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

Alexander Novoselsky and Eugene Kagan *

Posted Date: 6 February 2025

doi: 10.20944/preprints202409.0613.v2

Keywords: multivariate one-sample problem; multivariate means test; distance-based statistic

 doi:10.20944/preprints202409.0613.v2

Article

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

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

[1] BMC Software, 10 Ha-Barzel St., Tel-Aviv 6158101, Israel

[2] Ariel University, Kiryat a-Mada, Ariel 4070000, Israel

* Correspondence: evganyk@ariel.ac.il

Abstract: In this note, we suggest a one-sample version of the recently proposed two-sample test for comparison of the means of multivariate samples. The test checks a hypothesis that the mean of multivariate sample is equal to a certain value with the alternative hypothesis that the mean does not equal to this value. The suggested test is illustrated by analysis of real-world and simulated data.

Keywords: multivariate one-sample problem; multivariate means test; distance-based statistic

Mathematics Subject Classification: 62H15 – Hypothesis testing in multivariate analysis

---

1. Introduction

Recently proposed two-sample test allows comparison of the means of multivariate samples with unknown distributions [5]. It utilizes 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 centroids. These distances are considered as random variables, and the test compares distributions of these variables using the Wilcoxon signed-rank test.

In this note, we propose one-sample version of this test and illustrate its application to the simulated data and the real-world datasets - the women's nutrition data [2] and the perspiration data [4].

In the discourse below we use the same notation and terms as in the two-sample test in the paper [5].

2. Problem formulation

Let $x = (x_1, x_2, \dots, x_m)$ be n-dimensional sample such that each observation x_i is a random vector $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$, $i = 1,2, \dots, m$. The sample x is represented by random matrix

$x = (x_{ik})_{1 \le i \le m, 1 \le k \le n}$.

The problem is formulated as follows: given the multivariate sample x it is required to check, whether the mean of the population from which this sample was drawn is equal to certain value, or not.

For univariate sample drawn from normally distributed population, the problem is solved by one-sample t-test, and for multivariate sample with normal distribution it is solved by the one-sample Hotelling T^2-test [3]. For the data with non-normal but close to normal distributions the extended one-sample Hotelling T^2-test was suggested [1]. However, for arbitrary distributed data the problem was not solved yet.

Denote the vector of the means by

$y = (y_k)_{1 \le k \le n}$,

and expectation of the sample x by $E(x)$. Then, according to the formulated problem, the null and alternative hypotheses are

$H_0: E(x) = y$ and $H_1: E(x) \neq y$.

 doi:10.20944/preprints202409.0613.v2

Below, we also assume that the distribution F_x of sample x is continuous and that the expectation $E(x)$ is finite.

3. Suggested Solution

Similar to the two-sample test [5] the proposed test reduces the multivariate data to the univariate arrays and then considers these arrays as realizations of certain random variables.

Given the n-dimensional random sample x drawn from the population $\mathcal{X}$ with distribution F_x and finite expectation $E(x)$, the vector

$a = (a_1, a_2, \dots, a_m),\ a_i = \|x_i - E(x)\|,\ i = 1,2, \dots, m,$

of the distances between the observations x_i and its expectation $E(x)$, and the vector

$b = (b_1, b_2, \dots, b_m),\ b_i = \|x_i - y\|,\ i = 1,2, \dots, m,$

of the distances between the observations x_i and elements of the means vector y, are created.

From the equivalence of the expectation $E(x)$ and y it follows that the vectors a and b are equivalent and vice versa. Hence, to check the hypothesis $H_0: E(x) = y$ it is enough to check whether the vectors a and b are statistically equivalent.

The algorithm of the test is outlined as follows.

Algorithm: one-sample test of difference between means of multivariate dataset and given means vector.

Input: n-dimensional sample $x = (x_{ik})_{1 \le i \le m, 1 \le k \le n}$ that is random matrix and means vector $y = (y_k)_{1 \le k \le n}$ that is 1-dimensional array.

Output: conclusion about difference between the expectation $E(x)$ and y.

---

1. Compute the multivariate mean $\bar{x}$ (n-dimentional vector) of the sample x.
2. Compute the distance between each element x_i of the sample x and its mean $\bar{x}$ and combine them into vector a.
3. Compute the distance between each element x_i of the sample x and means vector y and combine them into vector b.
4. Apply the Wilcoxon signed-rank test to compare the vectors a and b.
5. If the vectors a and b are statistically equivalent, then
6. Accept the hypothesis $H_0: E(x) = y$,
7. else
8. Accept the hypothesis $H_1: E(x) \neq y$,
9. end if.
10. Return accepted hypothesis.

Similar to the two-sample test [5], in the calculations we use Euclidian distances, but for comparison of the vectors a and b, in contrast, we apply the Wilcoxon signed-rank test.

4. Verification of the Method

The suggested method was verified using real-world and simulated data. The algorithm was implemented in MATLAB®. The significance level in the Wilcoxon signed-rank test is $\alpha = 0.05$.

4.1. Trials on the Simulated Data

In trials on the simulated data, we compared the activity of the one-sample Hotelling T^2-test [3] with the activity of the proposed test. The implementation of the Hotelling T^2-test was downloaded from the MATLAB Central File Exchange [6].

 doi:10.20944/preprints202409.0613.v2

For verification, we compared the samples drawn from normally distributed population with several values of standard deviation $\sigma(x)$ with several means' vectors y. Results of the trials are summarized in **Error! Reference source not found.**.

Table 1. Results of the one-sample Hotelling T^2-test and the suggested test for bivariate normally distributed sample with different standard deviations ($m = 100,\ n = 2,\ E(x) = (0, 0)$).

		Hotelling T^2 test			Suggested test		
y		$\sigma = 0.5$	$\sigma = 1.2$	$\sigma = 1.9$	$\sigma = 0.5$	$\sigma = 1.2$	$\sigma = 1.9$
0	0	H_0	H_0	H_0	H_0	H_0	H_0
0.15	0	H_1	H_0	H_0	H_1	H_0	H_0
0.30	0	H_1	H_1	H_0	H_1	H_0	H_0
0.45	0	H_1	H_1	H_1	H_1	H_1	H_0
0.60	0	H_1	H_1	H_1	H_1	H_1	H_1
0.75	0	H_1	H_1	H_1	H_1	H_1	H_1
0.90	0	H_1	H_1	H_1	H_1	H_1	H_1
1.05	0	H_1	H_1	H_1	H_1	H_1	H_1
1.20	0	H_1	H_1	H_1	H_1	H_1	H_1
1.35	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 it was expected, it is less sensitive.

Thus, if it is known that the samples were drawn from the populations with normal distributions, then the Hotelling T^2-test is preferable, but if the distributions of the populations are not normal or unknown, then the suggested test can be used.

4.2. Trials on the Real-World Data

In addition, the algorithm was applied on two known datasets. The first is the women's nutrition dataset [2], which contains $m = 737$ records. Five nutritional components were measured: calcium, iron, protein, vitamin A, and vitamin C ($n = 5$).

Question of interest was whether women meet the federal nutritional intake guidelines. To answer this question, we test the null hypothesis

$H_0: E(x) = [1000, 15, 60, 800, 75]$.

The suggested test correctly rejected the null hypothesis H_0 with significance level $\alpha = 0.05$ and p-value close to zero. The same result was reported for the one-sample Hoteling T^2 test [2].

The second dataset contains perspiration data from 20 healthy females and includes $m = 20$ records with $n = 3$ parameters: sweat rate, sodium content, and potassium content [4, pp. 214-215].

Null hypothesis

$H_0: E(x) = [4, 50, 10]$

was checked.

As a result, with significance level $\alpha = 0.05$ the one-sample Hoteling T^2 test accepted null hypotheses with p-value 0.065, and the suggested test accepted the null hypothesis with p-value 0.502.

5. Conclusion

The proposed test compares the mean of multivariate sample with unknown distributions and given means vector. It correctly identifies statistical equivalence or difference between them.

Since the test utilizes the Wilcoxon signed-rank test, it is not limited by the type of data distribution and is applicable to any reasonable data.

The method was verified on simulated and real-world data and resulted in correct decisions.

Funding: This research has not received any grant or funding.

 doi:10.20944/preprints202409.0613.v2

Data availability statement: The data have been 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. Bulut H. A robust Hotelling test statistic for one sample case in high dimensional data. *Communications in Statistics - Theory and Methods*, 2021, 52(13), 4590–4604.
2. Inferences Regarding Multivariate Population Mean. In the course notes *Applied Multivariate Statistical Analysis*. Eberly College of Science, Pennsylvania State University. The women's nutrition dataset was downloaded from the page https://online.stat.psu.edu/stat505/lesson/7/7.1/7.1.4 (accessed 1 February 2025).
3. Hotelling H. The generalization of Student's ratio. *Annals of Mathematical Statistics*, 1931, 2(3), 360-378.
4. Johnson R.A., Wichern D.W. *Applied Multivariate Statistical Analysis.* 6th ed. Pearson Education: Upper Saddle River, NJ, 2007.
5. Novoselsky A., Kagan E. A distance based two-sample test of means difference for multivariate datasets. *Statistical Papers*, 2024, 1-14.
6. Trujillo-Ortiz A. HotellingT2, 2024. MATLAB Central File Exchange, https://www.mathworks.com/matlabcentral/fileexchange/2844-hotellingt2 (accessed 1 February 2025).