Behavior Research Methods
2009, 41 (3), 664-667
doi:10.3758/BRM.41.3.664

Randomization tests and the unequal-*N*/unequal-variance problem

D. J. K. Mewhort and Matthew Kelly
Queen's University, Kingston, Ontario, Canada

and

Brendan T. Johns
Indiana University, Bloomington, Indiana

When both the variance and the *N* are unequal in a two-group design, the probability of a Type I error shifts from the nominal 5% error rate. The probability is too liberal when the small cell has the larger variance and too conservative when the large cell has the larger variance. We present an algorithm to circumvent the problem when the smaller group has the larger variance and show, by simulation, that the algorithm brings the error rate back to the nominal value without sacrificing the ability to detect true effects.

Normal-distribution tests for significance (such as the *F* and *t* tests) assume that samples are taken from populations with equal variances. The *F* and *t* tests are both robust to violation of the equal-variance assumption, provided that the sample sizes are equal. When both the variances and the sample sizes are unequal, however, both tests are unreliable: They are too liberal when the small cell has the larger variance and too conservative when the large cell has the larger variance.

Unlike the standard normal-distribution tests for significance, the randomization test depends for its validity on random distribution of the subjects to cells, rather than on sampling from a population with known characteristics (e.g., Edgington, 1995; Manly, 1997). Hence, the statistical question refers to confounding of subject variability with the treatment of interest when the subjects are assigned to conditions. Because the randomization test's chance model provides a closer match to the procedure of a comparative experiment than the sampling model used in survey or quality-control studies, randomization tests are often preferred over parametric tests (e.g., Ludbrook & Dudley, 1998). In addition, because the randomization test's chance model does not depend on a symmetrical error distribution, a randomization test can be more sensitive to true differences between the cells than would the corresponding *F* test (e.g., Mewhort, 2005).

Because its validity depends on random assignment during the conduct of the experiment rather than on the characteristics of the underlying distributions, one might hope that the randomization test would escape the problems associated with heterogeneity of variance and unequal *N*s. Unfortunately, like the *F* and *t* tests, the randomization test is also sensitive to heterogeneity of variance when the cells are unequal in size (e.g., Box & Andersen, 1955; Hayes, 2000). In this article, we present a way to circumvent the problems associated with heterogeneity of variance when the cells are unequal.

Examining the Unequal-*N*/Heterogeneity Problem for the Randomization Test

Hayes (2000) tested unequal groups that differed in variance by a ratio from 1:1 to 1:10. To keep computation to a manageable level, he used an approximation of the full randomization test. With two groups of 10 subjects, 184,756 combinations are required for a full randomization test. Instead of computing the full test, Hayes used 5,000 combinations. With a true null hypothesis, the Type I error rate varied from less than .001 to greater than .35.

The approximate randomization test has the advantage of limiting the computational load, but it introduces a potential problem. Hayes (2000) varied the number of observations per group but fixed the number of combinations in the approximate test at 5,000. As a result, he also manipulated the proportion of the combinations from the full randomization test involved in each approximate test. For example, in a randomization test involving 10 subjects in each of two cells, 5,000 samples are about 3% of the 184,756 combinations required by the full test; with 5 and 35 subjects, by contrast, 5,000 combinations represent less than 1% of the 658,008 combinations required by the full test. An approximate test should be more stable when it includes a larger percentage of the total number of combinations. Hence, when the group size is manipulated while holding the number of combinations constant, the stability of the test is varied systematically.

An approximate test is unsatisfying for a second reason. As Pagano and Tritchler (1983) put the issue, "an unappealing feature of this method is the possibility of differ-

D. J. K. Mewhort, mewhortd@queensu.ca

Table 1
Rate of Rejecting a True Null Hypothesis As a Function of Sample Size and Variance: Randomization Test

				Variance Ratio						
N	n_1	n_2	C(N, n_1)	1:10	1:4	1:2	1:1	2:1	4:1	10:1
16	8	8	12,870	.0744	.0585	.0594	.045	.0616	.0646	.0665
20	8	12	125,970	.0312	.03	.0319	.058	.0921	.0984	.1152
24	8	16	735,471	.0156	.0158	.0181	.0468	.1222	.1304	.1618
28	8	20	3,108,105	.0072	.0095	.0104	.052	.1414	.1577	.1946
32	8	24	10,518,300	.0042	.0052	.0094	.058	.1631	.2024	.2133

ent investigators obtaining different results with the same data" (p. 435).

We replicated Hayes's (2000) demonstration using full randomization tests (i.e., complete enumeration) to ensure that the approximate test did not bias our examination of the *R* test. Table 1 presents the Type I error rate for the randomization test as a function of sample size and the ratio of the variances. In addition, the table shows the number of combinations associated with each combination of sample sizes. The error rates were obtained by computing the randomization test with a true H_0 using Gaussian data. We averaged across 5,000 independent replications of the full randomization test.

As is shown in Table 1, the full randomization test confirmed the basic trends documented by Hayes (2000): When the groups were equal, the Type I error rate was maintained at the 5% it should have been, but when the two groups differed in size, the Type I error rate shifted from the nominal 5%. Specifically, it was too small when the small cell had lower variance and too large when the large cell had the lower variance.

When the larger cell has the larger variance, the test becomes too conservative. Only a very large effect may yield a significant difference, but if the difference is significant, there is little worry that the test has been misleading. By contrast, when the smaller cell has the larger variance, the test is too liberal. As a result, one cannot trust the test's estimate of significance.

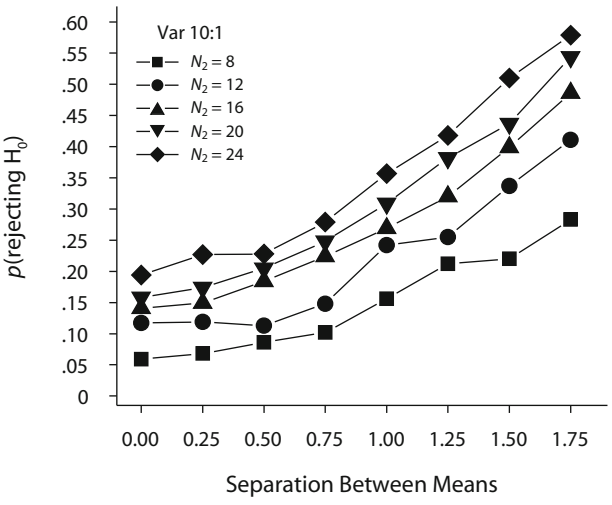

Figure 1. A Monte Carlo simulation showing the ability of the randomization test to detect a true difference as a function of the separation between means when the *N*s are unequal. The data were sampled from a Gaussian distribution, and the variance of the small cell was 10 times bigger than the variance of the larger cell.

Figure 1 extends the example by showing the randomization test's ability to detect a true effect when the smaller cell has the larger variance. For the example shown in Figure 1, N_1 was always 8 and N_2 varied between 8 and 24. The data were Gaussian, and the variance of the smaller cell was 10; the variance of the larger cell was 1. As before, we averaged across 5,000 independent replications of the full-enumeration randomization test.

As is shown in Figure 1, when the H_0 was true and the *N*s were equal, the probability of rejecting H_0 was approximately 5% level. As the difference in *N*s increased, the probability of rejecting a true H_0 also increased; with $N_2 = 24$, the probability reached 20%. When H_0 was false, the ability to detect a true difference in means increased; the rate of increase was greater the larger the difference in the *N*s.

The situation illustrated in Figure 1 presents a frightening prospect: As the difference in the *N*s increases, the test becomes increasingly liberal and the increase exaggerates a small true difference between the means. From an empiricist's perspective, the test for significance appears to be set to deceive systematically.

Given the risk of systematic deception, the best practice is to keep the *N*s equal. Such advice, however, may not always be practical. In some cases, the experimenter may want to minimize the number of subjects in one cell. For example, an experimental drug may be so costly that it can be administered to only a small number of animals, or a manipulation may be so painful that it should be administered to a minimum number of animals. Hence, when designing the study, an experimenter may want to keep one cell smaller than the other. Although the examples shown in Figure 1 may seem extreme, when they occur, their danger is a serious threat to the study's validity.

Circumventing the Unequal-*N*/Heterogeneity Problem When the Large Variance Is in the Small Cell

With $N = 32$ ($n_1 = 8$ and $n_2 = 24$) and variances in the ratio 10:1, the randomization test rejected a true null hypothesis at a rate of about 20%—a full 15% greater than the nominal 5%. To bring the rejection rate back to the nominal 5%, we used a bootstrap-like procedure. Specifically, we took scores at random (without replacement) from the larger group to create a sample of size equal to the smaller group, and computed a standard randomiza-

tion test on the two groups. We repeated the procedure independently 100 times, each time noting whether the equal-*N* test had rejected the H_0 at the 5% level. Finally, we took the proportion of cases (out of 100) in which the equal-*N* tests rejected the H_0 as an estimate of probability of rejecting H_0 for the test as a whole.

Figure 2 shows the probability of rejecting the H_0 as a function of difference in variance and the difference in *N*s. The data were drawn from a Gaussian distribution, and, as before, we averaged over 5,000 independent examples. As is shown in the figure, with variances in the ratio 1:1, the probability of rejecting the null took an increasing ogive-like function as the separation between means increased. The increase was independent of the differences in *N*; that is, the curves for $N_2 = 8$, 12, 16, 20, and 24 fell on top of each other. The curves for variance ratio 3.16:1 increased to about .5, whereas the data for variance ratio 4.47:1 increased to about .3. Taken together, the Monte Carlo results indicate that the resampling technique gives us control over the nominal alpha value: When the H_0 is true, alpha remains at the 5% it should, regardless of the difference in *N*s and variances. Furthermore, the test's sensitivity to real differences between cells depends on the difference in variance—not on the *N*s. In effect, the *N* for the larger cell does not matter—the ability to detect a true difference depends on the *N* of the smaller cell.

Although the resampling technique corrects the too-liberal behavior of the standard test, we do not recommend using it as a matter of course: If the *N*s are different but the variances are roughly equal, the resampling procedure is too conservative. The conservative nature of the resampling technique means that one can trust a significant result, but the trust is gained at the cost of sensitivity to true differences.

We have also examined the efficacy of the resampling algorithm using data drawn from Gaussian distributions

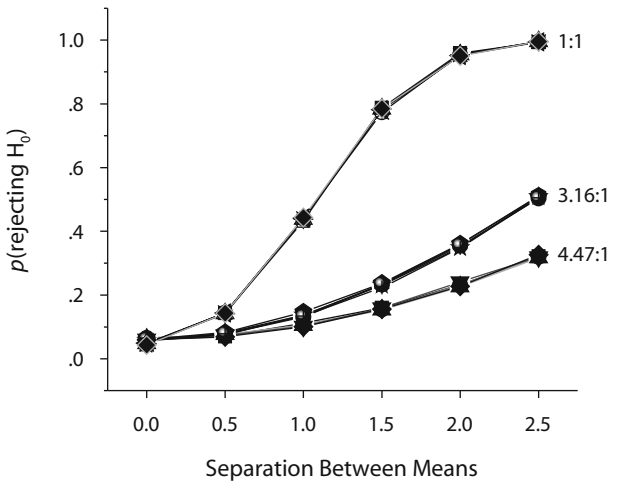

Figure 2. A Monte Carlo simulation showing the ability of the resampling technique to detect a true difference as a function of the separation between means and the ratio of the variances when the *N*s are unequal. The data were sampled from a Gaussian distribution.

using the *F* test—circumstances that meet the distributional assumptions of the *F* test. The results were essentially the same as those we have described for the randomization test. That said, the randomization test is preferred, because the distributional assumptions of the *F* test are hard to guarantee in practice.

Finally, we have also examined cases in which the parent distributions are not Gaussian. When the distributions are skewed and the skew is correlated with the treatment, Mewhort (2005) showed that the randomization test is better able to find a true effect than the corresponding *F* test. When we applied the resampling algorithm to examples based on that situation, the randomization test remained more sensitive than the corresponding *F* test, but the advantage was smaller than Mewhort documented (on the order of 4%–12%). The important point is not the difference in sensitivity favoring the randomization test but the fact that the resampling algorithm works with skewed data.

In summary, the resampling technique gives us control over the nominal alpha level, and the test remains sensitive to true effects. The downside of the procedure is that it requires considerable computing time.

Bringing the Computational Cost Under Control

To make the computing cost clear, consider an example in which the *N*s for the two cells are 10 and 16. If the variances were roughly equal, one would compute a single randomization test. It requires us to examine the difference between cells for each of the $C(26, 10) = 26!/(16! \times 10!) = 5{,}311{,}735$ combinations. If the variances are unequal (with larger variance in the smaller cell), the resampling technique is possible. The resampling technique requires 100 randomization tests; each of the 100 tests involves $C(20, 10) = 184{,}756$ combinations. Hence, the technique requires 18,475,600 recombinations of the data. That number of combinations is substantial, even for a fast desktop computer.

Fortunately, Gill (2007) invented an extremely clever algorithm that brings the computing cost into manageable proportions. His method uses a Fourier expansion to count extreme cases. Briefly, under H_0 all combinations of the data in a randomization test are equally likely. The idea is to compute the proportion of cases that is as extreme as, or more extreme than, the data observed. Gill defined a statistic *T* with an observed value *t*. Hence, the one-tailed probability of interest can be defined as $p(T > t) + p(T = t)/2$.

To compute the probability, Gill (2007) exploited the Heaviside function, *H*,

$$H(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2} & x = 0 \\ 1 & x > 0. \end{cases}$$

Using the Heaviside function, the one-tailed alpha can be defined as

$$\alpha = \frac{1}{N} \sum_{r=1}^{N} H(t_r - t),$$

where t_r is the value on the rth combination. To evaluate alpha, Gill used the Fourier expansion—that is,

$$H(x) = \frac{1}{N} + \frac{2}{\pi} F \sum_{k'=1}^{\infty} \frac{\exp(ikx)}{k},$$

where $k = 2k' - 1$, and $F(a)$ is the imaginary part of a. To ensure the validity of the expansion (i.e., to ensure $a < \pi$), he scaled the data so that the $\max|t - t_r| = 9\pi/10$. The $\max|t - t_r|$ is easy to compute by first ranking the data to obtain the most extreme combination.

Using Gill's (2007) algorithm, the computational cost of computing a randomization test can be brought to a practical level on a newer PC[1]; it is a little more costly than computing an F or t, but it is vastly faster than computing the full enumeration of all combinations.

Conclusions

In the present article, we describe a simple algorithm that avoids the systematic liberal bias associated with standard t, F, and R tests for data with unequal Ns and the larger variance in the smaller cell. Even though the circumstances under which data with these conditions may be created are rare, the experiment may be too costly to rerun, and standard statistical significance tests could easily lead to incorrect conclusions. Hence, we offer the resampling option to experimenters when circumstances require it. Unfortunately, however, the algorithm is not symmetrical: We have not yet discovered an algorithm with which to help reduce the conservative bias when the larger cell also has the larger variance.

AUTHOR NOTE

This research was supported by Grant AP-130 from the Natural Science and Engineering Research Council of Canada to the first author. The junior authors were supported by NSERC summer research scholarships.
M.K. is now at the School of Computing, Queen's University, Kingston, Ontario, Canada. Correspondence concerning this article should be addressed to D. J. K. Mewhort, Department of Psychology, Queen's University, Kingston, ON, K7L 3N6 Canada (e-mail: mewhortd@queensu.ca).

REFERENCES

BOX, G. E. P., & ANDERSEN, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society B*, **17**, 1-34.

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

GILL, P. M. W. (2007). Efficient calculation of p-values in linear-statistic permutation significance tests. *Journal of Statistical Computation & Simulation*, **77**, 55-61.

HAYES, A. F. (2000). Randomization tests and the equality of variance assumption when comparing group means. *Animal Behaviour*, **59**, 653-656. doi:10.1006/anbe.1999.1366

LUDBROOK, J., & DUDLEY, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *American Statistician*, **52**, 127-132.

MANLY, B. F. J. (1997). *Randomization, bootstrap, and Monte Carlo methods in biology* (2nd ed.). London: Chapman & Hall.

MEWHORT, D. J. K. (2005). A comparison of the randomization test with the F test when error is skewed. *Behavior Research Methods*, **37**, 425-435.

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

NOTE

1. Code to compute the randomization test using Gill's (2007) method is available from the first author. The code, in the form of a Fortran-90 module, includes routine for both completely randomized (between-subjects) and repeated measures (within-subjects) tests.

(Manuscript received November 21, 2008;
accepted for publication January 5, 2009.)