SOME REMARKS ABOUT THE NUMBER OF PERMUTATIONS ONE SHOULD CONSIDER TO PERFORM A PERMUTATION TEST

Marco Marozzi

1. INTRODUCTION

To calculate the *p*-value of a permutation test, in theory one should calculate the test statistic for all possible permutations and then compute the proportion of permutations that have test statistic greater than or equal to the observed one. In practice, since the number of all possible permutations is generally impractically large, the *p*-value is usually estimated by taking a random sample of size *B* from all possible permutations. Let us consider, for example, the two-sample location problem. Let $(X_{11}, X_{12}, ..., X_{1n_1})$ and $(X_{21}, X_{22}, ..., X_{2n_2})$ be random samples taken from populations that may differ only in their locations μ_1 and μ_2. Let $n = n_1+n_2$ and $\vartheta = \mu_1-\mu_2$. A permutation test for testing

H_0: $\vartheta = 0$ against H_1: $\vartheta > 0$

is based on the statistic

$$T^* = \sum_{i=1}^{n_1} Y_i^* ,$$

where Y_i^* denotes the *i*-th element of a permutation Y^* of the pooled sample

$$Y = (X_{11}, X_{12}, ..., X_{1n_1}, X_{21}, X_{22}, ..., X_{2n_2}) = (Y_1, ..., Y_{n_1}, Y_{n_1+1}, ..., Y_n) .$$

The observed value of the test statistic is $T_0 = \sum_{i=1}^{n_1} X_{1i} = \sum_{i=1}^{n_1} Y_i$. The *p*-value L_T that would have been obtained if we had considered all *n*! permutations of *Y* is consistently estimated by

$$\hat{L}_T = \frac{1}{B} \sum_{b=1}^{B} I(T_b \geq T_0) ,$$

where T_b denotes the value of T in the b-th (b = 1, …, B) permutation of Y and $I(.)$ denotes the indicator function: $I(T_b \geq T_0) = 1$ when $T_b \geq T_0$ and $I(T_b \geq T_0) = 0$ otherwise. To estimate L_T we use the conditional Monte Carlo method (CMC) described in Pesarin (1992, 2001)[1]. It should be noted that we use the term conditional Monte Carlo (CMC) to underline that we perform an ordinary Monte Carlo simulation with conditioning on Y. Since T^* is not affected by changing the order of added elements, the permutation distribution of T^* consists of $C = \binom{n}{n_1}$ elements instead of $n!$. However, except for very small sample sizes, C remains a rather unmanageable number. For example, when $n_1 = n_2 = 10$ $C = 184756$, but when $n_1 = 20$ $n_2 = 10$ $C > 3$ millions and when $n_1 = n_2 = 20$ $C = 137846$ millions. We would like to mention that for some univariate problems and some statistics based on data sums, there are available fast methods for the exact calculation of tail distribution of permutation tests (Mehta and Patel 1980; Pagano and Tritchler 1983; Mehta et al. 1988). These methods have been coded by Cytel® Software Corporation. We will use the corresponding package, called StatXact®, in section 4. According to the theory, and in particular to the Glivenko-Cantelli theorem, $\hat{L}_T$ is a strong-consistent estimator of L_T. But in practice, when we want to perform a permutation test, how many permutations should be used to obtain a reliable estimate of L_T? Although now fast and relatively cheap computing facilities are at our disposal, this problem is still interesting, in particular for applied statisticians. The problem should be addressed when a Monte Carlo study for estimating the power of a permutation test is performed as well.

The main idea is that it is not necessary to compute all possible permutations to obtain a good p-value estimate. This paper deals just with the choice of B. First, we discuss the approach of Edgington (1995) to this problem and the choice of B made by many authors in their papers. Then, we perform a Monte Carlo study of the power of T for many values of B and discuss the simulation results. Finally, we present two applications in which L_T is estimated for many values of B and then calculated exactly using all permutations.

2. THE APPROACH OF EDGINGTON (1995) AND THE VALUE OF B USED IN MANY PAPERS

For tackling the problem of fixing B, Edgington (1995) suggests to calculate the limits within which the estimated significance value will lie 99% of the time for a given significance level p from the full permutation distribution. When B is large, the p-value estimator is approximately normally distributed with mean p and variance $p(1-p)/B$. Then $(1-\gamma)100\%$ of estimated significance levels will be within $p - z_{1-\gamma/2}\sqrt{p(1-p)/B}$ and $p + z_{1-\gamma/2}\sqrt{p(1-p)/B}$, where $0 < \gamma < 1$ and $z_{1-\gamma/2}$ denotes the $(1-\gamma/2)100$-th percentile of the standard normal distribution.

[1] The method is commercially available from Methodologica® via NPC Test® package. In this paper we do not use NPC Test®, we just coded the CMC algorithm in R Environment.

From tables 1 and 2 it can be seen that if the observed result is just significant at the 5% level compared with the full permutation distribution then, 1000 permutations will "almost surely" give a significant or near-significant result. On the other hand, when the result is just significant at the 1% level, 5000 may be a realistic number. It seems therefore that 1000 is a reasonable number of permutations for a test at the 5% level of significance, whereas 5000 permutations are reasonable when the level is 1%. Edgington (1995) obtained the same results.

TABLE 1

95% probability intervals for estimated p-values

B	$p = 0.01$	$p = 0.05$
100	0.00000-0.02950	0.00728-0.09272
200	0.00000-0.02379	0.01979-0.08021
500	0.00128-0.01872	0.03090-0.06910
1000	0.00383-0.01617	0.03646-0.06351
2000	0.00564-0.01436	0.04045-0.05955
5000	0.00724-0.01276	0.04396-0.05604
10000	0.00805-0.01195	0.04573-0.05427
20000	0.00862-0.01138	0.04698-0.05302
50000	0.00913-0.01087	0.04809-0.05191
100000	0.00938-0.01062	0.04865-0.05135
200000	0.00956-0.01044	0.04904-0.05096
500000	0.00972-0.01028	0.04940-0.05060
1000000	0.00980-0.01020	0.04957-0.05043

TABLE 2

99% probability intervals for estimated p-values

B	$p = 0.01$	$p = 0.05$
100	0.00000-0.03567	0.00000-0.10623
200	0.00000-0.02815	0.01024-0.08976
500	0.00000-0.02148	0.02485-0.07515
1000	0.00188-0.01812	0.03222-0.06778
2000	0.00426-0.01574	0.03743-0.06257
5000	0.00637-0.01363	0.04205-0.05795
10000	0.00743-0.01257	0.04438-0.05562
20000	0.00818-0.01182	0.04602-0.05398
50000	0.00885-0.01115	0.04749-0.05251
100000	0.00919-0.01081	0.04822-0.05178
200000	0.00943-0.01057	0.04874-0.05126
500000	0.00964-0.01036	0.04920-0.05080
1000000	0.00974-0.01026	0.04944-0.05056

Keller-McNulty and Higgins (1987) concluded, on the basis of their simulation study, that there is a little reason to base a permutation test on all possible permutations. They suggest to use 800 permutations to estimate the power of the test and 1600 permutations in actual applications (at $\alpha = 5\%$).

We reviewed the number of permutations considered in many interesting papers. We noted that in the majority of them, 1000 permutations are used for estimating the power of a permutation test: see Anderson and Legendre (1999), Bailer (1989), Hayes (1997), Kennedy (1995), Marozzi (2002), Neuhaus and Zhu (1999), Pesarin (1997), Shipley (2000), Smith (1998) and Wan et al. (1997). Ernst and Schucany (1999) and Pesarin (1994) used 500 permutations, while O'Gorman (2001) used $B = 2000$ and Cade and Richards (1996) used $B = 5000$.

Kim et al. (1991), McQueen (1992) and Venkatraman and Begg (1996) used $B = 1000$ in their practical studies but the number of permutations computed in actual applications is often greater than that usually used in simulations experiments. For example, Ernst and Schucany (1999) used $B = 2000$, Kazi-Aoual (1995) used $B = 5000$ and Kennedy (1995) and Cade and Richards (1996) used 10000 permutations.

3. A MONTE CARLO POWER STUDY

We perform a Monte Carlo study of the power of the test described in section 1 to investigate how the choice of B affects the estimation procedure. To this end, we consider four sample size settings $(n_1, n_2) = (10, 10)$, (20, 20), (10, 20) and (20, 10), and four positive values of the location shift ϑ. One value is determined by the null hypothesis ($\vartheta = 0$) and the other values are specified so that the power of T (at $\alpha = 5\%$) is near 25%, 50% and 75%. For each sample size configuration, 50000 datasets are generated from the standard normal distribution. It should be noted that there is no loss of generality in using normal sampling for what concerns the problem addressed in this section. The permutation tests were based on subsets of 100, 200, 500, 1000, 2000, 5000 and 10000 permutations.

TABLE 3

Percent power estimates of T for $\alpha = 1\%$

				B			
ϑ	100	200	500	1000	2000	5000	10000
				n_1=10 n_2=10			
0.000	1.050	1.054	0.988	0.972	0.972	0.956	0.988
0.450	7.452	7.872	7.780	8.082	8.064	8.008	7.962
0.765	19.574	20.732	21.878	22.422	23.062	23.054	23.054
1.080	39.292	42.586	45.048	45.216	46.150	45.592	46.382
				n_1=20 n_2=20			
0.000	0.992	0.984	1.009	0.992	0.926	0.998	0.952
0.315	7.758	8.038	8.342	8.452	8.320	8.612	8.616
0.530	20.488	22.384	22.982	23.094	23.142	23.814	23.854
0.745	41.142	44.214	46.204	46.714	47.152	47.970	47.532
				n_1=10 n_2=20			
0.000	0.998	0.914	0.998	1.012	0.926	0.964	1.026
0.385	7.400	7.852	8.246	8.350	8.058	8.114	8.244
0.655	20.140	21.502	23.034	23.136	23.390	23.758	23.424
0.925	41.028	44.020	46.548	46.988	47.458	47.554	47.546
				n_1=20 n_2=10			
0.000	0.988	1.004	0.948	1.032	0.984	0.994	1.032
0.385	7.336	7.956	8.204	8.278	8.144	8.288	8.316
0.655	20.312	22.066	22.522	23.186	23.264	23.340	23.300
0.925	41.056	44.206	46.670	47.104	47.548	48.076	47.728

By inspecting table 3 we conclude that to estimate the size and power of T at the $\alpha = 1\%$ significance level 2000 permutations suffice. Table 4 shows that 500 iterations appear to be enough for size/power estimation at $\alpha = 5\%$. These results are consistent with those obtained in section 2, in which the focus were on the estimate of p-values in actual applications rather than on power estimations which clearly requires a minor number of permutations.

TABLE 4

Percent power estimates of T for $\alpha = 5\%$

				B			
ϑ	100	200	500	1000	2000	5000	10000
				n_1=10 n_2=10			
0.000	4.958	4.934	4.942	5.162	4.924	4.974	5.026
0.450	24.430	24.654	24.590	24.736	24.832	24.662	25.006
0.765	47.970	48.712	49.592	49.838	50.036	50.376	50.312
1.080	73.196	74.082	74.880	74.762	75.194	74.742	75.076
				n_1=20 n_2=20			
0.000	4.968	5.040	5.006	4.886	4.796	5.132	4.876
0.315	24.768	24.826	25.294	25.048	24.844	25.188	25.262
0.530	48.302	49.394	49.458	49.430	49.494	50.196	50.026
0.745	72.686	73.584	74.436	74.332	74.482	74.916	74.556
				n_1=10 n_2=20			
0.000	4.886	4.870	5.008	5.064	4.648	4.876	5.026
0.385	23.728	24.226	24.960	25.038	24.990	24.408	24.618
0.655	47.960	49.230	49.992	49.962	49.864	50.232	50.090
0.925	73.194	74.314	74.910	74.956	75.012	75.194	75.402
				n_1=20 n_2=10			
0.000	4.934	4.884	4.950	4.980	5.056	5.160	4.930
0.385	24.128	24.312	24.932	24.862	24.820	24.814	24.810
0.655	48.396	49.656	49.462	49.714	50.166	50.182	50.130
0.925	72.920	74.154	75.144	75.180	75.442	75.546	75.200

4. TWO ACTUAL APPLICATIONS

We applied *T* to two actual data sets. In order to explore how stable the *p*-values of *T* are to the number of considered permutations, we tried *B* values of 100, 200, 500, and of 1000 to 20000 with an increment of 1000. Then we also compute the exact *p*-value through considering all possible permutations.

Hollander and Wolfe (1999) report the data of a study conducted to see if children who watched TV violence were significantly more tolerant of "real-life" violence than children who instead watched a nonviolent TV show. 21 children (first group) were shown a violent TV show, whereas 21 children (second group) were shown a nonviolent TV show. Then each child was shown an act of "real life" violence (two younger children fighting). Toleration of violence was measured by the time (in seconds) each children stopped watching the fight. The data are reported in table 5. Were the children who viewed the violent TV show more tolerant of violence than those who viewed the nonviolent TV show? We report in table 6 the estimated *p*-values of *T*.

As shown in table 6, where *m* stands for 1000, the estimated *p*-values are between 10% and 13%. Moreover, they are between 10.25% and 11.23% for *B* greater than 500. It should be noted that the estimated *p*-values are comparable among themselves and they suggest the same decision about H_0. Consideration of all permutations leads to an exact *p*-value of 10.65% (we used StatXact® package from Cytel®).

TABLE 5

Toleration of violence

First group										
37	39	30	7	13	139	45	25	16	146	94
16	23	1	290	169	62	145	36	20	13	
Second group										
12	44	34	14	9	19	156	23	13	11	47
26	14	33	15	62	5	8	0	154	146	

Source: Hollander & Wolfe (1999).

TABLE 6

First application: resulting estimated percent p-values

B	100	200	500	m	$2m$	$3m$	$4m$	$5m$	$6m$	$7m$	$8m$	$9m$
$\hat{L}_T$	13.00	11.00	10.00	10.70	10.70	10.98	11.23	11.14	10.30	10.60	10.79	10.43
B	$10m$	$11m$	$12m$	$13m$	$14m$	$15m$	$16m$	$17m$	$18m$	$19m$	$20m$	
$\hat{L}_T$	10.63	10.28	10.50	10.25	10.56	10.46	10.79	10.49	10.71	11.06	10.83	

Eidelman et al. (1991) measured the score of an index of lung destruction over 16 smokers (first group) and over 9 nonsmokers (second group).

TABLE 7

Lung destruction index scores

First group							
16.6	13.9	11.3	26.5	17.4	15.3	15.8	12.3
18.6	12.0	24.1	16.5	21.8	16.3	23.4	18.8
Second group							
18.1	6.0	10.8	11.0	7.7	17.9	8.5	13.0
18.9							

Source: Eidelman et al. (1991).

Of course, Eidelman et al. (1991) assumed that the scores of smokers tended to be greater than those of nonsmokers. To tackle this problem we use T test. Table 8 contains the resulting estimated p-values.

As shown in table 8, the estimated p-values ranged from 0.8% and 1.1%. Moreover, they ranged from 0.877% to 0.956% for B greater than 3000. We note again that the estimated p-values are comparable among themselves and they suggest the same decision about H_0. Consideration of all permutations leads to an exact p-values of 0.88%.

TABLE 8

Second application: resulting estimated percent p-values

B	100	200	500	m	$2m$	$3m$	$4m$	$5m$	$6m$	$7m$	$8m$	$9m$
$\hat{L}_T$	1.000	1.000	0.800	1.100	0.950	1.067	0.900	0.980	0.900	0.986	0.950	0.944
B	$10m$	$11m$	$12m$	$13m$	$14m$	$15m$	$16m$	$17m$	$18m$	$19m$	$20m$	
$\hat{L}_T$	0.920	0.936	0.950	0.900	0.886	0.960	0.956	0.877	0.883	0.947	0.930	

5. CONCLUDING REMARKS

By using the approach of Edgington (1995), one can conclude that for permutation testing at $\alpha = 5\%$ 1000 permutations substantially suffice, and that 5000 permutations suffice when $\alpha = 1\%$. Keller-McNulty and Higgins (1987) suggested to use 800 permutations to Monte Carlo estimate the power of a permutation test and 1600 in actual applications (at $\alpha = 5\%$). By reviewing the number of permutations used by various authors, we concluded that generally 1000 permutations are used in simulation study of the power, while a larger value (i.e. 2000 or 5000) is used in actual applications. The results of our Monte Carlo study of the power of T suggest that 500 permutations are substantially enough for testing at $\alpha = 5\%$, and that 2000 iterations are enough when $\alpha = 1\%$.

Summarizing, on the basis of this research, we conclude that for estimating the power of a permutation test at $\alpha = 5\%$ one should use a number of permutations between 500 and 1000 and a number of permutations between 2000 and 5000 when $\alpha = 1\%$. As regards actual applications, we concluded that 5000 permutations are enough for testing at the significance level of 5%, and that 10000 are enough when $\alpha = 1\%$.

Dipartimento di Scienze statistiche "Paolo Fortunati" MARCO MAROZZI
Università di Bologna

RIFERIMENTI BIBLIOGRAFICI

M.J. ANDERSON, P. LEGENDRE (1999), *An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model*, "Journal of Statistical Computation and Simulation", 62, pp. 271-303.

A.J. BAILER (1989), *Testing variance equality with randomization tests*, "Journal of Statistical Computation and Simulation", 31, pp. 1-8.

B.S. CADE, J.D. RICHARDS (1996), *Permutation tests for least absolute deviation regression*, "Biometrics", 52, pp. 886-902.

E.S. EDGINGTON (1995), *Randomization tests*, 3rd Ed., Dekker, New York.

D.H. EIDELMAN, H. GHEZZO, W.D. KIM, M.G. COSIO (1991), *The destructive index and early lung destruction in smokers*, "American Review of Respiratory Disease", 144, pp. 156-159.

M.D. ERNST, W.R. SCHUCANY (1999), *A Class of Permutation Tests of Bivariate Interchangeability*, "Journal of the American Statistical Association", 94, pp. 273-284.

A.F. HAYES (1997), *Cautions in testing variance equality with randomization tests*, "Journal of Statistical Computation and Simulation", 59, pp. 25-31.

M. HOLLANDER, D.A. WOLFE (1999), *Nonparametric statistical methods*, Wiley, New York.

F. KAZI-AOUAL, S. HITIER, R. SABATIER, J.D. LEBRETON (1995), *Refined approximations to permutation tests for multivariate inference*, "Computational Statistics and Data Analysis", 20, pp. 643-656.

S. KELLER-MCNULTY, J.J. HIGGINS (1987), *Effect of tail weight and outliers on power and type-I error of robust permutation tests for location*, "Communications in Statistics – Computation and Simulation", 16, pp. 17-35.

P.E. KENNEDY (1995), *Randomization tests in econometrics*, "Journal of Businness and Economic Statistics", 13, pp. 85-94.

M.J. KIM, C.R. NELSON, R. STARTZ (1991), *Mean revision in stock prices? A reappraisal of the empirical evidence*, "Review of Economic Studies", 58, pp. 515-528.

M. MAROZZI (2002), *A bi-aspect nonparametric test for the two-sample location problem*, "Computational Statistics and Data Analysis", forthcoming.

G. MCQUEEN (1992), *Long-horizon mean-reverting stock prices revisited*, "Journal of Financial and Quantitative Analysis", 27, pp. 1-17.

C.R. MEHTA, N.R. PATEL (1980), *A network algorithm for the exact treatment of the 2xk contingency table*, "Communications in statistics – Computation and Simulation", 9, pp. 649-664.

C.R. MEHTA, N.R. PATEL, P. SENCHAUDURI (1988), *Importance sampling for estimating exact probabilities in permutational inference*, "Journal of the American Statistical Association", 83, pp. 999-1005.

G. NEUHAUS, L.X. ZHU (1999), *Permutations tests for multivariate location problems*, "Journal of Multivariate Analysis", 69, pp. 167-192.

T.W. O'GORMAN (2001), *An adaptive permutation test procedure for several common tests of significance*, "Computational Statistics and Data Analysis", 35, pp. 335-350.

M. PAGANO, D. TRITCHLER (1983), *On obtaining permutation distributions in polynomial time*, "Journal of the American Statistical Association", 78, pp. 435-440.

F. PESARIN (1992), *A resampling procedure for nonparametric combination of several dependent tests*, "Journal of the Italian Statistical Society", 1, pp. 87-101.

F. PESARIN (1994), *Goodness of fit for ordered discrete distributions by resampling techniques*, "Metron", pp. 57-71.

F. PESARIN (1997), *An almost exact solution for the multivariate behrens-fisher problem*, "Metron", pp. 85-100.

F. PESARIN (2001), *Multivariate permutation tests with applications in biostatistics*, Wiley, Chichester.

B. SHIPLEY (2000), *A permutation procedure for testing the equality of pattern hypotheses across groups involving correlation or covariance matrix*, "Statistics and Computing", 10, pp. 253-257.

E.B. SMITH (1998), *Randomization methods and the analysis of multivariate ecological data*, "Environmetrics", 9, pp. 37-51.

E.S. VENKATRAMAN, C.B. BEGG (1996), *A distribution-free procedure for comparing receiver characteristic curves from a paired experiment*, "Biometrika", 83, pp. 835-848.

Y. WAN, J. COHEN, R. GUERRA (1997), *A permutation test for the robust sib-pair linkage method*, "Annals of Human Genetics", 61, pp. 79-87.

RIASSUNTO

Qualche annotazione sulla scelta del numero di permutazioni da considerare per effettuare un test di permutazione

Il principale inconveniente pratico dei test di permutazione è che, eccetto per campioni molto piccoli, il numero di tutte le possibili permutazioni è in genere troppo alto. Sebbene siano adesso a nostra disposizione computer veloci e relativamente poco costosi, questo problema è ancora di interesse, soprattutto per gli statistici applicati. L'idea principale è che non sia necessario calcolare tutte le possibili permutazioni per ottenere una buona stima del *p-value* del test. Il problema viene affrontato approssimando il *p-value* dopo aver considerato un campione casuale di permutazioni. Lo scopo del lavoro è rispondere a questa domanda: quante permutazioni è bene considerare nella procedura di stima del *p-value*? Si suggerisce di usare dalle 500 alle 1000 permutazioni quando si effettua una stima Monte Carlo della potenza di un test di permutazione con un livello di significatività α del

5% e un numero compreso tra 2000 e 5000 quando α = 1%. Si suggerisce inoltre di usare 5000 permutazioni nelle applicazioni empiriche quando α = 5% e 10000 quando α = 1%. Queste indicazioni sono fornite sulle basi di una rassegna di diversi articoli pubblicati, di uno studio di simulazione e di due applicazioni a dati reali.

SUMMARY

Some remarks about the number of permutations one should consider to perform a permutation test

The main practical drawback of permutation testing is that, except for very small sample sizes, the number of all possible permutations is usually impractically large. Although now fast and relatively cheap computing facilities are at our disposal, this problem is still interesting, in particular for applied statisticians. The main idea is that it is not necessary to compute all possible permutations to obtain a reliable *p*-value estimate of the test. To deal with this problem, one may approximate the exact *p*-value of the test by using a random sample from all permutations. The aim of this paper is to reply to this question: how many permutations should be considered in the *p*-value estimation procedure? We suggest to use 500-1000 permutations to estimate the size and power of a permutation test, via Monte Carlo simulations, at the α significance level of 5% and 2000-5000 when α = 1%. Moreover, we suggest to use 5000 permutations in actual applications when α = 5% and 10000 when α = 1%. These suggestions are based on a review of many papers, a simulation study and two applications to actual data sets.