A Distance Based Two-Sample Test of Means Difference for Multivariate Datasets

Alexander Novoselsky and Eugene Kagan *

Posted Date: 28 November 2023

doi: 10.20944/preprints202311.1732.v1

Keywords: multivariate two-sample problem; multivariate means test; distance-based statistic; two-sample Kolmogorov-Smirnov test

Article

A Distance Based Two-Sample Test of Means Difference for Multivariate Datasets

Alexander Novoselsky [1] and Eugene Kagan [2,*]

[1] Weizmann Institute of Science, 234 Herzl St., Rehovot 76100, Israel
[2] Ariel University, Kiryat a-Mada, Ariel 40700, Israel; evganyk@ariel.ac.il
* Correspondence: evganyk@ariel.ac.il

Abstract. In the paper we present a new test for comparison of the means of multivariate samples with unknown distributions. The test is based on the comparison of the distributions of the distances between the samples' elements and their means using univariate two-sample Kolmogorov-Smirnov test. The activity of the suggested method is illustrated by numerical analysis of the real-world and simulated data.

Keywords: multivariate two-sample problem; multivariate means test; distance-based statistic; two-sample Kolmogorov-Smirnov test

1. Introduction

The problem of comparison of two samples obtained in different measurements appears in a wide range of tasks starting from physical research and ending with social and political studies. The comparison includes the tests of the samples' distributions and their parameters, and the result of the comparison specifies whether the samples were drawn from the same population or not.

For univariate samples, the problem is solved by different methods: the two-sample Student t-test and the Welch t-test (both for comparison of the means for normal distributions), the Fisher F-test (for comparison of variances for normal distributions), the Wilcoxon rank sum test and the paired permutation test (for comparison of the locations which differ from the means), the Kolmogorov-Smirnov test (comparison of the continuous distributions), the Tukey-Duckworth test (comparison of the samples' shift), and so on [1].

For multivariate samples, the problem is less studied and was solved for several specific cases. If the samples are drawn from populations with multivariate normal distributions with equivalent variances, then the comparison of the multivariate means is provided by the extension of the Student t-test that is the two-sample Hotelling T^2-test [2]. If the variances of the populations are different, then comparison of the multivariate means of the samples can be conducted by the family of the tests, which implement the same extension of the Hotelling statistics [3], or its different versions including the test with missing data [4].

Finally, there exists a small number of methods that address the multivariate two-sample problem in which the samples are drawn from the populations with the unknown or differ from normal distributions. The review of the methods based on the interpoint distances appears in the thesis [5], and of the non-parametric methods – in the thesis [6].

In particular, the mostly applicable Baringhaus-Franz test [7] implements the Euclidean distances between the elements of the samples (inter-sample distances) and the distances between the elements in each sample (intra-sample distances). The resulting statistic is a normalized difference between the sum of the inter-sample distances and the average of the intra-sample distances. Since this statistic is not distribution free, critical values are defined by the bootstrapping techniques [8].

In the paper, we follow the line of using the inter- and intra-sample distances and propose a distribution free test for comparison of the means of multivariate samples with unknown distributions. The proposed test implements the distances between the elements of the samples and the centroid of both samples and the distances between the elements of the samples and their

 doi:10.20944/preprints202311.1732.v1

centroids. These distances are considered as random variables, and the test compares distributions of these variables. Acceptance of null hypothesis about the equivalence of the distributions indicates that the populations from which the samples were drawn are equivalent (by equivalence of the means and forms of the distributions) and rejection of the null hypotheses indicates that the samples are drawn from the populations with different means.

Thus, in the proposed test the multivariate data samples are reduced to univariate samples of distances and then the distributions of the univariate samples are compared. If, similar to the Baringhaus-Franz test, the proposed test uses the Euclidian metrics, then the distances are interpreted as deviations of the samples' elements; however, the choice of the metric function is not crucial and can depend on the nature of the data. Comparison between the distances' samples is conducted using the standard two-sample Kolmogorov-Smirnov test.

The proposed test is illustrated by its application to the simulated data and the real-world Iris flowers [9] and Swiss banknotes [10] datasets, which are widely accepted for benchmark tasks.

2. Problem formulation

Let $x = (x_1, x_2, \dots, x_{m_x})$ and $y = (y_1, y_2, \dots, y_{m_y})$ be two n-dimensional samples such that each observation x_i is a random vector $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$, $i = 1,2, \dots, m_x$, and each observation y_j, is a random vector $y_j = (y_{j1}, y_{j2}, \dots, y_{jn})$, $j = 1,2, \dots, m_y$. In the other words, the samples are represented by random matrices

$$x = (x_{ik})_{1 \le i \le m_x, 1 \le k \le n} \quad \text{and} \quad y = (y_{ik})_{1 \le i \le m_y, 1 \le k \le n}.$$

We assume that the numbers m_x and m_y of observations appearing in the considered samples x and y are equal or at least are rather close.

Denote F_x and F_y the multivariate distributions on the populations $\mathcal{X}$ and $\mathcal{Y}$, respectively.

The question is: whether the samples x and y were drawn from the same population or populations $\mathcal{X}$ and $\mathcal{Y}$, from which the samples x and y were, respectively, drawn, are statistically different.

If populations $\mathcal{X}$ and $\mathcal{Y}$ are univariate, the samples x and y are random vectors, and the problem is solved by the standard two-sample tests for different known or arbitrary unknown distributions F_x and F_y. However, for multivariate populations complete analytical solution – the two-sample Hotelling T^2-test [2] – was suggested only for normal F_x and F_y. Together with that, in the last decade were suggested several multivariate two-sample tests [11,12] based on the multivariate version of the Kolmogorov-Smirnov test [13], but these and similar solutions either implement bootstrapping techniques or have certain limitations. For other directions in considering the problem see, e.g., the work [14] and references herein.

In the paper, we assume that the distributions F_x and F_y are continuous with finite expectations $E(x)$ and $E(y)$, respectively, and consider the null hypothesis and alternative hypotheses

$$H_0: E(x) = E(y) \quad \text{and} \quad H_1: E(x) \neq E(y).$$

From the construction of the test it follows that acceptance of null hypothesis indicates that the populations $\mathcal{X}$ and $\mathcal{Y}$ have equivalent expectations and rejection of the null hypotheses indicates that these populations are statistically different by the difference of their means.

The test of statistical equivalence of the populations $\mathcal{X}$ and $\mathcal{Y}$ requires additional test which is conducted after acceptance of the null hypothesis and considers the hypotheses

$$H_0: F_x = F_y \quad \text{and} \quad H_1: F_x \neq F_y$$

given $E(x) = E(y)$. Acceptance of the null hypothesis indicates that the populations $\mathcal{X}$ and $\mathcal{Y}$ are equivalent, and rejection of this hypothesis indicates that the populations $\mathcal{X}$ and $\mathcal{Y}$ are different with equivalent expectations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1732.v1

3. Suggested solution

The proposed test includes two stages: first, the test reduces the multivariate data to the univariate arrays, and second, it studies these arrays as realizations of certain random variables. For the univariate data the first stage is avoided, and the analysis includes the second stage only.

Let x and y be independent n-dimensional random samples respectively drawn from the populations $\mathcal{X}$ and $\mathcal{Y}$ with distributions F_x and F_y and finite expectations $E(x)$ and $E(y)$. Denote by $x \sqcup y$ concatenation of the samples such that if $x = (x_1, x_2, \dots, x_{m_x})$ and $y = (y_1, y_2, \dots, y_{m_y})$, then

$$z = x \sqcup y = (x_1, x_2, \dots, x_{m_x}, y_1, y_2, \dots, y_{m_y}).$$

Expectation of the concatenated sample $x \sqcup y$ is

$$E(z) = \frac{1}{2}(E(x) + E(y)).$$

Now we introduce four univariate random vectors which represent the distances between the observations x_i and y_j and the corresponding expectations. The first two vectors

$$a = (a_1, a_2, \dots, a_{m_x}), \quad a_i = \|x_i - E(x)\|, \quad i = 1,2, \dots, m_x,$$

$$b = (b_1, b_2, \dots, b_{m_y}), \quad b_j = \|y_j - E(y)\|, \quad j = 1,2, \dots, m_y,$$

are the vectors of distances between the observations and the expected values of these observations. The third vector is the concatenation of these two vectors a and b

$$c = a \sqcup b = (c_1, \dots, c_{m_x}, c_{m_x+1}, \dots, c_{m_x+m_y}),$$

in which $c_i = a_i$, $i = 1,2, \dots, m_x$, and $c_{m_x+j} = b_j$, $j = 1,2, \dots, m_y$. Finally, the fourth vector is the vector of distances between the observations and the expectation $E(z)$ of the concatenation $z = x \sqcup y$ of the vectors of observations

$$d = (d_1, \dots, d_{m_x}, d_{m_x+1}, \dots, d_{m_y+m_y}),$$

where $d_i = \|x_i - E(z)\|$, $i = 1,2, \dots, m_x$, and $d_{m_x+j} = \|y_j - E(z)\|$, $j = 1,2, \dots, m_y$.

It is clear that from the equivalence of the expectations $E(x)$ and $E(y)$ follows the equivalence of the vectors c and d and vice versa. Hence, to check the hypothesis that $H_0: E(x) = E(y)$ it is enough to check whether the vectors c and d are statistically equivalent.

Similar to the Baringhaus-Franz test [7], assume that the indicated distances are the Euclidian distances. Then the estimated distances are

$$a_i = \sqrt{\sum_{k=1}^{n}(x_{ik} - \bar{x}_k)^2}, \quad i = 1,2, \dots, m_x,$$

$$b_j = \sqrt{\sum_{k=1}^{n}(y_{jk} - \bar{y}_k)^2}, \quad j = 1,2, \dots, m_y,$$

$$c_l = \begin{cases} a_i, & i = 1,2, \dots, m_x, \quad l = 1,2, \dots, m_x, \\ b_j, & j = 1,2, \dots, m_y, \quad l = m_x + 1, m_x + 2, \dots, m_x + m_y, \end{cases}$$

$$d_l = \begin{cases} \sqrt{\sum_{k=1}^{n}(x_{ik} - \bar{z}_k)^2}, & i = 1,2, \dots, m_x, \quad l = 1,2, \dots, m_x, \\ \sqrt{\sum_{k=1}^{n}(y_{jk} - \bar{z}_k)^2}, & j = 1,2, \dots, m_y, \quad l = m_x + 1, m_x + 2, \dots, m_x + m_y, \end{cases}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1732.v1

where $\bar{x}_k = \frac{1}{m_x}\sum_{i=1}^{m_x} x_{ik}$, $\bar{y}_k = \frac{1}{m_y}\sum_{i=1}^{m_y} y_{ik}$ and $\bar{z}_k = \frac{1}{m_x+m_y}\left(\sum_{i=1}^{m_x} x_{ik} + \sum_{j=1}^{m_y} y_{jk}\right)$ are the elements of the multivariate estimated centers of distributions $\bar{x} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$, $\bar{y} = (\bar{y}_1, \bar{y}_2, \dots, \bar{y}_n)$ and $\bar{z} = (\bar{z}_1, \bar{z}_2, \dots, \bar{z}_n)$, respectively.

For comparison of the vectors c and d we apply the two-sample Kolmogorov-Smirnov test. Then for the considered vectors c and d and their empirical distributions F_c and F_d the Kolmogorov-Smirnov statistic

$$D_{m_x+m_y, m_x+m_y} = \sup_{\xi} |F_c(\xi) - F_d(\xi)|$$

is defined by the difference between the estimated centers of distributions $\bar{x}$, $\bar{y}$ and $\overline{x \sqcup y}$.

Note that acceptance of the hypothesis $H_0: F_c = F_d$ does not indicate the equivalence of the distributions F_x and F_y. To finalize the test and to check the hypothesis $H_0: F_x = F_y$ (after acceptance of $H_0: F_c = F_d$) we propose to apply the Kolmogorov-Smirnov test and compare the distances vectors a and b. Here the Kolmogorov-Smirnov statistic

$$D_{m_x, m_y} = \sup_{\xi} |F_a(\xi) - F_b(\xi)|$$

is defined by the difference between the distributions of the vectors F_a and F_b. Acceptance of the hypothesis $H_0: F_a = F_b$, together with the accepted hypothesis $H_0: F_c = F_d$, indicates that distributions F_x and F_y are statistically equivalent and the samples x and y were drawn from the same population or two statistically equivalent populations.

4. Examples of univariate and bivariate samples

To clarify the suggested method let us consider two simple examples. We start with the univariate two-sample problem.

Let the samples

$$x = (41, 21, 28, 30, 11, 35, 30, 13, 23, 11) \quad \text{and} \quad y = (3, 7, 2, 6, 5, 15, 10, 12)$$

of the lengths $m_x = 10$ and $m_y = 8$ be drawn from the population with normal distribution with the expected value $E(x) = 20$ (and standard deviation $\sigma(x) = 10$) and exponential distribution with $E(y) = 10$, respectively. For simplicity, we rounded the values in the samples.

Then the distances vectors are

$a = (16.7, 3.3, 3.7, 5.7, 13.3, 10.7, 5.7, 11.3, 1.3, 13.3)$,

$b = (4.5, 0.5, 5.5, 1.5, 2.5, 7.5, 2.5, 4.5)$,

$c = (16.7, 3.3, 3.7, 5.7, 13.3, 10.7, 5.7, 11.3, 1.3, 13.3,$
$4.5, 0.5, 5.5, 1.5, 2.5, 7.5, 2.5, 4.5)$,

$d = (24.2, 4.2, 11.2, 13.2, 5.8, 18.2, 13.2, 3.8, 6.2, 5.8,$
$13.8, 9.8, 14.8, 10.8, 11.8, 1.8, 6.8, 4.8)$.

The Kolmogorov-Smirnov test with significance level $\alpha = 0.05$ rejects the hypothesis $H_0: F_c = F_d$. Thus, it can be concluded that the expectations $E(x)$ and $E(y)$ are different and the samples x and y were drawn from different populations or, at least, are significantly shifted.

The same result is obtained by direct comparison of the samples x and y. The Kolmogorov-Smirnov test with significance level $\alpha = 0.05$ rejects the hypothesis $H_0: F_x = F_y$.

Now let both samples

$$x = (18, 28, 15, 14, 26, 19, 21, 31, 15, 22) \quad \text{and} \quad y = (28, 8, 21, 18, 25, 23, 20, 12)$$

of the lengths $m_x = 10$ and $m_y = 8$ be drawn from the population with normal distribution with the expected value $E(x) = 20$ (and standard deviation $\sigma(x) = 10$).

As expected, the Kolmogorov-Smirnov test with significance level $\alpha = 0.05$ accepts the hypothesis $H_0: F_c = F_d$, and then accepts the hypothesis $H_0: F_a = F_b$. Thus, it can be concluded that

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1732.v1

samples x and y were drawn from the same population, and direct comparison of the samples x and y confirms this conclusion.

Now let us consider an example of the bivariate two-sample problem. Assume that the samples are represented by the matrices

$$x = \begin{pmatrix} 17 & -4 & 29 & 25 & 20 & 27 & 26 & 20 & 13 & 16 \\ 22 & 17 & 2 & 15 & 28 & 17 & -4 & 7 & 22 & 17 \end{pmatrix},$$

$$y = \begin{pmatrix} 3 & 3 & 27 & 20 & 1 & 6 & 16 & 14 \\ 2 & 5 & 19 & 12 & 13 & 1 & 2 & 0 \end{pmatrix}.$$

The first matrix was drawn from normally distributed population with $E(x) = \begin{pmatrix} 20 \\ 20 \end{pmatrix}$ (and standard deviation $\sigma(x) = \begin{pmatrix} 10 \\ 10 \end{pmatrix}$) and the second matrix was drawn from exponential distribution with $E(y) = \begin{pmatrix} 10 \\ 10 \end{pmatrix}$. The mean vectors for these samples are

$$\bar{x} = \begin{pmatrix} 18.9 \\ 14.3 \end{pmatrix} \quad \text{and} \quad \bar{y} = \begin{pmatrix} 11.25 \\ 6.75 \end{pmatrix}.$$

The values of the samples and their centers are shown in Figure 1.

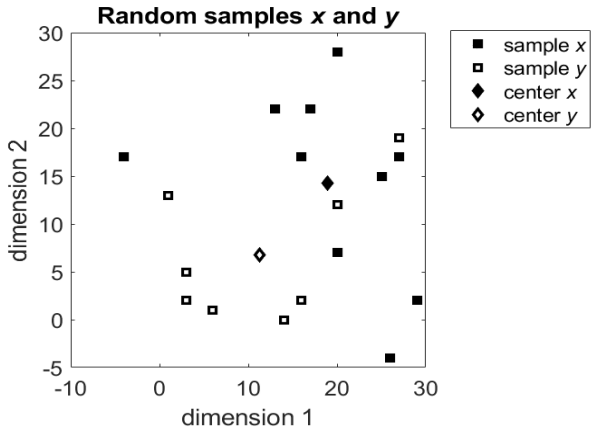

Figure 1. The bivariate samples x and y and their centers $\bar{x}$ and $\bar{y}$.

Then the distances vectors are

$a = (7.9, 23.1, 15.9, 6.1, 13.7, 8.5, 19.6, 7.4, 9.7, 4.0),$

$b = (9.5, 8.4, 20.0, 10.2, 12.0, 7.8, 6.7, 7.3),$

$c = (7.9, 23.1, 15.9, 6.1, 13.7, 8.5, 19.6, 7.4, 9.7, 4.0,$
$9.5, 8.4, 20.0, 10.2, 12.0, 7.8, 6.7, 7.3),$

$d = (11.2, 20.4, 16.2, 10.3, 17.6, 13.0, 18.3, 6.0, 11.3, 6.1,$
$15.4, 13.8, 14.0, 4.6, 14.6, 13.8, 9.0, 11.0).$

The Kolmogorov-Smirnov test with significance level $\alpha = 0.05$ rejects the hypothesis $H_0: F_c = F_d$. Thus, it can be concluded that the expectations $E(x)$ and $E(y)$ are different and the samples x and y were drawn from different populations or, at least, are significantly shifted.

Note that direct comparison of the vectors a and b results in acceptance of the hypothesis $H_0: F_a = F_b$, which, however, does not lead to additional conclusions about the expectations $E(x)$ and $E(y)$ and about the distributions F_x and F_y.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1732.v1

5. The algorithm of two-sample test

For convenience, let us formulate the proposed test in the algorithmic form.

Algorithm: two-sample test of means difference for multivariate datasets

Input: two n-dimensional samples that are independent random matrices $x = (x_{ik})_{1\le i\le m_x, 1\le k\le n}$ and $y = (y_{ik})_{1\le i\le m_y, 1\le k\le n}$.

Output: conclusions about difference between the expectations $E(x)$ and $E(y)$ and about difference between the distributions F_x and F_y of the samples.

1. Compute the multivariate mean $\bar{x}$ (n-dimentional vector) of the sample x.
2. Compute the multivariate mean $\bar{y}$ (n-dimentional vector) of the sample y.
3. Compute the distance between each element x_i of the sample x and its mean $\bar{x}$ and combine them into vector a.
4. Compute the distance between each element y_j of the sample y and its mean $\bar{y}$ and combine them into vector b.
5. Concatenate the vectors a and b of the distances into the vector c.
6. Concatenate the samples x and y into the sample z.
7. Compute the multivariate mean $\bar{z}$ (n-dimensional vector) of the sample z.
8. Compute the distance between each element z_l of the sample z and its mean $\bar{z}$ and combine them into vector d.
9. Apply the two-sample Kolmogorov-Smirnov test for the distributions F_c and F_d of the vectors c and d.
10. If the hypothesis $H_0: F_c = F_d$ is accepted, then
11. Accept the hypothesis $H_0: E(x) = E(y)$,
12. Apply the two-sample Kolmogorov-Smirnov test for the distributions F_a and F_b of the vectors a and b,
13. If the hypothesis $H_0: F_a = F_b$ is accepted, then
14. Accept the hypothesis $H_0: F_x = F_y$,
15. else
16. Accept the hypothesis $H_1: F_x \ne F_y$,
17. end if,
18. else
19. Accept the hypothesis $H_1: E(x) \ne E(y)$,
20. end if.
21. Return the accepted hypotheses.

Note again that the numbers m_x and m_y of observations in the samples x and y should be equal or at least rather close.

Application of the squared differences between the samples' elements and the means leads to certain similarity between the suggested method and the one-way analysis of variance [15] but without crucial requirement of the normal distribution of the samples.

6. Verification of the method

The suggested method was verified using real-world and simulated data. For verifications, we implemented the algorithm in MATLAB® and used the appropriate functions from its statistical toolbox. The significance level in the two-sample Kolmogorov-Smirnov tests is $\alpha = 0.05$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1732.v1

6.1. Trials on the simulated data

For the first trials we generated two multivariate samples with different distributions and parameters and then applied the suggested algorithm to these samples. Each sample includes $m_x = m_y = 100$ elements. Examples of the simulated bivariate normally distributed samples are shown in Figure 2. In the figure, the samples have different predefined means and different standard deviations. For simplicity here we show the samples with difference in their means only in one dimension; in the other dimension the means are equal to zero.

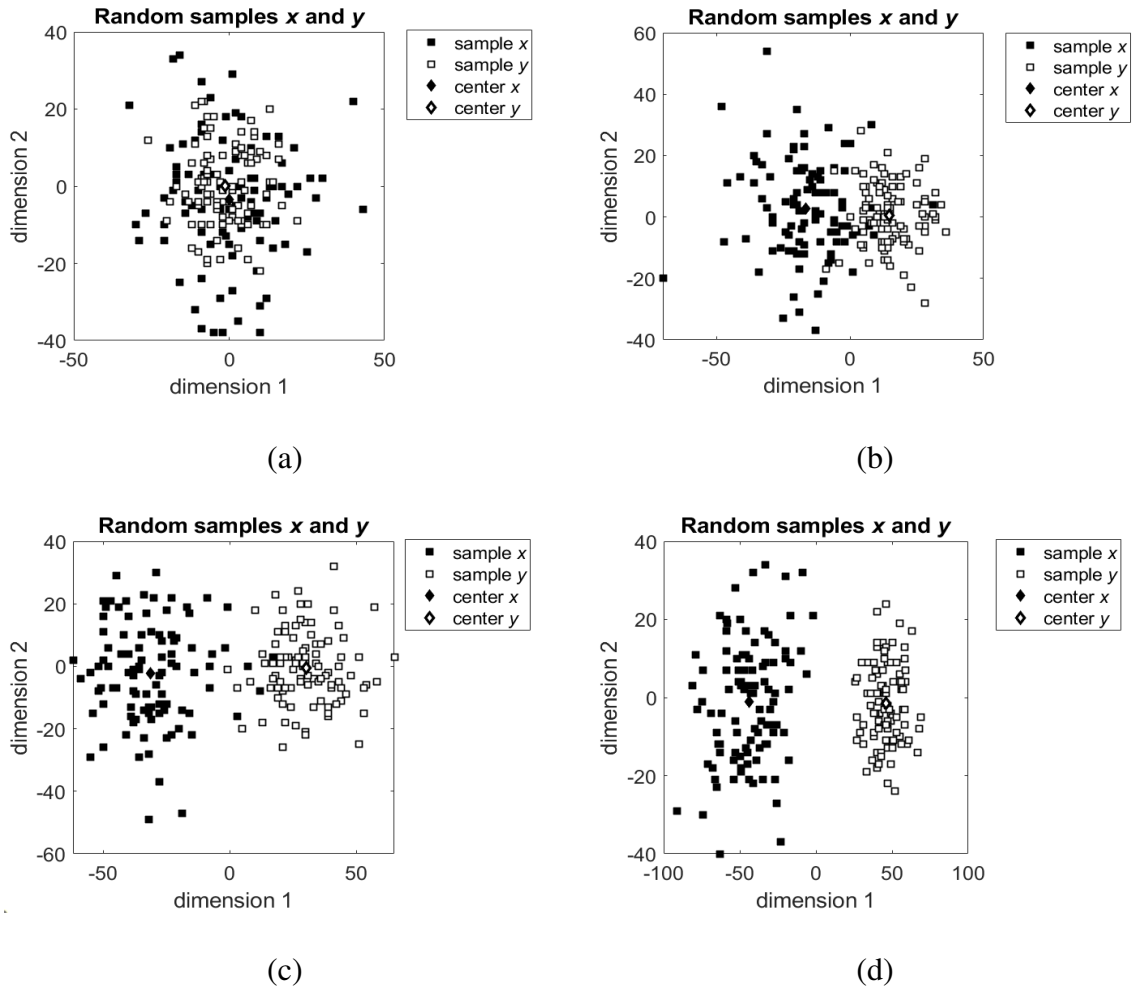

Figure 2. Randomly generated bivariate normally distributed samples with standard deviations $\sigma(x) = 15$ and $\sigma(y) = 10$: (a) expectations $E(x) = E(y) = 0$; (b) expectations $E(x) = -15$, $E(y) = 15$; (c) expectations $E(x) = -30$, $E(y) = 30$; (d) expectations $E(x) = -45$, and $E(y) = 45$.

The results of the tests of these samples by the suggested algorithm are summarized in Table 1.

Table 1. Results of the tests of the illustrative bivariate normally distributed samples with standard deviations $\sigma(x) = 15$ and $\sigma(y) = 10$.

Sample x	Sample y	$H_0: E(x) = E(y)$	$H_0: F_x = F_y$
$E(x) = 0$	$E(y) = 0$	Accepted	Rejected
$E(x) = -15$	$E(y) = 15$	Rejected	Rejected*
$E(x) = -30$	$E(y) = 30$	Rejected	Rejected*
$E(x) = -45$	$E(y) = 45$	Rejected	Rejected*

*Hypothesis was rejected by the rejection of the hypothesis $H_0: E(x) = E(y)$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1732.v1

As it was expected, for the first two samples the test accepted the hypothesis $H_0: E(x) = E(y)$ for equivalent expectations and rejected the hypothesis $H_0: F_x = F_y$ because of different standard deviations. In the next three cases, the test rejected the hypothesis $H_0: E(x) = E(y)$ since the expectations were indeed different and because of this difference the hypothesis $H_0: F_x = F_y$ was also rejected.

In the next trials we compared the activity of the Hotelling T^2-test [2] with the activity of the proposed test. The implementation of the Hotelling T^2-test was downloaded from the MATLAB Central File Exchange [16].

Following the requirement of the Hotelling T^2-test, in the trials, we compared two samples drawn from normally distributed populations with varying standard deviations $\sigma(x) = \sigma(y)$ and the expectations $E(x)$ and $E(y)$ such that the difference between them changes from zero (equivalent expectations) to the values for which the samples are separated with certainty. Results of the trials are summarized in Table 2.

Table 2. Results of the Hotelling T^2-test and the suggested test for bivariate normally distributed samples with different expected values and standard deviations.

		Hotelling T^2 test			Suggested test		
$E(x)$	$E(y)$	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 1.5$
0	0	H_0	H_0	H_0	H_0	H_0	H_0
0	0.5	H_1	H_1	H_0	H_0	H_0	H_0
0	1.0	H_1	H_1	H_1	H_1	H_0	H_0
0	1.5	H_1	H_1	H_1	H_1	H_1	H_0
0	2.0	H_1	H_1	H_1	H_1	H_1	H_1

The obtained results demonstrate that for normally distributed samples the suggested test recognizes the differences between the samples as correct as the Hotelling T^2-test, but as expected, it is less sensitive than the Hotelling T^2-test. Thus, if it is known that the samples were drawn from the populations with normal distributions, then the Hotelling T^2-test is preferable, and if the distributions of the populations are not normal or unknown, then the suggested test can be applied.

For validation of the suggested test on the samples drawn from the populations with not normal distributions it was trialed on several pair of samples with different distributions. For example, in Table 3 we summarized the results of the test on the samples with uniform distributions.

Table 3. Results of the suggested test for bivariate uniformly distributed samples with different expected values and interval widths.

$E(x)$	$E(y)$	$\lvert b-a\rvert = 0.5$	$\lvert b-a\rvert = 1.0$	$\lvert b-a\rvert = 1.5$
0	0	H_0	H_0	H_0
0	0.15	H_1	H_0	H_0
0	0.30	H_1	H_1	H_0
0	0.45	H_1	H_1	H_1
0	0.60	H_1	H_1	H_1

In addition, from the results presented in Table 2 and Table 3 it follows that similarly to any other statistical test, the sensitivity of the test is as lower as the spreading of the data (standard deviation σ in Table 2 and interval widths $|b-a|$ in Table 3) is higher.

6.2. Trials on the real-world data

For additional verification, we applied the suggested algorithm on two widely known datasets. The first is the Iris flower dataset [9], which contains three samples of Iris plant: *Iris setosa*, *Iris versicolour* and *Iris virginica*. The plants are described by $n = 4$ numerical parameters: sepal length, sepal width, petal length and petal width. Each sample includes $m = 50$ elements.

The sample representing the *Iris setosa* is linearly separable from the other two samples, the *Iris versicolour* and the *Iris virginica*, but these two samples are not linearly separable.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2023 doi:10.20944/preprints202311.1732.v1

The trial includes six independent two-sample tests. The first three tests consider the samples and compare each of them with each of two others. In these tests it was expected that the suggested method will identify that the samples represent different populations.

The second three tests compared each of the samples with itself. We compared the subsample of the first 25 elements with the subsample of the last 25 elements. In these tests, we certainly expected that the method will identify that the compared parts of the same sample are statistically equivalent.

Results of the tests are summarized in Table 4.

Table 4. Results of the tests of the Iris plant samples.

Sample x	Sample y	$H_0: E(x) = E(y)$	$H_0: F_x = F_y$
Iris setosa	*Iris versicolor*	Rejected	Rejected*
Iris setosa	*Iris virginica*	Rejected	Rejected*
Iris versicolor	*Iris virginica*	Rejected	Rejected*
Iris setosa	*Iris setosa*	Accepted	Accepted
Iris versicolor	*Iris versicolor*	Accepted	Accepted
Iris virginica	*Iris virginica*	Accepted	Accepted

*Hypothesis was rejected by the rejection of the hypothesis $H_0: E(x) = E(y)$.

As expected, the method correctly identified that the samples representing different types of Iris plants are statistically different. In all comparisons the hypotheses $H_0: E(x) = E(y)$ was rejected. Note that the method correctly identified the difference between two linearly non separable samples.

Also, the method correctly identified statistical equivalence of the subsamples taken from the same samples. In these comparisons the methods correctly accepted both the hypothesis $H_0: E(x) = E(y)$ and the hypothesis $H_0: F_x = F_y$.

The second dataset is the dataset of Swiss banknotes [10], which includes $m = 200$ records about 100 genuine and 100 counterfeit banknotes included in the samples x and y, respectively. Each banknote is characterized by $n = 6$ numerical parameters specifying their geometrical sizes.

The suggested test correctly rejected the null hypothesis $H_0: E(x) = E(y)$ and separated the records about genuine and counterfeit banknotes with significance level $\alpha = 0.05$ and p-value close to zero.

Note that the same result was reported for the two-sample Hoteling T^2 test which also rejected the null hypothesis about the equivalence of the samples and separated the records with p-value close to zero.

7. Conclusion

The proposed test for comparison of the means of multivariate samples with unknown distributions correctly identifies statistical equivalence and difference between the samples.

Since the test implements the Kolmogorov-Smirnov statistic, it does not require specific distributions of the samples and can be applied to any reasonable data.

In addition, the proposed method, in contrast to the existing tests, does not consider the pairwise relations between all elements of the samples and so it requires less computation power.

The method was verified on simulated and real-world data and in all trials it demonstrated correct results.

Funding: This research has not received any grant from funding agencies in the public, commercial, or non-profit sectors.

Data Availability Statement: The data obtained from open access repositories; the links appear in the references.

Conflicts of Interest: The authors declare no conflict of interest.

Competing Interests: The authors declare no competing interests.

References

1. Moore D.S., McCabe G.P., Craig B. *Introduction to the Practice of Statistics*. W. H. Freeman: New York, 2014.
2. Hotelling H. The generalization of Student's ratio. *Annals of Mathematical Statistics*, 1931, 2(3), 360-378.
3. Coombs W.T., Algina J., Oltman D.O. Univariate and multivariate omnibus hypothesis tests selected to control type I error rates when population variances are not necessarily equal. *Review of Educational Research*, 1996, 66(2), 137-179.
4. Wu Y., Genton M.G., Stefanski L.A. A multivariate two-sample mean test for small sample size and missing data. *Biometrics*, 2006, 62(3), 877-885.
5. Siluyele I.J. *Power Studies of Multivariate Two-Sample Tests of Comparison*. MSc Thesis. University of the Western Cape, Cape Town, SA, 2007.
6. Lhéritier A. *Nonparametric Methods for Learning and Detecting Multivariate Statistical Dissimilarity*. PhD Thesis. Université Nice Sophia Antipolis, Nice, France, 2015.
7. Baringhaus L., Franz C. On a new multivariate two-sample test. *J. Multivariate Analysis*, 2004, 88, 190-206.
8. Efron B. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 1979, 7(1), 1-26.
9. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 1936, 7(2), 179-188. The Iris flower dataset was downloaded from the UCI Machine Learning Repository, https://archive.ics.uci.edu/dataset/53/iris (accessed 25 Nov 2023).
10. The Two-Sample Hotelling's T-Square Test Statistic. In the course notes *Applied Multivariate Statistical Analysis*. Eberly College of Science, Pennsylvania State University. The Swiss banknotes dataset was downloaded from the page https://online.stat.psu.edu/stat505/lesson/7/7.1/7.1.15 (accessed 25 Nov 2023).
11. Sadhanala V., Wang Y.-X., Ramdas A., Tibshirani R.J. A Higher-order Kolmogorov-Smirnov test. In *Proc. 22nd Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2019, Naha, Okinawa, Japan, 89, 2621-2630.
12. Kesemen O., Tiryaki B.K., Tezel O., Özku E. A new goodness of fit test for multivariate normality. *Hacet. J. Math. Stat.*, 2021, 50(3), 872 – 894.
13. Justel A., Peña D., Zamar R. A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & Probability Letters*, 1997, 35, 251-259.
14. Qiu Z., Chen J., Zhang J.-T. Two-sample tests for multivariate functional data with applications. *Computational Statistics and Data Analysis*, 2021, 157, 107160, 1-14.
15. Tabachnick B.G., Fidell L.S., Ullman J.D. *Using Multivariate Statistics*. Pearson Education: London, UK, 2018.
16. Trujillo-Ortiz A. Hotelling T^2. MATLAB Central File Exchange, https://www.mathworks.com/matlabcentral/fileexchange/2844-hotellingt2 (accessed 25 Nov 2023).