

RANDOMIZATION TESTS FOR A MULTIVARIATE TWO-SAMPLE PROBLEM

J. H. CHUNG AND D. A. S. FRASER

University of Toronto

With few observations involving a large number of variables the T^2 test for the multivariate two-sample problem may not exist. Some alternative tests based on randomization methods are suggested and two of these are applied to an example. Also, valid randomization tests can be obtained by using subgroups of permutations; this provides a simple method for reducing computation which is desirable when the sample sizes are not small.

INTRODUCTION

A FAMILIAR two-sample problem is to test whether two samples have come from identical populations. A frequently considered alternative is slippage—that the populations differ in location. Many tests for the case of observations on a single variable may be found in the literature. However, for the case of observations on several variables there are only a few tests such as Hotelling's T^2 -test for normal theory and the Wald-Wolfowitz modification using the T^2 statistic with permutations of observations [7]. These T^2 -tests require the inversion of a $k \times k$ matrix (k is the number of variables) and in consequence the labour of computation increases rapidly with k . Also, if $k+2$ exceeds the size of the combined samples, the matrix is singular and the test does not exist.

Tests based on permutations of observations are non-parametric tests. For the two-sample problem they require no more than the basic assumption of the problem—that under the null hypothesis the two samples behave as a single sample from a population. Actually, they require less—that, under the null hypothesis, the probability distribution is symmetric under all permutations of the observations. If the populations correspond to different "treatments," this symmetry can be assured by randomly assigning the treatments to the experimental units. As a result, these tests are often called randomization tests. For some examples and references, see Wald and Wolfowitz [7].

In this paper several randomization tests are proposed for the multivariate two sample problem. They were developed primarily for the normal-theory two-sample problem having no T^2 -test, but they are valid more generally; they are nonparametric tests. In Section 3 two of these tests are applied to an example involving 62 measurements on each of 12 people, 4 alcoholics and 8 non-alcoholic controls. In Section 4 a method is proposed for avoiding the prohibitive amount of computation that is necessary if these tests are applied directly to large samples.

THE TWO-SAMPLE PROBLEM

Consider the two-sample problem and suppose that each observation provides measurements on k variables. Let the first sample, consisting of m observations, be designated by (x_{1j}, \dots, x_{kj}) for $j=1, \dots, m$, and let the second sample consisting of n observations be designated by (y_{1j}, \dots, y_{kj}) for $j=1, \dots, n$,

We propose some randomization tests of the null hypothesis that the populations are identical against the alternative hypothesis that there is a difference of "location" for some of the variables.

When searching for good test statistics we often try to maximize power. Here, we do not consider power. Rather we choose statistics intuitively with a view to obtaining statistics that are sensitive towards the type of outcome to be expected under the alternative. For the randomization tests the distribution under the null hypothesis is in a sense always available, but the test statistic needs to be evaluated many times. As a result, the only property other than sensitivity that we shall consider is the ease with which the test statistic can be evaluated.

First, we consider the distribution of a test statistic under the null hypothesis. If the two samples are from the "same population" the joint probability distribution is symmetric under any permutation of the $m+n$ observations. Each permutation thus has the same probability which must therefore be $1/(m+n)!$. Accordingly, the distribution of a test statistic is discrete and it has probability $1/(m+n)!$ at each of the values of the statistic as obtained from the $(m+n)!$ permutations of observations.

The above distribution is really a *conditional* distribution. For, if we are given the information that there are a specific $m+n$ observations in the combined sample, then the conditional distribution is concerned with the division of these observations into a "first sample" of m and a "second sample" of n . If we think of an *ordered* first sample of m and an *ordered* second sample of n , then the conditional distribution has equal probability $1/(m+n)!$ for each of the $(m+n)!$ permutations.

As an example, consider a first sample of $m=2$ and a second sample of $n=1$ and suppose the combined sample contains the three observations $\{1.7, 1.2, 2.5\}$. Then the conditional probability is $1/6$ for each of the 6 permutations: $(1.7, 1.2; 2.5)$, $(1.2, 1.7; 2.5)$, $(1.7, 2.5, 1.2)$, $(1.2, 2.5; 1.7)$, $(2.5, 1.2; 1.7)$, $(2.5, 1.7; 1.2)$.

For a distribution not under the null hypothesis the conditional probabilities will in general not be all equal. Ordinarily, there will be more probability for those permutations that produce a typical "outcome" of the alternative distribution and less for the others. A randomization test can be obtained by taking any reasonable statistic and choosing a critical value on the basis of the conditional distribution above. If a test has a certain significance level conditionally, it has of course the same significance level with respect to the marginal distribution.

We now construct some test statistics that are manageable when k is large. Our approach will be to take a statistic suitable for the single variable case, apply it to each of the k variables and add the resulting expressions. To do more than this, we would seemingly have to take account of sample covariances, as for example in Hotelling's T^2 , and this would require considerably more computation.

For measuring slippage the absolute value of the difference in sample means is simple. However, to prevent gross differences in the scaling of the variables from unbalancing the sum function, it is reasonable to divide this by some scale

function such as the within-sample standard deviation or mean deviation. We then obtain the test statistic

$$\sum_{i=1}^k \frac{|\bar{x}_i - \bar{y}_i|}{s_i} \quad (1)$$

where

$$\bar{x}_i = \frac{1}{m} \sum_{j=1}^m x_{ij}, \quad \bar{y}_i = \frac{1}{n} \sum_{j=1}^n y_{ij}$$

and s_i is the within-sample mean or standard deviation for the i th variable; the latter is given by

$$s_i^2 = \frac{1}{m+n-2} \left[\sum (x_{ij} - \bar{x}_i)^2 + \sum (y_{ij} - \bar{y}_i)^2 \right].$$

A similar statistic that would better emphasize a large value for one of the variables is

$$\sum_{i=1}^k \frac{|\bar{x}_i - \bar{y}_i|^2}{s_i^2}. \quad (2)$$

If the variables are independent and if the variances are known and replaced by their estimates in (2), then the resulting test is most stringent for normal alternatives—among nonparametric tests; see Lehmann and Stein [3].

It is reasonable to have a test statistic independent of the order of the observations in the first sample and independent of the order of the observations in the second sample. Such a test statistic need be evaluated only for the $\binom{m+n}{m}$ different divisions of the $m+n$ observations into a "first sample" of m and a "second sample" of n . The statistics just introduced are of this type. Even for samples of moderate size, however, this number of divisions $\binom{m+n}{m}$ can be very large. This weighs heavily against (1) and (2) because of the calculation of the s_j . A considerable simplification can be obtained by using ranks or coded ranks for each variable. Let r_{ij} , s_{ij} be the ranks of x_{ij} , y_{ij} among the values for the i th variable:

$$\{x_{i1}, \dots, x_{im}, y_{i1}, \dots, y_{in}\}.$$

A modified form of the statistic (1) is

$$\sum_{i=1}^k |\bar{r}_i - \bar{s}_i|. \quad (3)$$

By coding the ranks to have mean zero for each variable,

$$r_{ij}' = r_{ij} - \frac{m+n+1}{2}, \quad s_{ij}' = s_{ij} - \frac{m+n+1}{2},$$

we obtain an equivalent but simpler statistic,

$$\sum_{i=1}^k \left| \sum_{j=1}^m r_{ij}' \right|. \quad (4)$$

A modification of the statistic (2) is

$$\sum_{i=1}^k \left(\sum_{j=1}^m r_{ij}' \right)^2. \quad (5)$$

We might prefer (5) to (4) by analogy with the usual analysis-of-variance test statistic. For the one-variable case ($k=1$) the tests based on (4) and (5) are just the Wilcoxon-Mann-Whitney test.

For the one-variable case, the rank test locally most powerful for a normal alternative of slippage is obtained by recording ranks as shown below and using the difference in means as a test statistic; see Terry [5]. Suppose the $m+n$ rank values

$$1, 2, \dots, m+n$$

are replaced respectively by the mean values of the order statistics for a sample of $m+n$ from the standardized normal distribution, that is, by

$$Ez_{(1)}, Ez_{(2)}, \dots, Ez_{(m+n)}$$

where $z_{(1)}, \dots, z_{(m+n)}$ designate the order statistic for a sample of $m+n$ from the standardized normal. These are tabulated in [2], page 66 for samples up to size 30. Let r_{ij}^* , s_{ij}^* be the modified values for r_{ij} , s_{ij} ; then the statistic (4) becomes

$$\sum_{i=1}^k \left| \sum_{j=1}^m r_{ij}^* \right| \quad (6)$$

and the statistic (5) becomes

$$\sum_{i=1}^k \left(\sum_{j=1}^m r_{ij}^* \right)^2. \quad (7)$$

All of these test statistics should give reasonable tests against the alternative of a difference of location; they are consistent for the multivariate normal with a difference of location. The choice among them would be based on convenience, personal preference, or perhaps a Monte Carlo evaluation of power for some particular alternative distributions.

AN EXAMPLE

The researchers [1] have reported on a preliminary investigation of the metabolic characteristics of compulsive alcoholics (during non-drinking periods) as opposed to those of persons who show no evidence of alcoholic tendencies. A purpose of the experiment was to discover differences that could be subject to further investigation. Another was to reach a simple conclusion as to whether the data did or did not support the contention that compulsive drinkers have different metabolic characteristics. The statistical analysis has been criticized by Popham [4]. A careful survey of statistical techniques applicable to the problem has been made by Tukey [6].

In the experiment sixty-two metabolic characteristics were measured on four alcoholics and eight non-alcoholic controls. For each of the characteristics a two-

sample t -statistic was calculated and from it the corresponding probability level obtained from tables of the t -distribution (two-sided). It was observed that six of these sixty-two probability values were in the 5% zone, whereas approximately three are expected when the characteristics are independent of drinking compulsion. The conclusion was made "that compulsive drinkers do have individualistic metabolic characteristics".

In [4] Popham criticized the conclusion on the grounds that the occurrence of six extreme probabilities was not significant when three of the six were expected without individualistic metabolic characteristics.

In [6] Tukey observes that the characteristics may not be independent of each other and that this would invalidate Popham's quantitative evaluation of the probabilities. Tukey then discusses techniques of analysis both with and without the assumption of independence.

The anticipated type of difference between populations is a shift of mean for some of the characteristics. The number of characteristics or variables being measured is large, $k=64$. This is the type of problem for which we proposed some randomization tests in the preceding section.

Two of the tests were computed on a Ferranti computer, Ferut, at the University of Toronto. To conserve machine time the two simplest tests were chosen, (4) and (6).

The first sample (alcoholics) has $m=4$ and the second sample (non-alcoholics) has $n=8$. Given the $m+n=12$ observations, there are $\binom{12}{4}=495$ different ways of dividing these into a set of 4, thought of as a "first sample," and a set of 8, thought of as a "second sample". For each such division the value of the statistic (4) was calculated. The first value calculated corresponded to the *observed* division into first and second sample observations. This value was found to be at the 91.7% point of the 495 values in the conditional distribution. Using the statistics (6) the 93% point was obtained. Thus the results by either statistic are significant at the 10% level, but not at the 5% level. These probability levels are substantially the same as those obtained from a number of other tests by Professor Tukey.

The computation took approximately one day for programming and two minutes of machine time. For larger samples the required machine time would increase very rapidly—faster than the function $\binom{m+n}{m}$ increases!

FOR LARGE SAMPLES

In the example the sample sizes were small, 4 and 8. If we increase the sample sizes the number of values of the statistic that need to be calculated, $\binom{m+n}{m}$, increases *very* rapidly. Even for samples of the modest size 10, the number of values is in the neighbourhood of 190,000 and an exceptional computer would be needed.

There is no essential reason why all the permutations or combinations need be used. Suppose we have some rules for permuting the sequence of $m+n$ observations. Successive application of these permutations may produce new permutations, but eventually there will be no new permutations produced. The resulting collection of different permutations of the sequence of $m+n$ observations is than a *group*.

For the example let 1, 2, . . . , 12 designate the twelve observations, the first sample contained the observations 1, 2, 3, 4 and the second sample contained the remaining observations. Consider the following permutations

(1, 2, 3, 4; 5, 6, 7, 8, 9, 10, 11, 12)

(5, 6, 7, 8; 9, 10, 11, 12, 1, 2, 3, 4)

(9, 10, 11, 12; 1, 2, 3, 4, 5, 6, 7, 8).

The second, as a permutation of the first, if applied twice produces the third. If any of these are applied repeatedly or successively, no new permutation is produced—they form a *group* of permutations.

Suppose each of the above permutations is applied to the original sequence of 12 observations. Suppose then that we are given only the information that the observed permutation is one of the three. The conditional probability for each of the three, from symmetry, must be $\frac{1}{3}$ under the null hypothesis. There is a corresponding distribution for the values of any test statistic.

Consider now the statistic (6). Its values for the three permutations are 92.1, 72.5, 80.5. The only reasonable significance level available is $33\frac{1}{3}\%$. The *observed* value is 92.1; it is significant at the $33\frac{1}{3}\%$ level and at that level the null hypothesis would be rejected.

Any group of permutations of the $m+n$ observations may be used to construct a test.† *Given* the information that the observed permutation is one of such a group, then under the null hypothesis the conditional probability for any particular permutation is the reciprocal of the number of permutations in the group.

Consider the example again. For a significance level close to 5% we need more permutations. By pairing the observations successively from numbers 1 to 12 and taking all permutations of the six pairs, we obtain 15 different "first samples":

(1, 2, 3, 4), (1, 2, 5, 6), (1, 2, 7, 8), (1, 2, 9, 10), (1, 2, 11, 12)

(3, 4, 5, 6), (3, 4, 7, 8), (3, 4, 9, 10), (3, 4, 11, 12), (5, 6, 7, 8)

(5, 6, 9, 10), (5, 6, 11, 12), (7, 8, 9, 10), (7, 8, 11, 12), (9, 10, 11, 12).

Under the hypothesis each arrangement had equal probability of being the observed arrangement. The corresponding values of the statistics are

92.1 64.7 79.5 67.5 88.0

73.9 50.0 73.9 77.9 72.5

89.4 70.7 84.9 59.7 80.5

The observed value 92.1 is the largest and the hypothesis would be rejected at the $6\frac{2}{3}\%$ level.

† Given the set of 12 observations, there are $12!$ possible outcomes for performing a full randomization test. This collection of possible outcomes can be partitioned in any way, and a conditional test performed, *given* the set in which the actual outcome occurs. If the numerical values of coordinates of observations are not used in forming the sets (and it seems reasonable not to use them), then the "possible outcomes" in a set must result from applying a group of permutations to any one "possible outcome." The reason is that after any permutation has been applied we obtain a "possible outcome" which is as entitled to receive a further permutation as was the original outcome. The collection of permutations so generated will necessarily form a group.

Note that if the tests are to be sensitive we must use permutations that mix up the original first and second sample observations.

For the two-sample problem with larger samples, a group containing 100 to 500 permutations would be reasonable. The labour of computation would then be roughly proportional to the total sample size.

The methods in this paper extend simply to the r -sample problem, and also to any problem where the distribution has a symmetry under the hypothesis that is not found under the alternative.

REFERENCES

- [1] Beerstecher, Ernest, Jr., Sutton, H. Eldon, Berry, Helen Kirby, Brown, William Duane, Reed, Janet, Rich, Gene B., Berry, L. Joe, and Williams, Roger J. "Biochemical individuality v. explorations with respect to the metabolic patterns of compulsive drinkers," *Archives of Biochemistry*, 29 (1950), pp. 27-40.
- [2] Fisher, R. A. and Yates, Frank. *Statistical Tables for Biological, Agricultural, and Medical Research*. New York: Hafner Publishing Company, 1949.
- [3] Lehmann, E. L. and Stein C. "On the theory of some nonparametric hypotheses," *Annals of Mathematical Statistics*, 20 (1949), 28.
- [4] Popham, Robert E. "A critique of the genetotrophic theory of the etiology of alcoholism," *Quarterly Journal Studies on Alcohol*, 14 (1953), 228-37.
- [5] Terry, M. E. "Some rank-order tests which are most powerful against specific parametric alternatives," *Annals of Mathematical Statistics*, 23 (1952), 346.
- [6] Tukey, John W. "Comparing two small samples on many items," *Memorandum Report 54*, Statistical Research Group, Princeton University.
- [7] Wald, A. and Wolfowitz, J. "Statistical tests based on permutations of the observations," *Annals of Mathematical Statistics*, 15 (1944), 358.