Tutorials in Quantitative Methods for Psychology
2011, Vol. 7(1), p. 15-18.

DOI: 10.20982/tqmp.07.1.p015

Randomization test of mean is compuationally inaccessible when the number of groups exceeds two

Denis Cousineau
Université de Montréal

With the advent of fast computers, the randomization test of mean (also called the permutation test) received some attention in the recent years. Here we show that the randomization test is possible only for two-group design; comparing three groups requires a number of permutations so vast that even three groups of ten participants is beyond the current capabilities of modern computers. Further, we show that the rate of increase in the number of permutation is so large that simply adding one more participant per group to the data results in a computation time increased by at least one order of magnitude (in the three-group design) or more. Hence, the exhaustive randomization test may never be a viable alternative to ANOVAs.

With the advent of fast computers, an alternative to the Analysis of Variance (ANOVA) test of mean is receiving an increased amount of attention. This test, the randomization test (also called the permutation test), was proposed by Fisher in 1935 (Fisher, 1935/1951). It evaluates the significance of the results by examining the way the data might have been if there had been no effect of the conditions. To do so, the data are shuffled across groups and for each permutation, the effect size is computed. Finally, the probability of the observed effect size is assessed with regard to all the possible effect sizes.

The randomization test is a wonderful test because it does not require that the distribution of the population(s) from which the data are sampled be known. In particular, it does not require that the populations be normally distributed, as is the case for the ANOVA test. Further, it does not require that the variance be homogeneous across conditions, another prerequisite for an ANOVA test. Hence, by choosing to perform a randomization test, a researcher is spared two preliminary tests (a test of normality, e.g. the Kolmogorov-Smirnov/Lilliefors test, and a test of homogeneity of variances, e.g. the Levene test, Siegel and Castellan, 1988) whose statistical power are uncharted.

The type-I error rate and the power of the randomization tests were examined in two-group designs using Monte Carlo simulations. Mewhort (2005) varied the asymmetry of the data and found the randomization test to be more powerful than the ANOVA test while maintaining the same type-I error rate. Armstrong, Bors & Cheng (2007) examined the impact of heterogeneous variances and unequal sample sizes and found that the randomization test is both powerful and reliable except when the smaller of the two groups had the largest variance. Because the last condition was extreme (a ratio of 2:1 between the sample sizes and a ratio of 9:1 between the variances), the overall pattern of results is favorable to the randomization test.

Facing all these advantages, the randomization test comes with one difficulty: all the possible permutations of the data between the groups must be examined. The number of permutations increases rapidly with the number of participants. For two groups, it involves picking data to be placed in group 1, the remaining data being placed in the second group. The following computes the number of

Request for reprint should be addressed to Denis Cousineau, Département de psychologie, Université de Montréal, C. P. 6128, succ. Centre-ville, Montréal (Québec) H3C 3J7, CANADA, or using e-mail at Denis.Cousineau@Umontreal.CA. This research was supported in part by le Conseil pour la Recherche en Sciences Naturelles et en Génie du Canada.

Table 1. Number of permutations that have to be examined as a function of the number of groups (2 to 5) and as a number of participants per group (2 to 15).

Number of data per group	Number of groups			
	2	3	4	5
2	6	90	2 520	113 400
3	20	1 680	369 600	168 168 000
4	70	34 650	63 063 000	305 540 235 000
5	252	756 756	11 732 745 024	623 360 743 125 120
6	924	17 153 136	2 308 743 493 056	1 370 874 167 589 326 400
7	3 432	399 072 960	472 518 347 558 400	3 177 459 078 523 411 968 000
8	12 870	9 465 511 770	99 561 092 450 391 000	7 656 714 453 153 197 981 835 000
9	48 620	227 873 431 500	21 452 752 266 265 320 000	19 010 638 202 652 030 712 978 200 000
10	184 756	5 550 996 791 340	4 705 360 871 073 570 227 520	48 334 775 757 901 219 912 115 629 238 400
11	705 432	136 526 995 463 040	1 047 071 828 879 079 131 681 280	125 285 878 026 462 826 569 986 857 692 288 000
12	2 704 156	3 384 731 762 521 200	235 809 301 462 142 612 780 721 600	329 981 831 728 425 465 309 559 251 123 033 960 000
13	10 400 600	84 478 098 072 866 400	53 644 737 765 488 792 839 237 440 000	880 904 182 555 008 823 696 060 440 775 388 083 200 000
14	40 116 600	2 120 572 665 910 728 000	12 309 355 935 372 581 458 927 646 400 000	2 378 829 279 642 818 668 557 063 558 238 537 401 024 000 000
15	155 117 520	53 494 979 785 374 631 680	2 845 616 726 065 971 560 165 538 537 369 600	6 488 042 236 961 891 255 293 961 040 027 911 906 585 223 168 000

permutations:

$$\frac{N!}{(N-n_1)!n_1!} \tag{1}$$

where n_1 is the number of data in group 1 and N is the total number of data ($N = \sum n_i$). For example, with two groups of 10 participants, the number of permutations is 184 756, a number clearly within the grasp of actual computers based on the von Newman architecture.

The general impression is therefore that the randomization test is the test of mean to use with small sample sizes and that they will soon be used routinely.

As we show here, this impression is wrong and based on the fact that only two-group designs were examined. Adding just one more group results in a dramatic increase in the number of permutations, and the numbers are so dramatically high that they will forever be out of reach of computers.

Computing the number of permutations for p groups of n_i data (i= 1, ..., p) involves first selecting n_1 data for group 1, then among the remaining N - n_1 data, selecting n_2 data, and so on. The general formula is

$$\frac{N!}{(N-n_1)!n_1!} \times \frac{(N-n_1)!}{(N-n_1-n_2)!n_2!} \times \cdots \times \frac{(N-n_1-\cdots-n_{p-1})!}{(N-n_1-\cdots-n_p)!n_p!} \tag{2}$$

which contains p factors, but the last one simplifies to 1 as there is only one way to select n_p data among remaining n_p data. This formula simplifies to:

$$\frac{N!}{n_1!n_2!\cdots n_p!} \tag{3}$$

The first factor of Equation (2) is equivalent to Equation (1) in a two-group design. In case where all the groups are of equal size (n_1 = n_2 = … = n_p = n), this formula can be simplified to:

$$\frac{N!}{(n!)^p}. \tag{4}$$

As an illustration, adding a third group of 10 participants brings the number of permutations from 184 756 to more than 5 thousand billion (5.5×10^{12}). It represents an increase in the number of permutations by a factor of 3 million.

If we accept to run permutation test when the number of permutations does not exceed 10 millions (requiring less than an hour on a typical computer) (or 200 millions; requiring less than a day), we would be able to compare (a) two groups of 12 (14) participants, (b) three groups of 5 (8) participants, (c) four groups of 3 (4) participants, (d) five groups of 2 (2) participants. Clearly, with such small sample sizes, performing any test of means is questionable in the first place.

Table 1 lists the number of permutations as a function of

Table 2. The impact of adding one more data in each group on the number of permutation as a function of the number of groups (2 to 5) and the number of data per group including the one added (2 to ∞). A result of 4 means that 4 time more permutations exists.

Final number of data per group	Number of groups					
	2	3	4	5	⋯	p
2	3	15	105	945		
4	3.5	20.625	170.625	1 816.88		
8	3.75	23.718	210.703	2 409.70		
16	3.875	25.335	232.682	2 751.16		
32	3.937	26.162	244.171	2 933.92		
64	3.968	26.579	250.043	3 028.41		
⋯						
∞	4	27	256	3 125	⋯	p^p

the number of groups and of group size (all groups assumed equal) and. By comparison, the number of seconds elapsed since the beginning of the universe is believed to be about 300 000 000 000 000 000 (3×10^{17}).

With the improvement of computers, maybe these figures will soon be accessible? It is possible to show that it is not the case. Suppose that a computer can perform a randomization test with p groups of n participants in an acceptable amount of time. What would be the impact of adding just one more participant in each group? In the limit, adding one more participant to two groups has a consequence to increase the number of permutation by four, so that the computation time will likewise increase by a factor of four. However, for a four-group design, adding one more participant in each group increases the number of permutation by a factor of 256. Following the Moore's law (computers double their processing speed every year and a half), it will take twelve years before this extra participant can again be computed in an acceptable amount of time.

Table 2 lists the factor of increase in the number of permutations when one extra participant is added to each group as a function of the number of data per group (all groups assumed to be of equal size). At the limit, adding one more participant in each of the p groups increases the number of permutations by a factor of p^p.

Discussion

If permutation test is to become an alternative to ANOVA test, we need to reconsider seriously the necessity to explore all the permutations. Hayes, 2000, 1998, 1996, proposed to use only a sample of permutations chosen randomly. He proposed to limit the number of permutations to 5000, but this number could now easily be increased to 50,000, a safer sample size to infer decision thresholds for small probability (e.g. and α of 0.01).

The situation reported above pertains to independent-group designs. In repeated-measure designs, those figures may change drastically. Indeed, to test the significance of a within-subject factor, data need not be moved between participants. This restriction reduces considerably the number of possible permutations as they now increase as a function of the number of participants (n), not the total number of measures (N). The total number of within-subject permutation is given by:

$$(p!)^n \tag{4}$$

where p is the number of repeated measures. For example, for 10 participants measured in three conditions, there would be 60,466,176 possible permutations. This number is large, but not inaccessible to actual computers. It is also 91,803.3 times smaller than if independent groups had been used. Further explorations are required to assess the number of permutations in factorial designs and in designs involving both within and between subject factors.

There is still the possibility that permutation results can be computed efficiently. For example, Gill (2007) found that the 2-group permutation test could be decomposed using Fourier transform into a single difference statistic which can be computed in linear time. Likewise, Mewhort, Johns and Kelly (2010) showed how the Fourier transform could be used with factorial designs in which the number of levels is always 2 (e.g., a 2 × 2 design). However, it seems that a similar result cannot be achieved regarding sums of squares statistics. Hence, in the absence of a similar decomposition for multi-group designs, randomization tests may be forever inaccessible.

References

Armstrong, B. C., Bors, D. A., Cheng, B. (2007, June). *Issues of score distribution: Should the randomization test be employed in all psychological investigations?*. Proceedings of the Canadian Society for Brain, Behaviour and Cognitive Science, Victoria.

Fisher, R. A. (1935/1951). *The design of experiments*. New York: Hafner publishing co.

Gill, P. M. W. (2007). *Efficient calculation of p-values in linear-*

statistic permutation significance tests. *Journal of Statistical Computation and Simulation*, 77, 55-61.

Hayes, A. F. (1996). *PERMUSTAT: randomization tests for the MacIntosh*. *Behavior Research Methods, Instruments, & Computers*, 28, 473-475.

Hayes, A. F. (1998). *SPSS procedures for approximate randomization tests*. *Behavior Research Methods, Instruments, & Computers*, 30, 536-543.

Hayes, A. F. (2000). Randomization tests and the equality of variance assumption when comparing group means. *Animal Behaviour*, 59, 653-656.

Mewhort, D. J. K. (2005). A comparison of the randomization test with the F test when error is skewed. *Behavior Research Methods*, 37, 426-435.

Mewhort, D. K. J., Johns, B. T., & Kelly, M. (2010). Applying the permutation test to factorial designs. *Behavior Research Methods*, 42, 366-372.

Siegel, S. & Catellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw Hill.

Manuscript received 20 December 2010.