the *KS*. Its sensitivity to deviations is equal for all values of the underlying variable *z*. It means that the *Ku* statistic sensitivity at the tails is the same as in the middle of the range of *z*, which also makes it invariant to cyclic transformations of the variable. Unlike it, the *KS* statistic's values vary and are difficult to identify at the ends of the interval of *z* [Jin et al., 2015].

In the general case of continuous CDFs, it is not possible to investigate the nature of the supremum so that to make proper calculations of (1.5) and (1.6) precisely and with certainty. In (1.5) and in (1.6), we need to solve two continuous optimization problems, which are further complicated by the fact that they require to find suprema, not maxima. Any numerical solution of this problem is uncertain and time consuming. It is not a coincidence that even sophisticated statistical text do replace suprema with maxima with no proper justification [Press et al., 2007, p. 732, p. 737].

In this paper, we will prove that if we use the FECDFs in (1.3) and (1.4) for the sample distributions CDF_1 and CDF_2 in (1.5) and (1.6), then the two criteria can be simplified to:

$$KS = \max \left\{ \begin{array}{l} \max\limits_{k=1,2,\ldots,n_1} \left(FECDF_1\left(z_k^1\right) - FECDF_2\left(z_k^1\right) \right), \\ \max\limits_{k=1,2,\ldots,n_2} \left(FECDF_2\left(z_k^2\right) - FECDF_1\left(z_k^2\right) \right) \end{array} \right\} \tag{1.7}$$

$$Ku = \max\limits_{k=1,2,\ldots,n_1} \left(FECDF_1\left(z_k^1\right) - FECDF_2\left(z_k^1\right) \right) + \max\limits_{k=1,2,\ldots,n_2} \left(FECDF_2\left(z_k^2\right) - FECDF_1\left(z_k^2\right) \right) \tag{1.8}$$

By doing so, the two continuous optimization problems in (1.5) may be replaced with two optimizations over discrete sets with power n_1 and n_2 in (1.7). Similarly, the two

continuous optimization problems in (1.6) may be replaced with two optimizations over discrete sets with power n_1 and n_2 in (1.8). The discrete optimization problems in (1.7) and (1.8) can be calculated quickly and easily and the result will be guaranteed.

Substituting (1.3), (1.4) in (1.7) and (1.8), we can arrive at the value of *KS* and *Ku* directly from the observations in the samples without the need to construct the FECDFs.

$$KS = \max\left\{\begin{array}{l} \max\limits_{i=1,2,\ldots,n_1}\left(\sum\limits_{\substack{k=1\\ z_k^1\le z_i^1}}^{n_1}\mu_k^1 \Big/ \sum_{k=1}^{n_1}\mu_k^1 - \sum\limits_{\substack{k=1\\ z_k^2\le z_i^1}}^{n_2}\mu_k^2 \Big/ \sum_{k=1}^{n_2}\mu_k^2\right), \\ \max\limits_{i=1,2,\ldots,n_2}\left(\sum\limits_{\substack{k=1\\ z_k^2\le z_i^2}}^{n_2}\mu_k^2 \Big/ \sum\limits_{\substack{k=1\\ z_k^2\le z_i^2}}^{n_2}\mu_k^2 + \sum\limits_{\substack{k=1\\ z_k^1\le z_i^2}}^{n_1}\mu_k^1 \Big/ \sum_{k=1}^{n_1}\mu_k^1\right) \end{array}\right\} \tag{1.9}$$

$$\begin{aligned} Ku = &\max_{i=1,2,\ldots,n_1}\left(\sum\limits_{\substack{k=1\\ z_k^1\le z_i^1}}^{n_1}\mu_k^1 \Big/ \sum_{k=1}^{n_1}\mu_k^1 - \sum\limits_{\substack{k=1\\ z_k^2\le z_i^1}}^{n_2}\mu_k^2 \Big/ \sum_{k=1}^{n_2}\mu_k^2\right) + \\ &+ \max_{i=1,2,\ldots,n_2}\left(\sum\limits_{\substack{k=1\\ z_k^2\le z_i^2}}^{n_2}\mu_k^2 \Big/ \sum\limits_{\substack{k=1\\ z_k^2\le z_i^2}}^{n_2}\mu_k^2 + \sum\limits_{\substack{k=1\\ z_k^1\le z_i^2}}^{n_1}\mu_k^1 \Big/ \sum_{k=1}^{n_1}\mu_k^1\right) \end{aligned} \tag{1.10}$$

As a result, the use of (1.9) and (1.10) allows to bring down the calculation of *KS* and *Ku* statistics to a finite number of FECDF calculations in given data points.

2. General outline of the proof

The main result of the paper in given in section 5. It contains the theorem for calculation of Kolmogorov-Smirnov and Kuiper criteria over fuzzy samples. It proves that: a) both criteria always exist; b) both criteria belong to the interval [0; 1]; c) the suprema in (1.5) and (1.6) are maxima; d) the criteria can be calculated using (1.9) and (1.10) in not more than $(n_1+ n_2)$ points.

To prove this theorem, we use two lemmas that could also exist separately and have their importance.

The first is a lemma for the bounded suprema, given in section 3. It proves that for arbitrary distribution functions CDF_1 and CDF_2, there always exist three functionals F_1,

F_2, and F, that are non-negative and not greater than 1. F_1 is the supremum of the difference of CDF_1 and CDF_2. F_2 is the supremum of the difference of CDF_2 and CDF_1. F is the sum of suprema of the two differences of CDF_1 and CDF_2.

The second is a lemma for the discrete maximum, given in section 4. It treats some properties of the function D, which is the difference between: a) FECDFs as in (1.3), derived from Fuzzy Sample 1; b) an arbitrary distribution function. The lemma proves that the function D has a global non-negative maximum, not higher than 1.

A corollary of the proven theorem for calculation of the Kolmogorov-Smirnov and Kuiper criteria over rigid samples is proven in section 6. In that section, the results of the theorem are trivially applied on samples, where all degrees of membership are equal to 1, which makes the samples rigid:

$$Z^1 = \left\{ z_1^1, z_2^1, ..., z_{n_1}^1 \right\} \tag{2.1}$$

$$Z^2 = \left\{ z_1^2, z_2^2, ..., z_{n_2}^2 \right\} \tag{2.2}$$

In that case, FECDFs are replaced by ECDFs:

$$ECDF_1(z) = \sum_{\substack{k=1 \\ z_k^1 \le z}}^{n_1} 1 \Big/ n_1, \quad \text{for } z \in (-\infty; +\infty) \tag{2.3}$$

$$ECDF_2(z) = \sum_{\substack{k=1 \\ z_k^2 \le z}}^{n_2} 1 \Big/ n_2, \quad \text{for } z \in (-\infty; +\infty) \tag{2.4}$$

The corollary from section 6 proves that for the Kolmogorov-Smirnov and Kuiper criteria calculated over rigid samples: a) both criteria always exist; b) both criteria belong to the interval [0; 1]; c) the suprema in (1.5) and (1.6) are maxima; d) the criteria can be calculated using (2.5) and (2.6) in not more than $(n_1 + n_2)$ points:

$$KS = max \left\{ \max_{i=1,2,...,n_1} \left(\sum_{\substack{k=1 \\ z_k^1 \le z_i^1}}^{n_1} 1 \Big/ n_1 - \sum_{\substack{k=1 \\ z_k^2 \le z_i^1}}^{n_2} 1 \Big/ n_2 \right), \max_{i=1,2,...,n_2} \left(\sum_{\substack{k=1 \\ z_k^2 \le z_i^2}}^{n_2} 1 \Big/ n_2 - \sum_{\substack{k=1 \\ z_k^1 \le z_i^2}}^{n_1} 1 \Big/ n_1 \right) \right\} \tag{2.5}$$

$$Ku = \max_{i=1,2,\ldots,n_1} \left(\sum_{\substack{k=1 \\ z_k^1 \le z_i^1}}^{n_1} 1 \Big/ n_1 - \sum_{\substack{k=1 \\ z_k^2 \le z_i^1}}^{n_2} 1 \Big/ n_2 \right) + \max_{i=1,2,\ldots,n_2} \left(\sum_{\substack{k=1 \\ z_k^2 \le z_i^2}}^{n_2} 1 \Big/ n_2 - \sum_{\substack{k=1 \\ z_k^1 \le z_i^2}}^{n_1} 1 \Big/ n_1 \right) \tag{2.6}$$

In what follows, section 3 provides the setup and proof of a lemma for the bounded suprema. Section 4 presents the setup and proof of the lemma for the discrete maximum. Section 5 presents the setup and proof of the theorem for calculation of Kolmogorov-Smirnov and Kuiper criteria over fuzzy samples, whereas the Corollary for the calculation of Kolmogorov-Smirnov and Kuiper criteria over fuzzy samples is given in section 6.

3. Lemma for the bounded suprema

Setup of the lemma for the bounded suprema

Let $CDF_1(z)$ be a real function of a numeric argument, defined for any real z, with the following properties:

- $CDF_1(.)$ is increasing:

If $z_1 > z_2$, then $CDF_1(z_1) \ge CDF_1(z_2)$ (3.1)

- When z approaches $-\infty$, the limit of $CDF_1(z)$ exists and is equal to 0:

$$\lim_{z \to -\infty} CDF_1(z) = 0 \tag{3.2}$$

- When z approaches $+\infty$, the limit of $CDF_1(z)$ exists and equals to 1:

$$\lim_{z \to +\infty} CDF_1(z) = 1 \tag{3.3}$$

Let $CDF_2(z)$ be a real function of a numeric argument, defined for any real z, with the following properties:

- $CDF_2(.)$ is increasing:

If $z_1 > z_2$, then $CDF_2(z_1) \ge CDF_2(z_2)$ (3.4)

- When z approaches $-\infty$, the limit of $CDF_2(z)$ exists and is equal to 0:

$$\lim_{z\to-\infty} CDF_2(z) = 0 \tag{3.5}$$

- When z approaches $+\infty$, the limit of $CDF_2(z)$ exists and is equal to 1:

$$\lim_{z\to+\infty} CDF_2(z) = 1 \tag{3.6}$$

Then:

- the functional F_1 of the functions $CDF_1(.)$ and $CDF_2(.)$, defined as the supremum of the difference of $CDF_1(.)$ and $CDF_2(.)$, always exists, is non-negative and is not greater than 1:

$$\text{there exists } F_1 \text{ such that } F_1(CDF_1, CDF_2) = \sup_R(CDF_1 - CDF_2) \in [0;1] \tag{3.7}$$

In (3.7), and throughout the section, the set of real numbers is denoted as R.

- the functional F_2 of the functions $CDF_1(.)$ and $CDF_2(.)$, defined as the supremum of the difference of $CDF_2(.)$ and $CDF_1(.)$, always exists, is non-negative and is not greater than 1:

$$\text{there exists } F_2 \text{ such that } F_2(CDF_1, CDF_2) = \sup_R(CDF_2 - CDF_1) \in [0;1] \tag{3.8}$$

- the functional F of the functions $CDF_1(.)$ and $CDF_2(.)$, defined as the sum of the suprema of the differences of those functions, always exists, is non-negative and is not greater than 1:

there exists F such that

$$F(CDF_1, CDF_2) = \left(\sup_R(CDF_1 - CDF_2) + \sup_R(CDF_2 - CDF_1)\right) \in [0;1] \tag{3.9}$$

Proof of the lemma for the bounded suprema

1) Introduce auxiliary functions

Let $f_1(z)$ be a real function of a numeric argument, defined for each real z as a difference between $CDF_1(.)$ and $CDF_2(.)$:

$$f_1(z) = CDF_1(z) - CDF_2(z), \text{ for } z \in R \tag{3.10}$$

Let $f_2(z)$ be a real function of a numeric argument, defined for each real z, and the difference of $CDF_2(.)$ and $CDF_1(.)$ is:

$$f_2(z) = CDF_2(z) - CDF_1(z) = -f_1(z), \text{ for } z \in R \tag{3.11}$$

Let $f(z)$ be a real function of a numeric argument, defined for each real ordered pair (z_1, z_2), where the sum of the differences of $CDF_1(.)$ and $CDF_2(.)$ is:

$$f(z_1, z_2) = CDF_1(z_1) - CDF_2(z_1) + CDF_2(z_2) - CDF_1(z_2) = f_1(z_1) + f_2(z_2), \quad \text{for} (z_1, z_2) \in R^2 \tag{3.12}$$

2) Proof of statement (3.7)

2.1) Bound of the function $f_1(.)$ from above

$f_1(z)$

$= CDF_1(z) - CDF_2(z)$ (according to (3.10))

$\le 1 - CDF_2(z)$ (since $CDF_1(z) \le 1$, according to (3.1) and (3.3))

$\le 1 - 0$ (since $CDF_2(z) \ge 0$, according to (3.4) and (3.5))

$=1$

It follows then that

$$f_1(z) \le 1, \text{ for } z \in R \tag{3.13}$$

2.2) Existence of the functional F_1

$F_1(CDF_1, CDF_2)$

$= \sup_R (CDF_1 - CDF_2)$ (according to the definition of F_1 in (3.7))

$= \sup_R (f_1)$ (according to (3.10))

$= sup\{f_1(z)|z \in R\}$ (according to the definition for supremum of a function [Apostol, 1981])

The set $\{f_1(z)|z \in R\}$ is a non-empty subset of real numbers, which is bounded from above according to (3.13). According to the principle of continuity of real numbers, such sets always have a supremum, which also is a real number [Royden, Fitzpatrick, 2010]. Then:

there exists F_1 such that

$$F_1(CDF_1, CDF_2) = \sup_R(CDF_1 - CDF_2) = \sup_R(f_1) = sup\{f_1(z)|z \in R\} \quad (3.14)$$

2.3) Upper bound of the supremum F_1

According to (3.13), the upper bound of the function $f_1(.)$ is 1, and hence this is also true for the set of real numbers $\{f_1(z)|z \in R\}$. According to (3.14), the supremum F_1 of the function $f_1(.)$ always exists. According to the definition for supremum of a numerical set, the supremum is the smallest upper bound of the set [Rudin, 1976]. Therefore:

$$F_1(CDF_1, CDF_2) = \sup_R(CDF_1 - CDF_2) = sup\{f_1(z)|z \in R\} \leq 1 \quad (3.15)$$

2.4) Lower bound of the supremum F_1

According to (3.14), the supremum F_1 of the function $f_1(.)$ always exists. Let us assume that F_1 is a negative number a:

$$F_1(CDF_1, CDF_2) = \sup_R(CDF_1 - CDF_2) = \sup_R(f_1) = a < 0 \quad (3.16)$$

The limit of $f_1(.)$ in $+\infty$ may be easily calculated:

$$\lim_{z \to +\infty} f_1(z)$$

$$= \lim_{z \to +\infty} (CDF_1(z) - CDF_2(z)) \quad \text{(according to (3.10))}$$

$= \lim_{z\to+\infty} CDF_1(z) - \lim_{z\to+\infty} CDF_2(z)$ (according to the theorem for the limit of differences [Stewart, 2016])

$=1 - 1$ (according to (3.3) and (3.6))

$=0$

It follows that:

$$\lim_{z\to+\infty} f_1(z) = 0 \tag{3.17}$$

From (3.17) and from the definition of function limit in $+\infty$ [Thomas, Weir, Hass, 2014] it follows that if we select $\varepsilon = -a/2 > 0$, then there exists z^* such that

$$|f_1(z)| = |f_1(z) - 0| < \varepsilon = -a/2 \text{ for all } z>z^* \tag{3.18}$$

Since $(z^*+1)>z^*$, from (3.18) it follows that:

$$a/2 = -(-a/2) < f_1(z^*+1) < -a/2 \tag{3.19}$$

Then:

$f_1(z^*+1)$

$> a/2$ (according to (3.19))

$> a$ (since a is negative according to the assumption (3.16))

$= \sup_R(f_1)$ (according to the assumption (3.16))

$\geq f_1(z^*+1)$ (the supremum of a function is not lower than any arbitrary value of the function)

It follows that $f_1(z^*+1) > f_1(z^*+1)$, which is a contradiction. Hence, the assumption (3.16) is not correct. Since according to (3.14) the supremum F_1 of $f_1(.)$ always exists, then F_1 is a non-negative number a:

$$F_1\left(CDF_1, CDF_2\right) = \sup_R\left(CDF_1 - CDF_2\right) = \sup_R\left(f_1\right) = a \geq 0 \tag{3.20}$$

2.5) Generalization for the supremum F_1

Taking into account (3.14), (3.15) and (3.20) it follows that (3.7) is correct:

$$\text{there exists } F_1 \text{ such that } F_1\left(CDF_1, CDF_2\right) = \sup_R\left(CDF_1 - CDF_2\right) \in [0;1]$$

which we needed to prove.

3) Proof of statement (3.8)

3.1) Bound of the *function $f_2(.)$ from above*

$f_2\left(z\right)$

$= CDF_2\left(z\right) - CDF_1\left(z\right)$ (according to (3.11))

$\leq 1 - CDF_1\left(z\right)$ (since $CDF_2\left(z\right) \leq 1$, according to (3.4) and (3.6))

$\leq 1 - 0$ (since $CDF_1\left(z\right) \geq 0$, according to (3.1) and (3.2))

$=1$

It follows that

$$f_2\left(z\right) \leq 1, \text{ for } z \in R \tag{3.21}$$

3.2) Existence of the functional F_2

$F_2\left(CDF_1, CDF_2\right)$

$= \sup_R\left(CDF_2 - CDF_1\right)$ (according to the definition of F_2 in (3.8))

$= \sup_R\left(f_2\right)$ (according to (3.11))

$= \sup\left\{f_2\left(z\right) \middle| z \in R\right\}$ (according to the definition of supremum

of a function [Apostol, 1981])

The set $\left\{ f_2\left(z\right) \middle| z \in R \right\}$ is non-empty subset of the real numbers, which is bounded from above according to (3.21). According to the principle for continuity of real numbers, such sets always have a supremum, which is also a real number [Royden, Fitzpatrick, 2010]. Hence:

there exists F_2 such that

$$F_2\left(CDF_1, CDF_2\right) = \sup_R\left(CDF_2 - CDF_1\right) = \sup_R\left(f_2\right) = sup\left\{ f_2\left(z\right) \middle| z \in R \right\} \tag{3.22}$$

3.3) Upper bound of *the supremum F_2*

According to (3.21), the upper limit of the function $f_2(.)$ is 1, and hence this is also true for the set of real numbers $\left\{ f_2\left(z\right) \middle| z \in R \right\}$. According to (3.22), the supremum F_2 of the function $f_2(.)$ always exists. According to the definition for supremum of a numerical set, the supremum is the smallest upper bound of the set [Rudin, 1976]. Hence:

$$F_2\left(CDF_1, CDF_2\right) = \sup_R\left(CDF_2 - CDF_1\right) = sup\left\{ f_2\left(z\right) \middle| z \in R \right\} \le 1 \tag{3.23}$$

3.4) Lower bound of the supremum F_2

According to (3.22), the supremum F_2 of the function $f_2(.)$ always exists. Let us assume that F_1 is a negative number a:

$$F_2\left(CDF_1, CDF_2\right) = \sup_R\left(CDF_2 - CDF_1\right) = \sup_R\left(f_2\right) = a < 0 \tag{3.24}$$

The limit of $f_2(.)$ in $+\infty$ may be easily found to be:

$$\lim_{z\to+\infty} f_2\left(z\right)$$

$$= \lim_{z\to+\infty}\left(CDF_2\left(z\right) - CDF_1\left(z\right)\right) \qquad \text{(according to (3.11))}$$

$$= \lim_{z\to+\infty} CDF_2\left(z\right) - \lim_{z\to+\infty} CDF_1\left(z\right) \qquad \text{(according to the theorem for the limit of a differences [Stewart, 2016])}$$

$=1-1$ (according to (3.3) and (3.6))

$=0$

It follows that

$$\lim_{z\to+\infty} f_2(z) = 0 \tag{3.25}$$

From (3.25) and from the definition for the limit of a function in $+\infty$ [Thomas, Weir, Hass, 2014] it follows that if you choose $\varepsilon = -a/2 > 0$, then there exists z^*, such that

$$\left|f_2(z)\right| = \left|f_2(z) - 0\right| < \varepsilon = -a/2 \text{ for all } z>z^* \tag{3.26}$$

Since $(z^*+1)>z^*$, from (3.26) it follows that:

$$a/2 = -(-a/2) < f_2(z^*+1) < -a/2 \text{ for all } z>z^* \tag{3.27}$$

Then:

$f_2(z^*+1)$

$> a/2$ (according to (3.27))

$>a$ (since a is negative according to the assumption (3.24))

$= \sup_R(f_2)$ (according to the assumption (3.24))

$\geq f_2(z^*+1)$ (the supremum of the function is not smaller than an arbitrary value of the function)

It follows that $f_2(z^*+1) > f_2(z^*+1)$, which is a contradiction. Hence, the assumption (3.24) is not correct. Since according to (3.22), the supremum F_2 of the function $f_2(.)$ always exists, then F_2 is a non-negative number a:

$$F_2\left(CDF_1, CDF_2\right) = \sup_R\left(CDF_2 - CDF_1\right) = \sup_R\left(f_2\right) = a \geq 0 \tag{3.28}$$

3.5) Generalization for the supremum F_2

Taking into account (3.22), (3.23) and (3.28) we can say that (3.8) is correct:

$$\text{there exists } F_2 \text{ such that } F_2\left(CDF_1, CDF_2\right) = \sup_R\left(CDF_2 - CDF_1\right) \in [0;1]$$

which we had to prove.

4) Proof of statement (3.9)

4.1) Bound of the function f(.,.) above

Let z_1 be an arbitrary real value in the interval $(-\infty;+\infty)$. To find the upper bound of the real function $f(.,.)$ defined in (3.12), we have to analyse three cases.

Case 1: $z_2 \in (z_1;+\infty)$

$f(z_1, z_2)$

$= CDF_1(z_1) - CDF_2(z_1) + CDF_2(z_2) - CDF_1(z_2)$ (according to (3.12))

$\leq CDF_1(z_1) - CDF_2(z_1) + 1 - CDF_1(z_2)$ (since $CDF_2(z_2) \leq 1$, according to (3.4) and (3.6))

$\leq CDF_1(z_1) - CDF_2(z_1) + 1 - CDF_1(z_1)$ (since $CDF_1(z_2) \geq CDF_1(z_1)$, according to (3.1) and $z_2 > z_1$)

$\leq -0 + 1$ (since $CDF_2(z_1) \geq 0$, according to (3.4) and (3.5))

≤ 1

It follows that

$$f(z_1,z_2)=CDF_1(z_1)-CDF_2(z_1)+CDF_2(z_2)-CDF_1(z_2)\le 1,$$
$$\text{for } z_1\in(-\infty;+\infty) \text{ and } z_2\in(z_1;+\infty) \tag{3.29}$$

Case 2: $z_2\in(-\infty;z_1)$

$f(z_1,z_2)$

$=CDF_1(z_1)-CDF_2(z_1)+CDF_2(z_2)-CDF_1(z_2)$ (according to (3.12))

$=CDF_1(z_1)-CDF_2(z_1)+CDF_2(z_2)-0$ (since $CDF_1(z_2)\ge 0$, according to (3.1) and (3.3))

$=CDF_1(z_1)-CDF_2(z_1)+CDF_2(z_1)$ (since $CDF_2(z_1)\ge CDF_2(z_2)$, according to (3.4) and $z_2<z_1$)

≤ 1 (since $CDF_1(z_1)\le 1$, according to (3.1) and (3.3))

It follows that

$$f(z_1,z_2)=CDF_1(z_1)-CDF_2(z_1)+CDF_2(z_2)-CDF_1(z_2)\le 1,$$
$$\text{for } z_1\in(-\infty;+\infty) \text{ and } z_2\in(-\infty;z_1) \tag{3.30}$$

Case 3: $z_2=z_1$

$f(z_1,z_2)$

$=f(z_1,z_1)$ (since $z_2=z_1$)

$=CDF_1(z_1)-CDF_2(z_1)+CDF_2(z_1)-CDF_1(z_1)$ (according to (3.12))

$=0\le 1$

It follows that:

$$f(z_1,z_2)=CDF_1(z_1)-CDF_2(z_1)+CDF_2(z_2)-CDF_1(z_2)=0\le 1,$$
$$\text{for } z_1\in(-\infty;+\infty) \text{ and } z_2=z_1 \tag{3.31}$$

If we combine the results (3.29), (3.30) and (3.31) from the three cases, it shows that 1 is the upper bound of the function $f(.,.)$:

$$f\left(z_1, z_2\right) = CDF_1\left(z_1\right) - CDF_2\left(z_1\right) + CDF_2\left(z_2\right) - CDF_1\left(z_2\right) \le 1, \text{ for} \left(z_1,\ z_2\right) \in R^2 \qquad (3.32)$$

4.2) Existence of the functional F

$$F\left(CDF_1, CDF_2\right)$$

$$= \sup_R\left(CDF_1 - CDF_2\right) + \sup_R\left(CDF_2 - CDF_1\right) \text{ (according to the definition (3.9) of } F)$$

$$= F_1\left(CDF_1, CDF_2\right) + F_2\left(CDF_1, CDF_2\right) \qquad \text{(according to the proven (3.7) and (3.8))}$$

Then

$$F\left(CDF_1, CDF_2\right) = F_1\left(CDF_1, CDF_2\right) + F_2\left(CDF_1, CDF_2\right) \qquad (3.33)$$

According to the proven (3.7) and (3.8), the suprema F_1 and F_2 exist for arbitrary functions $CDF_1(.)$ and $CDF_2(.)$ and correspond to conditions (3.1) - (3.6). Since according to (3.33), the functional $F(CDF_1,\ CDF_2)$ is a sum of suprema F_1 and F_2, then the functional F exists under the same conditions:

$$\text{there exists } F \text{ such that } F\left(CDF_1, CDF_2\right) =$$

$$= \left(\sup_R\left(CDF_1 - CDF_2\right) + \sup_R\left(CDF_2 - CDF_1\right) \right) = \qquad (3.34)$$

$$= F_1\left(CDF_1, CDF_2\right) + F_2\left(CDF_1, CDF_2\right)$$

4.3) Connection between the functional F and the function f(.,.)

The supremum of the real numerical function $f(.,.)$ (if it exists) may be replaced with the supremum of a real numerical set according to the definition for supremum of a function [Apostol, 1981]:

$$\sup_{R^2}\left(f\right) = \sup\left\{ f\left(z_1, z_2\right) \middle| \left(z_1, z_2\right) \in R^2 \right\} \qquad (3.35)$$

The set $\left\{ f\left(z_1,z_2\right) \middle| \left(z_1,z_2\right) \in R^2 \right\}$ is a non-empty subset of the set of real numbers, which is bounded from above according to (3.32). According to the principle for continuity of real numbers, such sets always have a supremum, which is also a real number [Royden, Fitzpatrick, 2010]. Then, taking into account (3.12):

$$\begin{aligned} &\text{there exists } F^* \text{ such that } F^* = \sup_{R^2}\left(f\right) = sup\left\{ f\left(z_1,z_2\right) \middle| \left(z_1,z_2\right) \in R^2 \right\} \\ &= sup\left\{ f_1\left(z_1\right) + f_2\left(z_2\right) \middle| \left(z_1,z_2\right) \in R^2 \right\} = \sup_{R^2}\left(f_1 + f_2\right) \end{aligned} \tag{3.36}$$

According to the proven statement (3.7), the functional $F_1\left(CDF_1,CDF_2\right)$ is the supremum of the real numerical function $f_1(.)$, defined in (3.10). According to the definition for supremum of a function [Apostol, 1981], for each $\varepsilon_1 > 0$ the following dependencies apply:

$$\text{for every } z_1 \in R,\ F_1\left(CDF_1,CDF_2\right) \geq f_1\left(z_1\right) \tag{3.37}$$

$$\text{there exists } z_1^* \in R \text{ such that } F_1\left(CDF_1,CDF_2\right) - \varepsilon_1 < f_1\left(z_1^*\right) \tag{3.38}$$

According to the proven statement (3.8) the functional $F_2\left(CDF_1,CDF_2\right)$ is the supremum of the real numerical function $f_2(.)$ defined in (3.11). According to the definition for supremum of the function [Apostol, 1981], for each $\varepsilon_2 > 0$ the following dependencies apply:

$$\text{for every } z_2 \in R,\ F_2\left(CDF_1,CDF_2\right) \geq f_2\left(z_2\right) \tag{3.39}$$

$$\text{there exists } z_2^* \in R \text{ such that } F_2\left(CDF_1,CDF_2\right) - \varepsilon_2 < f_2\left(z_2^*\right) \tag{3.40}$$

After adding the inequalities (3.37) and (3.39) it follows that the sum of the functionals (3.10) and (3.11) is an upper bound of the real numerical function $f(.,.)$ defined in (3.12):

$$\begin{aligned} &\text{for every } \left(z_1,z_2\right) \in R^2, \\ &F_1\left(CDF_1,CDF_2\right) + F_2\left(CDF_1,CDF_2\right) \geq f_1\left(z_1\right) + f_2\left(z_2\right) = f\left(z_1,z_2\right) \end{aligned} \tag{3.41}$$

After adding the inequalities (3.38) and (3.40) and substituting $\varepsilon = (\varepsilon_1 + \varepsilon_2) > 0$ it follows that no real number, smaller than the sum of the functionals (3.10) and (3.11) is the upper bound of the real numerical function *f*(.,.) defined in (3.12):

$$\begin{aligned} &\text{there exists } (z_1^*, z_2^*) \in R^2 \text{ such that} \\ &F_1(CDF_1, CDF_2) + F_2(CDF_1, CDF_2) - \varepsilon < f_1(z_1^*) + f_2(z_2^*) = f(z_1^*, z_2^*) \end{aligned} \tag{3.42}$$

From (3.41) and (3.42) it follows that the sum of the functionals (3.10) and (3.11) is the smallest upper bound of the function *f*(.,.), which is the supremum of *f*(.,.) according to the definition for supremum of a function [Apostol, 1981]:

$$\sup_{R^2}(f) = \sup_{R^2}(f_1 + f_2) = F_1(CDF_1, CDF_2) + F_2(CDF_1, CDF_2) \tag{3.43}$$

From the comparison of (3.34), (3.36) and (3.43) it follows that the functional *F* and *F**coincide and are equal to the supremum of real numerical function *f*(.,.) defined in (3.12):

$$F(CDF_1, CDF_2) = F^*(CDF_1, CDF_2) = F_1(CDF_1, CDF_2) + F_2(CDF_1, CDF_2) = \sup_{R^2}(f) \tag{3.44}$$

4.4) Upper bound of the supremum F

According to (3.32), the upper bound of the function *f*(.) is 1. According to (3.34) and (3.44), the supremum *F* of the function *f*(.) always exists. According to the definition for supremum of a numerical function, the supremum is the smallest upper bound of the function [Apostol, 1981]. Hence

$$F(CDF_1, CDF_2) = \left(\sup_R (CDF_1 - CDF_2) + \sup_R (CDF_2 - CDF_1) \right) \le 1 \tag{3.45}$$

4.5) Lower bound of the supremum F

$$\begin{aligned} &F(CDF_1, CDF_2) \\ &= \sup_{R^2}(f) \qquad \text{(according to (3.44))} \end{aligned}$$

$\geq f(z_1, z_1)$ (the supremum is not less than any value of the function)

$= CDF_1(z_1) - CDF_2(z_1) + CDF_2(z_1) - CDF_1(z_1)$ (according to (3.12))

=0

It follows that

$$F(CDF_1, CDF_2) = \left(\sup_R (CDF_1 - CDF_2) + \sup_R (CDF_2 - CDF_1) \right) \geq 0 \tag{3.46}$$

4.6) Generalization for the supremum F

Taking into account (3.34), (3.45) and (3.46) we can say that (3.9) is correct:

there exists F, such that $F(CDF_1, CDF_2) = \left(\sup_R (CDF_1 - CDF_2) + \sup_R (CDF_2 - CDF_1) \right) \in [0;1]$

which we needed to prove.

4. Lemma for the discrete maximum

Setup of the lemma for the discrete maximum

Let Z be a sample of n observations from a given population of a one-dimensional random variable, and the degree of membership to that population of the i-th observation is $\mu_i \in (0;1]$:

$$Z = \{(z_1, \mu_1), (z_2, \mu_2), ..., (z_n, \mu_n)\} \tag{4.1}$$

Let the distribution of the random variable be approximated on the sample data according to a fuzzy empirical distribution function (FECDF) as follows:

$$FECDF(z) = \sum_{\substack{k=1 \\ z_k \leq z}}^{n} \mu_k \Big/ \sum_{k=1}^{n} \mu_k \text{ , for } z \in R \tag{4.2}$$

In (4.2), and throughout the section, the set of real numbers is denoted as R.

Let $F(z)$ be a real function of a numeric argument, defined for each real z with the following properties

- $F(.)$ is increasing:

If $z_1 > z_2$, then $F(z_1) \geq F(z_2)$ (4.3)

- When z approaches $-\infty$, the limit of $F(z)$ exists and is equal to zero:

$$\lim_{z \to -\infty} F(z) = 0 \tag{4.4}$$

- When z approaches $+\infty$, the limit of $F(z)$ exists and is equal to one:

$$\lim_{z \to +\infty} F(z) = 1 \tag{4.5}$$

Let $D(z)$ be a real function of a numeric argument, defined for each real z as a difference between $FECDF(.)$ and $F(.)$:

$$D(z) = FECDF(z) - F(z), \text{ for } z \in R \tag{4.6}$$

Then, in at least one of the points z_i (for i=1, 2, …, n) the function $D(.)$ has a global non-negative maximum M, not higher than 1, i.e.:

$$\text{there exists } j_{max} \in \{1,2,\ldots,n\}, \text{ such that } D(z) \leq D(z_{j_{max}}) = M \in [0;1], \text{ for } z \in R \tag{4.7}$$

Proof of the lemma for the discrete maximum

Let the observations from the sample Z from (4.1) and their degrees of membership be sorted in ascending order and renumbered in the sample Z^{sort}:

$$Z^{sort} = \left\{ \left(z_1^{sort}, \mu_1^{sort} \right), \left(z_2^{sort}, \mu_2^{sort} \right), ..., \left(z_n^{sort}, \mu_n^{sort} \right) \right\} \tag{4.8}$$

where $z_1^{sort} \le z_2^{sort} \le ... \le z_n^{sort}$. Then the fuzzy empirical cumulative distribution function constructed on (4.8) shall be:

$$FECDF^{sort}\left(z\right) = \sum_{\substack{k=1 \\ z_k^{sort} \le z}}^{n} \mu_k^{sort} \Bigg/ \sum_{k=1}^{n} \mu_k^{sort} \text{, for } z \in R \tag{4.9}$$

Let an arbitrary $\left(z_j - \mu_j \right)$ from (4.1) be renumbered as $\left(z_i^{sort}, \mu_i^{sort} \right)$ from (4.8). The denominators (4.9) and (4.2) are equal, because the sum of a finite number of addends are commutative. Since $z_i^{sort} = z_j$, then the condition $\left(z_j \le z \right)$ in the sum of the numerator of (4.2) is satisfied by $\left(z_j, \mu_j \right)$, if and only if the condition $\left(z_i^{sort} \le z \right)$ in the sum of the numerator of (4.9) is satisfied by $\left(z_i^{sort}, \mu_i^{sort} \right)$. Then, bearing in mind that $\mu_i^{sort} = \mu_j$, it follows that the numerators of (4.9) and (4.2) are equal. It follows that the values of the functions (4.2) and (4.9) coincide for the entire domain:

$$FECDF^{sort}\left(z\right) = FECDF\left(z\right), \text{ for } z \in R \tag{4.10}$$

It follows from here that the values of the function $D(.)$, derived from the non-sorted sample (4.1) shall coincide with the values of the same function, but derived from the sorted sample (4.8):

$$D^{sort}\left(z\right) = FECDF^{sort}\left(z\right) - F\left(z\right) = FECDF\left(z\right) - F\left(z\right) = D\left(z\right), \text{ for } z \in R \tag{4.11}$$

Let us analyse the finite set S of values of $D^{sort}(.)$ in the points z_i^{sort}:

$$S = \left\{ D^{sort}\left(z_1^{sort} \right), D^{sort}\left(z_2^{sort} \right), \ldots, D^{sort}\left(z_n^{sort} \right) \right\}, \tag{4.12}$$

where

$$D^{sort}\left(z_j^{sort}\right)=\sum_{\substack{k=1\\ z_k^{sort}\le z_j^{sort}}}^{n}\mu_k^{sort}\Bigg/\sum_{k=1}^{n}\mu_k^{sort}-F\left(z_j^{sort}\right),\ \text{for } j=1, 2, \ldots, n \tag{4.13}$$

The set S contains at least one or n at most different real numbers. Hence it always has a largest element M. Let i_{max} be the smallest running number of the element from (4.12), whose value is M:

$$\begin{cases} M=D^{sort}\left(z_{i_{max}}^{sort}\right)\ge D^{sort}\left(z_i^{sort}\right), for\ i=i_{max}+1, i_{max}+2,\ \ldots n \\ M=D^{sort}\left(z_{i_{max}}^{sort}\right)> D^{sort}\left(z_i^{sort}\right), for\ i=1,\ 2,\ ..,i_{max}-1 \end{cases} \tag{4.14}$$

Let us assume that there exists a real z^*, different from the points z_i^{sort} (i=1, 2, …, n) such that the value of $D^{sort}(.)$ is larger than M:

$$\text{there exists } z^*\in R\setminus\left\{z_1^{sort}, z_2^{sort},..., z_n^{sort}\right\}, \text{ such that } D^{sort}(z^*)>M \tag{4.15}$$

Case 1: $z^*\in\left(z_n^{sort};+\infty\right)$

Then

$M<D^{sort}(z^*)$	(according to the assumption (4.15))
$=FECDF^{sort}(z^*)-F(z^*)$	(according to (4.11))
$=FECDF^{sort}(z_n^{sort})-F(z^*)$	(because $FECDF^{sort}(z_n^{sort})=FECDF^{sort}(z^*)$ for $z^*>z_n^{sort}$ according to (4.9))
$\le FECDF^{sort}(z_n^{sort})-F(z_n^{sort})$	(because $F(z_n^{sort})\le F(z^*)$ for $z^*>z_n^{sort}$ according to (4.3))
$=D(z_n^{sort})$	(according to (4.11))
$\le D(z_{i_{max}}^{sort})=M$	(according to (4.14))

It follows that $M<M$, hence for Case 1, the assumption (4.15) is not true.

Case 2: $z^*\in\left(z_j^{sort};z_{j+1}^{sort}\right)$, for j=1, 2, …, n– 1

Case 2.1: $z_j^{sort} = z_{j+1}^{sort}$

Then the open interval $\left(z_j^{sort}; z_{j+1}^{sort}\right)$ is empty, and hence z^* does not exist. So, in Case 2.1, the assumption (4.15) is not true.

Case 2.2: $z_j^{sort} < z_{j+1}^{sort}$

Then

$M<D^{sort}(z^*)$	(according to the assumption (4.15))
$=FECDF^{sort}(z^*) - F(z^*)$	(according to (4.11))
$=FECDF^{sort}(z_j^{sort}) - F(z^*)$	(because $FECDF^{sort}(z_j^{sort}) = FECDF^{sort}(z^*)$ for $z_j^{sort} < z^* < z_{j+1}^{sort}$ according to (4.9))
$\le FECDF^{sort}(z_j^{sort}) - F(z_j^{sort})$	(because $F(z_j^{sort}) \le F(z^*)$ for $z_j^{sort} < z^*$ according to (4.3))
$=D(z_n^{sort})$	(according to (4.11))
$\le D(z_{i_{max}}^{sort}) = M$	(according to (4.14))

It follows that $M<M$, hence for Case 2.2, the assumption (4.15) is not true.

Case 3: $z^* \in \left(-\infty; z_1^{sort}\right)$

Then

$M<D^{sort}(z^*)$	(according to the assumption (4.15))
$=FECDF^{sort}(z^*) - F(z^*)$	(according to (4.11))
$=0 - F(z^*)$	(because $FECDF^{sort}(z^*) = 0$ according to (4.9), since the condition in the numerator $z_k^{sort} \le z^* < z_1^{sort}$ is never true as (4.8) is sorted))
$\le 0 - 0$	(because $F(z^*) \ge 0$ for any z^*, according to (4.3) and (4.4))
$=0=1 - 1$	
$\le 1 - F(z_n^{sort})$	(because $F(z_n^{sort}) \le 1$ for any z^*,

according to (4.3) and (4.5))

$=FECDF^{sort}(z_n^{sort}) - F(z_n^{sort})$ (because $FECDF^{sort}(z_n^{sort})=1$ according to (4.9), since the condition in the numerator $z_k^{sort} \le z_n^{sort}$ is true as (4.8) is sorted)

$=D(z_n^{sort})$ (according to (4.11))

$\le D(z_{i_{max}}^{sort})=M$ (according to (4.14))

It follows that $M<M$, hence for Case 3, the assumption (4.15) is not true.

From Cases 1, 2.1, 2.2 and 3 it follows that the assumption (4.15) is not true. Then

there does not exist $z^* \in R \setminus \left\{z_1^{sort}, z_2^{sort}, \ldots, z_n^{sort}\right\}$, such that $D^{sort}(z^*)>M$ (4.16)

Then

$$D^{sort}(z) \le D^{sort}\left(z_{i_{max}}^{sort}\right) = M, \text{ for } z \in (-\infty; +\infty) \quad (4.17)$$

The lower bound of the maximum M is:

$M= D\left(z_{i_{max}}^{sort}\right)$ (according to (4.17))

$\ge D\left(z_n^{sort}\right)$ (according to (4.14))

$=FECDF^{sort}(z_n^{sort}) - F(z_n^{sort})$ (according to (4.11))

$=1- F(z_n^{sort})$ (because $FECDF^{sort}(z_n^{sort})=1$ according to (4.9))

$\ge 1-1$ (because $F(z_n^{sort}) \le 1$ according to (4.3) and (4.5))

$=0$

It follows that

$$M \geq 0 \tag{4.18}$$

The upper bound of the maximum M is

$M = D\left(z_{i_{max}}^{sort}\right)$ (according to (4.17))

$= FECDF^{sort}(z_{i_{max}}^{sort}) - F(z_{i_{max}}^{sort})$ (according to (4.11))

$\leq 1 - F(z_{i_{max}}^{sort})$ (because $FECDF^{sort}(z_{i_{max}}^{sort}) \leq 1$ according to (4.9))

$\leq 1 - 0$ (because $F(z_{i_{max}}^{sort}) \geq 0$ according to (4.3) and (4.4))

$= 1$

It follows that

$$M \leq 1 \tag{4.19}$$

Let j_{max} be the number of the observations in the non-sorted sample (4.1), which corresponds to the number of observation i_{max} in the sorted sample (4.8). Then according to (4.17), (4.18) and (4.19)

$$\text{there exists } j_{max} \in \{1,2,\ldots,n\}, \text{ such that } D^{sort}(z) \leq D^{sort}\left(z_{j_{max}}\right) = M \in [0;1], \text{ for } z \in R \tag{4.20}$$

But the functions $D^{sort}(.)$ and $D(.)$ coincide according to (4.11). Then, taking into account (4.20) it follows that

$$\text{there exists } j_{max} \in \{1,2,\ldots,n\}, \text{ such that } D(z) \leq D\left(z_{j_{max}}\right) = M \in [0;1], \text{ for } z \in R$$

which had to be proven.

5. Theorem for calculation of Kolmogorov-Smirnov and Kuiper criteria over fuzzy samples

Setup of the theorem for calculation of Kolmogorov-Smirnov and Kuiper criteria over fuzzy samples

Let Z^1 and Z^2 be two samples containing n_1 and n_2 observations from two populations of a one-dimensional random variable respectively. The degree of membership to the first population of the i-th observation from the first sample is $\mu_i^1 \in (0;1]$, and the degree of membership to the second population of the i-th observation from the second sample is $\mu_i^2 \in (0;1]$:

$$Z^1 = \left\{\left(z_1^1, \mu_1^1\right), \left(z_2^1, \mu_2^1\right), ..., \left(z_{n_1}^1, \mu_{n_1}^1\right)\right\} \tag{5.1}$$

$$Z^2 = \left\{\left(z_1^2, \mu_1^2\right), \left(z_2^2, \mu_2^2\right), ..., \left(z_{n_2}^2, \mu_{n_2}^2\right)\right\} \tag{5.2}$$

Let the distributions of the one-dimensional random variable in both populations be approximated on the sample data with fuzzy empirical distribution functions $CDF_1(.)$ and $CDF_2(.)$ as follows:

$$CDF_1(z) = \sum_{\substack{k=1 \\ z_k^1 \le z}}^{n_1} \mu_k^1 \Big/ \sum_{k=1}^{n_1} \mu_k^1 \text{ ,for z} \in R \tag{5.3}$$

$$CDF_2(z) = \sum_{\substack{k=1 \\ z_k^2 \le z}}^{n_2} \mu_k^2 \Big/ \sum_{k=1}^{n_2} \mu_k^2 \text{ , for z} \in R \tag{5.4}$$

In (5.4), and throughout the section, the set of real numbers is denoted as R.

Let the Kolmogorov-Smirnov criterion and the Kuiper criterion for identity of the distributions of both populations be defined as:

$$KS = \sup_R \left(\left|CDF_1 - CDF_2\right|\right) \tag{5.5}$$

$$Ku = \sup_R \left(CDF_1 - CDF_2\right) + \sup_R \left(CDF_2 - CDF_1\right) \tag{5.6}$$

Then *KS* and *Ku* always exist and belong to the interval [0; 1], and the suprema in (5.5) and (5.6) are maximums. Each of those may be identified by calculating the criterion in not more than $(n_1+ n_2)$ points:

$$KS = \max\left\{\begin{array}{l}\max\limits_{i=1,2,\ldots,n_1}\left(\sum\limits_{\substack{k=1\\ z_k^1\le z_i^1}}^{n_1}\mu_k^1\Big/\sum_{k=1}^{n_1}\mu_k^1-\sum\limits_{\substack{k=1\\ z_k^2\le z_i^1}}^{n_2}\mu_k^2\Big/\sum_{k=1}^{n_2}\mu_k^2\right),\\ \max\limits_{i=1,2,\ldots,n_2}\left(\sum\limits_{\substack{k=1\\ z_k^2\le z_i^2}}^{n_2}\mu_k^2\Big/\sum_{k=1}^{n_2}\mu_k^2-\sum\limits_{\substack{k=1\\ z_k^1\le z_i^2}}^{n_1}\mu_k^1\Big/\sum_{k=1}^{n_1}\mu_k^1\right)\end{array}\right\}\in[0;1] \tag{5.7}$$

$$Ku = \max_{i=1,2,\ldots,n_1}\left(\sum\limits_{\substack{k=1\\ z_k^1\le z_i^1}}^{n_1}\mu_k^1\Big/\sum_{k=1}^{n_1}\mu_k^1-\sum\limits_{\substack{k=1\\ z_k^2\le z_i^1}}^{n_2}\mu_k^2\Big/\sum_{k=1}^{n_2}\mu_k^2\right)+ \tag{5.8}$$

$$+\max_{i=1,2,\ldots,n_2}\left(\sum\limits_{\substack{k=1\\ z_k^2\le z_i^2}}^{n_2}\mu_k^2\Big/\sum_{k=1}^{n_2}\mu_k^2-\sum\limits_{\substack{k=1\\ z_k^1\le z_i^2}}^{n_1}\mu_k^1\Big/\sum_{k=1}^{n_1}\mu_k^1\right)\in[0;1]$$

Proof of the theorem for calculation of Kolmogorov-Smirnov and Kuiper criteria over fuzzy samples

1) Properties of the functions $CDF_1(.)$ and $CDF_2(.)$, defined in (5.3) and (5.4).

Let $z_1>z_2$ where z_1 and z_2 are two real numbers and $CDF_1(.)$ are calculated at both z_1 and z_2:

$$CDF_1(z_1)=\sum\limits_{\substack{k=1\\ z_k^1\le z_1}}^{n_1}\mu_k^1\Big/\sum_{k=1}^{n_1}\mu_k^1\,, \tag{5.9}$$

$$CDF_1(z_2)=\sum\limits_{\substack{k=1\\ z_k^1\le z_2}}^{n_1}\mu_k^1\Big/\sum_{k=1}^{n_1}\mu_k^1 \tag{5.10}$$

Then for arbitrary $\left(z_k^1,\mu_k^1\right)$ from (5.1) the conditions in the numerators of (5.9) and (5.10) are either satisfied when $z_k^1\le z_2<z_1$, or are not satisfied when $z_2<z_1<z_k^1$, or only condition (5.9) is satisfied, when $z_2<z_k^1\le z_1$. It then follows that μ_k^1 is either part of the

numerators of (5.9) and (5.10), or is not part of the numerators of (5.9) and (5.10), or is only part of the numerator of (5.9). Taking into account the denominators of (5.9) and (5.10) it follows that the function $CDF_1(.)$ is increasing:

$$\text{If } z_1 > z_2, \text{ then } CDF_1(z_1) \geq CDF_1(z_2) \tag{5.11}$$

On the other hand, for an arbitrary $\left(z_k^1, \mu_k^1\right)$ from (5.1) the condition in the numerator of (5.10) would not be satisfied when $z_2 < min\left\{z_1^1, z_2^1, \ldots, z_{n_1}^1\right\}$. Then the numerator of (5.10) will be 0, hence (5.10) will be 0 and thus the limit of $CDF_1(.)$ in $-\infty$ will exist and shall be 0:

$$\lim_{z_2 \to -\infty} CDF_1\left(z_2\right) = 0 \tag{5.12}$$

At the same time, for an arbitrary $\left(z_k^1, \mu_k^1\right)$ from (5.1) the condition in the numerators of (5.10) shall be satisfied when $z_1 \geq max\left\{z_1^1, z_2^1, \ldots, z_{n_1}^1\right\}$. Then the numerator of (5.10) coincides with the denominator and (5.10) equals to 1 and thus the limit of $CDF_1(.)$ in $+\infty$ exists and equals to 1:

$$\lim_{z_1 \to +\infty} CDF_1\left(z_1\right) = 1 \tag{5.13}$$

In the same fashion, let the values of the function $CDF_2(.)$ be calculated for the real numbers $z_1 > z_2$:

$$CDF_2\left(z_1\right) = \sum_{\substack{k=1 \\ z_k^2 \leq z_1}}^{n_2} \mu_k^2 \Bigg/ \sum_{k=1}^{n_2} \mu_k^2 , \tag{5.14}$$

$$CDF_2\left(z_2\right) = \sum_{\substack{k=1 \\ z_k^2 \leq z_2}}^{n_2} \mu_k^2 \Bigg/ \sum_{k=1}^{n_2} \mu_k^2 \tag{5.15}$$

Then for an arbitrary $\left(z_k^2, \mu_k^2\right)$ from (5.2) the conditions in the numerators of (5.14) and (5.15) are either satisfied, when $z_k^2 \leq z_2 < z_1$; or not satisfied with $z_2 < z_1 < z_k^2$; or only the condition in (5.14) is satisfied when $z_2 < z_k^2 \leq z_1$. It follows that μ_k^2 is either part

of the numerators of both (5.14) and (5.15); or is not part of any of the numerators of (5.14) and (5.15); or is only part of the numerator of (5.14). Taking into account the equal denominators of (5.14) and (5.15) it follows that the function $CDF_2(.)$ is increasing:

$$\text{If } z_1 > z_2, \text{ then } CDF_2(z_1) \geq CDF_2(z_2) \tag{5.16}$$

On the other hand, for an arbitrary $\left(z_k^2, \mu_k^2\right)$ from (5.2) the condition in the numerator of (5.15) would not be satisfied when $z_2 < min\left\{z_1^2, z_2^2, \ldots, z_{n_1}^2\right\}$. Then the numerator of (5.15) will be 0, hence (5.15) will be 0 and then the limit of $CDF_2(.)$ in $-\infty$ will exist and shall be 0:

$$\lim_{z_2 \to -\infty} CDF_2\left(z_2\right) = 0 \tag{5.17}$$

At the same time, for an arbitrary $\left(z_k^2, \mu_k^2\right)$ from (5.2) the conditions in the numerator of (5.14) shall be satisfied when $z_1 \geq max\left\{z_1^2, z_2^2, \ldots, z_{n_1}^2\right\}$. Then the numerator of (5.14) coincides with the denominator and (5.14) equals to 1 and the limit of $CDF_2(.)$ in $+\infty$ exists and equals to 1:

$$\lim_{z_1 \to +\infty} CDF_2\left(z_1\right) = 1 \tag{5.18}$$

2) Defining auxiliary functions

Let $D_1(.)$ be a real function of a numeric argument, defined for each real z as the difference of $CDF_1(.)$ from (5.3) and $CDF_2(.)$ from (5.4):

$$D_1\left(z\right) = CDF_1\left(z\right) - CDF_2\left(z\right), \text{ for } z \in R \tag{5.19}$$

Let $D_2(.)$ be a real function of a numeric argument, defined for each real z as the difference of $CDF_2(.)$ from (5.4) and $CDF_1(.)$ from (5.3):

$$D_2\left(z\right) = CDF_2\left(z\right) - CDF_1\left(z\right) = -D_1\left(z\right), \text{ for } z \in R \tag{5.20}$$

The lemma for the bounded suprema holds: the properties (5.11), (5.12) and (5.13) for $CDF_1(.)$ from (5.3) are the conditions (3.1), (3.2) and (3.3) that the lemma for the bounded suprema requires; properties (5.16), (5.17) and (5.18) for $CDF_2(.)$ from (5.4) are the conditions (3.4), (3.5) and (3.6) from the lemma for the bounded suprema. Then for an arbitrary function $CDF_1(.)$ and $CDF_2(.)$, corresponding to condition (3.1)-(3.6) of the lemma for the bounded suprema, it can be proven that the suprema of $D_1(.)$ and $D_2(.)$, as well as their sum exist (according to the lemma for the bounded suprema):

$$\text{there exists } F_1 \text{ such that } F_1\left(CDF_1, CDF_2\right) = \sup_R\left(CDF_1 - CDF_2\right) = \sup_R\left(D_1\right) \in [0;1] \tag{5.21}$$

$$\text{there exists } F_2 \text{ such that } F_2\left(CDF_1, CDF_2\right) = \sup_R\left(CDF_2 - CDF_1\right) = \sup_R\left(D_2\right) \in [0;1] \tag{5.22}$$

there exists F such that

$$F\left(CDF_1, CDF_2\right) = \left(\sup_R\left(CDF_1 - CDF_2\right) + \sup_R\left(CDF_2 - CDF_1\right)\right) \tag{5.23}$$

$$= \left(\sup_R\left(D_1\right) + \sup_R\left(D_2\right)\right) \in [0;1]$$

The auxiliary function $D_1(.)$ from (5.19) complies with the lemma for the discrete maximum, since $CDF_1(.)$ from (5.3) is the very same fuzzy empirical distribution function, constructed over fuzzy sample data from Z^1 defined in (5.1), defined in the lemma for the discrete maximum by the function (4.2), constructed over fuzzy sample data (4.1). The properties (5.16), (5.17) and (5.18) for $CDF_2(.)$ from (5.4) are the properties (4.3), (4.4) and (4.5) of the lemma for the discrete maximum. Therefore the maximum of $D_1(.)$ exists and coincides with the value of the function in at least one of the points $z_1^1, z_2^1, \ldots, z_{n_1}^1$ of the fuzzy sample Z^1 defined in (5.1):

$$\text{there exists } j_{max} \in \{1,2,\ldots,n_1\}, \text{ such that } D_1\left(z\right) \le D_1\left(z_{j_{max}}^1\right) = M_1 \in [0;1], \text{ for } z \in R \tag{5.24}$$

Since the supremum of an arbitrary function, if it exists, is unique whereas the maximum of each function (if the function exists) is the supremum, then from (5.21) and (5.24) it follows that the supremum F_1 of $D_1(.)$ is always its maximum M_1. Then:

$$\text{there exists } F_1 \text{ such that } F_1 = \sup_R\left(D_1\right) = M_1 = \max_{i=1,2,\ldots,n_1}\left(D_1\left(z_i^1\right)\right) \in [0;1] \tag{5.25}$$

Taking into account (5.19), (5.14) and (5.15), from (5.25) it follows:

$$\text{there exists } F_1 \text{ such that } F_1 = \sup_R\left(D_1\right) = \max_{i=1,2,\ldots,n_1}\left(CDF_1\left(z_i^1\right) - CDF_2\left(z_i^1\right)\right)$$
$$= \max_{i=1,2,\ldots,n_1}\left(\sum_{\substack{k=1 \\ z_k^1 \le z_i^1}}^{n_1} \mu_k^1 \Big/ \sum_{k=1}^{n_1} \mu_k^1 - \sum_{\substack{k=1 \\ z_k^2 \le z_i^1}}^{n_2} \mu_k^2 \Big/ \sum_{k=1}^{n_2} \mu_k^2\right) \tag{5.26}$$

The auxiliary function $D_2(.)$ from (5.20) complies with the lemma for the discrete maximum, since $CDF_2(.)$ from (5.4) is the very same fuzzy empirical distribution function, constructed over fuzzy sample data from Z^2 defined in (5.2), defined in the lemma for the discrete maximum by the function (4.2), constructed over fuzzy sample data (4.1). The properties (5.11), (5.12) and (5.13) for $CDF_1(.)$ from (5.3) are the properties (4.3), (4.4) and (4.5) of the lemma for the discrete maximum2. Therefore the maximum of $D_2(.)$ exists and coincides with the value of the function in at least one of the points $z_1^2, z_2^2, \ldots, z_{n_2}^2$ of the fuzzy sample Z^2 defined in (5.2):

$$\text{there exists } j_{max} \in \left\{1,2,\ldots,n_2\right\}, \text{ such that } D_2\left(z\right) \le D_2\left(z_{j_{max}}^2\right) = M_2 \in \left[0;1\right], \text{ for } z \in R \tag{5.27}$$

Since the supremum of an arbitrary function, if it exists, is unique whereas the maximum of each function (if the function exists) is the supremum, then from (5.22) and (5.27) is follows that the supremum F_2 of $D_2(.)$ is always its maximum M_2. Then:

$$\text{there exists } F_2 \text{ such that } F_2 = \sup_R\left(D_2\right) = M_2 = \max_{i=1,2,\ldots,n_2}\left(D_2\left(z_i^2\right)\right) \in \left[0;1\right] \tag{5.28}$$

Taking into account (5.19), (5.20) and (5.15) and (5.28), from (5.6) it follows:

$$\text{there exists } F_2 \text{ such that } F_2 = \sup_R\left(D_2\right) = \max_{i=1,2,\ldots,n_2}\left(CDF_2\left(z_i^2\right) - CDF_1\left(z_i^2\right)\right)$$
$$= \max_{i=1,2,\ldots,n_2}\left(\sum_{\substack{k=1 \\ z_k^2 \le z_i^2}}^{n_2} \mu_k^2 \Big/ \sum_{k=1}^{n_2} \mu_k^2 - \sum_{\substack{k=1 \\ z_k^1 \le z_i^2}}^{n_1} \mu_k^1 \Big/ \sum_{k=1}^{n_1} \mu_k^1\right) \tag{5.29}$$

3) Proof of assumption (5.7)

The module in the Kolmogorov-Smirnov criterion (5.5) may be interpreted by the auxiliary functions $D_1(.)$ from (5.19):

$$\left|CDF_1(z)-CDF_2(z)\right|=\left|D_1(z)\right|=\begin{cases} D_1(z) & \text{for} \quad z\in\left\{R\middle|D_1(z)>0\right\} \\ 0 & \text{for} \quad z\in\left\{R\middle|D_1(z)=0\right\} \\ -D_1(z) & \text{for} \quad z\in\left\{R\middle|D_1(z)<0\right\} \end{cases} \tag{5.30}$$

Equation (5.30) may be better represented if we use the auxiliary functions $D_2(.)$ from (5.19)

$$\left|CDF_1(z)-CDF_2(z)\right|=\begin{cases} D_1(z) & \text{for} \quad z\in\left\{R\middle|D_1(z)\geq 0\right\} \\ D_2(z) & \text{for} \quad z\in\left\{R\middle|D_2(z)\geq 0\right\} \end{cases} \tag{5.31}$$

Equation (5.31) is not traditional because if for some z* both distribution functions are equal $CDF_1(z^*)=CDF_2(z^*)$, then $D_1(z^*)=0\geq 0$, and $D_2(z^*)=0\geq 0$. However, it is evident that even though the two conditions in (5.31) are true, the value of the function $\left|CDF_1(z^*)-CDF_2(z^*)\right|$ in z* is 0, regardless of the functional expressions, since $D_1(z^*)=D_2(z^*)=0$.

The set of real numbers R may be divided into two disjoint sets $R_1^{0+}=\left\{R\middle|D_1(z)\geq 0\right\}$ and $R_1^{-}=\left\{R\middle|D_1(z)<0\right\}$:

$$R=R_1^{0+}\cup R_1^{-}=\left\{R\middle|D_1(z)\geq 0\right\}\cup\left\{R\middle|D_1(z)<0\right\} \tag{5.32}$$

From (5.25) it follows that $\underset{i=1,2,\ldots,n_1}{max}\left(D_1\left(z_i^1\right)\right)\in[0;1]$. Then $D_1(z)\geq 0$ in at least one of the points $z_1^1,z_2^1,\ldots,z_{n_1}^1$ of the fuzzy sample Z^1 defined in (5.1). Hence the set R_1^{0+} is non-empty. For the set R_1^{-} we can analyse two cases.

Case 1: $CDF_1(z)\geq CDF_2(z)$ for $z\in R$

Here R_1^{-} is an empty set because $D_1(z)\geq 0$ for $z\in R$. Then:

$$R=R_1^{0+}\cup R_1^{-}=R_1^{0+}\cup\varnothing=R_1^{0+} \tag{5.33}$$

Evidently, the suprema of $D_1(.)$ in the sets R and R_1^{0+} coincide because according to (5.33) the sets coincide:

$$F_1 = \sup_{R}\left(D_1\right) = \sup_{R_1^{0+}}\left(D_1\right) \text{ when for every } z \in R, CDF_1\left(z\right) \geq CDF_2\left(z\right) \tag{5.34}$$

Case 2: there exists z^* such that $CDF_1\left(z^*\right) < CDF_2\left(z^*\right)$

Here R_1^- is not an empty set because it contains at least z^* since $D_1(z^*)<0$. Then:

$$D_1\left(z_1\right) \geq 0 > D_1\left(z_2\right) \quad \text{for every } z_1 \in R_1^{0+} \text{ and every } z_2 \in R_1^- \tag{5.35}$$

According to (5.25) the supremum F_1 of $D_1(.)$ in the set R always exists and coincides with the maximum M_1 of the function, which also always exists. Evidently, the suprema of $D_1(.)$ in the sets R and R_1^{0+} coincide, because according to (5.35) the maximum M_1 can never be in a point from R_1^-:

$$F_1 = \sup_{R}\left(D_1\right) = \sup_{R_1^{0+}}\left(D_1\right) \text{ when there exists } z^* \in R \text{ , where } CDF_1\left(z^*\right) \geq CDF_2\left(z^*\right) \tag{5.36}$$

Combining the results of (5.34) and (5.36) gives

$$F_1 = \sup_{R}\left(D_1\right) = \sup_{R_1^{0+}}\left(D_1\right) \tag{5.37}$$

In the same fashion as (5.32), the set of real numbers R may be divided into two disjoint sets $R_2^{0+} = \left\{R \middle| D_2\left(z\right) \geq 0\right\}$ and $R_2^- = \left\{R \middle| D_2\left(z\right) < 0\right\}$:

$$R = R_2^{0+} \cup R_2^- = \left\{R \middle| D_2\left(z\right) \geq 0\right\} \cup \left\{R \middle| D_2\left(z\right) < 0\right\} \tag{5.38}$$

From (5.28) it follows that $\max\limits_{i=1,2,\ldots,n_2}\left(D_2\left(z_i^2\right)\right)\in[0;1]$. Then $D_2\left(z\right)\geq 0$ in at least one of the points $z_1^2, z_2^2, \ldots, z_{n_2}^2$ of the fuzzy sample Z^2 defined in (5.2). Hence the set R_2^{0+} is non-empty. For the set R_2^- we can analyse two cases.

Case 1: $CDF_2\left(z\right)\geq CDF_1\left(z\right)$ for $z\in R$

Here R_2^- is an empty set because $D_2\left(z\right)\geq 0$ for $z\in R$. Then:

$$R = R_2^{0+}\cup R_2^- = R_2^{0+}\cup\varnothing = R_2^{0+} \tag{5.39}$$

Evidently, the suprema of $D_2(.)$ in the sets R and R_2^{0+} coincide because according to (5.39) the sets coincide:

$$\text{if } CDF_2\left(z\right)\geq CDF_1\left(z\right) \text{ for all } z\in R, \text{ then } F_2 = \sup_{R}\left(D_2\right) = \sup_{R_2^{0+}}\left(D_2\right) \tag{5.40}$$

Case 2: there exists z^* such that $CDF_2\left(z^*\right) < CDF_1\left(z^*\right)$

Here R_2^- is not an empty set because it contains at least z^* since $D_2(z^*)<0$. Then:

$$D_2\left(z_1\right)\geq 0 > D_2\left(z_2\right) \quad \text{for every } z_1\in R_2^{0+} \text{ and every } z_2\in R_2^- \tag{5.41}$$

According to (5.28) the supremum F_2 of $D_2(.)$ in the set R always exists and coincides with the maximum M_2 of the function, which also always exists. Evidently, the suprema of $D_2(.)$ in the sets R and R_2^{0+} coincide, because according to (5.41) the maximum M_2 can never be in a point from R_2^-:

$$\text{if } CDF_2\left(z^*\right)\geq CDF_1\left(z^*\right) \text{ for some } z^*\in R, \text{ then } F_2 = \sup_{R}\left(D_2\right) = \sup_{R_2^{0+}}\left(D_2\right) \tag{5.42}$$

Combining the results of (5.40) and (5.42) gives

$$F_2 = \sup_{R}\left(D_2\right) = \sup_{R_2^{0+}}\left(D_2\right) \tag{5.43}$$

Using the sets introduced in (5.32) and (5.38), from (5.5) and (5.31) it follows that if the Kolmogorov-Smirnov criterion exists, then it is:

$$KS = \sup_{R}\left(\left|CDF_1 - CDF_2\right|\right) = max\left\{\sup_{R_1^{0+}}\left(D_1\right), \sup_{R_2^{0+}}\left(D_2\right)\right\} \tag{5.44}$$

From (5.37) and (5.43) it follows that

$$KS = max\left\{\sup_{R_1^{0+}}\left(D_1\right), \sup_{R_2^{0+}}\left(D_2\right)\right\} = max\left\{\sup_{R}\left(D_1\right), \sup_{R}\left(D_2\right)\right\} = max\left\{F_1, F_2\right\} \tag{5.45}$$

According to (5.21) the supremum F_1 of $D_1(.)$ always exists, and according to (5.22) the supremum F_2 of $D_2(.)$ always exists. The set $\{F_1, F_2\}$ contains two real numbers and hence there always exists a maximum element, as it is partially ordered by the relation "smaller or equal" [Richmond, Richmond, 2009]. Hence the Kolmogorov-Smirnov criterion always exists according to (5.44) and (5.45):

$$\text{there exists } KS = \sup_{R}\left(\left|CDF_1 - CDF_2\right|\right) = max\left\{F_1, F_2\right\} \tag{5.46}$$

According to (5.21) and (5.22) the supremum F_1 of $D_1(.)$ and the supremum F_2 of $D_2(.)$ are always non-negative and not greater than 1. Therefore, the larger value of F_1 and F_2 is also non-negative and not greater than 1:

$$\text{there exists } KS = max\left\{F_1, F_2\right\} \in [0;1] \tag{5.47}$$

Taking into account (5.26) and (5.29), from (5.47) it follows:

there exists KS

$$= max\left\{\begin{array}{l} \underset{i=1,2,\ldots,n_1}{max}\left(\sum\limits_{\substack{k=1\\ z_k^1\le z_i^1}}^{n_1}\mu_k^1 \Big/ \sum\limits_{k=1}^{n_1}\mu_k^1 - \sum\limits_{\substack{k=1\\ z_k^2\le z_i^1}}^{n_2}\mu_k^2 \Big/ \sum\limits_{k=1}^{n_2}\mu_k^2\right), \\ \underset{i=1,2,\ldots,n_2}{max}\left(\sum\limits_{\substack{k=1\\ z_k^2\le z_i^2}}^{n_2}\mu_k^2 \Big/ \sum\limits_{k=1}^{n_2}\mu_k^2 - \sum\limits_{\substack{k=1\\ z_k^1\le z_i^2}}^{n_1}\mu_k^1 \Big/ \sum\limits_{k=1}^{n_1}\mu_k^1\right) \end{array}\right\} \in [0;1]$$

which had to be proven.

4) Proof of assumption (5.8)

Taking into account (5.23), (5.20) and (5.22), from (5.6) it follows that the Kuiper criterion always exists:

$$\begin{array}{l} \text{there exists } Ku = \underset{R}{sup}\left(CDF_1 - CDF_2\right) + \underset{R}{sup}\left(CDF_2 - CDF_1\right) = \\ = \underset{R}{sup}\left(D_1\right) + \underset{R}{sup}\left(D_2\right) = F_1 + F_2 \in [0;1] \end{array} \tag{5.48}$$

Taking into account (5.26) and (5.29) it follows:

$$\begin{array}{l} F_1 + F_2 \\ = \underset{i=1,2,\ldots,n_1}{max}\left(\sum\limits_{\substack{k=1\\ z_k^1\le z_i^1}}^{n_1}\mu_k^1 \Big/ \sum\limits_{k=1}^{n_1}\mu_k^1 - \sum\limits_{\substack{k=1\\ z_k^2\le z_i^1}}^{n_2}\mu_k^2 \Big/ \sum\limits_{k=1}^{n_2}\mu_k^2\right) + \\ + \underset{i=1,2,\ldots,n_2}{max}\left(\sum\limits_{\substack{k=1\\ z_k^2\le z_i^2}}^{n_2}\mu_k^2 \Big/ \sum\limits_{k=1}^{n_2}\mu_k^2 - \sum\limits_{\substack{k=1\\ z_k^1\le z_i^2}}^{n_1}\mu_k^1 \Big/ \sum\limits_{k=1}^{n_1}\mu_k^1\right) \in [0;1] \end{array} \tag{5.49}$$

Finally, from (5.48) and (5.49) it follows that

there exists Ku

$$= \max_{i=1,2,\ldots,n_1} \left(\sum_{\substack{k=1 \\ z_k^1 \le z_i^1}}^{n_1} \mu_k^1 \Big/ \sum_{k=1}^{n_1} \mu_k^1 - \sum_{\substack{k=1 \\ z_k^2 \le z_i^1}}^{n_2} \mu_k^2 \Big/ \sum_{k=1}^{n_2} \mu_k^2 \right) + \max_{i=1,2,\ldots,n_2} \left(\sum_{\substack{k=1 \\ z_k^2 \le z_i^2}}^{n_2} \mu_k^2 \Big/ \sum_{k=1}^{n_2} \mu_k^2 - \sum_{\substack{k=1 \\ z_k^1 \le z_i^2}}^{n_1} \mu_k^1 \Big/ \sum_{k=1}^{n_1} \mu_k^1 \right) \in [0;1]$$

which we had to prove.

6. Corollary for calculation of Kolmogorov-Smirnov and Kuiper criteria over rigid samples

Setup of the corollary for calculation of Kolmogorov-Smirnov and Kuiper criteria over rigid samples

Let Z^1 and Z^2 be two samples containing n_1 and n_2 observations from two populations of a one-dimensional random variable respectively:

$$Z^1 = \left\{ z_1^1, z_2^1, \ldots, z_{n_1}^1 \right\} \tag{6.1}$$

$$Z^2 = \left\{ z_1^2, z_2^2, \ldots, z_{n_2}^2 \right\} \tag{6.2}$$

Let the distributions of the one-dimensional random variable in both populations be approximated on the sample data with empirical distribution functions $CDF_1(.)$ and $CDF_2(.)$ as follows:

$$CDF_1(z) = \sum_{\substack{k=1 \\ z_k^1 \le z}}^{n_1} 1 \Big/ n_1, \quad \text{for } z \in R \tag{6.3}$$

$$CDF_2(z) = \sum_{\substack{k=1 \\ z_k^2 \le z}}^{n_2} 1 \Big/ n_2 \quad \text{, for } z \in R \tag{6.4}$$

In (6.4), and throughout the section, the set of real numbers is denoted as R.

Let the Kolmogorov-Smirnov and the Kuiper criteria for identity of distributions of both populations be defined as:

$$KS = \sup_R \left(\left| CDF_1 - CDF_2 \right| \right) \tag{6.5}$$

$$Ku = \sup_R \left(CDF_1 - CDF_2\right) + \sup_R \left(CDF_2 - CDF_1\right) \tag{6.6}$$

Then *KS* and *Ku* always exists and belong to the interval [0; 1], and the suprema in (6.5) and (6.6) are maximums. Each of those may be identified after calculating in not more than $(n_1 + n_2)$ points:

$$\text{there exists } KS = max\left\{\begin{array}{l} \max\limits_{i=1,2,\ldots,n_1}\left(\sum\limits_{\substack{k=1 \\ z_k^1 \le z_i^1}}^{n_1} 1 \Big/ n_1 - \sum\limits_{\substack{k=1 \\ z_k^2 \le z_i^1}}^{n_2} 1 \Big/ n_2\right), \\ \max\limits_{i=1,2,\ldots,n_2}\left(\sum\limits_{\substack{k=1 \\ z_k^2 \le z_i^2}}^{n_2} 1 \Big/ n_2 - \sum\limits_{\substack{k=1 \\ z_k^1 \le z_i^2}}^{n_1} 1 \Big/ n_1\right) \end{array}\right\} \in [0;1] \tag{6.7}$$

$$\begin{aligned} \text{there exists } Ku = & \max_{i=1,2,\ldots,n_1}\left(\sum_{\substack{k=1 \\ z_k^1 \le z_i^1}}^{n_1} 1 \Big/ n_1 - \sum_{\substack{k=1 \\ z_k^2 \le z_i^1}}^{n_2} 1 \Big/ n_2\right) + \\ & + \max_{i=1,2,\ldots,n_2}\left(\sum_{\substack{k=1 \\ z_k^2 \le z_i^2}}^{n_2} 1 \Big/ n_2 - \sum_{\substack{k=1 \\ z_k^1 \le z_i^2}}^{n_1} 1 \Big/ n_1\right) \in [0;1] \end{aligned} \tag{6.8}$$

Proof of the corollary for the calculation of Kolmogorov-Smirnov and Kuiper criteria over rigid samples

The observations in the rigid sample Z^1 in (6.1) from the first population of the one-dimensional random variable may be interpreted as fuzzy, but with a degree of membership to the first population $\mu_i^1 = 1$, for i=1,2,…, n_1 :

$$Z^1 = \left\{z_1^1, z_2^1, \ldots, z_{n_1}^1\right\} = \left\{\left(z_1^1, 1\right), \left(z_2^1, 1\right), \ldots, \left(z_{n_1}^1, 1\right)\right\} = \left\{\left(z_1^1, \mu_1^1\right), \left(z_2^1, \mu_2^1\right), \ldots, \left(z_{n_1}^1, \mu_{n_1}^1\right)\right\} \tag{6.9}$$

The observations in the rigid sample Z^2 in (6.2) from the second population of the one-dimensional random variable may be interpreted as fuzzy, but with a degree of membership to the first population $\mu_i^2 = 1$, for i=1,2,…,n_2 :

$$Z^2 = \left\{z_1^2, z_2^2, \ldots, z_{n_2}^2\right\} = \left\{\left(z_1^2, 1\right), \left(z_2^2, 1\right), \ldots, \left(z_{n_2}^2, 1\right)\right\} = \left\{\left(z_1^2, \mu_1^2\right), \left(z_2^2, \mu_2^2\right), \ldots, \left(z_{n_2}^2, \mu_{n_2}^2\right)\right\} \tag{6.10}$$

The empirical distribution function $CDF_1(.)$ from (6.3) may be interpreted as fuzzy empirical distribution function, but with a degree of membership to the first population $\mu_i^1 = 1$, for i=1,2,…, n_1 :

$$CDF_1(z) = \sum_{\substack{k=1 \\ z_k^1 \le z}}^{n_1} 1 \Big/ n_1 = \sum_{\substack{k=1 \\ z_k^1 \le z}}^{n_1} 1 \Big/ \sum_{k=1}^{n_1} 1 = \sum_{\substack{k=1 \\ z_k^1 \le z}}^{n_1} \mu_k^1 \Big/ \sum_{k=1}^{n_1} \mu_k^1 \text{ , for } z \in R \tag{6.11}$$

The empirical distribution function $CDF_2(.)$ from (6.4) may be interpreted as fuzzy empirical distribution function, but with a degree of membership to the first population $\mu_i^2 = 1$, for i=1,2,…, n_1 :

$$CDF_2(z) = \sum_{\substack{k=1 \\ z_k^2 \le z}}^{n_2} 1 \Big/ n_2 = \sum_{\substack{k=1 \\ z_k^2 \le z}}^{n_2} 1 \Big/ \sum_{k=1}^{n_2} 1 = \sum_{\substack{k=1 \\ z_k^2 \le z}}^{n_2} \mu_k^2 \Big/ \sum_{k=1}^{n_2} \mu_k^2 \text{ , for z} \in R \tag{6.12}$$

The theorem fr calculation of the Kolmogorov-Smirnov and Kuiper criteria over fuzzy samples holds according to (6.9), (6.10), (6.11) and (6.12). Then:

there exists KS

$$= max \left\{ \begin{array}{l} \max\limits_{i=1,2,\ldots,n_1} \left(\sum\limits_{\substack{k=1 \\ z_k^1 \le z_i^1}}^{n_1} \mu_k^1 \Big/ \sum\limits_{k=1}^{n_1} \mu_k^1 - \sum\limits_{\substack{k=1 \\ z_k^2 \le z_i^1}}^{n_2} \mu_k^2 \Big/ \sum\limits_{k=1}^{n_2} \mu_k^2 \right), \\ \max\limits_{i=1,2,\ldots,n_2} \left(\sum\limits_{\substack{k=1 \\ z_k^2 \le z_i^2}}^{n_2} \mu_k^2 \Big/ \sum\limits_{k=1}^{n_2} \mu_k^2 - \sum\limits_{\substack{k=1 \\ z_k^1 \le z_i^2}}^{n_1} \mu_k^1 \Big/ \sum\limits_{k=1}^{n_1} \mu_k^1 \right) \end{array} \right\} \in [0;1] \tag{6.13}$$

there exists Ku

$$= \max_{i=1,2,\ldots,n_1} \left(\sum_{\substack{k=1 \\ z_k^1 \le z_i^1}}^{n_1} \mu_k^1 \Big/ \sum_{k=1}^{n_1} \mu_k^1 - \sum_{\substack{k=1 \\ z_k^2 \le z_i^1}}^{n_2} \mu_k^2 \Big/ \sum_{k=1}^{n_2} \mu_k^2 \right) + \tag{6.14}$$

$$+ \max_{i=1,2,\ldots,n_2} \left(\sum_{\substack{k=1 \\ z_k^2 \le z_i^2}}^{n_2} \mu_k^2 \Big/ \sum_{k=1}^{n_2} \mu_k^2 - \sum_{\substack{k=1 \\ z_k^1 \le z_i^2}}^{n_1} \mu_k^1 \Big/ \sum_{k=1}^{n_1} \mu_k^1 \right) \in [0;1]$$

From (6.13) and (6.14), taking into account that $\mu_i^1 = 1$ for i=1,2,…, n_1 and that $\mu_i^2 = 1$, for i=1,2,…, n_2 it follows that:

$$\text{there exists } KS = max\left\{ \max_{i=1,2,\ldots,n_1} \left(\sum_{\substack{k=1 \\ z_k^1 \le z_i^1}}^{n_1} 1 \Big/ n_1 - \sum_{\substack{k=1 \\ z_k^2 \le z_i^1}}^{n_2} 1 \Big/ n_2 \right), \max_{i=1,2,\ldots,n_2} \left(\sum_{\substack{k=1 \\ z_k^2 \le z_i^2}}^{n_2} 1 \Big/ n_2 - \sum_{\substack{k=1 \\ z_k^1 \le z_i^2}}^{n_1} 1 \Big/ n_1 \right) \right\} \in [0;1]$$

$$\text{there exists } Ku = \max_{i=1,2,\ldots,n_1} \left(\sum_{\substack{k=1 \\ z_k^1 \le z_i^1}}^{n_1} 1 \Big/ n_1 - \sum_{\substack{k=1 \\ z_k^2 \le z_i^1}}^{n_2} 1 \Big/ n_2 \right) + \max_{i=1,2,\ldots,n_2} \left(\sum_{\substack{k=1 \\ z_k^2 \le z_i^2}}^{n_2} 1 \Big/ n_2 - \sum_{\substack{k=1 \\ z_k^1 \le z_i^2}}^{n_1} 1 \Big/ n_1 \right) \in [0;1]$$

which we had to prove.

7. Discussion

The paper uses Kolmogorov-Smirnov and Kuiper criteria to compared the difference between population distributions approximated over two fuzzy samples. The procedures to calculate those criteria prove that the suprema in the standard formulae of *KS* and *Ku* are maxima. Furthermore, it was demonstrated that these criteria are always between 0 and 1, and they may be calculated in a given number of data points.

The objective of the paper is quick and reliable calculation of the Kolmogorov-Smirnov (1.5) and Kuiper (1.6) criteria both in the case of rigid samples and fuzzy samples. In the case of fuzzy samples (1.1) and (1.2), the formulae for *KS* (1.7) and *Ku* (1.8) are proven in the case of FECDF (1.3) and (1.4). The calculation procedures are much faster than the one-dimensional continuous optimization problems, which the original formulae (1.5) and (1.6) suggest. On top, the offered procedure gives complete certainty of the calculation. The paper also presented how these formulae would transform to adapt to rigid samples. The rigid sample dependencies are very much intuitive, they are well known in literature, but lack formal definition [Press et al., 2007], which is another contribution of this paper.

Literature offers analytical techniques to construct the sample distribution functions. However, those techniques are only available for rigid samples, and are also asymptotic in nature. In the case of a single sample, they are able to test if that sample was derived from a predefined distribution. In all other cases it is necessary to use simulation modelling techniques, such as Bootstrap [Efron, Tibshirani, 1994], to construct the distribution of the test statistic. This is especially true for the case of fuzzy samples. Additionally, in the process of simulation modelling, the *KS* and *Ku* need to be

calculated in each pseudo reality. If the calculation procedures are slow, difficult and uncertain, then the whole simulation procedure to calculate p_{value} of the test is also compromised. The proposed procedures in this paper provide stability, speed and reliability of calculation, hence immensely contributing to an optimal simulation modelling procedure to conduct the tests.

References

Apostol, T., Mathematical Analysis. Second Edition. Addison-Wesley. p. 9, 1981.

Böhm, W., & Hornik, K. A Kolmogorov-Smirnov Test for r Samples. *Research Report Series / Department of Statistics and Mathematics*, **105**. WU Vienna University of Economics and Business, Vienna, 2010

Chernobai, A., Rachev, S. T. & Fabozzi, F. .Composite Goodness-of-Fit Tests for Left-Truncated Loss Samples, In Lee, C.-F. & Lee, J.-C. Handbook of Financial Econometrics and Statistics. pp. 575-596, 2014

Groebner, D.F., Shannon, P.W., Fly, Ph. C. & Smith, K.D. Business Statistics – A Decision-Making Approach. Eighth Edition, Prentice Hall, USA. pp. 770-788, 2011

Efron, B. & Tibshirani, R.J. An Introduction to the Bootstrap. CRC Press, 1994

Jin, X., Chow, T. W. S. Sun, Y., Shan, J. & Lau, B. C. P. Kuiper Test and Autoregressive Model-Based Approach for Wireless Sensor Network Fault Diagnosis. Wireless Networks **21**, pp. 829-839, 2015

Lemeshko, Yu. & Gorbunova, A. A. Application and Power of the Nonparametric Kuiper, Watson, and Zhang Tests of Goodness-of-Fit. Measurement Techniques **56(5)** (2013a) 465-475

Nikolova, N.D., Chai, S., Ivanova, S., Kolev, K. & Tenekedjiev, K., Bootstrap Kuiper Testing of the Identity of 1D Continuous Distributions using Fuzzy Samples. International Journal of Computational Intelligence Systems., **8(2)**, pp. 63-75, 2015

Press, W.H., Teukolski, S. A., Vetterling, W. T. & Flannery, B. P. Numerical Recipes – The Art of Scientific Computing. Third Edition. Cambridge University Press, 2007

Richmond, B. & Richmond, Th. A Discrete Transition to Advanced Mathematics, American Mathematical Society, 2009

Royden, H.L. & Fitzpatrick, P.M. Real Analysis. Fourth Edition. Pearson, p. 9, 2010

Rudin, W. Principles of Mathematical Analysis. Third Edition. Pp. 4, McGraw-Hill, 1976

Stewart, J. Calculus. 8th Edition. Metric Version. pp. 62, Cengage Learning, 2016.

Thomas, G.B., Weir, M.D. & Hass, J.R. Thomas' Calculus Early Transcendentals. 13th Edition. pp. 78, Pearson, 2014

Viertl, R. Statistical Methods for Fuzzy Data. John Wiley. UK, 2011