


5-1-2004

## Depth Based Permutation Test For General Differences In Two Multivariate Populations

Yonghong Gao

Center for Devices and Radiological Health Food and Drug Administration, yhg@cdrh.fda.gov

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

### Recommended Citation

Gao, Yonghong (2004) "Depth Based Permutation Test For General Differences In Two Multivariate Populations," *Journal of Modern Applied Statistical Methods*: Vol. 3 : Iss. 1 , Article 6.

DOI: 10.22237/jmasm/1083369960

Available at: <http://digitalcommons.wayne.edu/jmasm/vol3/iss1/6>

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

## Depth Based Permutation Test For General Differences In Two Multivariate Populations

Yonghong Gao

Center for Devices and Radiological Health  
Food and Drug Administration

---

For two  $p$ -dimensional data sets, interest exists in testing if they come from the common population distribution. Proposed is a practical, effective and easy to implement procedure for the testing problem. The proposed procedure is a permutation test based on the concept of the depth of one observation relative to some population distribution. The proposed test is demonstrated to be consistent. A small Monte Carlo simulation was conducted to evaluate the power of the proposed test. The proposed test is applied to some numerical examples.

Key words: Depth, Monte Carlo, permutation test, spatial rank

---

### Introduction

Let  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$  be independent random samples from continuous  $p$ -dimensional populations with cumulative distribution functions  $F(x)$  and  $G(y)$  respectively. The in question interest is in assessing whether there are any differences whatsoever between the  $X$  and  $Y$  probability distributions. Thus, the null hypothesis is tested (1.1) against the most general alternative possible (1.2):

$$H_0: F(t) = G(t), \text{ for any } t, \quad (1.1)$$

$$H_1: F(t) \neq G(t), \text{ for at least one } t. \quad (1.2)$$

In the univariate case, a popular statistic is the two-sided two-sample Kolmogorov-Smirnov statistic  $KS$ , which is

$$KS = (m n / d) \text{Sup}_x \{ | F_m(x) - G_n(x) | \} \quad (1.3)$$

---

Dr. Yonghong Gao may be reached at the Center for Devices and Radiological Health, Food and Drug Administration, 1350 Piccard Drive, Rockville, MD 20850 U.S.A. Telephone: (301)827-0953, fax: (301)443-8559. Email: yhg@cdrh.fda.gov

where  $d$  is the greatest common divisor of  $m$  and  $n$ ,  $F_m(x)$  and  $G_n(x)$  are the empirical distribution functions for the  $X$  and  $Y$  samples, respectively. Under the null hypothesis,  $KS$  is expected to be small, so the null hypothesis is rejected if  $KS > J_\alpha$ , where the constant  $J_\alpha$  is chosen to make the type I error probability equal to  $\alpha$ . When sample sizes are small, values of  $J_\alpha$  are given in tables, when sample sizes are large, where  $\min\{m, n\} \rightarrow \infty$ , Smirnov (1939) derived the asymptotic distribution of the standardized  $KS$  and the limiting distribution of  $KS$  is quite complex.

Another popular approach to the univariate testing problem is the density-based approach, where the two population density functions are estimated using kernel or spline estimation methods and then the test is defined as the distance (maximum distance or mean distance) between the two estimated density functions. Bowman (1985) uses the  $L_2$  distance and Allen (1997) uses the  $L_1$  distance. Allen (1997) conducts a comprehensive simulation study to compare the power of the  $KS$ -test,  $L_2$  distance density test,  $L_1$  density test and  $t$ -type permutation test, the simulation results show that there is no uniformly superior test.

In multivariate setting, two special cases of the testing problem (1.1) have been studied by many investigators. The first case (more extensively studied case) is the two-sample location problems:

$$H_0: \mu=0, \text{ where } G(x)=F(x-\mu). \quad (1.4)$$

The Hotelling's  $T^2$ -test is the usual normal theory test for this problem, it is well-known that the Hotelling's  $T^2$  is the best when distribution is multivariate normal. To free the constraint of normality and to gain the benefit of robustness, many sign-based and rank-based nonparameter tests are proposed using the multivariate versions of the Mood median test and Mann-Whitney test, see Marden's (1999) excellent review paper on this topic.

The second case is the testing of homogeneity of covariances problems:

$$H_0: \text{Var}(X) = \text{Var}(y). \quad (1.5)$$

The Box's M-test is the likelihood ratio test for this problem under multivariate normal distributions.

For the general testing problem (1.1), there is not much activity in existing literature. To develop the multivariate analog of Kolmogorov-Smirnov test, the first challenge faced is to define the empirical distribution based on multivariate data, and that challenge has not been met satisfactorily. Marden (1999) notices the association of  $F(x)$  and  $R(x, F)$  in univariate case:  $R(x, F)=2F(x)-1$ , where  $R(x, F)$  is the rank of  $x$  relative to distribution  $F$ :  $R(x, F)=E(\text{Sign}(x-X))$ , with  $X \sim F$ . Hence Marden (1999) suggests we could use  $KS_R$ ,

$$KS_R = \text{Sup}_x \{ | R_m(x, F) - R_n(x, G) | \} \quad (1.6)$$

where  $R_m(x, F)$  is the multivariate spatial rank of  $x$  relative to sample  $\{X_i\}$ , so far no research activity in investigating the performance of  $KS_R$  has been reported yet.

In this article a KS-test is examined from another aspect. The key idea of Kolmogorov-Smirnov's test is to compare the two distribution functions  $F(x)$  and  $G(x)$ . Noticed was that the distribution function  $F(x)$  is some sort of measure of the position of  $x$  relative to distribution  $F$ , for example, if  $F(x)$  is close to .5, then  $x$  is in the close neighbor of the center of distribution  $F$ , if  $F(x)$  is close to 0 or 1, then  $x$  is on the outskirts of distribution  $F$ , which leads to the idea of the depth of one observation relative to a distribution. It is believed that the depth

function  $D(x, F)$  of one observation  $x$  relative to some distribution  $F$  is some continuous function of  $F(x)$ :  $D(x, F)=g(F(x))$ . For example, in univariate setting, the rank-based depth  $D_r(x, F)$  and the simplex's depth  $D_s(x, F)$  are concave functions of  $F(x)$ :

$$\begin{aligned} D_r(x, F) &= 4 F(x) (1-F(x)), \\ D_s(x, F) &= 2 F(x) (1-F(x)). \end{aligned} \quad (1.7)$$

Unfortunately in higher dimensions there does not exist a similar explicit formula supporting the conjecture that  $D(x, F)$  is some continuous function of  $F(x)$ .

Given the association of  $D(x, F)$  and  $F(x)$ , we use the difference of  $D(x, F)$  and  $D(x, G)$  to measure the difference of  $F(x)$  and  $G(x)$ . While the depth function and the corresponding empirical version are well defined in multivariate settings.

### Methodology

Statistical depth functions have been used to measure the centrality of a multivariate data point with respect to a given data cloud, a center is usually given by a point of maximal depth. This center-outward ordering of the multivariate data provides a foundation for new nonparametric methods in multivariate estimation and inference.

For recent results of different versions of depth function and their applications, see Liu (1990), Liu and Singh (1993), Yeh and Singh (1997) and Zuo, Cui and He (2003). The depth functions usually seen in literature are Tukey's depth proposed by Tukey (1975), simplex depth introduced by Liu (1990), projection depth and Mahalanobis depth. They are all affine invariant and show great potential in multivariate analysis. Mahalanobis's depth is the simplest but least popular one, mainly because it is not robust. Projection depth, Tukey's depth and simplex depth can be quite robust, but the common disadvantage of these three depth functions is that the calculations of these depth functions are quite computationally intensive, especially in high dimensions. Gao (2003) proposes a robust yet easy to calculate depth function based on spatial ranks. In this paper we use this notion of the depth.

For a point  $x$  in  $R^p$  and a  $p$ -variate distribution  $F$ , the spatial rank of  $x$  relative to  $F$  is defined as

$$R(x, F) = E(\text{Sign}(x - Y)), Y \sim F, \quad (2.8)$$

where  $\text{Sign}(x)$  is a unit vector in the same direction of  $x$ . Then the depth of point  $x$  relative to distribution  $F$  is

$$D(x, F) = 1 - \|R(x, F)\|^2 \quad (2.9)$$

The sample version of  $R(x, F)$  and  $D(x, F)$  based on iid sample  $X_1, \dots, X_n$  are

$$R_n(x, F) = (\sum \text{Sign}(x - X_i)) / n \quad (2.10)$$

$$D_n(x, F) = 1 - \|R_n(x, F)\|^2 \quad (2.11)$$

Under the null hypothesis (1.1),  $D(x, F) = D(x, G)$  for any  $x$ , so the proposed test statistic is

$$T(m, n) = \text{Sup}_x \{|D_m(x, F) - D_n(x, G)|\} \quad (2.12)$$

and the null hypothesis is rejected when  $T(m, n) > t_\alpha$ , where  $t_\alpha$  is chosen to make the type I error probability equal to  $\alpha$ .

#### Proposition 1

Under the null hypothesis (1.1), when  $\min\{m, n\} \rightarrow \infty$ ,  $T(m, n) \rightarrow 0$ . The proof of above proposition is based on the following result presented in Gao (2003) about the rank based depth,

$\lim_{n \rightarrow \infty} \text{Sup}_x \{|D(x, F) - D_n(x, F)|\} = 0$ , for any  $x, F$ . Note that test  $T(m, n)$  and test  $KS_R$  are closely related and produces the following result:

$$T(m, n) \leq 2 KS_R.$$

It is not easy to get the distribution (exact or asymptotic) of  $T(m, n)$  under the null hypothesis, bootstrap and permutation resampling methods provide the attractive alternative approaches to determine a critical point for the test. Permutation approach usually shows slightly higher power than the bootstrap approach, hence we use permutation in this paper. The procedure is implemented as the following.

The original two samples are pooled into one large sample  $\{X_1, \dots, X_m, Y_1, \dots, Y_n\}$ . Two resampled data sets are drawn without replacement from the pooled forming the permuted samples  $\{X_1^*, \dots, X_m^*\}$  and  $\{Y_1^*, \dots, Y_n^*\}$ . Each pair of resampled datasets gives a permuted value of the statistic  $T^*(m, n)$ . We repeat this process  $B$  times, yielding  $B$  permuted values of  $T^*(m, n)$ , for a specified level of significance  $\alpha$ , the hypothesis (1.1) is rejected if  $\#\{T^*(m, n) \geq T(m, n)\} + 1 \leq (B+1)\alpha$ .

#### Example 1: Iris data

The Iris dataset was introduced by R. A. Fisher as an example for discriminate analysis. The data report four characteristics (sepal width, sepal length, petal width and petal length) of three species of Iris flower: Setosa, Versicolor and Virginica. From the scatter plot of the any two variables it can be seen that Setosa is quite different from the other two species. The proposed test is applied,  $T(m, n)$ , Marden's rank-based test  $KS_R$ , Box's  $M$ -test  $T_M$  and the Hotelling  $T^2$  test on the three pairs of dataset: (Setosa and Versicolor), (Versicolor and Virginica), and (Setosa and Virginica). The values of the test statistics and the  $p$ -values (the values within the parenthesis) are shown in table 1. From the table we can see that the three species are all significantly different from each other using any of the three tests.

Table 1. Analysis of Iris Data.

Test	Setosa and Versicolor	Versicolor and Virginica	Setosa and Virginica
$T(m, n)$	.9756 (0)	.9885 (0)	.8843 (0)
$KS_R$	1.8807 (0)	1.942 (0)	1.372 (0)
$T_M$	71.302 (0)	116.648 (0)	37.392(0)
$T^2$	2580.8 (0)	4879.6 (0)	355.4(0)

#### Example 2: Hotdogs

The Hotdogs (1989) data file contains data on the sodium and calories contained in each of 54 major hot dog brands. The hotdogs are classified by type: beef, poultry, and meat

(mostly pork and beef, but up to 15% poultry meat), the two variables are Sodium (Milligrams of sodium per hot dog) and Calories (Calories per hot dog). Corresponding to three different type of hot dog produces three data sets, the proposed test is used to determine if these three datasets have the same distribution in terms of the two variables being considered. The analysis result is shown in Table 2.

It is shown in Table 2 that the four tests agree on the following conclusions: there is no significant evidence to say that the beef hotdogs and the meat hotdogs are different, but the beef hotdogs and the poultry hotdogs are significantly different. For meat hotdogs and poultry hotdogs, there is some disagreement among the four tests, both depth test and rank test show some but not that strong evidence to say that these two types of hotdogs are different, while Hotelling's  $T^2$ -test and Box's M-test show significant evidence of difference. To explain this disagreement, the data is further analyzed. One outlier is found (with extreme low sodium value) for the Meat-type hotdogs, because of that one observation, the poultry hotdogs look more like part of the meat hotdogs family (the range of meat hotdogs covers the range of poultry hotdogs). The outlier is deleted and compared with the poultry hotdogs again. The result is in Table 2, where MeatN means the new meat hotdogs data set. Then the four test procedures give us the same conclusion that the meat hotdogs and poultry hotdogs are different.

From this example it is seen that the depth-based permutation test is not powerful when the range of one data set covers the range of another data set, and we should always check the data first, clean the data if possible before implementing any formal testing procedure.

Results

Two simulation experiments were conducted studying the empirical power of the proposed test. The first experiment investigates the sensitivity of the test to the mean effect, the second investigates the sensitivity of the tail mass effect (characterized by variance matrix). For comparison purpose we estimate powers of the Hotelling's  $T^2$ -test, Box's M-test  $T_M$  and Marden's  $KS_R$  test as well in the conducted experiments. For every trial, two samples are generated, one from distribution F and one from G, the hypothesis (1.1) is tested independently using each of the four testing statistics mentioned above. The level of significance is 5%, the bootstrap size B is 199, the sample size is  $m=n=30$ , and dimension is  $p=2$ . The trial was repeated 1000 times for each case (corresponding to different pairs of (F, G)), the empirical power (the number of times the null hypothesis was rejected divided by 1000) is recorded for each test and the results are summarized in Table 3 and Table 4.

Let  $N_2(\mu, \sigma^2 I_2)$  denote the bivariate normal distribution with mean vector as  $\mu$  and covariance matrix as  $\sigma^2 I_2$ . For experiment 1, use  $F = N_2((0,0), I_2)$ ,  $G = N_2((a,a), I_2)$ , with  $a=0, .2, .4, .6$  and  $.8$ . For experiment 2, use  $F = N_2((0,0), I_2)$ ,  $G = N_2((0,0), bI_2)$ , with  $b=1, 1.2, 1.4, 1.6$  and  $1.8$ . When the case is the location problem in multivariate normal distribution (corresponding to experiment 1), the Hotelling's  $T^2$  has the highest power as it should be, the permutation test  $T(m,n)$  has power as much as 80% of the Hotelling's  $T^2$  test, the Box's M-test has no power in this case since it is location invariant, Marden's  $KS_R$  test has some power but lower than  $T(m,n)$  test.

Table 2. Analysis of Hotdogs Data.

Test	Beef vs. Meat	Beef vs. Poultry	Meat vs. Poultry	MeatN vs. Poultry
$T(m,n)$	.2208 (.71)	.7712 (.005)	.2183 (.1)	.6301 (0)
$KS_R$	.6260 (.73)	.9382 (.006)	.8208 (.13)	.8976 (0)
$T_M$	1.696 (.73)	5.011 (.003)	2.454 (0)	5.411 (0)
$T^2$	.506 (.78)	119.1 (0)	87.96 (0)	81.82 (0)

Table 3. Simulation Study 1: Study of the sensitivity to the mean effect.

Test	a=0	a=.2	a=.4	a=.6	a=.8
T(m,n)	.047	.148	.287	.509	.781
KS <sub>R</sub>	.051	.121	.145	.241	.422
T <sub>M</sub>	.052	.048	.049	.049	.051
T <sup>2</sup>	.051	.149	.493	.764	.983

Table 4. Simulation Study 2: Study of the sensitivity to the tail mass effect.

Test	b=1	b=1.2	b=1.4	b=1.6	b=1.8
T(m,n)	.047	.089	.101	.149	.356
KS <sub>R</sub>	.051	.069	.06	.09	.067
T <sub>M</sub>	.052	.099	.11	.198	.511
T <sup>2</sup>	.051	.069	.06	.054	.066

For the case of the homogeneity of covariance matrices (corresponding to experiment 2), the Box's M-test has the highest power, the proposed test T(m,n) is the second best the Hotelling's T<sup>2</sup> test and Marden's Marden's KS<sub>R</sub> test have no power. From this small simulation study it is determined that the proposed test is competitive at least in those two cases and further research is needed to investigate its properties under other situations.

#### References

- Anderson, A., Hall, P., & Titterton, D. W. (1994). Two-Sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50, 41-54.
- Allen, D. (1997). Hypothesis testing using an [trial mode]-distance bootstrap. *The American Statistical*, 51, 145-150.
- Bowman, A. W. (1985). A comparative study of some kernel-based nonparametric density estimators. *Journal of Statistical Computation and Simulation*, 21, 313-327.
- Gao, Y. (2003). Data depth based on spatial ranks. *Statistics and Probability Letters*, 65, 217-225.

Liu, R. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 18, 405-414.

Liu, R. & Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *The Journal of the American Statistical Association*, 88, 252-260.

Marden, J. (1999). Multivariate rank tests. multivariate analysis, design of experiments, and survey sampling. *Marcel Dekker*, 401-432.

Moore, D. S., & Macabe, G. (1989). Introduction to the practice of statistics. original source. *Consumer Reports*, June 1986, pp. 366-367.

Tukey, J. W. (1975). Mathematics and picturing data, proceedings of international congress of mathematics. *Vancouver*, 2, 523-531.

Yeh, A., & Singh, K. (1997). Balanced confidence regions based on Tukey's depth and the bootstrap. *Journal of Royal Statistical Society*, B, 59, 639-652.

Zuo, Y., Cui, H., & He, X. (2003). On the Stahel-Donoho estimator and depth-weighted means of multivariate data, *The Annals of Statistics*, (In Press).