Anim. Behav., 1996, **51**, 733–738

Using randomization techniques to analyse behavioural data

DEAN C. ADAMS* & CARL D. ANTHONY†

*Department of Ecology and Evolution, State University of New York at Stony Brook, U.S.A.

†Department of Biology, University of Southwestern Louisiana, Lafayette, U.S.A.

(Received 1 February 1995; initial acceptance 17 May 1995;
final acceptance 22 July 1995; MS. number: A7250*)*

Abstract. Data from behavioural studies are frequently non-normally distributed and cannot be analysed with traditional parametric statistics. Instead, behaviourists must rely on rank-transformation tests, which lose potentially valuable information present in the data. Recently, however, biologists in other disciplines have resolved similar statistical difficulties by using resampling methods. Results from Kruskal–Wallis non-parametric ANOVA and randomization tests were compared for two behavioural data sets. It was found that randomization tests were more powerful than Kruskal–Wallis, and could thus detect smaller effect sizes present in the data. In addition, the variance was calculated around the *P*-value at eight levels of replication ranging from 500 to 10 000, to determine the optimal number of replications for the randomization test. The variance around the *P*-value decreased as the number of replications increased. The *P*-value stabilized at 5000 replications, and thus it is recommended that at least 5000 replications be used for randomization tests on behavioural data.

Animal behaviourists frequently use non-parametric statistical methods when their data are not normally distributed (Martin & Bateson 1986). Most of these methods involve ranking the data sequentially and performing statistics on the ranks (Sokal & Rohlf 1995). Although non-parametric approaches do not require data to be normally distributed, they usually assume that the distributions being compared are symmetrical (Day & Quinn 1989). One drawback of these methods, however, is that some information is invariably lost in the rank-transformation stage, thus making non-parametric approaches less powerful and less desirable than traditional parametric statistics. In recent years, alternative tests have become available which are based on resampling methods. These tests do not make distributional assumptions, and are often more powerful than other non-parametric approaches (Manly 1991). Here we briefly describe resampling approaches and apply randomization tests and non-parametric analysis of variance (Kruskal–Wallis) to two behavioural data sets for comparison.

Resampling statistical techniques are computer-intensive methods that take many samples from the original data and analyse the data based on these samples (Crowley 1992). They can be used to determine confidence intervals for a given statistic (e.g. bootstrap, jackknife and permutation tests), for determining the effect of a stochastic process (Monte Carlo) or even for hypothesis testing (permutation and Monte Carlo). As microcomputers have become faster and less expensive, resampling techniques have been used more and more frequently in biology (for a complete review, see Crowley 1992). All of these techniques can be used by running simple computer programs, many of which have routines that are readily available (see Manly 1991).

For biologists, probably the most widely known of the resampling methods is the bootstrap, which generates a confidence interval for a given statistic by taking *N* samples (of size *N*) with replacement from a data set, and calculates an estimate of the statistic. This procedure allows for the possibility that some values will be sampled more than once, but that others are not

Correspondence: D. C. Adams, Department of Ecology and Evolution, SUNY at Stony Brook, Stony Brook, NY 11794-5245, U.S.A. (email: DCA2879@LIFE.BIO.SUNYSB.EDU). C. D. Anthony is at the Department of Biology, University of Southwestern Louisiana, Lafayette, U.S.A.

0003–3472/96/040733+06 $18.00/0

sampled at all (Manly 1991; Crowley 1992). This resampling is done many times to generate multiple estimates of the statistic to provide a variance for that statistic.

The jackknife also generates an estimate of the test statistic, but it first removes one observation from the data and then calculates the statistic (Crowley 1992). This process is repeated, each time removing a different observation from the data set. From these, a mean value and variance are calculated which are used as the estimate of the test statistic (Manly 1991).

In Monte Carlo methods, a confidence interval is generated based on a particular random or stochastic process (Crowley 1992). A sample statistic is then calculated based on this underlying process (Manly 1991). This process is sampled multiple times to generate a distribution of statistics. If the goal of the researcher is hypothesis testing, the actual test statistic from the raw data is compared to the distribution of sample statistics generated from the Monte Carlo simulation (Crowley 1992).

The permutation test creates a distribution of possible outcomes by reshuffling and sampling one's own data (Edgington 1987; Manly 1991). First, a test statistic is calculated from the raw data. The data are then randomly re-assigned to treatments and another test statistic is calculated (Edgington 1987). By performing this procedure many times, a distribution of possible outcomes is generated (Manly 1991). The test statistic calculated from the raw data is then compared to the frequency distribution of possible outcomes to determine the likelihood that the pattern observed in the data could be found by chance, which yields a probability for the test from which biological inferences can be drawn (Edgington 1987). If the number of possible permutations is small, all permutations can be calculated. This is called an exact permutation test (Sokal & Rohlf 1995). If the number of possible permutations is large, however, a subset is used (Manly 1991). This test is called a randomization test (Manly 1991). Similar inferences can be drawn from both tests.

Although some of the statistical properties of randomization methods are not entirely understood (Crowley 1992), there are several good reasons to consider them in place of traditional methods. First, the assumptions of normality and homoscedasticity are eliminated because randomization tests create their own frequency distributions. Thus, when sample sizes are small or when the underlying distribution is unknown, randomization techniques are especially useful. Randomization techniques also permit inferences on data sets that were not collected completely at random (Edgington 1987). For example, the goal of obtaining 50 random samples from a population of salamanders in a given area is unlikely to be met by locating all possible individuals in that area and then choosing 50; rather, one is more likely to choose the first 50 individuals found. Although this example, like most biological samples, is not of a completely random sample (Manly 1991), randomization tests can still be used to analyse these data. Randomization tests also have higher power than other non-parametric techniques, because they use the actual data rather than ranks (Edgington 1987; Manly 1991). Finally, when there are peculiarities in the data, such as mixed-distribution data sets (where the data for one of the treatments are normally distributed), randomization tests are particularly powerful (Manly 1991).

Although these methods have been used in many areas of biology (Elgar & Harvey 1987; Burt 1989; Harvey & Pagel 1991; reviewed in Crowley 1992), their use in behavioural studies has been surprisingly limited. They have been used to calculate confidence intervals (Ritchie 1990) and probabilities of behaviour (Kiepenheuer et al. 1993), but have not been used in hypothesis testing. Randomization tests were also used by Cantoni (1993) because of frequent repeated measures of the same individual. We suggest, however, that randomization tests should be an important tool in the behavioural ecologist's repertoire, so we demonstrate its use on two behavioural data sets. The first is a mixed distribution data set, where the data in one treatment are normally distributed but the other data are not. The second is a data set that was not significant using Kruskal–Wallis, and is used to show how small effect sizes may not be detected when lower powered, rank-transformation approaches are used.

METHODS

The two data sets used were taken from a larger study on the behavioural interactions of two sympatric species of salamander (Anthony 1995).

These data sets were used to determine whether there were differences in territorial behaviour between the species. In the first experiment, we randomly paired 25 adult males of each species with 25 conspecific and 25 congeneric opponents. We recorded the number of bites delivered by the resident salamander during each trial.

In the second experiment, we recorded the time an individual spent in burrows that had been marked by (1) a conspecific individual, (2) a congeneric individual or (3) the individual being tested. In each trial, the individual had the choice of spending time in burrows marked by either: themselves or a conspecific, themselves or a congeneric, or two burrows marked by themselves (control). We addressed inferences about interspecific territoriality by determining the burrow in which the individual spent the most time.

We tested the assumption of normality using a goodness-of-fit test. We could not use parametric ANOVA because most of the treatments were not normally distributed. In addition, the data contained outliers and were heteroscedastic, which further demonstrated that ANOVA was inappropriate. We therefore used a non-parametric ANOVA, the Kruskal–Wallis test, to determine whether differences between treatments existed.

In addition to performing the Kruskal–Wallis test, we analysed the data using randomization tests. We wrote a program in the SAS programming language that calculated the sums-of-squares treatments (ss_t) for the data, which was stored in an output file. The data were randomly shuffled into the treatment groups, and another sums-of-squares treatments generated. This process was repeated 5000 times to create a frequency distribution of ss_t (the command sequence is available through electronic mail from D. C. Adams). We compared the ss_t for the actual data to the frequency distribution to determine the significance of the test. Because the total sums-of-squares does not change for each iteration, comparing ss_t is equivalent to comparisons of *F*-ratios from ANOVA.

RESULTS

Bite Data

We found a significant difference between treatments using Kruskal–Wallis ($H = 49.12$, $P < 0.001$). The frequency distribution of possible ss_t from the randomization test ranged from 0.25 to 648.59. The raw data ss_t was 648.59, which was larger than 4999 of the randomly calculated ss_t, yielding a *P*-value of $P = 0.0002$ (Fig. 1). This result suggests that the differences between treatments in the original data were not likely to have occurred by chance.

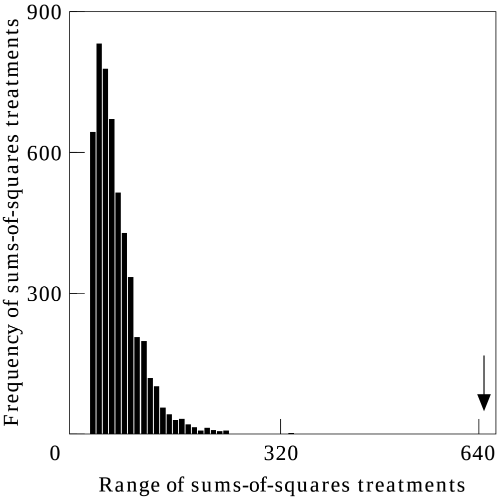

Figure 1. Frequency distribution of sums-of-squares treatment for 5000 iterations based on the number of bite data. The arrow represents the value found from the raw data.

Site Occupation Data

We found no differences between treatments using Kruskal–Wallis ($H = 10.67$, $P = 0.058$). Using a randomization test, however, we found that the raw data ss_t was larger than 4861 of the 5000 randomly calculated ss_t values (Fig. 2), making the test significant at $P = 0.0278$. Thus there were significant differences between treatments.

DISCUSSION

Both the randomization test and the non-parametric Kruskal–Wallis test detected significant differences between treatments in the first data set. We believe the randomization test is more appropriate in this case, however, because the data set was of mixed distribution. Converting the data in the normally distributed treatment to ranks certainly loses information. By using a randomization test, however, such information is

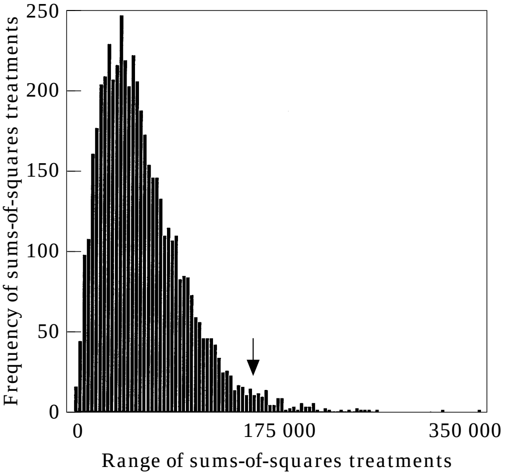

Figure 2. Frequency distribution of sums-of-squares treatment for 5000 iterations based on the site occupation data. The arrow represents the value found from the raw data.

not lost, and the test is more sensitive to treatment effects and thus, more powerful.

In the second data set, randomization tests detected significant treatment effects but Kruskal–Wallis did not. This difference may concern some who believe that using randomization tests will increase the number of published studies that have significant results based on small effect sizes. Although a discussion of effect sizes and the problems of reporting basic statistics is beyond the scope of this paper (see Gurevitch et al. 1992), it is important to realize why non-parametric methods may fail to detect significance. We used Cohen's (1988) power analysis tables to determine the power of both Kruskal–Wallis and randomization tests for small, medium and large effect sizes. The power of the Kruskal–Wallis test was lower than the power of the randomization test at all effect sizes (Table I). Thus, the randomization test is better able to detect differences between treatments, especially when the effect sizes are small. This result is not unexpected, because when data are transformed to ranks, small effects may be masked because of a decrease in power (Sokal & Rohlf 1995).

Another concern when using randomization tests is determining the number of iterations of the randomization procedure. Manly (1991) suggested that 1000 iterations of the data are sufficient to test hypotheses at a 5% level of significance, and 5000 iterations are sufficient to test at a 1% level.

Table I. Power analysis of both Kruskal–Wallis and randomization tests based on small, medium and large effect sizes

Effect size	Kruskal–Wallis power	Randomization power
Small	0.115	0.999
Medium	0.535	0.999
Large	0.897	0.999

Potvin & Roff (1993) also recommend 5000 replications. Because accurate estimates of probability values are crucial for drawing meaningful biological conclusions from the data, we investigated the effect of number of iterations on *P*-values. We performed 10 randomization tests at each of the following replication levels (treatments): 500, 1000, 2000, 3000, 4000, 5000, 7500 and 10 000 replications. For each randomization test a probability value was determined, and a mean and variance around that value were calculated for each treatment. As the number of replications increased, the variance first dropped rapidly and then decreased more slowly (Fig. 3). Because there is so little variance remaining around the *P*-value by 5000 replications, we recommend that at least 5000 iterations be used in randomization tests.

One final concern with using randomization tests is that the frequency distributions used for hypothesis testing are created with the data being tested. This concern is irrelevant for several reasons. First, if the null hypothesis states that there are no differences between treatment means, then random assignment of values to treatments is perfectly valid (Edgington 1987). In fact, this procedure was first proposed by Fisher (1935) who assigned values to treatments in all possible ways and used these values to determine the significance of the test. Second, if one assumes that their sample represents the population in some way (as is always the case), then multiple permutations of the sample data will yield an approximation of the population distribution itself (Manly 1991). Therefore, the generated distribution is one that is theoretically possible for the population, and testing against it rather than the defined, normal distribution, is legitimate.

As microcomputers become faster and more powerful, resampling methods will be used more and more frequently. In many areas of biology, such as phylogenetics, resampling methods are

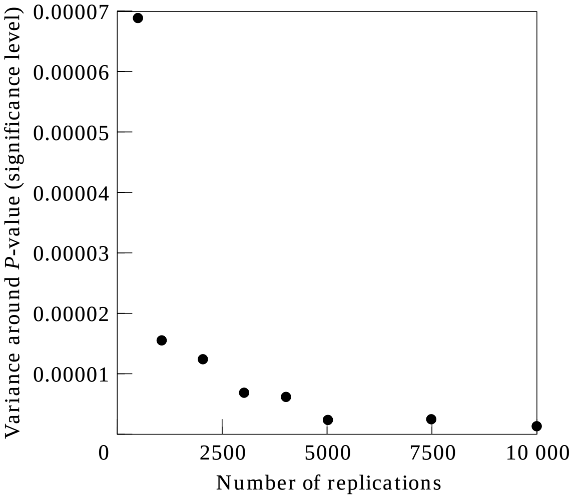

Figure 3. Variance around *P*-values (significance level) based on 10 randomization tests at each level of replication.

already a standard tool in the biologist's repertoire (Crowley 1992). These methods do not rely on a knowledge of the underlying distribution of the data, so they are more versatile than traditional statistical methods. Because behavioural data are usually non-normally distributed, resampling methods could be especially useful to the behavioural ecologist. When effect sizes are small, the coarseness of rank-transformation tests can mask effects present in the data. Randomization tests are able to detect these small effects because of their high power. Randomization techniques also test the hypothesis that there are differences in mean responses between treatments, while nonparametric Kruskal–Wallis tests the hypothesis that the distributions of the treatments are the same (Martin & Bateson 1986). Thus randomization techniques allow one to draw stronger inferences from their statistical conclusions by more directly testing the hypothesis of interest. For these reasons, we recommend that randomization techniques be considered as an alternative test to analyse behavioural data. Also, we recommend that researchers use at least 5000 iterations in their randomization tests to reduce variation around their significance levels.

ACKNOWLEDGMENTS

We thank Paul Leberg for his help in programming SAS to perform the analysis. We also thank F. James Rohlf, Robert Jaeger and three anonymous referees for their comments and suggestions on versions of the manuscript. This work was supported in part by NSF grant DEB-9314081 to Robert Jaeger. This is contribution number 943 from the Program in Ecology and Evolution at the State University of New York at Stony Brook.

REFERENCES

Anthony, C. D. 1995. Competitive interactions within and between two species of *Plethodon* in the Ouachita Mountains: effects of territoriality and parasitism. Ph.D. thesis, University of Southwestern Louisiana, Lafayette.

Burt, A. 1989. Comparative methods using phylogenetically independent contrasts. *Oxf. Surv. evol. Biol.*, **6,** 33–53.

Cantoni, D. 1993. Social and spatial organization of free-ranging shrews, *Sorex coronatus* and *Neomys fodiens* (Insectivora, Mammalia). *Anim. Behav.*, **45,** 975–995.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd edn. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Crowley, P. H. 1992. Resampling methods for computation intensive data analysis in ecology and evolution. *A. Rev. Ecol. Syst.*, **23,** 405–447.

Day, R. W. & Quinn, G. P. 1989. Comparisons of treatments after an analysis of variance in ecology. *Ecol. Monogr.*, **59,** 433–463.

Edgington, E. S. 1987. *Randomization Tests*. 2nd edn. New York: Marcel Dekker.

Elgar, M. A. & Harvey, P. H. 1987. Basal metabolic rates in mammals: allometry, phylogeny and ecology. *Funct. Ecol.*, **1,** 25–36.

Fisher, R. A. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Gurevitch, J., Morrow, L. L., Wallace, A. & Walsh, J. S. 1992. A meta-analysis of competition in field experiments. *Am. Nat.*, **140,** 539–572.

Harvey, H. P. & Pagel, M. D. 1991. *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press.

Kiepenheuer, J., Neumann, M. F. & Wallraff, H. G. 1993. Home-related and home-independent orientation of displaced pigeons with and without olfactory access to environmental air. *Anim. Behav.*, **45,** 169–182.

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

Martin, P. & Bateson, P. 1986. *Measuring Behavior*. New York: Cambridge University Press.

Potvin, C. & Roff, D. A. 1993. Distribution-free and robust statistical methods: viable alternatives to parametric statistics? *Ecology*, **74,** 1617–1628.

Ritchie, M. E. 1990. Optimal foraging and fitness in Colombian ground squirrels. *Oecologia* (*Berl.*), **82,** 56–67.

Sokal, R. R. & Rohlf, F. J. 1995. *Biometry*. 3rd edn. New York: W. H. Freeman.