ANIMAL BEHAVIOUR, 2000, **59**, 653–656
doi:10.1006/anbe.1999.1366, available online at http://www.idealibrary.com on IDEAL®

COMMENTARY

Randomization tests and the equality of variance assumption when comparing group means

ANDREW F. HAYES
Amos Tuck School of Business Administration and Department of Psychology, Dartmouth College

(Received 24 June 1999; initial acceptance 22 September 1999; final acceptance 4 November 1999; MS. number: AS-1243)

Randomization tests are often advocated as an alternative data analysis method when assumptions of more commonly used inferential statistical procedures are violated (e.g. Edgington 1966; Wampold & Worsham 1986; Blair & Karniski 1993; May & Hunter 1993; Adams & Anthony 1996; Thomas & Poulin 1997). Unlike parametric tests, like Student's *t* test and the traditional analysis of variance (ANOVA) that rely on a mathematically known but assumption-constrained sampling distribution to derive probabilities, randomization tests generate probabilities by repeated 'resampling' of the data and evaluating the obtained result with reference to an empirically derived distribution, called the randomization distribution. This procedure allows the investigator to relax one assumption that can in some contexts invalidate tests on group means such as the *t* test and ANOVA: the assumption of normally distributed parent populations. However, in contrast to the statement made by Adams & Anthony (1996, page 734), and later reinforced by Thomas & Poulin (1997), randomization tests do not necessarily allow one to relax the assumption of conditional equality of population variances (often called 'homoscedasticity') when the hypothesis of interest concerns mean differences. Here, I provide results from a series of simulations showing that when the population variances differ, the use of a randomization test described by Adams & Anthony (1996) to compare group mean differences (using either the mean difference in the two-group case or related statistics such as the treatment sum of squares in the general case) often falsely rejects a true null hypothesis at a rate sometimes far greater than the level of significance chosen for the test.

Because there are a number of existing articles and books that describe in considerable detail the randomization test procedure, Adams & Anthony (1996) and Thomas & Poulin (1997) among them, only a brief introduction to the logic and computation of randomization tests is presented here. For more detail, see, for example, Noreen (1989), Manly (1991), May & Hunter (1993) or Edgington (1995).

Classical tests based on the normal theory model derive *P* values by comparing an obtained test statistic (such as *t* or *F*) to the sampling distribution of that statistic when the null hypothesis is true. If the obtained result yields a small *P* value, this means that assuming the null hypothesis is true, the obtained result or one more discrepant from the null hypothesis is a rather unlikely event. This leads the researcher to reject the null hypothesis as a reasonable description of 'reality' in favour of the alternative hypothesis. The appropriate use of the sampling distribution requires that the assumptions inherent in the mathematics that generate it are at least approximately met. When these assumptions are violated, the sampling distribution (such as the *t* or *F* distribution) may not accurately reflect the realm of possible results assuming a null hypothesis is true, and a decision error can result.

Randomization tests are conceptually identical, but they differ from tests based on the normal theory model in how the *P* value is computed. Instead of relying on a mathematically defined but assumption-constrained sampling distribution, *P* values are derived empirically by a form of 'resampling' of the data without replacement. The obtained result is quantified in some fashion with a test statistic sensitive to the hypothesis of interest. The obtained scores on the dependent variable are then randomly reassigned to groups and the test statistic computed in this new 'sample'. Repeated many times, it is possible to determine how frequently a random reassignment of the observed scores to groups yields a result equal to or more extreme from the null hypothesis than the originally obtained result. This frequency divided by the total number of times this reassignment procedure is undertaken (ideally, 5000 or more times) gives the *P* value

Correspondence: A. F. Hayes, Amos Tuck School of Business Administration, Dartmouth College, 100 Tuck Hall, Hanover, NH 03755, U.S.A. (email: andrew.f.hayes@dartmouth.edu).

Table 1. Type I error rates for the randomization test when sampling from a normal population

Sample size	n_1	n_2	Group 1–Group 2 population variance ratio						
			1:10	1:4	1:2	1:1	2:1	4:1	10:1
40	5	35	<0.001	0.003	0.012	0.048	0.120	0.231	0.358
	10	30	0.002	0.010	0.020	0.052	0.090	0.153	0.224
	20	20	0.054	0.056	0.052	0.048	0.049	0.055	0.054
80	10	70	<0.001	0.002	0.010	0.047	0.141	0.235	0.346
	20	60	0.004	0.010	0.023	0.050	0.093	0.150	0.221
	40	40	0.049	0.050	0.051	0.052	0.052	0.048	0.048
160	20	140	<0.001	0.002	0.011	0.049	0.130	0.231	0.344
	40	120	0.002	0.010	0.022	0.053	0.092	0.148	0.210
	80	80	0.054	0.048	0.049	0.052	0.052	0.049	0.050

Note: 5000 randomizations per sample and 5000 replications.

for the obtained result. If $P \leq \alpha$, the null hypothesis can be rejected. While this might seem like a computationally impractical procedure, there are many statistical programs available that can conduct randomization tests, some of which are in the public domain (see e.g. May et al. 1993; Hayes 1996b, 1998).

However, the use of a randomization test does not necessarily mean that we can relax all assumptions when comparing group means. The procedure that Adams & Anthony (1996) describe allows us to relax only the normality assumption inherent in many tests based on the normal theory model. Importantly, when the groups are originally sampled from populations with different variances, the procedure they describe can yield a liberal test, meaning that it will falsely reject a true null hypothesis at a rate greater than α.

Heteroscedasticity and the Performance of the Randomization Test: Monte Carlo Results

There are mathematical arguments that show that a randomization test can be invalid when the population variances are unequal (e.g. Box & Anderson 1955), but these arguments are largely technical and not especially convincing to the nonmathematician. To illustrate this point differently, we used a set of Monte Carlo studies. The simulations were conducted using the GAUSS program (Aptech Systems 1997). In all, 126 simulations were conducted in a design which orthogonally manipulated four variables: total sample size (three levels: 40, 80, 160) group 1 to group 2 sample size ratio (three levels: 1:1, 1:3, 1:7), group 1 to group 2 population variance ratio (seven levels: 1:10, 1:4, 1:2, 1:1, 2:1, 4:1, 10:1), and population distribution shape (two levels: normal and exponential). In all simulations, the null hypothesis $\mu_1=\mu_2$ was true. In each simulation, two random samples of sizes n_1 and n_2 were taken from a population with a mean of zero that was distributed as either normal or exponential (an extremely skewed distribution). The normal samples were generated with the GAUSS *rndn* function, and the exponential samples were generated with the function $Y = -\ln(U)$, where U is a random uniform variate generated with the *rdnu* function (Ross 1989). To simulate differences in population variance, each score in group 1 was multiplied by the square root of v, where v equalled either 0.1, 0.25, 0.5, 1, 2, 4 or 10, depending on the amount of variance inequality desired for that condition. (When the scores in a population are multiplied by a constant, the variance of the population increases by the square of the constant.) In each simulation, the test statistic used to quantify the obtained result was the sum-of-squares treatment (SS_t), as used in Adams & Anthony (1996). The sum-of-squares treatment is an equivalent test statistic to F and t or the mean difference in the two group case (Edgington 1995). Thus, the results generalize to the use of these test statistics as well. After SS_t was computed, the obtained scores were randomly reassigned to different groups and SS_t recomputed. Over 5000 randomizations, the P value for that repetition was then computed as the number of times that SS_t in a resample equalled or exceeded SS_t in the original sample. In reality, only 4999 randomizations were actually undertaken because the original data is considered a randomization and should be used in the computation of the P value (cf. Ongenha & May 1995). This entire procedure was itself repeated 5000 times in each of the 126 conditions. The type I error rate was computed in each condition as the proportion of times that the randomization test yielded as P value equal to or less than 0.05.

Tables 1 and 2 give the results from the simulations. There are four notable findings. First, when the groups had equal population variances, the randomization test was valid, yielding a type I error rate near 0.05 regardless of whether the population shape was normal or exponential. This reflects the fact that randomization tests, unlike normal theory tests, require no assumptions about distribution shape. Second, so long as the sample sizes were equal, type I errors were kept in control, or nearly so, regardless of whether the population variances were equal or different. Third, the combination of sample-size inequality and variance inequality produced a test that was either liberal or conservative. When the larger group had the larger variance, the randomization test was actually conservative, meaning that the computed P value was too large and therefore the probability of false rejection smaller than the alpha level. However, when the smaller

Table 2. Type I error rates for the randomization test when sampling from an exponential population

Sample size	n_1	n_2	Group 1–Group 2 population variance ratio						
			1:10	1:4	1:2	1:1	2:1	4:1	10:1
40	5	35	0.007	0.012	0.021	0.050	0.116	0.224	0.354
	10	30	0.015	0.020	0.029	0.049	0.110	0.163	0.254
	20	20	0.085	0.063	0.058	0.048	0.048	0.064	0.080
80	10	70	0.002	0.006	0.020	0.045	0.134	0.239	0.349
	20	60	0.008	0.017	0.021	0.053	0.096	0.168	0.227
	40	40	0.067	0.057	0.055	0.052	0.050	0.059	0.067
160	20	140	0.001	0.006	0.015	0.053	0.130	0.233	0.354
	40	120	0.005	0.011	0.022	0.048	0.102	0.157	0.218
	80	80	0.062	0.056	0.056	0.053	0.054	0.050	0.059

Note: 5000 randomizations per sample and 5000 replications.

group had the larger variance, the randomization test was invalid, in many cases with actual type I error rates substantially higher than 0.05 when α=0.05 is used. Fourth, this invalidity was not influenced by absolute sample size (n_1+n_2). Whether the absolute sample size was large or small, type I error rates were inflated when the smaller group was sampled from a population with a larger variance.

Discussion

Tests based on data permutation require an assumption called 'exchangeability'. Exchangeability is very similar to the assumption known to behavioural scientists as the assumption of independently and identically distributed variables, or IID (see for example Draper et al. 1993). When two or more means are being compared, we must assume that deviations between scores and the overall mean are distributed independently and identically across groups (and in many cases, that those discrepancies follow a normal distribution) in order to apply many of the commonly used hypothesis-testing procedures. The randomization test also requires this assumption because in order to compute the *P* value using this procedure, we have to assume that any random reassignment of scores to groups was as likely to have been the obtained result as any other reassignment if the null hypothesis is true. If the scores are independently and identically distributed across groups, and therefore exchangeable, every randomization was equally likely under the null hypothesis. But if this is not true, some randomizations of the data would have been more or less likely if the null hypothesis is true (see Hayes 1996a, for an example) so scores cannot just be randomly reassigned to the groups when the test is conducted. The result can be an inaccurate *P* value and therefore potentially an incorrect decision about the null hypothesis.

Anyone familiar with the performance of the pooled-variance independent groups *t* test in the presence of violations of the homoscedasticity assumption will recognize the results in Tables 1 and 2. With the exception of the invariance of type I error rate to absolute sample size and population shape, the *t* test performs identically. So long as the sample size in the two groups is similar, the independent groups *t* test is valid except when the test variable is extremely skewed. If there are differences in variance combined with differences in sample size, the *t* test is conservative when the larger group has the larger variance, and liberal when the smaller group has the larger variance (see e.g. Boneau 1960). Thus, the randomization test described by Adams & Anthony (1996) offers nothing over the *t* test except for the elimination of the normality assumption. (Hayes 1996a, illustrates the same phenomenon when testing a null hypothesis about zero correlation with a randomization test.) This is of course one important advantage of randomization tests. Unlike Adams & Anthony (1996) and Thomas & Poulin (1997) suggest, however, the randomization test is not necessarily a complete solution to problems produced by the violation of assumptions in the *t* or ANOVA context.

The primary problem with the randomization test procedure described by Adams & Anthony (1996) is that the test statistic used is sensitive to effects other than the one of interest. When comparing means, SS_t, *F*, or *t* or the mean difference (when there are only two groups) all are sensitive to differences in group variance and distribution shape as well as differences in mean. As a result, a randomization test using one of these test statistics tests the null hypothesis that the group distributions are identical in shape and variability as well as location, other than expected random variation. Mielke & Berry (1994) suggest a randomization-based method that they report can validly test for mean differences in the presence of group variance heterogeneity, but little is known about how their procedure actually performs. Currently, there are no perfect solutions to the difficulties produced by variance heterogeneity when comparing group means, although several methods are promising (e.g. Alexander & Govern 1994; Efron & Tibshirani 1998, page 223; Zimmerman & Zumbo 1992; unpublished data).

Still, there are many advantages of randomization tests. They are still quite useful and offer some conceptual and practical advantages when sample sizes are small and the distributions highly skewed. Randomization tests, like other resampling-based methods such as the bootstrap, do not require an assumption-constrained sampling

distribution in order to generate probabilities, and they are useful in situations where no known sampling distribution exists for the statistic used to quantify the effect of interest. And unlike traditional inferential procedures, randomization tests are not based on the traditional concept of random sampling from known populations (cf. Edgington 1966). They thus seem conceptually better suited to many of the kinds of research designs that experimental scientists use most frequently. But they are not a panacea to assumption violations.

References

Adams, D. C. & Anthony, C. D. 1996. Using randomization techniques to analyse behavioural data. *Animal Behaviour*, **51,** 733–738.

Alexander, R. A. & Govern, D. M. 1994. A new and simpler approximation for ANOVA under variance heterogeneity. *Journal of Educational Statistics*, **19,** 91–101.

Aptech Systems. 1997. *GAUSS.* Version 3.2. Maple Valley, Washington: Aptech Systems.

Blair, R. C. & Karniski, W. 1993. An alternative method for significance testing of waveform difference potentials. *Psychophysiology*, **30,** 518–524.

Boneau, C. A. 1960. The effects of violating the assumptions underlying the *t* test. *Psychological Bulletin*, **57,** 49–64.

Box, G. E. P. & Anderson, S. L. 1955. Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society, Series B*, **17,** 1–26.

Draper, D., Hodges, J. S., Mallows, C. L. & Pregibon, D. 1993. Exchangeability and data analysis. *Journal of the Royal Statistical Society, Series A*, **156,** 9–37.

Edgington, E. S. 1966. Statistical inference and nonrandom samples. *Psychological Bulletin*, **66,** 485–487.

Edgington, E. S. 1995. *Randomization Tests*. 3rd edn. New York: Dekker.

Efron, B. & Tibshirani, R. J. 1998. *An Introduction to the Bootstrap*. Boca Raton, Florida: Chapman & Hall.

Hayes, A. F. 1996a. Permutation test is not distribution-free: testing H_o: $\rho=0$. *Psychological Methods*, **1,** 184–198.

Hayes, A. F. 1996b. PERMUSTAT: randomization tests for the Macintosh. *Behavior Research Methods, Instruments and Computers*, **28,** 473–475.

Hayes, A. F. 1998. SPSS Procedures for approximate randomization tests. *Behavior Research Methods, Instruments and Computers*, **30,** 536–543.

Manly, B. F. J. 1991. *Randomization and Monte Carlo Methods in Biology*. London: Chapman & Hall.

May, R. B. & Hunter, M. A. 1993. Some advantages of permutation tests. *Canadian Psychology*, **34,** 401–407.

May, R. B., Hunter, M. & Masson, M. E. J. 1993. *NPStat.* Version 3.7. Victoria, British Columbia: University of Victoria, Department of Psychology.

Mielke, P. W. & Berry, K. J. 1994. Permutation tests for common locations among samples with unequal variances. *Journal of Educational and Behavioral Statistics*, **19,** 217–226.

Noreen, E. 1989. *Computer-intensive Methods for Testing Hypotheses: an Introduction*. New York: J. Wiley.

Onghena, P. & May, R. B. 1995. Pitfalls in computing and interpreting randomization test *p* values: a commentary on Chen and Dunlap. *Behavior Research Methods, Instruments and Computers*, **27,** 408–411.

Ross, S. 1988. *A First Course in Probability*. New York: MacMillan.

Thomas, F. & Poulin, R. 1997. Using randomization techniques to analyse fluctuating asymmetry data. *Animal Behaviour*, **54,** 1027–1029.

Wampold, B. E. & Worsham, N. L. 1986. Randomization tests for multiple baseline designs. *Behavioral Assessment*, **8,** 135–143.

Zimmerman, D. W. & Zumbo, B. D. 1992. Parametric alternatives to the student *t* test under violation of normality and homogeneity of variance. *Perceptual and Motor Skills*, **74,** 835–844.