ANIMAL BEHAVIOUR, 2001, **61**, 79–86
doi:10.1006/anbe.2000.1576, available online at http://www.idealibrary.com on IDEAL®

Assessing the robustness of randomization tests: examples from behavioural studies

PEDRO R. PERES-NETO* & JULIAN D. OLDEN*†
*Department of Zoology, University of Toronto
†Colorado State University, Department of Biology

(Received 8 July 1999; initial acceptance 7 October 1999; final acceptance 21 June 2000; MS. number: A8537R)

Behavioural studies are commonly plagued with data that violate the assumptions of parametric statistics. Consequently, classic nonparametric methods (e.g. rank tests) and novel distribution-free methods (e.g. randomization tests) have been used to a great extent by behaviourists. However, the robustness of such methods in terms of statistical power and type I error have seldom been evaluated. This probably reflects the fact that empirical methods, such as Monte Carlo approaches, are required to assess these concerns. In this study we show that analytical methods cannot always be used to evaluate the robustness of statistical tests, but rather Monte Carlo approaches must be employed. We detail empirical protocols for estimating power and type I error rates for parametric, nonparametric and randomization methods, and demonstrate their application for an analysis of variance and a regression/correlation analysis design. Together, this study provides a framework from which behaviourists can compare the reliability of different methods for data analysis, serving as a basis for selecting the most appropriate statistical test given the characteristics of data at hand.

Behavioural studies often contain data that violate the statistical assumptions of parametric tests (i.e. normality, homogeneity of variances, independence of errors and balanced designs). Consequently, nonparametric approaches have been widely applied in the behavioural sciences rather than parametric approaches. These tests usually impose rank transformations in order to relax some assumptions and obtain the probability distribution for a given test statistic under the null hypothesis (e.g. U, F, χ^2). However, although not readily acknowledged, classic nonparametric tests are also constrained by some assumptions. For small sample sizes, exact distributions can be obtained by finding all possible combinations of ranks, whereas with larger sample sizes asymptotic approximations are necessary (see Mundry & Fisher 1998). In both cases, nonparametric approaches assume that observations are independent and although the samples do not have to follow any particular population, they are assumed to all have the same form or shape (Kruskal & Wallis 1952; Motulsky 1995). When these assumptions are not met there is generally a loss in the power of the test (Kruskal & Wallis 1952).

Correspondence: P. R. Peres-Neto, Department of Zoology, University of Toronto, Toronto, Ontario M5S 3G5, Canada (email: pperes@zoo.utoronto.ca). J. D. Olden is at the Colorado State University, Department of Biology, Fort Collins, Colorado, U.S.A. (email: olden@lamar.colostate.edu).

As a result of the problems associated with parametric and classic nonparametric tests, a great deal of attention has recently been focused on the possible enhanced statistical power that novel distribution-free methods, such as randomization tests, can provide researchers (Potvin & Roff 1993; Adams & Anthony 1996; Manly 1997). Randomization tests are a class of distribution-free methods where the test statistic is contrasted against a null distribution that is empirically constructed using the data at hand. In the broadest sense, a randomization test begins with choosing a test statistic reflecting the question of interest and calculating it for the original data. Next the observed test statistic is contrasted against a null distribution, which is generated by randomly allocating the data and calculating the test statistic a larger number of times in order to nullify the hypothesis in question. Under the null hypothesis the observed test statistic is just one possible value from the null distribution and its likelihood can be evaluated as the proportion of permuted values that are equal to or more extreme than the observed. Since null distributions are generated empirically, they do not make any assumptions regarding the type of population from which the samples were drawn, and the original data is used rather than their ranks (Manly 1995). Interestingly, randomization tests can also use rank-transformed data in order to avoid some of the assumptions associated with classic

nonparametric tests. Furthermore, due to their empirical nature (i.e. entirely based on the sample data), randomization methods offer great versatility, permitting the investigator to assess the statistical significance (i.e. the degree of departure from random) of nearly any parameter of interest. This even includes parameters for which the probability distribution has yet to been investigated. Examples of such plasticity can be found in a variety of biological studies (e.g. Roff & Bentzen 1989; Solow 1990; Jackson 1995; Manly 1998).

Although randomization tests commonly involve large computational effort, they are now being increasingly used due to the advancement and availability of appropriate computer software. As a result, behaviourists are starting to incorporate these methods in their studies (e.g. Johnston & Johnson 1989; Hemelrijk 1990; Dagosto 1994; Adams & Anthony 1996; Thomas et al. 1996; Goldberg & Wrangham 1997; Thomas & Poulin 1997; Mundry 1999). However, the robustness of such methods in terms of statistical power and probability of type I error have seldom been evaluated, being generally assumed that randomization tests are better buffered against nonnormality (e.g. bimodal and highly skewed distributions), homogeneity of variances and problems associated with small sample sizes. The rationale behind this belief is based on a limited number of comparative studies. For instance, Kempthorne & Doerfler (1969) showed that under certain circumstances, randomization methods are more powerful than classic nonparametric methods (also see Mundry 1999 and references therein). Romano (1989) showed that under common statistical assumptions, parametric and randomization tests produce similar levels of power for large sample sizes (Romano 1989), and Bradbury (1987) and Routledge (1997) found that randomization tests can be more robust for small sample sizes. Nevertheless, classic tests can be still more robust than randomization tests when assumptions are only slightly violated (Lindman 1974; Chen & Chen 1998), although the power and type I error levels will be different from what are expected based on the original assumptions (Heilizer 1964). Although randomization tests are essentially free of distributional assumptions, it is not to say that they are equally robust in all situations where statistical assumptions are violated. For instance, Manly (1995) shows that randomization approaches to compare means are sensitive to sample size, unequal variation and the type of distribution. Randomization tests also can be affected by the degree of dependence of observations (Manly 1998). Anderson & Legendre (1999) also show that randomization protocols for multiple regression are sensitive to the type of distribution of the data.

Altogether, there is a lack of sufficient evidence supporting the commonly held notion that randomization tests are more robust than other statistical methods when statistical assumptions are not met (Manly 1997). This probably reflects the fact that power and type I error rates estimates for randomization tests have to be evaluated empirically, requiring customized computer routines and frequently an overwhelming amount of computational effort (e.g. Peres-Neto & Marques, in press). With these constraints, behaviourists are commonly unable to evaluate the robustness of their randomization test and thus are often forced to rely on the results of simulation studies, which might be based on types of data that do not follow the characteristics of their own data. Consequently, the aim of this paper is to address these issues and provide a framework for evaluating the robustness of randomization tests. Specifically, our objective is two-fold: (1) outline a statistical protocol for estimating the power and the probability of type I error of randomization tests; (2) illustrate the implementation of these protocols using behavioural examples. These protocols will provide researchers with the ability to assess and compare the robustness of statistical approaches, thus serving as a basis for selecting the most appropriate test given the characteristics of their data.

STATISTICAL POWER AND DISTRIBUTIONAL ASSUMPTIONS

Statistical power is defined as the probability of a statistical test rejecting the null hypothesis, when the null hypothesis is truly false. It is an important component in the process of statistical hypothesis testing, specifically when designing studies and/or to provide a degree of confidence when a null hypothesis is not rejected. There are two types of power analysis: a priori (or prospective) analysis and a posteriori (or retrospective) analysis. A priori power analysis is conducted before starting a study or experiment to determine the sample size required to obtain acceptable levels of power, or to estimate how large an effect size would have to exist for acceptable power to be achieved. A posteriori power analysis is used mostly in interpreting results of a statistical test that has already failed to reject the null hypothesis. In both cases, a meaningful biological or behavioural effect size has to be established in order to perform power calculations. Although not directly relevant to this study, it is important to emphasize that the goal of a posteriori power analysis is not to verify if the null hypothesis was not rejected because the experiment lacked sufficient power to detect an effect based on the original data. That is uninformative because if the test already failed to reject the null hypothesis, it is inevitable that the statistical test exhibits low power. More appropriately, a researcher should be concerned with the question that, given the original sample size and variance, would it be possible to detect a behaviourally meaningful effect? If the answer is no, then it is possible that the experiment lacked sufficient power to detect an effect due to inadequate sample size or large sample variability (for a more detailed discussion see Peterman 1990; Thomas 1997).

As with any other statistical measure, the power of any test is an estimation because it is based solely on sample values. Moreover, power estimates are also affected by the degree to which the data fail to meet the underlying assumptions on which the particular test was built (Peterman 1990). Since the results from randomization tests are based solely on the characteristics of the data, it is inappropriate to use parametric procedures for measuring statistical power because these procedures

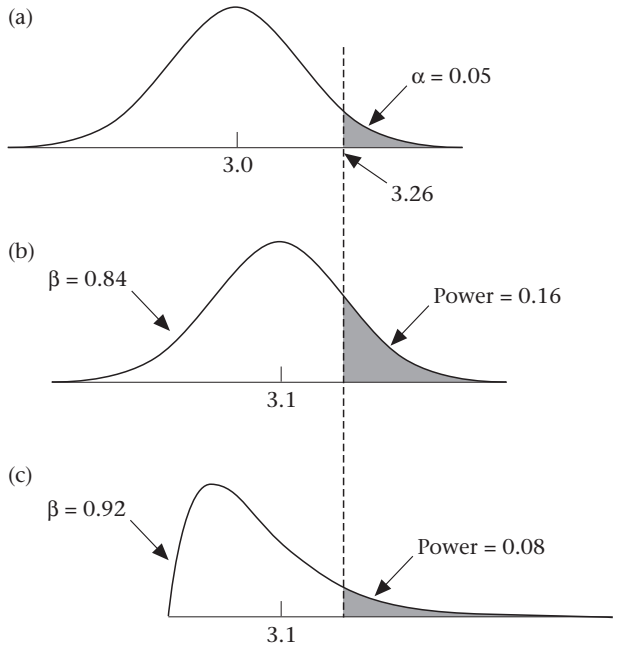

Figure 1. Sampling distribution for (a) testing the null hypothesis that a sample comes from a normal population with a mean equal to 3.0, (b) an alternative (i.e. rejection of the null hypothesis) normal population with a mean of 3.1 and (c) a fictional non-normal population with a mean of 3.1. All populations exhibit the same variance. This example shows how classical statistical power is dependent on the underlying assumptions of the test. Here, power would have been overestimated if a classic approach was applied to the non-normal population (i.e. sampling distribution shown in c).

make assumptions about the data. To illustrate this point, we present a simple example (one-tailed, one-sample hypothesis test) demonstrating how standard procedures for assessing power cannot be applied to data that do not conform to the assumptions of the classic statistical methods. Let us suppose an investigator is studying the mating behaviour of a water strider species, and hypothesizes that the premating struggle (i.e. between the male and female) should be constrained to 3.0 s or less due to increased risk of predation to both individuals during this activity (null hypothesis). A sample of 27 struggle events is observed in order to determine whether the mean struggle time is greater than 3.0 s. Let us say that the mean struggle time for this sample was 3.1 s with a standard deviation of 0.8 s. According to the sampling distribution of means of a normal population (i.e. *t* distribution), based on this data the null hypothesis should not be rejected (Fig. 1a). This is because the critical value (converted from a *t* critical value of 1.706, alpha=0.05) is larger than the sample mean struggle time. Now suppose that the null hypothesis is in fact false and that the true population struggle time is 3.1 s (Fig. 1b). In this case, only sample means greater than 3.26 s would be sufficient to reject the null hypothesis, thus resulting in a statistical power equal to 0.16. In other words, there is a 16% chance that we would reject the null hypothesis, when the null hypothesis is actually false. Finally, suppose again that a sample had a mean struggle time of 3.1 s with a standard deviation of 0.8 s, however, with a positively skewed sampling distribution (Fig. 1c). The smaller right tail of the distribution (i.e. acceptance area) results in a power level equal to 0.08, which is twice as small as the case with a normal sampling distribution. Therefore, based on the parametric procedure for estimating power, we would have overestimated the power of the test for the asymmetric population. This example clearly illustrates why it is invalid to employ parametric methods for assessing the power of studies that examine data that do not meet the underlying statistical assumptions on which the parametric tests are based. The same reasoning applies to estimating type I error rates (i.e. probability of a statistical test erroneously rejecting a null hypothesis), where parametric tests can become more or less conservative (i.e. the true type I error of the test is smaller or larger than the alpha established a priori) depending on how the data depart from the underlying assumptions.

APPROACHES FOR ESTIMATING STATISTICAL POWER

Here, we feel it is important to make a distinction between what we refer to as analytical and empirical methods for estimating the power of a statistical test. Analytical methods are based on the same probability theory and assumptions that are used to identify the appropriate statistical distribution for any traditional statistical method. Several prefabricated tables (e.g. Cohen 1988) and computer software packages (see Thomas & Krebs 1997) based on numerical solutions are available for estimating the power of most commonly used statistical tests. However, when analytical formulae for estimating power have not been derived (e.g. Thomas & Juanes 1996) or when there is interest in assessing the power of a test in which statistical assumptions have been violated, power tables can still be generated using a Monte Carlo approach (e.g. Stephens 1974). In this case, one simulates statistical populations and manipulates them in order to introduce a desirable effect size (e.g. difference between means) or sample variability (e.g. variance). Following this, a large number of samples are taken and the test statistic is calculated each time (Oden 1991). If the effect size is manipulated to be zero (i.e. the null hypothesis is true), the probability of committing a type I error is estimated as the fraction of tests that erroneously rejected the null hypothesis. If the effect size is set different from zero, the proportion of cases in which the null hypothesis was correctly rejected is used as an estimate of statistical power. A comprehensive simulation study, where a large number of scenarios are constructed by manipulating several different combinations of effect sizes, sample variation, alpha levels and sample size, provides a basis for understanding the behaviour of any particular test and for comparing different tests. This aids in identifying the most appropriate statistical test for a particular scenario (i.e. combinations of factors that can be influential to the statistical test).

Since randomization tests are uniquely related to the particular data set being analysed, their statistical power has to be estimated using empirical methods. When conducting Monte Carlo simulations, one has to provide a sampling scheme where the characteristics of the statistical populations being generated are defined. However, even for empirical approaches, it would be infeasible to generate all types of distributions underlying all possible populations found in nature, especially since the true distribution of the population is not known. Moreover, because of their flexibility, different randomization procedures can be used to test the same null hypothesis, but depending on the nature of the data, the results can be conflicting (e.g. Kennedy & Cade 1996; Manly 1997). For these reasons, we argue that appropriate procedures for addressing the robustness of randomization tests should be devised specifically for the characteristics of the data under investigation. To accomplish this we see two possible approaches. In the first approach, one must construct plausible distributions according to the nature of the data being studied, using known distributions with parameter values equivalent to the ones observed in real situations (e.g. normal, uniform, log-normal, exponential and bimodal). Following this, a Monte Carlo experiment is performed. If the test being studied demonstrates reasonable performance in a large number of scenarios, one can assume that it will exhibit a similar performance when applied to the data set of interest. The second approach involves constructing empirical distributions using the data at hand or from previous studies of similar nature. The advantage of this approach is that the analysis is being performed using the specific type of distribution, regardless of what that might be, eliminating the need of generating several plausible scenarios to approximate the distributional properties of the samples. Similarly, when applying empirical distributions, effect sizes and other characteristics can be manipulated and examined using a Monte Carlo experiment.

The difference between the two approaches described above relates in part to the distinction between a priori and a posteriori power analysis. Since the distribution of the population being studied is assumed to be unknown, the most appropriate way for conducting a priori power analysis for randomization tests is to use a large number of different distributions, hoping that at least one will approximate the distribution being studied. On the other hand, in a posteriori analysis, if the null hypothesis is not rejected, one can verify whether the sample size and the alpha level used would be sufficient to provide enough power for detecting a behaviourally meaningful effect size, given the original variation in the data. Nevertheless, much more can be achieved by conducting a posteriori Monte Carlo experiments. By manipulating distributional parameters (e.g. variance and difference between means), a sensitivity analysis can be also conducted to acquire some understanding about the behaviour of the statistical test being applied (e.g. Taylor & Gerrodette 1993). In addition, if the distributional parameters in the data are changed in order to make the null hypothesis true, it can be verified if the probability of type I error corresponds to the alpha level established a priori. Altogether, a well-planned Monte Carlo experiment can provide the researcher with the confidence that the most appropriate test is being applied, as well as aiding in the planning of future studies.

In the following section, we use two empirical examples to illustrate protocols for conducting Monte Carlo experiments to evaluate the power and type I error rates of randomization tests. As a first step, it seems to us that it would be more clear and useful to behaviourists if we restrict our presentation and examples to protocols that address the robustness of randomization tests based on a posteriori analysis. In addition, a priori analysis for randomization procedures usually involves a great deal of computational time and effort in terms of developing computer routines to generate a multitude of different scenarios and combination of factors that might be influential (e.g. Manly 1995), and therefore is beyond the scope of this study. More generally, these protocols not only apply to randomization tests, but also are appropriate for estimating power and type I error rates for any statistical test that do not meet parametric assumptions.

EMPIRICAL EXAMPLES AND MONTE CARLO PROTOCOLS

Here we present in detail the steps involved in a Monte Carlo experiment for evaluating the robustness of a statistical test, which are based on modifications of procedures suggested by Manly (1997). Since any Monte Carlo protocol is highly coupled with the type of the statistical test being conducted, we decided to present examples for an analysis of variance (ANOVA) design and for regression and correlation analyses since they are widely used in behavioural studies. Whatever the statistical test, the general idea is to modify the original data to introduce certain effects. When the effect is set to zero, type I error rates can be estimated, whereas when the effect is set at larger than zero, the power of the test can be assessed. Although we will provide all the details of these Monte Carlo protocols, we feel that they might present some computational difficulties when implemented. For this reason, we developed a flexible computer program that estimates the statistical power and type I error rates based on the protocols described here. The software is available from the authors upon request.

Example 1: Analysis of Variance

The following details a Monte Carlo experiment for assessing the robustness of different tests applicable to an ANOVA design. We use the data of Pitcher & Stutchbury (2000) to address the question of whether foray rates (per hour) differ between fertile, incubating and nestling stages of hooded warblers, *Wilsonia citrina*. Foray rate was measured as the number of times/h that an individual left its territory and entered the core area of an adjacent defended territory. The original sample size, mean and variance for each stage is presented in Table 1. We compared the efficiency of the randomized ANOVA, parametric ANOVA and the Kruskal–Wallis test based on

Table 1. Means and variances (in parentheses) for the original data and the simulation scenarios used for estimating type I error (H_0=true) and power (H_0=false) for testing differences in foray rates (per hour) of fertile, incubating and nestling stages of hooded warblers

Nesting stage	*N*	Original data*	Scenario A H_0=true	Scenario B H_0=false	Scenario C H_0=false
Fertile	15	0.821 (0.187)	0.892	0.821	0.821
Incubation	13	0.841 (0.430)	0.892	0.841	0.841
Nestling	14	1.013 (0.316)	0.892	1.300	1.500

Variances in all scenarios were held equal to the variance in the original data.
*Pitcher & Stutchbury (2000).

the chi-square approximation (Zar 1999). When these tests were applied to the original data, none produced significant results. As is standard practice in conducting power analyses, we evaluated the robustness of these tests by altering the sample means, whereas the variances were held constant according to the original data. To assess type I error rates, we modified the means to be the same as the mean for all samples combined (scenario A: Table 1). Although counterintuitive, due to male parental care in hooded warblers, it seems that males abandon their territories slightly more often during the nestling stage (but not significantly more) than do individuals in the other stages (Table 1). To verify how large the mean nestling foray rate should be to provide adequate power to reject the null hypothesis, we kept the means for the fertile and incubation stages as observed, but modified the mean of the nestling stage to be 1.3 and 1.5 (i.e. scenarios B and C: Table 1), as well as applying the procedure to the original sample means and variances. The Monte Carlo protocol is as follows.

(1) Calculate the mean $\bar{X}_i$ and the standard deviation s_i for each sample.
(2) Standardize each sample (i.e. nesting stage) separately so that their means equal 0 and variances equal 1 using the equation $t_j=(x_j-\bar{X}_i)/s_i$, where t_j is the j_{th} standardized observation and x_j is the j_{th} original observation.
(3) Randomly permute the standardized observations with respect to one another, redistributing them by the samples, respecting their original sample sizes.
(4) Modify the samples in order to change their means (i.e. effect size) according to the planned scenarios (Table 1). This can be accomplished by using the standardization process in step 2, but solving for x_j using the equation $x_j=\bar{X}+t_j s$. For instance, for scenario A, the fertile stage would be modified by $x_j=0.892+t_j$ 0.187 (Table 1).
(5) Conduct the statistical test (i.e. parametric ANOVA, randomized ANOVA and the Kruskal–Wallis test).
(6) Repeat steps (3) to (5) a large number of times (in this study we repeated 1000 times), recording the number of significant outcomes for each test.

In essence, by randomizing standardized observations among samples our protocol mimics the standard protocol for conducting a priori Monte Carlo experiments to assess power as described before, but using the distributional characteristics of the data. In fact, if samples are drawn from normal populations with equal variances, our protocol would provide equivalent results to the ones obtained in standard power tables. However, it is important to note that our protocol makes two assumptions: samples are randomly drawn and all samples follow populations with the same form of distribution (i.e. shape). The rationale and relevance of these assumptions is related to the way observations are standardized and then randomized among samples. If samples do not follow the same distributional type, standardized observations cannot be interchanged without modifying sample properties and hence the analytical outcomes. Since both parametric ANOVA and nonparametric Kruskal–Wallis tests have the same assumptions made here, these assumptions are reasonable for comparing the robustness of these approaches to a randomization test. The randomization test is conducted as follows.

(1) Calculate the original *F* ratio for the standardized data set (F_{obs}).
(2) Randomly permute the observations with respect to one another, recalculating the *F* ratio for the randomized data set (F_{rnd}).
(3) Repeat the randomization a large number of times (in this study we used 9999), where the probability of rejecting the null hypothesis is calculated as: (number of F_{rnd} equal to or larger than F_{obs}+1)/(number of randomizations+1). The addition of 1 in the numerator and the denominator represents the observed *F* ratio for the original data, which is considered as a possible value of the randomized distribution.

For scenario A (Table 1: i.e. no difference among means), we estimated type I error rates as the number of times that a statistical test erroneously reported a significant outcome, and for scenarios B, C and the original data, we estimated statistical power as the number of times that a test correctly rejected the null hypothesis. We calculated both for a variety of alpha levels (i.e. 0.2, 0.1, 0.05, 0.01, 0.005 and 0.001).

Table 2 contains the comparisons of the three statistical approaches for each scenario. The randomization test produced the best type I error rates, with the parametric ANOVA rejecting the null hypothesis less frequently and the Kruskal–Wallis test rejecting it more frequently than expected by the pre-established alpha level. In addition, the Kruskal–Wallis test presented the highest power, followed by the randomization test and then the parametric ANOVA. Considering a power level of 0.8 as

Table 2. Results from the Monte Carlo simulation study evaluating type I error rates (scenario A) and power (scenarios B and C) when testing differences between nesting stages of hooded warblers

	Original data*			Scenario A (H_0=true)			Scenario B (nestling mean=1.3)			Scenario C (nestling mean=1.5)		
Alpha	RN	*F*	KW	RN	*F*	KW	RN	*F*	KW	RN	*F*	KW
0.20	0.343	0.308	0.390	0.211	0.182	0.254	0.868	0.848	0.936	0.991	0.990	0.997
0.10	0.192	0.157	0.219	0.115	0.070	0.131	0.746	0.660	0.848	0.977	0.959	0.992
0.05	0.121	0.077	0.139	0.054	0.039	0.070	0.584	0.490	0.720	0.934	0.893	0.977
0.01	0.034	0.011	0.034	0.013	0.002	0.015	0.341	0.177	0.379	0.801	0.557	0.845
0.005	0.018	0.004	0.020	0.007	0.000	0.006	0.247	0.086	0.264	0.671	0.410	0.737
0.001	0.006	0.000	0.003	0.000	0.000	0.000	0.104	0.013	0.097	0.451	0.142	0.571

RN: Randomization ANOVA; *F*: parametric ANOVA; KW: Kruskal–Wallis test.
*Pitcher & Stutchbury (2000).

adequate (Cohen 1988), the Kruskal–Wallis and the randomization test would be satisfactory to detect a significant foray rate of 1.5 times/h at alpha levels of 0.05 and 0.01, but not for a foray rate of 1.3 times/h. Since forays are believed to be related to extrapair copulation effort in the hooded warbler, rather than searching for food (Pitcher & Stutchbury 2000), it would be beneficial to acquire a larger sample size to verify if in fact the nestling stage has a larger foray rate. This would be particularly desirable since this observation contradicts theoretical expectations that during the nestling stage males should spend less time out of their territory (Westneat et al. 1990). Considering these results, it appears that a randomization test might be a better approach since it exhibits adequate power to the Kruskal–Wallis test (for a rate of 1.5 forays/h), but produces more consistent type I error rates.

Example 2: Regression and Correlation Analysis

Our second empirical example is from Gibbons (1987), who examined the role of juveniles in parental reproductive success of the common moorhen, *Gallinula chloropus*. To test whether a relationship existed, Gibbons calculated the correlation between the number of juveniles to feed the second brood (X) and the number of second-brood chicks reared to independence (Y). Although this study was interested in examining the strength of this correlation, we use his data to address the robustness of the parametric t test and a randomization test to assess the statistical significance of the slope of the regression line. Also, it is important to note that since the correlation coefficient is a standardized form of the simple regression slope, our simulation protocol and results are also relevant for the Pearson product moment correlation coefficient. Furthermore, we found that the data contained a high level of heteroscedasticity, supporting the notion that a nonparametric test might be more appropriate. We propose the following Monte Carlo protocol to determine the most appropriate statistical method to be employed.

(1) Using the original data, construct the regression model and calculate the residuals e (i.e. actual minus predicted number of second-brood chicks reared to independence).

(2) Randomize the residuals e in relation to X and add them to X to create a 'new' set of Y values based on the equation $Y_i=bX_i+e_i$. The slope (b) of the regression line is set as zero to assess type I error rates, and is set to a nonzero value to assess power.

(3) Construct the regression model for the randomized data set and conduct the t test and the randomization test to determine whether the slope of the regression line is significantly different from zero.

(4) Repeat steps (2) and (3) a large number of times (in this study we repeated 1000 times), recording the number of significant outcomes for each test.

Note that at each iteration, the residuals are being reallocated to different X values, and Y values are recalculated each time. The only assumption made is that regardless of the distribution of the residuals, they are independent and therefore interchangeable among the X values. Since independence of residuals is also a parametric assumption, it seems a very appropriate way of assessing the robustness of a parametric compared to a nonparametric approach. If in fact residuals are independent, normally distributed and homoscedastic, parametric tests are not adversely affected. For each generated data set, the randomization test is conducted as follows.

(1) Calculate the slope (b_{obs}) of the regression line based on the original data.

(2) Randomly permute the new values of Y with respect to X, recalculating the slope for the randomized dataset (b_{rnd}).

(3) Repeat the randomization a large number of times (in this study we used 9999), where the probability of rejecting the null hypothesis is calculated as: (number of b_{rnd} equal to or larger than b_{obs}+1)/(number of randomizations+1). The addition of 1 in the numerator and the denominator represents the observed slope for the original data, which is considered as a possible value of the randomized distribution.

We estimated type I error rates as the number of times that a statistical test reported a significant outcome, and we estimated statistical power as the number of times that

Table 3. Results from Monte Carlo simulation study evaluating type I error rates (b=0) and power (b=0.64 and 1.28) for a randomization test (RN) and the parametric t test when testing the statistical significance of the regression slope relating the number of second-brood chicks reared to independence to the number of juveniles to feed the second brood in common moorhens*

	b=0		b=0.64		b=1.28	
Alpha	RN	t test	RN	t test	RN	t test
0.200	0.187	0.184	0.742	0.556	0.992	0.971
0.100	0.095	0.094	0.552	0.391	0.972	0.912
0.050	0.046	0.056	0.389	0.285	0.912	0.827
0.010	0.015	0.015	0.155	0.103	0.654	0.53
0.005	0.005	0.009	0.101	0.064	0.512	0.397
0.001	0.001	0.004	0.026	0.016	0.228	0.172

*Original data from Gibbons (1987).

a test correctly rejected the null hypothesis for a slope of 0.64 and the original slope of 1.28. We calculated both for a variety of alpha levels (i.e. 0.001, 0.005, 0.01, 0.05, 0.1 and 0.2). Table 3 contains the estimates of type I error rates and statistical power. The randomization procedure and the parametric t test have similar type I error rates, being both relatively close to the values expected by the alpha levels. Although both tests presented lower performance in terms of power, the randomization procedure performed slightly better.

Although not directly related to the present study, we would also like to point out that randomization tests based on different test statistics can provide different levels of power. For instance, Adams & Anthony (1996) used the sum of squares between treatments as the test statistic in their randomization test, whereas others have used the F ratio (e.g. this study), mean of squares between treatments and residuals values (see Manly 1997). Another example includes regression analysis where the original residuals are often used in generating random slopes, rather than the original dependent variable (Manly 1997). Finally, the number of permutations are also important, where Manly (1997) suggests at least 1000 for testing a hypothesis using an alpha=0.05 and 5000 permutations when an alpha=0.01 is used. These considerations are important when designing Monte Carlo simulations for assessing the power of randomization tests, and must not be overlooked.

CONCLUSIONS

In the present study we have shown that analytical methods for assessing the robustness of statistical tests are not appropriate when assumptions on which the particular test was built are violated. We have described protocols for an ANOVA design and a regression/correlation analysis that detail the mechanics involved in conducting Monte Carlo experiments to estimate type I error rates and statistical power using parametric, classic nonparametric and randomization approaches. The selection of empirical examples illustrated in this study was deliberate to provide the researcher with a broad demonstration of the application of Monte Carlo protocols presented here. Specifically, the first example examined foray rates in different nestling stages of hooded warblers and was chosen to represent a case where the null hypothesis was not rejected. Consequently, we described a posteriori power analysis to evaluate which test would be more appropriate given the characteristics of the data. Initially, the Kruskal–Wallis seemed to be the best option, but in order to detect a behaviourally meaningful effect, the randomization ANOVA proved to be more appropriate because it not only provides adequate statistical power, but it also exhibits smaller rates of type I error compared with the Kruskal–Wallis test. The second example, which examined the role of parental reproductive success in common moorhens, was chosen to illustrate that power is not the only important component when evaluating the appropriateness of a particular statistical test. As already emphasized, it is important to check whether the type I error probability corresponds to the pre-established alpha level since the power of a test can be adequate, but at the expense of a high probability of committing a type I error. In this case, it was shown that the randomization test and the t test have similar type I error rates, but the randomization test is more powerful.

In conclusion, we believe that researchers should not have any reason to anticipate which statistical test is the most appropriate and powerful for their data. Since parametric and classic nonparametric tests are affected by the degree to which the data do not meet the assumptions (Motulsky 1995; Zar 1999), and randomization tests, by the degree of variation in the data (Manly 1995), it is always desirable to compare different tests to evaluate their relative costs in terms of type I and type II errors (Peterman 1990). Here, we have provided a framework to conduct such comparisons, thus enabling bahaviourists to select the most robust statistical methods for analysing their data. When comparing different tests we stress that type I error rates and statistical power should be computed prior to conducting and interpreting test results so that the investigator is not influenced by any preconception of how the methods will perform according to any particular expectation. This issue is equivalent to establishing alpha levels prior to conducting statistical tests and it should be taken seriously. The statistical method showing the best combinations of type I error rates and power should be conducted and the results reported and interpreted accordingly. Furthermore, in cases where the investigator has already chosen a specific statistical test to analyse the data, our protocols can be used to assess the robustness and reliability of the results. Together, although randomization tests appear to be a powerful alternative to parametric and classic nonparametric statistics, this is not a general rule and their appropriateness should be judged and compared to alternatives. Given that a great deal of effort is spent in collecting data, researchers should dedicate more effort to comparing different statistical methods and choosing the method that is best suited to the particular characteristic of the data at hand.

Acknowledgments

We thank Don Jackson, Len Thomas and two anonymous referees for their insightful comments on this paper; Bryan Manly for his insights regarding the manipulation of effect sizes; and Trevor Pitcher and Bridget Stutchbury for providing the hooded warbler data set. Funding for this project was provided by a CNPq Doctoral Fellowship to P.R.P-N., an NSERC Graduate Scholarship to J.D.O. and an NSERC Research grant to D. A. Jackson.

References

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

Anderson, M. J. & Legendre, P. 1999. An empirical comparison of permutation methods for tests of partial regression coefficient in a linear model. *Journal of Statistical and Computational Simulation*, **62,** 271–303.

Bradbury, I. 1987. Analysis of variance versus randomization tests: a comparison. *British Journal of Mathematics and Statistical Psychology*, **40,** 177–187.

Chen, S-Y. & Chen, H. J. 1998. Single-stage analysis of variance under heteroscedasticity. *Communication in Statistics—Simulation and Computation*, **27,** 641–666.

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

Dagosto, M. 1994. Testing positional behaviour of Malagasy lemurs: a randomization approach. *American Journal of Physical Anthropology*, **94,** 189–202.

Gibbons, D. W. 1987. Juvenile helping in the moorhen, *Gallinula chloropus*. *Animal Behaviour*, **35,** 170–181.

Goldberg, T. L. & Wrangham, R. W. 1997. Genetic correlates of social behaviour in wild chimpanzees: evidence from mitochondrial DNA. *Animal Behaviour*, **54,** 559–570.

Heilizer, F. 1964. A note on variance heterogeneity in the analysis of variance. *Psychological Reports*, **14,** 532–534.

Hemelrijk, C. K. 1990. A matrix partial correlation test used in investigations of reciprocity and other social interaction pattern at group level. *Journal of Theoretical Biology*, **143,** 405–420.

Jackson, D. A. 1995. PROTEST: a procrustean randomization test of community environment concordance. *Écoscience*, **2,** 297–303.

Johnston, R. F. & Johnson, S. G. 1989. Nonrandom mating in feral pigeons. *Condor*, **91,** 23–29.

Kempthorne, O. & Doerfler, T. E. 1969. The behaviour of some significance tests under experimental randomization. *Biometrika*, **56,** 231–248.

Kennedy, P. E. & Cade, B. S. 1996. Randomization tests for multiple regressions. *Communications in Statistics—Simulation and Computation*, **25,** 923–936.

Kruskall, W. H. & Wallis, W. A. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, **47,** 583–621.

Lindman, H. R. 1974. *Analysis of variance in complex experimental designs*. San Francisco: W. H. Freeman.

Manly, B. J. F. 1995. Randomization tests to compare means with unequal variation. *Sankhyā, Series B*, **57,** 220–222.

Manly, B. J. F. 1997. *Randomization, Bootstrap and Monte Carlo methods in Biology*. London: Chapman & Hall.

Manly, B. J. F. 1998. Testing for latitudinal and other body-size gradients. *Ecology Letters*, **1,** 104–111.

Motulsky, H. 1995. *Intuitive Biostatistics*. New York: Oxford University Press.

Mundry, R. 1999. Testing related samples with missing values: a permutation approach. *Animal Behaviour*, **58,** 1143–1153.

Mundry, R. & Fisher, J. 1998. Use of statistical programs for nonparametric tests of small samples often leads to incorrect *P* values: examples from *Animal Behaviour*. *Animal Behaviour*, **56,** 256–259.

Oden, N. L. 1991. Allocation of effort in Monte Carlo simulation for power of permutation tests. *Journal of the American Statistical Association*, **86,** 1074–1076.

Peres-Neto, P. R. & Marques, F. In press. When are random data not random, or is the PTP test useful? *Cladistics*.

Peterman, R. M. 1990. Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Sciences*, **47,** 2–15.

Pitcher, T. E. & Stutchbury, B. J. M. 2000. Extraterritorial forays and male parental care in hooded warblers. *Animal Behaviour*, **59,** 1261–1269.

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

Roff, D. A. & Bentzen, P. 1989. The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. *Molecular Biology and Evolution*, **6,** 539–545.

Romano, J. P. 1989. Bootstrap and randomization tests of some nonparametric hypothesis. *Annals of Statistics*, **17,** 141–159.

Routledge, R. D. 1997. *P* values from permutation and *F* tests. *Computational Statistics and Data Analysis*, **24,** 379–386.

Solow, A. R. 1990. A randomization test for misclassification probability in discriminant analysis. *Ecology*, **71,** 2379–2382.

Stephens, M. A. 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, **69,** 730–737.

Taylor, B. L. & Gerrodette, T. 1993. The uses of statistical power in conservation biology: the vaquita and northern spotted owl. *Conservation Biology*, **7,** 489–500.

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

Thomas, F., Renaud, F. & Cézilly, F. 1996. Assortative pairing by parasitic prevalence in *Gammarus insensibilis* (Amphipoda): patterns and processes. *Animal Behaviour*, **52,** 125–142.

Thomas, L. 1997. Retrospective power analysis. *Conservation Biology*, **11,** 276–280.

Thomas, L. & Juanes, F. 1996. The importance of statistical power analysis: an example from *Animal Behaviour*. *Animal Behaviour*, **52,** 856–859.

Thomas, L. & Krebs, C. 1997. A review of statistical power analysis software. *Bulletin of the Ecological Society of America*, **78,** 126–140.

Westneat, D. F., Sherman, P. W. & Marten, M. L. 1990. The ecology and evolution of extra-pair copulations in birds. *Current Ornithology*, **7,** 331–369.

Zar, J. H. 1999. *Biostatistical Analysis*. Englewood Cliffs, New Jersey: Prentice Hall.