Statistica, anno LXII, n. 1, 2002

A TEST FOR COMPARING DISTRIBUTION FUNCTIONS WITH STRONGLY UNBALANCED SAMPLES

Piotr Kulczycki

1. INTRODUCTION

In statistical practice the problem of testing a hypothesis which states that two independent random variables have the same probability distribution occurs rather often. In typical applications, especially in the fields of economics and the life sciences, the sizes of the random samples obtained from populations with comparable distributions are similar, and this fact was the fundamental premise used in constructing the classical non-parametric tests (see *e.g.* the textbooks by Fisz, 1963, or Wilks, 1962). Yet with the expansion of computer technology in modern engineering, *e.g.* during work in the on-line regime in automatic control systems, the need has appeared for statistical inference regarding the equality of two distributions on the sole basis of one current value of the selected vector quantities, and thus in a case when one of the random samples is one-element. The present paper will be devoted to this issue. A new test is proposed based on the kernel estimators technique and the methodology of order statistics. Simulation investigations indicate that its properties are more advantageous than in the case of the classical tests familiar from the literature.

The paper is organized into 5 short parts. In section 2 the form of the proposed test will be presented. The next two sections, 3 and 4, contain descriptions of the kernel estimator of the density function and the order estimator of the quantile, which are used to determine the form of the statistic and the critical set. Final comments and the results of empirical verification of the test developed here are found in section 5.

2. THE TEST

Consider the independent n-dimensional random variables X and Y, assuming that the distribution of the first of them has a density function. Let the random samples $x_1, x_2, \ldots, x_m$ and y (therefore of sizes m and 1), obtained respectively from the above variables, be given. At significance level r, the hypothesis of the equality of the distribution functions of the variables X and Y will be verified.

In the proposed form of the test, the criterion for the above hypothesis is the statistic

$$S(x_1, x_2, \ldots, x_m, y) = \hat{f}_{x_1,\ldots,x_m}(y), \tag{1}$$

where $\hat{f}_{x_1,\ldots,x_m}$ denotes the kernel estimator of the density function of the random variable X generated on the basis of the sample $x_1, x_2, \ldots, x_m$. The falsehood of the hypothesis being verified is shown by small values of the statistic S, and so the critical interval will be accepted in the left-sided form:

$$A = (-\infty, a], \tag{2}$$

where the critical value a is accepted to be the quantile estimator of order r of the distribution of the statistic S, obtained from the sample $S(x_1)$, $S(x_2)$, ..., $S(x_m)$ (for the definition of the quantile, see *e.g.* Fisz (1963); in practice this means such a real value q that the probabilities of the intervals $(-\infty, q]$ and $[q, \infty)$ amount to r and $1 - r$, respectively).

The next two sections will briefly present the procedures for constructing the kernel estimator of the density function and the order estimator of the quantile, which completes the test proposed above.

3. KERNEL ESTIMATOR OF DENSITY FUNCTION

The kernel estimator of the density function of the n-dimensional random variable X, calculated on the basis of its m realizations $x_1, x_2, \ldots, x_m$, is defined by the dependence

$$\hat{f}(x) = \frac{1}{mh^n} \det(D) \sum_{i=1}^{m} K\left(D\frac{x - x_i}{h}\right), \tag{3}$$

where the measurable and symmetrical function $K : \mathbb{R}^n \to [0, \infty)$ with a unique integral and a weak global maximum in point 0 is called the kernel, the positive constant h is known as the smoothing parameter, and D means the $n \times n$-dimensional, diagonal matrix of the inverses of the variable X's standard deviations

$$D = [d_{j,k}] = \begin{cases} \left(\left(\frac{1}{m}\sum_{i=1}^{m}(x_i^j)^2\right) - \left(\frac{1}{m}\sum_{i=1}^{m} x_i^j\right)^2\right)^{-1/2} & \text{if } j = k, \\ 0 & \text{if } j \neq k \end{cases} \tag{4}$$

while x_i^j denotes the j-th coordinate of the vector x_i. Detailed information concerning the rules for choosing the function K and fixing the value of the parameter h is found *e.g.* in Silverman (1986). Most often used in practice is the exponential kernel

$$K(x) = (2\pi)^{-n/2} \exp\left(-\frac{\|x\|^2}{2}\right). \tag{5}$$

If the density function being estimated is of the class C^2, and moreover both this function and its second derivative are bounded, then the value of the smoothing parameter is most often determined using the mean square criterion. The approximate value can then be calculated by assuming the normal distribution of the random variable X; in this case, for the exponential kernel (5), one obtains

$$h = \left(\frac{4}{2n+1}\frac{1}{m}\right)^{1/(n+4)}. \tag{6}$$

In many applications, it proves to be particularly advantageous to introduce the concept of modification of the smoothing parameter. The construction of the estimator can then be done in the following manner:

A) the kernel estimator $\hat{f}$ is calculated in accordance with basic dependence (3);

B) the modifying parameters $s_i > 0$ $(i = 1, 2, \ldots, m)$ are stated as

$$s_i = \left(\frac{\hat{f}(x_i)}{c}\right)^{-1/2}, \tag{7}$$

where c denotes the geometric mean of the numbers $\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_m)$, given in the form of the logarithmic equation

$$\log(c) = \frac{1}{m}\sum_{i=1}^{m}\log(\hat{f}(x_i)); \tag{8}$$

C) the kernel estimator with the modified smoothing parameter is defined by the formula

$$\hat{f}(x) = \frac{1}{mh^n}\det(D)\sum_{i=1}^{m}\frac{1}{s_i^n}K\left(D\frac{x - x_i}{hs_i}\right). \tag{9}$$

The use of the modification procedure improves the quality of the estimator, but it also noticeably decreases its sensitivity to the exactness of the choice of the constant h. Accordingly, in practice this property gives grounds to accept the approximate value given by dependence (6).

An illustrative description of the above methodology can be found in Silverman (1986). In particular, proposals are contained there for other forms of the matrix D, different types of kernels diverging from the normal (5), as well as a methodology for determining the smoothing parameter when it proves impossible to apply the simplified formula (6). These issues are also presented in mathematicized form by Prakasa Rao (1983); a differing concept has been given by Devroe and Gyorfi (1985). A detailed review of the methods used to fix the smoothing parameter is also presented in the survey paper by Jones and Maron (1996).

4. ORDER ESTIMATOR OF QUANTILE

In view of the features of the test here proposed, a modified concept of the order estimator, based on the form recommended in the survey paper by Parrish (1990), will be proposed below for estimating the quantile.

Consider the 1-dimensional random variable Z, its m-element random sample z_1, z_2, ..., z_m and a non-zero natural number k no greater than the size of this sample: $1 \leq k \leq m$. Let Z_k denote the k-th order statistic of the random variable Z (for its definition see *e.g.* the textbook by Fisz, 1963; the value of the statistic Z means in practice the k-th element, in terms of magnitude, of the random sample z_1, z_2, ..., z_m). If the size of the random sample m and the order of the quantile R fulfill the condition

$$0,5 \leq mR \leq m - 0,5, \tag{10}$$

then the quantile estimator $\hat{q}$ may be defined by the following formula (Parrish, 1990):

$$\hat{q} = (0,5 + i - mR)\, Z_i + (0,5 - i + mR)\, Z_{i+1}, \tag{11}$$

given that

$$i = [mR + 0,5], \tag{12}$$

where $[b]$ constitutes an integral part of the number $b \in \mathbb{R}$. If the support of the random variable Z is of the form $[0, \infty)$, then in the case

$$mR < 0,5, \tag{13}$$

formula (11) can be supplemented by

$$\hat{q} = 2mRZ_1. \tag{14}$$

It is obvious, in fact, that in such case

$$\lim_{mR \to 0^+} \hat{q} = 0 \tag{15}$$

$$\lim_{mR \to 0,5^-} \hat{q} = Z_1 = \lim_{mR \to 0,5^+} \hat{q}. \tag{16}$$

Ultimately, the generalized order estimator of the quantile can be proposed in the form

$$\hat{q} = \begin{cases} 2mRZ_1 & \text{when } mR \leq 0,5 \\ (0,5 + i - mR)\, Z_i + (0,5 - i + mR)\, Z_{i+1} & \text{when } 0,5 \leq mR \leq m - 0,5 \end{cases}, \tag{17}$$

where the number i is given by formula (12).

For the purposes of the test designed here, the random variable Z introduced in this section should be defined as

$$Z \equiv \hat{f}_{x_1,\ldots,x_m} \circ X, \tag{18}$$

whereas its m-element random sample by

$$\begin{aligned} z_1 &= \hat{f}_{x_1,\ldots,x_m}(x_1) \\ z_2 &= \hat{f}_{x_1,\ldots,x_m}(x_2) \\ &\vdots \\ z_m &= \hat{f}_{x_1,\ldots,x_m}(x_m). \end{aligned} \tag{19}$$

The degree of the quantile is identical with the assumed level of significance of the test:

$$R = r. \tag{20}$$

Finally, the quantile estimator (17) obtained above is equal to the test's critical value introduced by formula (2):

$$a = \hat{q}. \tag{21}$$

In conclusion, it should be noted that, taking into account the levels of significance r applied in practice, typically 0,01, 0,05 or 0,1, which in the task at hand determine – through equality (20) – the degree of the quantile R, conditions (10) and (13) exhaust all the possibilities actually encountered.

Differing concepts of the quantile estimators can be found in the survey paper by Parrish (1990), which deals with order estimators, and by Sheather and Marron (1990), where kernel estimators are considered.

5. SIMULATION RESULTS AND FINAL COMMENTS

The material presented above provides the complete procedure necessary to construct a non-parametric test of significance regarding the equality of the distribution functions of two independent n-dimensional random variables, in a case where one sample is one-element. Having obtained the random samples $x_1, x_2, \ldots, x_m$ and y, after assuming the level of significance r, one should successively:

A) generate the kernel estimator of the density function $\hat{f}_{x_1,\ldots,x_m}$ in accordance with the instructions given in section 3:

a) calculate the value of the smoothing parameter, using formula (6);

b) give the basic form of the kernel estimator (3) along with (4)-(5);

c) introduce a modification of the smoothing parameter, on the basis of algorithm (7)-(9);

B) define the values of the random sample for the distribution of the statistic S using dependence (19);

C) calculate the quantile of order R (taking into account formula (20) equal to the assumed level of significance r), applying dependence (17) together with (12);

D) thanks to the results of item (A), define the form of test statistic (1) and – on the basis of item (C) – critical set (2), making use of equality (21);

E) from formula (1) calculate the value of the test statistic $\hat{f}_{x_1,\ldots,x_m}(y)$; if it be-

longs to the critical set, then the hypothesis of the equality of the distribution functions should be rejected; in the opposite case, there are no grounds to do so.

The correctness of the results obtained by using the above test has been positively verified by means of comprehensive simulation investigations. In order to assure a stable outcome with the precision presented below, the results here described were obtained by testing from 10.000 to 100.000 samples.

Accordingly, table 1 shows the comparative results of the power of the test here designed and the classic Smirnov test (see Fisz, 1963, or Wilks, 1962) for a typical case when the random variables X and Y are 1-dimensional and have normal distributions with unit variance and expectations $-s$ and s, that is, $N_X(-s, 1)$ and $N_Y(s, 1)$. (In the case of a one-element random sample, the Smirnov statistic has an asymptotic (with respect to size m) uniform distribution on the interval [0,5; 1]). In analyzing the contents of table 1, it should be noted that:

A) the test proposed here holds steady at the assumed level of significance for smaller sizes of the random sample $x_1, x_2, \ldots, x_m$ than does the classic Smirnov test;

B) the test designed in this paper has greater power.

(Even if in one case or another the number of type 2 errors proves to be less for the Smirnov test, this is always associated with a disproportionately large number of type 1 errors). Similar encouraging results were obtained for many other tested distributions, including also asymmetric and "long-tailed"; for an illustration, see table 2, given for exponential distribution with unit variance and expectations $-s$ and s, that is, $E_X(-s, 1)$ and $E_Y(s, 1)$. It should be emphasized that the Smirnov test has been characterized as the most useful among the classical tests in the case considered here of a one-element sample. Indeed, it can easily be observed that – for example – the series tests popular in the literature (see e.g. Fisz, 1963, or Wilks, 1962) are in this situation most often completely unsuitable for practical applications. In their basic form, when the value of the statistic constitutes the received number of the series, it can assume only the values 2 or 3, and so the information assembled by this method is exceedingly scanty. Also in the case of other series tests here examined, not so inordinately disadvantageous, e.g. Mann-Withney (Fisz, 1963; Wilks, 1962), the results were significantly worse than those presented in table 1. The lack of data which would enable the critical value to be fixed constitutes a serious hindrance to the widespread application of the popular tests of significance for the case of a one-element sample, since these tests do not take such cases into account; not infrequently the distribution of statistics is given in an asymptotic form with respect to the size of both random samples, and thus in a form that is here utterly useless.

The crucial problem in using kernel estimators is the selection of the smoothing parameter value. It should be underlined that the designed test proved to be only very slightly sensitive for that value. Table 3 shows the results obtained for the smoothing parameter h decreased and increased two- and four-fold (!) with respect to the value given by dependence (6). Thus the power of the test and the minimum sample size guaranteeing the assumed level of significance undergo only insignificant changes, for such an extreme differentiation. This very advantageous feature can be explained by the empirical method of fixing the critical value (by estimating

TABLE 1

Empirical power of: a) the test designed in this paper, b) the classical Smirnov test, for the 1-dimensional random variables X and Y with normal standard distributions shifted by – s and s, respectively, given in percentages; the cases marked with bold are those where the assumed level of significance r held steady with 50-percent precision

$r = 0{,}01$						
m	$s = 0$		$s = 1$		$s = 2$	
	(a)	(b)	(a)	(b)	(a)	(b)
5	3,33	32,75	31,48	76,12	81,29	99,15
10	1,68	17,49	28,77	65,62	83,16	98,23
20	1,52	8,78	30,56	55,25	87,56	97,14
50	**1,50**	3,59	**33,94**	41,13	**93,23**	94,67
100	**1,49**	1,81	**34,91**	32,42	**93,29**	91,99
200	**1,24**	**0,92**	**33,58**	**24,31**	**93,61**	**88,44**
500	**1,08**	**1,09**	**33,01**	**29,06**	**93,47**	**92,29**
1000	**1,05**	**1,07**	**32,61**	**27,37**	**93,54**	**91,71**

$r = 0{,}05$						
m	$s = 0$		$s = 1$		$s = 2$	
	(a)	(b)	(a)	(b)	(a)	(b)
5	10,82	32,75	51,61	76,12	94,06	99,15
10	11,04	17,49	58,55	65,62	97,45	98,23
20	**7,44**	8,78	**56,95**	55,25	**98,17**	97,14
50	**5,93**	7,67	**54,71**	54,61	**98,18**	97,93
100	**5,18**	**5,45**	**54,21**	**52,19**	**98,27**	**97,64**
200	**4,84**	**4,37**	**54,36**	**50,31**	**98,43**	**97,68**
500	**5,11**	**5,07**	**54,06**	**51,97**	**98,43**	**97,87**
1000	**5,01**	**4,89**	**54,00**	**51,76**	**98,44**	**97,92**

$r = 0{,}1$						
m	$s = 0$		$s = 1$		$s = 2$	
	(a)	(b)	(a)	(b)	(a)	(b)
5	23,94	32,75	68,78	76,12	98,41	99,15
10	**14,97**	17,49	**66,85**	65,62	**98,75**	98,23
20	**12,74**	**8,78**	**66,41**	**55,25**	**99,06**	**97,14**
50	**10,74**	**10,80**	**65,64**	**64,08**	**99,19**	**98,90**
100	**10,08**	**9,08**	**65,44**	**62,31**	**99,33**	**98,88**
200	**9,71**	**9,15**	**65,43**	**63,35**	**99,37**	**98,94**
500	**10,25**	**10,02**	**65,25**	**63,84**	**99,37**	**99,02**
1000	**9,69**	**9,52**	**65,03**	**63,92**	**99,39**	**99,09**

TABLE 2

Empirical power of: a) the test designed in this paper, b) the classical Smirnov test, for the 1-dimensional random variables X and Y with exponential standard distributions ($\lambda = 1$) shifted by $-s$ and s respectively, given in percentages; the cases marked with bold are those where the assumed level of significance r held steady with 50-percent precision

	$r = 0{,}01$					
m	$s = 0$		$s = 1$		$s = 2$	
	(a)	(b)	(a)	(b)	(a)	(b)
5	5,47	34,70	38,11	71,19	74,28	95,59
10	3,32	18,98	27,05	52,96	67,34	91,55
20	2,42	9,73	21,18	33,09	63,25	83,76
50	1,65	4,06	18,22	14,11	66,76	65,15
100	**1,48**	2,13	**15,28**	7,27	**65,70**	45,50
200	**1,31**	**1,06**	**13,25**	**3,72**	**63,19**	**26,58**
500	**1,15**	**1,16**	**12,02**	**4,22**	**60,03**	**32,57**
1000	**1,05**	**1,03**	**11,55**	**3,61**	**58,22**	**27,46**

	$r = 0{,}05$					
m	$s = 0$		$s = 1$		$s = 2$	
	(a)	(b)	(a)	(b)	(a)	(b)
5	10,89	34,70	54,82	71,19	87,26	95,59
10	9,82	18,98	55,28	52,96	94,11	91,55
20	**7,39**	9,73	**52,10**	33,09	**96,86**	83,76
50	**5,97**	7,91	**44,79**	28,83	**98,77**	90,93
100	**5,50**	**6,02**	**42,29**	**21,55**	**99,98**	**90,92**
200	**5,09**	**4,94**	**40,60**	**18,17**	**100,00**	**91,82**
500	**5,03**	**5,27**	**39,53**	**18,97**	**100,00**	**97,85**
1000	**5,04**	**5,04**	**39,34**	**18,43**	**100,00**	**99,04**

	$r = 0{,}1$					
m	$s = 0$		$s = 1$		$s = 2$	
	(a)	(b)	(a)	(b)	(a)	(b)
5	21,91	34,70	71,82	71,19	97,78	95,59
10	**14,06**	18,98	**74,89**	52,96	**99,08**	91,55
20	**12,18**	**9,73**	**74,97**	**33,09**	**99,96**	**83,76**
50	**10,74**	**11,94**	**75,23**	**42,80**	**100,00**	**98,24**
100	**10,37**	**9,94**	**75,40**	**36,36**	**100,00**	**99,21**
200	**10,17**	**9,79**	**75,76**	**36,78**	**100,00**	**99,95**
500	**10,12**	**10,00**	**75,16**	**36,99**	**100,00**	**100,00**
1000	**10,07**	**9,89**	**74,54**	**36,97**	**100,00**	**100,00**

TABLE 3

Empirical power of the test designed in this paper, for the 1-dimensional random variables X and Y with normal standard distributions shifted by – s and s respectively, for: a) the value given by dependence (6) decreased four-fold, b) the value given by dependence (6) decreased two-fold, c) the value given by dependence (6) (repetition of columns (a) from tab. 1), d) the value given by dependence (6) increased two-fold, e) the value given by dependence (6) increased four-fold, given in percentages; the cases marked with bold are those where the assumed level of significance r held steady with 50-percent precision

						$r = 0{,}01$									
m	$s = 0$					$s = 1$					$s = 2$				
	(a)	(b)	(c)	(d)	(e)	(a)	(b)	(c)	(d)	(e)	(a)	(b)	(c)	(d)	(e)
5	10,00	6,01	3,33	1,51	0,48	52,64	42,22	31,48	21,63	11,54	94,46	89,50	81,29	69,01	48,19
10	4,83	2,98	1,68	0,72	0,23	44,25	36,02	28,77	20,87	11,73	92,79	88,55	83,16	75,60	59,68
20	2,85	2,15	1,52	0,74	0,46	39,22	34,70	30,56	25,97	18,14	91,95	89,49	87,56	84,72	77,36
50	1,86	1,59	**1,50**	**1,39**	**1,41**	36,57	34,83	**33,94**	**33,43**	**33,41**	94,28	93,59	**93,23**	**92,99**	**92,91**
100	1,58	**1,48**	**1,49**	**1,46**	**1,39**	36,36	**35,51**	**34,91**	**34,42**	**33,89**	93,63	**93,44**	**93,29**	**93,22**	**92,92**
200	**1,36**	**1,30**	**1,24**	**1,25**	**1,24**	**34,30**	**33,88**	**33,58**	**33,40**	**33,34**	**93,83**	**93,63**	**93,61**	**93,50**	**93,46**
500	**1,10**	**1,11**	**1,08**	**1,08**	**1,08**	**33,37**	**33,16**	**33,01**	**32,80**	**32,79**	**93,49**	**93,49**	**93,47**	**93,43**	**93,43**
1000	**1,05**	**1,05**	**1,05**	**1,03**	**1,03**	**32,74**	**32,66**	**32,61**	**32,65**	**32,66**	**93,58**	**93,58**	**93,54**	**93,64**	**93,67**
						$r = 0{,}05$									
5	19,63	13,66	10,82	8,01	4,30	64,09	57,51	51,61	45,73	35,62	97,41	96,01	94,06	92,07	87,61
10	16,17	12,31	11,04	10,85	10,98	66,12	61,07	58,55	57,72	57,98	98,82	98,08	97,45	97,19	97,15
20	9,50	7,48	7,44	7,26	7,55	60,08	**58,20**	56,95	55,75	55,04	98,48	98,35	**98,17**	**97,89**	97,68
50	**6,31**	**5,98**	**5,93**	**5,87**	**5,88**	**55,61**	**55,01**	**54,71**	**54,42**	**54,43**	**98,23**	**98,16**	**98,18**	**98,12**	**98,10**
100	**5,29**	**5,23**	**5,18**	**5,17**	**5,14**	**54,83**	**54,42**	**54,21**	**53,94**	**53,91**	**98,41**	**98,35**	**98,27**	**98,19**	**98,16**
200	**4,94**	**4,81**	**4,84**	**4,76**	**4,76**	**54,63**	**54,49**	**54,36**	**54,18**	**54,19**	**98,41**	**98,37**	**98,43**	**98,37**	**98,37**
500	**5,12**	**5,07**	**5,11**	**5,11**	**5,13**	**54,12**	**54,06**	**54,06**	**54,09**	**54,14**	**98,43**	**98,44**	**98,43**	**98,43**	**98,43**
1000	**4,99**	**5,02**	**5,01**	**4,96**	**4,97**	**54,08**	**54,01**	**54,00**	**54,05**	**53,97**	**98,51**	**98,48**	**98,44**	**98,41**	**98,42**
						$r = 0{,}1$									
5	40,47	28,53	23,94	23,99	24,59	79,52	72,74	68,78	68,15	68,85	99,53	99,11	98,41	98,19	98,24
10	21,35	15,80	**14,97**	**14,21**	15,17	71,49	68,68	**66,85**	**65,74**	65,12	99,11	98,88	**98,75**	**98,61**	98,54
20	**14,87**	**13,38**	**12,74**	**12,54**	**12,47**	**68,45**	**67,05**	**66,41**	**65,96**	**65,82**	**99,20**	**99,08**	**99,06**	**98,99**	**98,99**
50	**11,48**	**10,95**	**10,74**	**10,67**	**10,62**	**66,49**	**65,93**	**65,64**	**65,61**	**65,55**	**99,11**	**99,16**	**99,19**	**99,15**	**99,16**
100	**10,37**	**10,16**	**10,08**	**10,03**	**10,02**	**65,55**	**65,57**	**65,44**	**65,29**	**65,32**	**99,28**	**99,31**	**99,33**	**99,34**	**99,35**
200	**9,80**	**9,74**	**9,71**	**9,56**	**9,58**	**65,37**	**65,42**	**65,43**	**65,46**	**65,41**	**99,29**	**99,32**	**99,37**	**99,36**	**99,35**
500	**10,33**	**10,29**	**10,25**	**10,23**	**10,25**	**65,26**	**65,31**	**65,25**	**65,27**	**65,26**	**99,36**	**99,35**	**99,37**	**99,38**	**99,38**
1000	**9,78**	**9,76**	**9,69**	**9,69**	**9,69**	**65,11**	**65,03**	**65,03**	**65,00**	**64,99**	**99,39**	**99,37**	**99,39**	**99,36**	**99,38**

the quantile), which prompts a sort of self-adaptation of the procedure. This positive property was precisely the reason for the earlier forcing of an approximate formula (6) instead of the highly sophisticated selection rules suitable for use in other applications of kernel estimators (see Jones and Maron, 1996).

The data displayed in table 1 can also be used to ascertain the minimal size of the random sample that should be obtained from the variable X before applying the test designed in this paper, i.e. the smallest permissible value m. It results from the above data, then, that the test proposed here guarantees the assumed level of significance at the minimal size of the sample $m_{\min}$ given by the approximate dependence

$$m_{\min} \cong \frac{1}{r}. \tag{22}$$

Table 4, in turn, shows the minimal size of the random sample $M(n)$ assuring the 10-percent precision of the kernel estimator of a density function at the zero point for the standard normal distribution, depending on the dimension of the investigated random variables n (Silverman, 1986, p. 93). Thus in cases of greater dimensionality, formula (22) may be generalized to

$$m_{\min} \cong \frac{1}{r}\frac{M(n)}{M(1)} = \frac{M(n)}{4r}. \tag{23}$$

Simulation tests have confirmed the correctness of the criterion formulated in this way. In interpreting the above dependence in conjunction with the contents of table 4, attention should be drawn to the very rapid increase in the requisite sample size in tandem with the increase of the dimensionality n. This property, as a general characteristic of kernel estimators, is of course transmitted to their applications – in the case of the test proposed, fortunately only in linear proportion.

The test conception here presented has been applied to the task of fault detection in automated systems working in the on-line regime (Kulczycki, 1998). In general outline, the compared random variables X and Y characterized respectively the

TABLE 4

Minimal size of the random sample assuring 10-percent precision of the kernel estimator of the density function at the zero point for the n-dimensional standard normal distribution (Silverman, 1986, p. 93)

n	$M(n)$
1	4
2	19
3	67
4	223
5	768
6	2790
7	10700
8	43700
9	187000
10	842000

correct and current operating conditions of the device being supervised. (In practice no limitations occur on the size of the random sample obtained from the variable X, whereas in view of the requisite speed of operation of the fault detection system, the random sample obtained from the second variable Y should be one-element.) In a case when the hypothesis of the equality of the distributions of the random variables X and Y is rejected, one infers device malfunction. The test designed in this paper was fully satisfactory under conditions of practical application in such a task of automatic control.

Polish Academy of Sciences, Systems Research Institute PIOTR KULCZYCKI
Warsaw, Poland

REFERENCES

L. DEVROE, L. GYORFI (1985), *Nonparametric density estimation: the L_1 view*, Wiley, New York.
M. FISZ (1963), *Probability theory and mathematical statistics*, Wiley, New York.
M.C. JONES, J.S. MARRON (1996), *A brief survey of bandwidth selection for density estimation*, "Journal of the American Statistical Association", 91, pp. 401-407.
P. KULCZYCKI (1998), *Fault detection in automated systems by statistical methods*, Alfa, Warsaw.
R.S. PARRISH (1990), *Comparison of quantile estimators in normal sampling*, "Biometrics", 46, pp. 247-257.
B.L.S. PRAKASA RAO (1983), *Nonparametric functional estimation*, Academic Press, Orlando.
S.J. SHEATHER, J.S. MARRON (1990), *Kernel quantile estimators*, "Journal of the American Statistical Association", 85, pp. 410-416.
B.W. SILVERMAN (1986), *Density estimation for statistics and data analysis*, Chapman and Hall, London.
S.S. WILKS (1962), *Mathematical statistics*, Wiley, New York.

RIASSUNTO

Un test per il confronto di funzioni di distribuzione in campioni fortemente sbilanciati

L'articolo presenta un test non parametrico per saggiare l'uguaglianza di funzioni di distribuzione di due variabili indipendenti n-dimensionali, sulla base di campioni di cui uno con un solo elemento. Per risolvere questo problema sono stati usati stimatori Kernel e statistiche d'ordine. Viene presentata una procedura per i calcoli numerici.

SUMMARY

A test for comparing distribution functions with strongly unbalanced samples

The paper presents a non-parametric test of significance regarding the equality of the distribution functions of two independent n-dimensional random variables, on the basis of samples, one of which is one-element. Kernel estimators and order statistics have been used to solve this problem. A fully elaborated procedure is provided for numerical computations.