PERSPECTIVE

Permutation tests for univariate or multivariate analysis of variance and regression

Marti J. Anderson

Abstract: The most appropriate strategy to be used to create a permutation distribution for tests of individual terms in complex experimental designs is currently unclear. There are often many possibilities, including restricted permutation or permutation of some form of residuals. This paper provides a summary of recent empirical and theoretical results concerning available methods and gives recommendations for their use in univariate and multivariate applications. The focus of the paper is on complex designs in analysis of variance and multiple regression (i.e., linear models). The assumption of exchangeability required for a permutation test is assured by random allocation of treatments to units in experimental work. For observational data, exchangeability is tantamount to the assumption of independent and identically distributed errors under a null hypothesis. For partial regression, the method of permutation of residuals under a reduced model has been shown to provide the best test. For analysis of variance, one must first identify exchangeable units by considering expected mean squares. Then, one may generally produce either (*i*) an exact test by restricting permutations or (*ii*) an approximate test by permuting raw data or some form of residuals. The latter can provide a more powerful test in many situations.

Résumé : La stratégie la plus appropriée pour générer une distribution de permutation en vue de tester les termes individuels d'un plan expérimental complexe n'est pas évidente à l'heure actuelle. Il y a souvent plusieurs options, dont la permutation restreinte et la permutation d'une quelconque forme des résiduels. On trouvera ici un résumé d'informations récentes empiriques et théoriques sur les méthodes disponibles, ainsi que des recommandations pour leur utilisation dans des applications unidimensionnelles et multidimensionnelles. L'emphase est mise sur les plans complexes d'analyse de variance et de régression multiple (i.e. les modèles linéaires). Dans un travail expérimental, la supposition d'échangeabilité requise pour un test par permutation est assurée par l'assignation au hasard à des unités des divers traitements. Dans le cas d'observations, l'échangeabilité équivaut à supposer que les erreurs, dans une hypothèse nulle, sont indépendantes et distribuées de façon identique. Pour la régression partielle, la méthode de permutation des résiduels dans un modèle réduit s'est avérée la meilleure. Pour l'analyse de variance, il faut d'abord identifier les unités échangeables à l'examen des carrés moyens attendus. Ensuite, il est généralement possible de produire (*i*) un test exact en restreignant les permutations ou alors (*ii*) un test approximatif en permutant les données brutes ou une forme quelconque des résiduels. Cette dernière méthode fournit, dans plusieurs situations, un test plus puissant.

[Traduit par la Rédaction]

Introduction

Biologists and ecologists are faced with increasingly complex circumstances for the statistical analysis of data. In experimental and observational studies, the assumptions that errors are independent and identically distributed as normal random variables with common variance and an expectation of zero, as required by traditional statistical methods, are no longer generally considered realistic in many practical situations (e.g., Clarke 1993; Gaston and McArdle 1994; Anderson 2001). The traditional approach relies on the assumptions to use the statistical distribution of a test statistic, such as *t*, χ^2, or *F*, that is known under a specified null hypothesis, for the calculation of a probability (i.e., a *P* value), commonly relying on tabulated values. An alternative to this traditional approach that does not rely on such strict assumptions is to use a permutation test. A permutation test calculates the probability of getting a value equal to or more extreme than an observed value of a test statistic under a specified null hypothesis by recalculating the test statistic after random re-orderings (shuffling) of the data.

The first descriptions of permutation (also called randomization) tests for linear statistical models (including analysis of variance and regression) can be traced back to the early half of this century in the work of Neyman (1923), Fisher

Received July 18, 2000. Accepted December 21, 2000. Published on the NRC Research Press Web site on March 7, 2001.
J15877

M.J. Anderson.[1] Centre for Research on Ecological Impacts of Coastal Cities, Marine Ecology Laboratories, A11, University of Sydney, Sydney, NSW 2006.

[1] Present address: Department of Statistics, The University of Auckland, Private Bag 92019, Auckland, New Zealand (e-mail: mja@stat.auckland.ac.nz).

DOI: 10.1139/cjfas-58-3-626

(1935), and Pitman (1937*a*, 1937*b*, and 1937*c*). Such tests are computationally intensive, however, and the use of these tests as opposed to the traditional normal-theory tests did not receive much attention in the natural and behavioral sciences until much later, with the emergence of widely accessible computer power (Edgington 1995; Manly 1997; Good 2000).

There is general agreement concerning an appropriate method of permutation for exact tests of hypotheses in one-way analysis of variance (ANOVA) or simple linear regression (or, more simply, tests for the relationship between two variables, e.g., Edgington (1995) and Manly (1997)). This is not the case, however, for tests of individual terms in the context of multiple linear regression (e.g., Kennedy and Cade 1996; Manly 1997) or multifactorial analysis of variance (ter Braak 1992; Edgington 1995; Gonzalez and Manly 1998). Thus, how to do permutations for complex experimental designs is not at all clear. Such complex designs are common, however, in biological and ecological studies, where, usually, several factors are of interest, concomitant environmental variables are measured, or nested hierarchies of sampling at different temporal and spatial scales are necessary.

Recent work in this area has resolved many of these arguments, demonstrating the differences (and (or) similarities) among various approaches, both theoretically (Anderson and Robinson 2001) and empirically (Gonzalez and Manly 1998; Anderson and Legendre 1999; M.J. Anderson and C.J.F. ter Braak, unpublished data). However, a unified treatment of the subject that provides an accessible summary of these recent results and indicates how and when different methods should be used does not currently exist. The purpose of this paper is, therefore, to review and consolidate recent findings and to provide some practical and accessible recommendations for ecologists to construct permutation tests in regression, multiple regression, and analysis of variance, for univariate or multivariate response data.

With the rediscovery and increasing use of permutation tests, praised for being "distribution free," there has sometimes been a failure to recognize assumptions still inherent in these tests for different contexts and statistical inferences. This paper begins, therefore, with a review of the origin and history of permutation tests, clarifying and describing their assumptions in terms of where their validity for statistical inference lies in different contexts. This is followed by a description of the exact permutation test for simple designs, namely, one-way ANOVA and simple linear regression, and a summary of the important considerations for these tests on multivariate response data. Ensuing sections deal in detail with constructing tests for complex designs, with practical recommendations concluding the paper.

Background and rationale for permutation tests

Experimental tests: validity through random allocation

A clear description of a statistical test using permutations of the original observations occurs in the book *Design of Experiments* by R.A. Fisher (1935). Fisher described an experiment by Darwin, first analyzed by Galton, to test the null hypothesis of no difference in the growth rates of *Zea mays* (maize) plants that were self-fertilized versus those that were cross-fertilized. Fisher justifies the validity of the statistical test by virtue of the randomization procedure that is carried out by the experimenter at the beginning of the experiment.

Suppose pairs of seeds, with each pair containing one self- and one cross-fertilized seed (designated, say, by the labels A and B, respectively), are randomly assigned to pairs of plots (in Darwin's experiment there were $n = 15$ such pairs). The determination of which kind of seed will be allocated to which plot in each pair is determined by a random process, such as the flip of a coin. The experimenter then plants the seeds accordingly and, after a period, measures the growth of the plants in each pair (i.e., $y_{\mathrm{A}i}$ and $y_{\mathrm{B}i}$ for each of $i = 1,\ldots,n$ pairs). (Fisher (1935) noted with dismay that such a random allocation procedure was, unfortunately, not in fact carried out by Darwin in his study).

Consider the sum of differences in the measurements obtained between the plants in each pair of plots ($D_{\mathrm{obs}} = \Sigma_{i=1}^{n} d_i$, where $d_i = (y_{\mathrm{A}i} - y_{\mathrm{B}i})$ and $i = 1,\ldots,n$). Consider also the null hypothesis H_0: the two series of seeds are random samples from the same population. Under H_0, the identity of the seeds randomly allocated to plots could have resulted in d_i for any pair being either positive or negative. In particular, what is "randomized" for the experiment a priori—the allocation of seed types in a particular order (A,B) or (B,A) on pairs of plots—is the same thing that can be randomized to realize all the alternative allocations (and associated outcomes) we could have obtained if H_0 were true. If we consider all possible 2^{15} realizations of the experiment consisting of all possible allocations of the seed pairs to plots, we obtain a distribution of D under the null hypothesis. By comparing the observed value D_{obs} with the distribution of D for all possible outcomes under H_0, we can calculate the probability under a true null hypothesis associated with the particular value we obtained: P = the number of values of $|D| \geq D_{\mathrm{obs}}$, divided by the total possible number of values of $|D|$.

This probability is only conditional on the absolute values of the observations (i.e., on the values obtained, not on their sign) and the random allocation of the seeds (A or B) to experimental units (plots). It is the experimental design that is considered random, while the observed values and their associated errors are considered fixed.

Note that the procedure of random allocation at the outset is what ensures the validity of the test. By validity, I mean that the probability of rejecting the null hypothesis when it is true is exactly the level of significance (α) chosen for the test a priori. The test is exact. No assumption concerning the nature of the distribution of errors associated with seeds (or measured outcomes) is necessary. Fisher's view of the validity of the permutation test, as for the familiar normal-theory tests, derived from randomization in experimental design (Fisher 1935, p. 51).

Fisher's idea that the random allocation of treatments in an experimental design ensures the validity of a test by permutation has been adopted and re-iterated by many authors throughout this century (e.g., Scheffé 1943; Kempthorne 1955; Still and White 1981). Kempthorne (1955) gave a useful summary and notation for the idea that the random component in a linear model for a permutation test is the design component, while the errors and observations are treated as fixed.

Not long after Fisher, the notion of a test by permutation was championed in a series of papers by Pitman (1937*a*, 1937*b*, and 1937*c*), who particularly emphasized that the tests were valid because they did not make reference to an unknown population. For purposes of the test, we consider that the sample we have obtained *is* the population of interest. This, he claimed, meant that the test of significance had no distributional assumptions. Kempthorne (1955) and Edgington (1995) have especially popularized Pitman's idea that statistical inferences from randomization tests may not be regarded as extending beyond the sample itself, except when the extra assumption of random sampling from a population holds.

Thus, for statistical inference in an experimental study using permutation of the observations, the only real requirement for the test is that the units have received a random allocation of experimental treatments. For this approach, we consider only the null hypothesis H_0: the treatments have no effect on the units. Together with the initial random allocation, this provides for exchangeability of units without reference to independence or to any distribution. Our inferences are restricted, in this case, to the result in terms of the sample itself. If we intend to make inferences from our sample to a wider population, then an added assumption is necessary, namely, any effect of treatments on the observed units is the same as the effect of treatments on the set of units in a wider specified population. Such an assumption is weaker than the assumption that we actually have a random sample from the population.

As an added note, it is known that, when the assumptions of the normal-theory test are fulfilled, the permutation test and normal-theory test converge (e.g., Fisher 1936). Also, the related approach of bootstrapping provides a test that is asymptotically equivalent to the permutation test (Romano 1988).

Observational tests: validity through exchangeability

The validity of the permutation test does not have the same origin for studies that are observational rather than experimental (e.g., Kempthorne 1966). By an observational study, I mean that the experimenter does not allocate experimental treatments to the units studied, but that a classification of units already exists about which one wishes to test hypotheses. For example, we may wish to test the null hypothesis H_0: there is no difference in the numbers of snails occurring in a midshore region and a high-shore region of a rocky intertidal shore. We can sample the numbers of snails in random representative units (such as quadrats) from each of these areas, but there is no way in which we may "allocate" the classification of being in a high-shore or midshore area to the units randomly. (The important issue of pseudoreplication in observational (mensurative) or experimental ecological field studies (Hurlbert 1984) is not treated in this simplified example.)

In such a situation, we must consider a conceptual notion of *exchangeability* among the units under H_0. If the numbers of snails in the high-shore and midshore areas are really not different, then the values obtained in quadrats labeled "high shore" and quadrats labeled "midshore" are exchangeable. Such units are certainly not exchangeable by virtue of the design, as would have been the case if we somehow could have experimentally randomly "allocated" the quality of being in either the midshore or the high-shore region. This classification exists in nature without our having experimental control over it. For observational data, we must assume exchangeability under H_0.

The assumption of exchangeability under H_0 is equivalent to the assumption of independent and identically distributed (iid) random errors. We do not need to assume *what kind* of population distribution our particular errors are obtained from (normal or otherwise), only that whatever that population is, the errors associated with the units we have are iid. Note that this means that a test by permutation does not avoid the assumption of homogeneity of error variances—the observation units must be exchangeable under the null hypothesis.

It is a common misconception that a permutation test has "no assumptions." However, exchangeability must either be assumed (e.g., for observational data) or it must be ensured by virtue of an a priori random allocation of units in an experiment. The reliance of the permutation test on exchangeable units has been clearly shown in empirical studies (Boik 1987; Hayes 1996). For example, a test using the *t* statistic for a difference in mean between two samples will be sensitive to differences in error variances, even if the *P* value is calculated using permutations (Boik 1987).

Although Kempthorne and Doerfler (1969) suggested distinguishing tests on observational data from experimental tests by calling them "permutation tests" and "randomization tests," respectively, I follow Manly (1997) in choosing not to make this distinction. It is the philosophical context, sampling design, and subsequent inferences that are distinguishable in these two cases, rather than any physical difference in the way the test itself is done. Note also, in this context, that, if we have a manipulative experimental design where random allocation did not form part of the procedure, then the iid assumption for the test by permutation still applies.

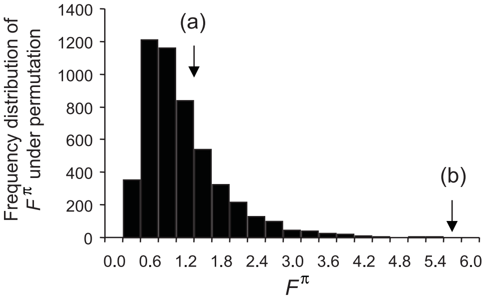
In what follows, I consider that the permutation tests discussed will assume iid random errors (not necessarily normal) that are exchangeable under the null hypothesis, as this provides for the most general application of ideas to either experimental or observational studies. This assumption can be relaxed to a more general assumption of experimental exchangeability under the null hypothesis (allowing the assumptions of independence and homogeneity to be ignored), if this is ensured by physical random allocation in the experimental design a priori.

Simple designs

One-way univariate ANOVA

Consider a one-way ANOVA design with *n* replicate observations (measurements of a response variable) from each of *k* groups. A reasonable statistic to test for differences among the groups is Fisher's *F* statistic. Suppose the null hypothesis is true and the groups are not really different (in terms of the measured variable). If this is the case, then the observations are exchangeable between the different groups. That is, the labels that are associated with particular values, identifying them as belonging to a particular group, could be randomly shuffled (permuted) and a new value of *F* could be obtained, which we will call, say, F^{π}. If we were to calculate

Fig. 1. Frequency distribution of values of the F statistic (F^{π}) for 5000 permutations of a data set with n = 10 replicates in each of three groups. Depending on the observed value of the F statistic obtained from the original data (F_{obs}) by reference to this distribution, the null hypothesis of no differences among groups may be accepted (where $F_{\text{obs}} < F_{\text{crit}}$) (*a*) or rejected (where $F_{\text{obs}} \geq F_{\text{crit}}$) (*b*), where α is the chosen significance level and F_{crit} is the value of F that is equal to or exceeded by 100α% of the values of F^{π} obtained by permutation.



F^{π} for all the different possible allocations of the labels to the observed values, this would give the entire distribution of the F statistic under a true null hypothesis, given the particular data set (Fig. 1).

To calculate a P value for the test, we compare the value of F calculated on the original data with the distribution of values F^{π} obtained for a true null by permuting the labels (Fig. 1). The empirical frequency distribution of F^{π} can be articulated entirely: that is, the number of possible ways that the data could have been re-ordered is finite. The probability associated with the null hypothesis is calculated as the proportion of the F^{π} greater than or equal to F. Thus,

$$P = \frac{(\text{no. of } F^{\pi} \geq F)}{(\text{total no. of } F^{\pi})} \tag{1}$$

In this calculation, the observed value is included as a member of the distribution. This is because one of the possible random orderings of the treatment labels is the ordering that was actually obtained. This P value gives an exact test of the null hypothesis of no differences among groups, that is, the rate of Type I error is exactly equal to the significance level chosen for the test a priori (Hoeffding 1952).

The usual scientific convention of an a priori significance level of α = 0.05 is generally used to interpret the significance of the result, as in other statistical tests. The P value can also be viewed as providing a level of confidence with which any particular opinion about the null hypothesis may be held (e.g., Fisher 1955; Freedman and Lane 1983). The only assumption of the permutation test is that the observations are exchangeable under a true null hypothesis.

With k groups and n replicates per group, the number of distinct possible outcomes for the F statistic in a one-way test is $(kn)!/[k!(n!)^k]$ (e.g., Clarke 1993). In reality, it is usually not practical to calculate all possible permutations, because with modest increases in n this becomes prohibitively time-consuming. A P value can also be obtained by taking a large random subset of all possible permutations to create the distribution (Hope 1968). Increasing the number of permutations increases the precision of the P value. Manly (1997) suggested using at least 1000 permutations for tests with an α level of 0.05 and at least 5000 permutations for tests with an α level of 0.01.

Sometimes the total number of permutations is greatly reduced in exact permutation tests for terms in complex ANOVA designs. This occurs when permutations are restricted to occur only within categories of other factors. If, for example, there are only 20 possible permutations, then the smallest P value that can be obtained is 0.05. Although such permutation tests still have exact Type I error, their power is extremely low (M.J. Anderson and C.J.F. ter Braak, unpublished data). If the total possible number of permutations is less than 100 for the exact test, it would be a good idea to consider using an approximate permutation test (as described in Partial regression and Factorial (orthogonal) designs, below).

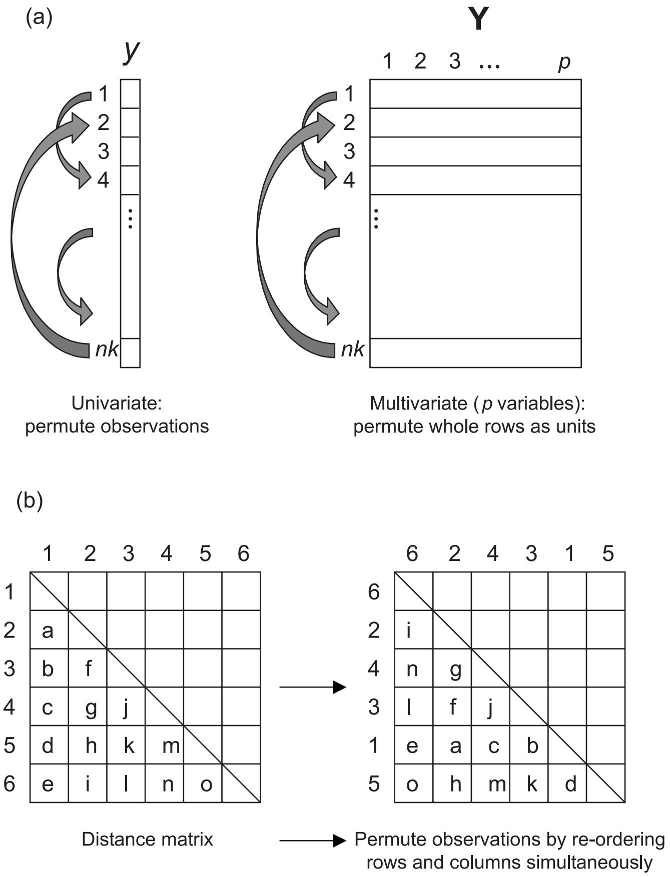
The one-way multivariate case

Consider the one-way analysis of a multivariate set of measurements on p variables (e.g., species) for each of the n replicate observations in each of the k groups. Let the data be organised so that the nk observations are rows and the p variables are columns in a matrix. For the multivariate test of differences among groups by permutation (e.g., using the R statistic of Clarke 1993), the assumption of exchangeability applies to observations (rows) not variables (columns). Permutation in the multivariate context means randomizing whole rows (Fig. 2*a*). The numbers in the raw data matrix are not shuffled just anywhere. An entire row of values is shuffled as a unit. This is important for two reasons. First, the units that are exchangeable under the null hypothesis of no treatment effect are the replicate observations, not the numbers of individual species in each observation. Second, in multivariate analysis, the species will probably have some non-zero correlation, that is, some relationship with one another. They are not likely to be independent, so they cannot be considered to be exchangeable in the sense of a permutation test.

If the multivariate analysis is based on a distance matrix (i.e., a measure of dissimilarity or distance calculated between every pair of values, as in the method of Clarke (1993)), then the permutation of observations can be achieved by permuting the rows and columns of the distance matrix simultaneously (Fig. 2*b*). As an example, consider a new ordering of six rows of an original data matrix from {1, 2, 3, 4, 5, 6}, the original ordering, to {6, 2, 4, 3, 1, 5}, a new ordering under permutation. One does not need to recalculate the (6 × 6) distance matrix, because distances (or dissimilarities) between pairs of observations have not changed with the permutation. What has changed is the ordering of the observations (i.e., their labels). In other words, the observations are exchangeable, not the distances. So the new distance matrix under this particular permutation has simultaneously re-ordered rows as {6, 2, 4, 3, 1, 5} and columns as {6, 2, 4, 3, 1, 5}.

A further consideration for permutation tests with multivariate data is that the assumption of iid observations (exchangeability) means that the test will be sensitive to differences in the multivariate dispersions of observations (e.g., Clarke 1993). Let each observation be a point in the space of p dimensions (variables or axes). Variability in the values taken along any of these axes corresponds to the dispersion

Fig. 2. (*a*) Schematic representation of the relationship between permutation tests for univariate versus multivariate data. In each case, the exchangeable units are the observations but, for multivariate data, the observations are whole rows of information for *p* variables. (*b*) Schematic representation of the permutation of multivariate observations done directly on the distance matrix. The numbers correspond to the original sampling units (rows and columns) and the letters correspond to distances between pairs of units.


(a)
y
1
2
3
4
nk
Univariate: permute observations
Y
1 2 3 ... p
1
2
3
4
nk
Multivariate (p variables): permute whole rows as units
(b)
1 2 3 4 5 6
1
2 a
3 b f
4 c g j
5 d h k m
6 e i l n o
6 2 4 3 1 5
6
2 i
4 n g
3 l f j
1 e a c b
5 o h m k d
Distance matrix
Permute observations by re-ordering rows and columns simultaneously

of points in multivariate space. Just as a univariate test by permutation will be sensitive to deviations from the assumption of homogeneity of variances (Boik 1987; Hayes 1996), so too will a multivariate test be sensitive to deviations from the assumption that the observations in different groups have similar dispersions (Clarke 1993; Anderson 2001). The interpretation of significant results must be treated with caution in this regard.

Equivalent test statistics under permutation

Some researchers have highlighted that simplified test statistics can give equivalent *P* values under permutation, thus saving computational time (Edgington 1995; Manly 1997). Simplification of the test statistic is done by considering components of the calculation that can be dropped because they remain the same for all permutations or by considering a simpler test statistic that is monotonically related to the original test statistic.

For example, in the case of the univariate one-way *F* ratio, the degrees of freedom do not change, no matter what the ordering of the data is, nor does the total sum of squares. Thus, either the among-group or within-group sum of squares (SS_A or SS_W, respectively) is monotonically related to the *F* ratio. This means that the position of the observed value of SS_A relative to the distribution of SS_A^π under permutation is the same as that of the observed value of *F* relative to the distribution of F^π. SS_A can therefore be used as an equivalent (and more simply calculated) statistic to obtain a *P* value.

In cases of more complex designs, the idea of an equivalent test statistic under permutation becomes less useful. In many cases, it is not possible to simplify the test statistic much beyond what is called a “pivotal” statistic. A pivotal statistic is defined as follows: in general, to find a confidence interval for a parameter θ, based on an estimator $T(\mathbf{y})$, where $\mathbf{y}$ denotes a vector of *n* independent random variables, a pivotal statistic τ has the following properties: (*i*) τ is a function of $T(\mathbf{y})$ and θ, (*ii*) τ is monotonic in θ, and (*iii*) τ has a known sampling distribution that does not depend on θ or on any other unknown parameters. Statistics like *t*, *F*, and the correlation coefficient, *r*, (or its square) are pivotal. A slope coefficient (in regression) or a sum of squares or mean square (in ANOVA) is not pivotal, because each depends on the value of some unknown parameter(s) for any particular data set, even if the null hypothesis is true. Approximate permutation tests in complex designs, described in more detail below, generally do not work, if a nonpivotal statistic is used (Anderson and Legendre 1999). In practice, it is always wise to use a pivotal statistic for a permutation test, especially for complex designs. Statistics like *t*, *F*, or r^2 also have the advantage of being interpretable in themselves: their value has some meaning that can be compared across different studies.

Simple linear regression

Consider *n* pairs of observations of a random variable *Y* with fixed values of a variable *X* collected to test the null hypothesis of no (linear) relationship between *Y* and *X*. I restrict attention to the case of Model 1 regression (sensu Sokal and Rohlf 1981). Given the linear model of $Y = \mu + \beta X + \varepsilon$, the null hypothesis is that the slope $\beta = 0$. An appropriate test statistic for the two-tailed test is the square of the least-squares correlation coefficient, r^2.

The rationale for the permutation test for simple linear regression follows the same general rationale as that for the one-way ANOVA case. Namely, if the null hypothesis of no relationship between the two variables is true, then the *n* observations of *Y* could have been observed in any order with respect to the *n* fixed values of *X*. An exact test is therefore given by recalculating the test statistic (called, say, $(r^2)^\pi$) for each of the possible re-orderings (permutations) of *Y*, with the order of values of *X* remaining fixed. The probability for the test is obtained by comparing the original observed value r^2 with the distribution of values of $(r^2)^\pi$ obtained for all permutations. The *P* value is that fraction of the permutations for which $(r^2)^\pi \geq r^2$.

In this case, there are *n*! unique possible permutations, which clearly gets to be very large with moderate increases in *n*. In practical terms, the same considerations are applied as for ANOVA: a random subset of all possible permutations is obtained and the *P* value calculated as in eq. 1, but for r^2. Note that the only assumption of the test is that the errors are iid or, more generally, under a true null hypothesis, the observations Y are exchangeable. The errors do not have to be normally distributed. In a similar fashion to the ANOVA case, the assumption of exchangeability is assured if the n units were randomly allocated to each of the values of X at the beginning of the experiment.

Complex designs—multiple and partial regression

Multiple regression

Let Y be a biological or ecological response variable (like the growth of bivalves) and X and Z be some independent variables (such as particulate organic matter (POM) and temperature, respectively) predicted under some hypothesis to affect Y. Say that the growth of bivalves is measured in each of n different combinations of POM (Z) and temperature (X), where these have either been fixed experimentally or, more generally, are simply measured in situ along with Y. The linear model is

$$(2) \qquad Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon'$$

If X and Z were measured with error, this is not strictly a Model I regression. Data can, however, be analyzed as a fixed case (Model I), provided the tests for significant relationships among the variables are interpreted as conditional on the values of X and Z actually observed (e.g., see Neter et al. 1996 for details).

To test for the relationship of Y versus X and Z together, the null hypothesis is H_0: $\beta_1 = \beta_2 = 0$. An appropriate test statistic is the coefficient of multiple determination, R^2 (e.g., Sokal and Rohlf 1981; Neter et al. 1996). To do the test by permutation requires careful consideration of what is exchangeable under the null hypothesis. If H_0 is true, the model becomes $Y = \beta_0 + \varepsilon'$. So, under the assumption that the errors are iid, the observations (Y) are exchangeable under the null hypothesis. That is, if Y has no relationship with X and Z together, the values obtained for Y could have been observed in any order relative to the fixed pairs of values of X and Z. So, an exact P value for the test in multiple regression can be obtained by randomly permuting Y, leaving X and Z fixed, and recalculating $(R^2)^\pi$ under permutation, as was done for simple linear regression.

Partial regression

When several independent variables are measured (or manipulated), the researcher is usually interested in something more specific than the general test of multiple determination. For the example involving bivalves, the purpose may be to investigate the relationship between the growth of bivalves and POM, given any effects of temperature. The null hypothesis is H_0: $\beta_2 = 0$. The hypothesis is that POM explains a significant proportion of the variability in bivalve growth over and above any effect of temperature. The effect of temperature (if it is there) must be “removed” to allow a test for any relationship between growth and POM. This is a partial regression with temperature as a covariable.

An appropriate test statistic for the relationship between Y and Z given X is the squared partial correlation coefficient

$$(3) \qquad r^2_{YZ,X} = \frac{(r_{YZ} - r_{XZ}\,r_{YX})^2}{(1 - r_{XZ})^2(1 - r_{YX})^2}$$

where r_{YZ} is the simple correlation coefficient between Y and Z, r_{XZ} is that between X and Z, and so on. For a permutation test in this situation, consider what the model would be if the null hypothesis were true:

$$(4) \qquad Y = \beta_0 + \beta X + \varepsilon$$

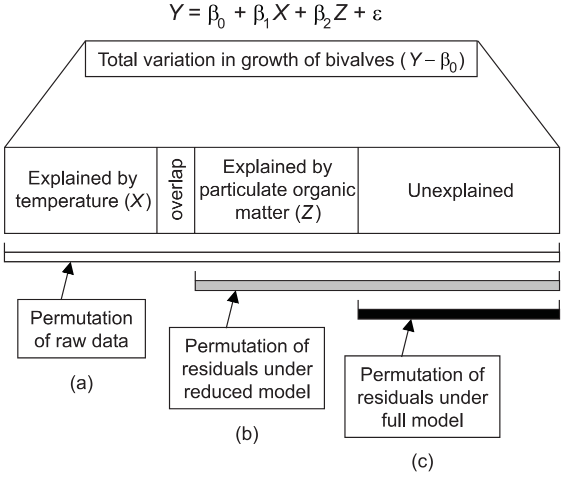
where β is the simple regression coefficient of Y versus X. Under the assumption of iid errors, what are exchangeable under the null hypothesis are not the observations Y, but rather the errors ε after removing the effect of X. That is, the exchangeable units under H_0 are $(Y - \beta_0 - \beta X)$. These are the units that should be shuffled for the exact permutation test. Unfortunately, the parameters β_0 and β are not known. Thus, for the test of a partial regression (e.g., a test of the relationship between Y and Z given X), no exact permutation test is possible (Anderson and Robinson 2001).

Several approximate permutation methods for this situation have been suggested (e.g., Freedman and Lane 1983; ter Braak 1992; Manly 1997). These methods (Fig. 3) have been compared theoretically (Anderson and Robinson 2001) and in empirical simulations (Anderson and Legendre 1999).

The idea of Freedman and Lane (1983; Fig. 3*b*) is the most intuitively appealing. The essence of their idea is this: although β_0 and β are unknown, they can be estimated by the intercept (b_0) and the regression coefficient (b) from the simple linear regression of Y versus X. The residuals of this regression $R_{Y,X} = (Y - b_0 - bX)$ approximate the errors (ε) that are exchangeable under the null hypothesis. That is, the residuals of the regression of bivalve growth versus temperature can represent the observations after “removing” the influence of temperature. These are exchangeable against fixed values of POM (Z) under the null hypothesis and are, therefore, permuted. For each of the n! possible re-orderings, the value of the squared partial correlation coefficient $(r^2_\mathrm{F})^\pi$ is calculated for the permuted residuals $R^\pi_{Y,X}$ versus Z given X (the subscript F indicates the Freedman and Lane (1983) method). Note that X and Z are not shuffled but remain in their original order. This ensures that the relationship (if any) between X and Z remains intact and does not affect the test for a partial effect of Z on Y given X (e.g., Anderson and Legendre 1999; Anderson and Robinson 2001). The P value is calculated as the proportion of values $(r^2_\mathrm{F})^\pi$ that equal or exceed the value of $r^2_{YZ,X}$ for the original data. This method of permutation is sometimes called permutation “under the null model” or “under the reduced model” (see Anderson and Legendre 1999). Another important aspect of this method to note is that it introduces a further assumption for this particular permutation test: that is, the relationship between Y (bivalve growth) and X (temperature, the covariable) conforms to a linear model. That is, the effect of temperature has only been “removed” in terms of its linear effect on Y.

This permutation test is not exact, but it is asymptotically exact. An asymptotically exact test has a Type I error (the probability of rejecting the null hypothesis when it is true) that asymptotically approaches the significance level chosen for the test (e.g., $\alpha = 0.05$), with increases in n. The larger the sample size, n, the closer the estimate b will be to the

Fig. 3. Diagram of the partitioning of variability of a response variable (*Y*, bivalve growth) in terms of two independent variables (temperature, a covariable *X*, and particulate organic matter, *Z*). The three methods of approximate permutation for a test of the partial regression of *Y* on *Z* given *X* correspond to permutation of different portions of the variability in the response variable.


Y = β0 + β1X + β2Z + ε
Total variation in growth of bivalves (Y − β0)
Explained by temperature (X)
overlap
Explained by particulate organic matter (Z)
Unexplained
Permutation of raw data
(a)
Permutation of residuals under reduced model
(b)
Permutation of residuals under full model
(c)

true parameter β, so the better the residuals $R_{Y,X}$ will be at estimating the exchangeable errors, ε. Anderson and Robinson (2001) have shown that this method (Freedman and Lane 1983) comes the closest to a conceptually exact permutation test and Anderson and Legendre (1999) have shown that it gives the best empirical results, in terms of Type I error and power, compared with other methods.

One alternative method is to permute the raw observations, *Y*, while keeping *X* and *Z* fixed (Fig. 3*a*; see Manly 1997). The partial correlation coefficient is then recalculated for each permutation. The essential requirement for this to work is that the distribution of the observations (*Y*) must be similar to the distribution of the errors (ε) under the null hypothesis with a fixed effect of *X* (Kennedy and Cade 1996; Anderson and Legendre 1999). This may not be true if there is an outlier in *X*, the covariable. For example, consider that one of the values of temperature (*X*) is very high relative to the others. Consider also that temperature does affect bivalve growth (say, β = 1.0), so that the growth of bivalves for that particular temperature is quite large relative to the others. Let the null hypothesis of no effect of POM on growth be true and say the errors follow a normal distribution. Since $Y = \beta_0 + \beta X + \varepsilon$, *Y* is not distributed like ε, because that one value of *X* is large relative to the others. When the *Y* are permuted, the outlying value of *Y* will no longer be paired with the outlier in *X*, so the permutation test starts to go awry. The units being shuffled (*Y*) are not distributed like the exchangeable units under the null hypothesis (ε), so permuting *Y* does not give an accurate approximate test of H_0.

Kennedy and Cade (1996) and Anderson and Legendre (1999) have shown that the presence of an outlier in the covariable destabilizes this test, often leading to inflated Type I error, so shuffling raw observations is not a method that should generally be used for tests of partial correlation. If, however, the sample size is quite small (i.e., $n < 10$), this method avoids having to calculate residuals as estimates of errors (unlike the method of Freedman and Lane (1983)) and so it may be used in this case, if it is known that there are no outliers in *X*.

A third method of permutation to consider is one proposed by ter Braak (1992). Like the method of Freedman and Lane (1983), it permutes residuals. In this case, what are shuffled, however, are the residuals of the full multiple regression (Fig. 3*c*). The rationale for this is, if the null hypothesis is true, then the distribution of errors of the full model (ε′ in eq. 2) should be like the distribution of errors under the null hypothesis (ε in eq. 4). The values of β_0, β_1, and β_2 are unknown, but we can obtain the least-squares estimates of these as b_0, b_1, and b_2, respectively. The residuals estimating the errors (ε′) are then calculated as $R_{Y,XZ} = Y - b_0 - b_1X - b_2Z$. These residuals are permuted and, for each of the *n*! re-orderings, a value is calculated for the squared partial correlation coefficient $(r^2_{\mathrm{T}})^{\pi}$ for $R^{\pi}_{Y,XZ}$ versus *Z* given *X* (the subscript T indicates the ter Braak (1992) method). Note once again that *X* and *Z* are not shuffled but remain in their original order. This method is sometimes called permutation "under the alternative hypothesis" or "under the full model" (ter Braak 1992; Anderson and Legendre 1999). Like the method of Freedman and Lane (1983), it assumes a linear model to calculate residuals.

As might be expected, this method of permutation can have the same kind of problem as permuting raw observations *Y*. Things start to go awry (i.e., Type I error does not remain at the chosen significance level), if the errors under the null hypothesis (ε) have a different kind of distribution than the errors of the full model (ε′). This can happen, for example, if the errors of the full model have a highly skewed distribution and the covariable contains an outlier (Anderson and Legendre 1999). Empirically, ter Braak's method of permutation (1992) for multiple regression gives results highly comparable with those obtained by the method of Freedman and Lane (1983). Very extreme situations need to be simulated before any destabilization occurs. Like the method of Freedman and Lane (1983), ter Braak's method (1992) is only asymptotically exact. It too must rely on estimates of regression coefficients (b_1 and b_2) to obtain residuals. If sample sizes are small ($n < 10$), then this method is not optimal.

Recommendations

In general, for tests in partial regression, the best method to use is that of Freedman and Lane (1983). It is the closest to a conceptually exact permutation test (Anderson and Robinson 2001). If sample sizes are very small ($n < 10$), it is better to use permutation of raw data, provided there are no outliers in the covariables (Anderson and Legendre 1999).

One computational advantage of ter Braak's method (1992) is for the situation when many partial regression coefficients are being tested in a single model. For example, say partial tests are to be done for the relationship between bivalve growth and each of POM, salinity, dissolved oxygen, and temperature. For each test, the other independent variables are covariables. With ter Braak's method (1992), only one set of residuals (the residuals of the full model) needs to be shuffled, regardless of which particular independent variable (or set of variables) in the model is being tested. So all the tests can be achieved by permuting one set of residuals. This differs from the situation for the method of Freedman and

Lane (1983), which would require a new set of residuals to be calculated for each particular null hypothesis being tested.

In some situations for multiple regression, one has several repeated values of an independent variable. For example, there could have been three different temperature regimes, with several replicate observations of bivalve growth and POM in each temperature regime (i.e., X takes three repeated fixed values). In that case, an exact permutation test of Y versus Z given X could be achieved by restricting the permutations to occur only among the observations from the same temperature regime (Brown and Maritz 1982). That is, individual measurements of bivalve growth within the different temperature regimes are exchangeable (as against POM) under the null hypothesis, but they are not exchangeable across different temperature regimes. Restricted permutations are considered in more detail below for multifactorial ANOVA.

Complex designs—multifactorial analysis of variance

The method of permutation required to obtain an exact test in ANOVA is not simple when there are several factors in the design. In fact, for some terms (i.e., interaction terms), there is no exact permutation test. So, for tests of interaction, an approximate permutation test that is asymptotically exact must be used. There are some general guidelines, however, that can help in making a decision about two important aspects of permutation tests for complex designs (M.J. Anderson and C.J.F. ter Braak, unpublished data). First, one must determine what is to be permuted (i.e., what units are exchangeable under the null hypothesis). Second, one must control for other factors not being tested by either (*i*) restricting permutations (which yields an exact test) or (*ii*) permuting residuals (which yields an approximate test). If the exchangeable units are the observations, then one can also obtain an approximate test by unrestricted permutation of raw data. I consider, here, nested and factorial two-way designs, to illustrate some of the important concepts for constructing permutation tests in multifactorial ANOVA.

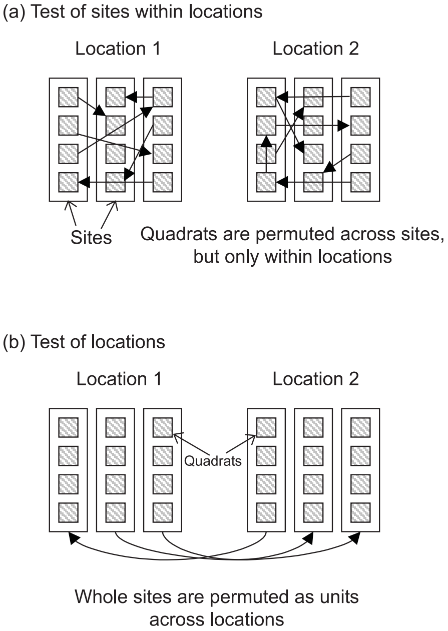
Nested (hierarchical) designs

The method of permutation used depends on the factor being tested. For an exact test of a nested factor, the permutations are done randomly across the units but are restricted to occur within each category of the higher-ranked factor in the hierarchy (Fig. 4*a*). This is a consequence of the logic of the experimental design. The categories of a nested factor are specific to each of the categories of the factor in which they are nested, so it is not logical to permute values across different categories of the upper-level factor. For example, consider a nested hierarchical design with two factors: two locations (scales of hundreds of metres) and three sites nested within each location (scales of tens of metres). The number of snails, for example, is counted within each of n = 4 replicate 1 m × 1 m quadrats at each site.

The linear model for the number of snails in quadrat k of site j within location i is

$$y_{ijk} = \mu + A_i + B(A)_{j(i)} + \varepsilon_{ijk} \tag{5}$$

Fig. 4. Schematic diagram of units that are permuted for a test of sites (*a*) or locations (*b*) in a two-way nested ANOVA design with observations from n = 4 replicate quadrats in each of three sites nested in each of two locations.


(a) Test of sites within locations
Location 1
Location 2
Sites
Quadrats are permuted across sites, but only within locations
(b) Test of locations
Location 1
Location 2
Quadrats
Whole sites are permuted as units across locations

where μ is the overall mean, A_i is the effect of location i, $B(A)_{j(i)}$ is the effect of site j within location i, and ε_{ijk} is the error associated with quadrat k in site j of location i.

For a test of differences in the numbers of snails among sites within locations, values obtained for replicate quadrats are exchangeable among the sites but they are not exchangeable across the different locations (Fig. 4*a*). Effects due to locations must be controlled in some way. In general, there are two ways of doing this. First, one can control for location differences by restricting the permutations to occur within each location. When permutations are restricted within each location, the sum of squares due to locations (SS_L) remains fixed throughout the permutations. The following is already known about the ANOVA model: (*i*) SS_T, the total sum of squares, remains constant across all permutations and (*ii*) $SS_T = SS_L + SS_{S(L)} + SS_R$, where $SS_{S(L)}$ is the sum of squares due to sites within locations and SS_R is the residual sum of squares. The F ratio for the test of sites within locations is

$$F_{S(L)} = \frac{SS_{S(L)}/2(3-1)}{SS_R/6(4-1)} \tag{6}$$

If the null hypothesis were true, the variability due to sites within locations would be similar to the residual variation. That is, variability due to sites is exchangeable with residual variation under the null hypothesis. If SS_T and SS_L remain constant under permutation, then the shuffling strategy is "mixing" variability only between $SS_{S(L)}$ and SS_R, which gives an exact test for differences among sites.

Second, one can control for location effects in another way by obtaining the residuals, which is done by subtracting means:

$$(7) \qquad r_{ijk}^{(B)} = y_{ijk} - \bar{y}_{i..}$$

where $\bar{y}_{i..}$ is the mean for location *i*. Permuting these residuals and recalculating $F_{\mathrm{S(L)}}$ under permutation gives a good approximate test. Simulations have shown that the exact test is more powerful in this particular situation, so there is no real advantage to be gained by permuting residuals, unless using restricted permutations yields too few possible permutations for a reasonable test (M.J. Anderson and C.J.F. ter Braak, unpublished data).

Next, for the test of the higher-ranked factor (locations), it is known that components of variability associated with the nested term (sites) are incorporated into its expected mean square (Scheffé 1959). Thus, the mean square of the nested factor (sites) is the appropriate denominator mean square for the *F* ratio of this test. If the null hypothesis were true, then variability from site to site would be exchangeable with variability due to locations. So, for a permutation test of locations, the four replicate quadrats in any one site are kept together as a unit. These units (i.e., the different sites, of which there are six in the above design) are permuted randomly across the two locations (Fig. 4*b*). The *F* ratio for the test is then:

$$(8) \qquad F_{\mathrm{L}} = \frac{\mathrm{SS_L}/1}{\mathrm{SS_{S(L)}}/2(3-1)}$$

This demonstrates the notion of exchangeable units for permutation tests in ANOVA. In general, for any term in any ANOVA model, the appropriate exchangeable units are identified by the denominator mean square in the *F* ratio for the test (M.J. Anderson and C.J.F. ter Braak, unpublished data). If quadrats within sites are permuted together as a unit (i.e., whole sites are permuted), then $\mathrm{SS_R}$ remains constant under permutation. As $\mathrm{SS_T}$ also remains constant under permutation, the shuffling of these units clearly exchanges variability only between $\mathrm{SS_L}$ and $\mathrm{SS_{S(L)}}$. This gives an exact permutation test for differences among locations, given that significant variability may occur across sites within locations.

No approximate tests are possible here. Random permutation of raw data or of any kind of residuals across all observations (ignoring the site units) results in inflated Type I error for this test, when significant variability due to sites is present (M.J. Anderson and C.J.F. ter Braak, unpublished data).

In the above example, there are only three sites within each location, which gives $6!/[2!(3!)^2] = 10$ total possible permutations. Thus, the smallest possible *P* value obtainable is 0.10. This is an important consequence of the choice of experimental design on the permutation test. If given the choice, a researcher would be well advised to increase the number of categories of the nested factor (sites), rather than increasing the number of replicates (quadrats). This will give greater ability to test the higher-ranked factor in the hierarchy (locations) using permutations.

One way to increase the possible number of permutations for the test of locations is to ignore the effect of sites. This can only logically be done when variation from site to site is not statistically significant. In this case, one may then consider all quadrats to be exchangeable for the test of differences among locations. Considering the consequences of ignoring sites for the test of locations is analogous to the consideration of whether "to pool or not to pool" in traditional analyses. Thus, the general rule of thumb is not to pool (i.e., do not permute individual quadrats in the test of locations) unless the *P* value associated with the test of site effects is larger than 0.25 (Winer et al. 1991). If in doubt about the possible effects of sites (e.g., $0.05 < P < 0.25$), stick with the exact test (i.e., permute the sites, keeping quadrats within each site together as a unit) in the test for effects of locations. The decision about what to permute in the test of locations (i.e., quadrats or site units) when sites do not differ significantly may also depend on the severity of the consequences of making Type I versus Type II errors. The recommended criterion for the pooling of $P > 0.25$ is merely a rule of thumb (Winer et al. 1991).

In general, one must permute appropriate exchangeable units. These are identifiable by reference to the denominator mean square of the *F* ratio (M.J. Anderson and C.J.F. ter Braak, unpublished data).

Factorial (orthogonal) designs

Consider a two-factor experiment examining the effects of predators and the effects of food (and their possible interaction) on the numbers of snails on a rocky shore. There are, for example, areas where predators have been experimentally removed and areas where predators have not been removed (factor *A*). In addition, for each of these states (presence or absence of predators), there are some areas where the amount of food (algae) has been reduced and other areas where food has been left intact (factor *B*). Say that, in each of these combinations of factors *A* and *B* (there are four—with or without food in the presence or absence of predators), the numbers of snails are recorded in $n = 5$ replicate quadrats.

The linear ANOVA model is

$$(9) \qquad y_{ijk} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{ijk}$$

where μ is the population mean, A_i is the effect of category *i* of factor *A*, B_j is the effect of category *j* of factor *B*, AB_{ij} is the interaction effect of the *ij*th combination of factors *A* and *B*, and ε_{ijk} is the error associated with observation y_{ijk}.

For the test of a significant interaction (*A* × *B*), there is no exact permutation test. The problem is to create a test for the interaction, controlling for the possibility of there being main effects of some kind. Restricting permutations to within categories of other factors is one method of controlling for these factors in the model. If, however, permutations were restricted to occur within each of the four combinations of predators and food (categories of factors *A* and *B*), there would be no other values of the *F* statistic for the test of interaction apart from the one obtained with the original data.

Thus, an approximate permutation test is necessary here. One may control for the main effects by estimating residuals. The main effects are unknown, but can be estimated by calculating the means (or, in multivariate analysis, the centroids, which consist of the means for each variable) for each group in each of the factors. Effects of the factors can then be "removed" by subtracting the appropriate mean from each observation to obtain residuals. For a single variable, let y_{ijk} be the *k*th observation from the *i*th category of factor

A and the *j*th category of factor *B*. Then the residuals removing the effects of *A* and *B* are

$$(10) \qquad r_{ijk}^{(AB)} = y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{\bar{y}}_{...}$$

where $\bar{y}_{i..}$ is the mean for category *i* of factor *A*, $\bar{y}_{.j.}$ is the mean for category *j* of factor *B*, and $\bar{\bar{y}}_{...}$ is the overall mean. If there were no significant interaction of these factors (i.e., the null hypothesis about interactions were true), then these residuals would be estimates of the errors associated with each sampling unit in the model (without interaction) and are iid. These residuals are thus exchangeable under the null hypothesis of no interaction and can be permuted to obtain a test. Note that it was necessary to assume an additive ANOVA model to get the residuals for this approach.

This method of permutation was described by Still and White (1981) and is the equivalent counterpart in ANOVA to the method of Freedman and Lane (1983) in partial regression. It is generally the best method to use of the approximate methods, because empirically it gives the best power and maintains Type I error for complex designs in the widest circumstances (M.J. Anderson and C.J.F. ter Braak, unpublished data). This is, however, only an asymptotically exact test, because the "true" cell means corresponding to effects of different factors are not known. Consequently, when the residuals are calculated from these, they do not correspond to the "true" errors. This can be problematic with small sample sizes, where estimates of means are not very precise. With increases in sample size, estimates of means get better, and permutation of residuals gets closer to being an exact test.

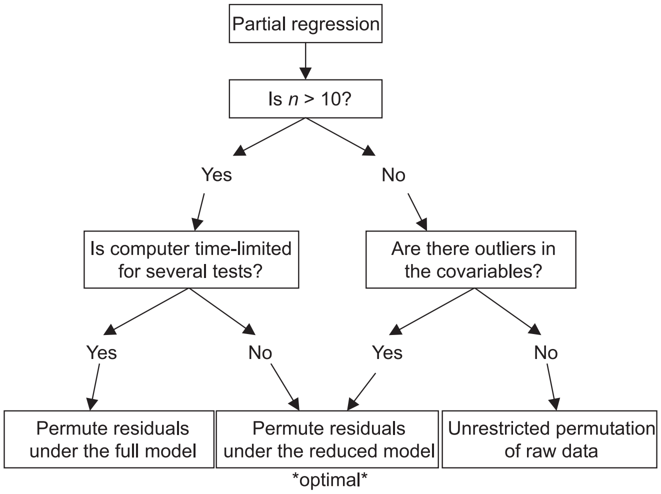
It is also possible to use the permutation method of ter Braak (1992) in the test for interaction in ANOVA. Here, the residuals that are permuted are those corresponding to the estimates of the errors for the full model, including the interaction. So these residuals are calculated as

$$(11) \qquad r_{ijk}^{(\mathrm{full})} = y_{ijk} - \bar{y}_{ij.}$$

where $\bar{y}_{ij.}$ is the mean of the *n* replicates in the cell corresponding to the *i*th category of factor *A* and the *j*th category of factor *B*. In practice, this method gives results that are highly comparable with those obtained using permutation of residuals under the reduced model (Anderson and Legendre 1999).

Another approach to obtain an approximate test of the interaction is simply to permute the raw data without restriction (Manly 1997). This can be used for tests in ANOVA designs where exchangeable units are the individual observations (i.e., any tests where the denominator mean square of the *F* ratio is the residual). Permutation of raw data will not have the same problem in ANOVA that it has when used for tests in partial regression. There are no such things as "outliers" in codes (or dummy variables) used to identify categories in the multiple regression model corresponding to an ANOVA design (e.g., Neter et al. 1996). Also, unrestricted permutation of raw data may be preferred over either method of permutation of residuals in the case of small sample sizes (Gonzalez and Manly 1998), because no estimates of means are required. However, simulations have shown that permutation of residuals does give a more powerful test (M.J. Anderson and C.J.F. ter Braak, unpublished data).

Fig. 5. Flow chart indicating appropriate permutation methods for tests of partial regression.


Partial regression
Is *n* > 10?
Yes
No
Is computer time-limited for several tests?
Are there outliers in the covariables?
Yes
No
Yes
No
Permute residuals under the full model
Permute residuals under the reduced model
Unrestricted permutation of raw data
optimal

If the interaction is not significant, then tests of main effects are the next logical step. The objective is to obtain a permutation test for, say, the effect of factor *A*, while controlling for the possible effect of factor *B*. An exact test (in the absence of any interaction) is achieved by restricting the permutations to occur within categories of factor *B*. So, for example, to test for the effects of predators, one would permute values within each of the food treatments. Alternatively, permutation of residuals (under either the reduced or full model) can also be used. Note that for permutation under the full model, the residuals to be permuted are the same regardless of the term being tested. The residuals to be permuted under the reduced model depend, however, on the term being tested. For example, the residuals that would be permuted for the test of factor *A*, removing the effect of factor *B*, if any (in the absence of an interaction), are

$$(12) \qquad r_{ijk}^{(A)} = y_{ijk} - \bar{y}_{.j.}$$

For testing factor *B* in the presence of *A* (without interaction), the residuals are

$$(13) \qquad r_{ijk}^{(B)} = y_{ijk} - \bar{y}_{i..}$$

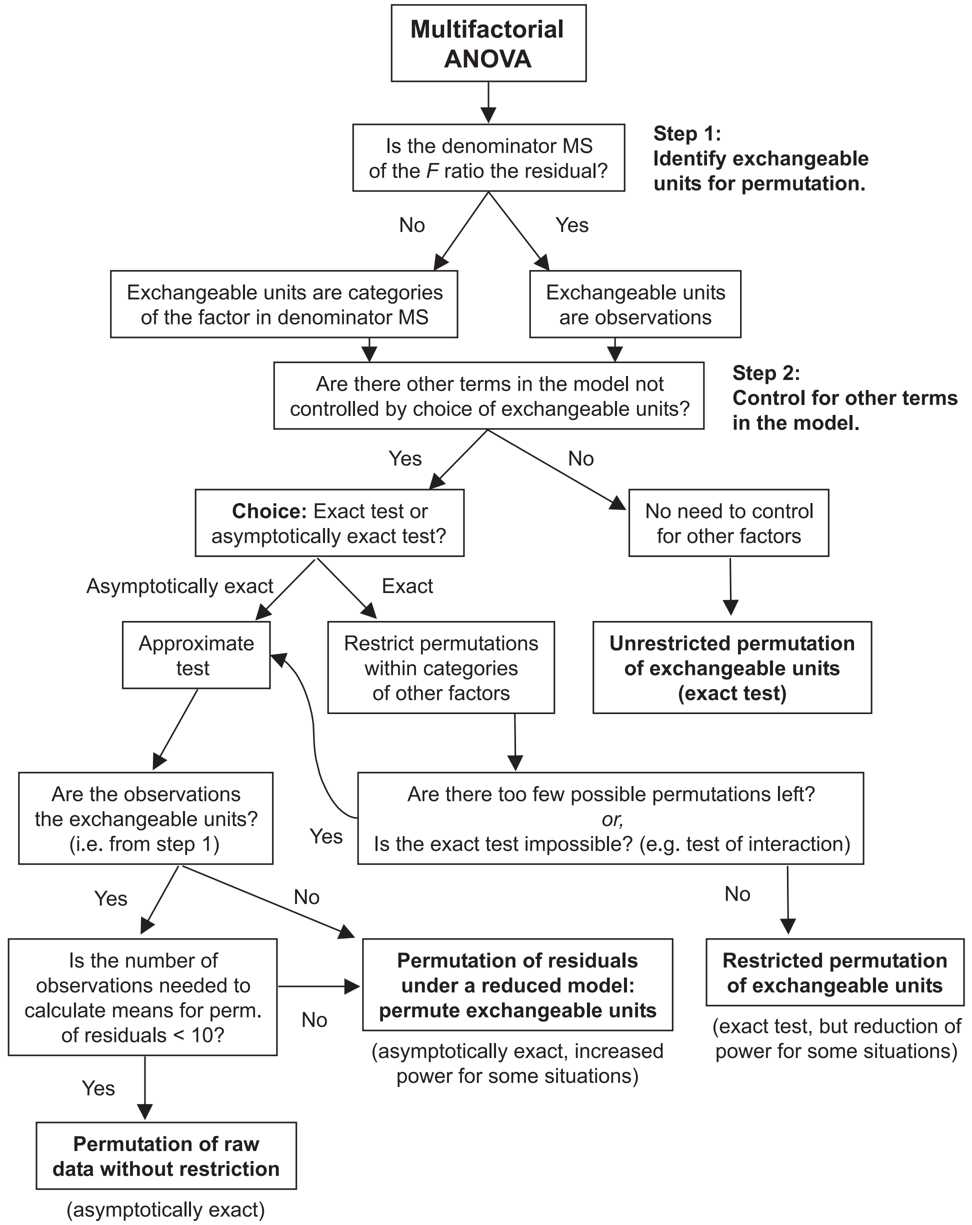
Simulations have demonstrated that permutation of residuals under the reduced model in this situation is more powerful than an exact test using restricted permutations (M.J. Anderson and C.J.F. ter Braak, unpublished data).

Note that the above strategies for permutation tests in factorial designs also apply to observational studies with factorial structures and to BACI (before-after, control-impact) studies.

Recommendations

Some general recommendations for practical situations are appropriate. Empirical simulations supporting these statements can be found in the unpublished data of M.J. Anderson and C.J.F. ter Braak and in Manly (1997) and Gonzalez and Manly (1998).

Fig. 6. Flow chart indicating appropriate permutation methods for tests of individual terms in multifactorial ANOVA.


Multifactorial ANOVA
Is the denominator MS of the F ratio the residual?
Step 1: Identify exchangeable units for permutation.
No
Yes
Exchangeable units are categories of the factor in denominator MS
Exchangeable units are observations
Are there other terms in the model not controlled by choice of exchangeable units?
Step 2: Control for other terms in the model.
Yes
No
Choice: Exact test or asymptotically exact test?
No need to control for other factors
Asymptotically exact
Exact
Approximate test
Restrict permutations within categories of other factors
Unrestricted permutation of exchangeable units (exact test)
Are the observations the exchangeable units? (i.e. from step 1)
Yes
Are there too few possible permutations left? or, Is the exact test impossible? (e.g. test of interaction)
Yes
No
No
Is the number of observations needed to calculate means for perm. of residuals < 10?
No
Permutation of residuals under a reduced model: permute exchangeable units
(asymptotically exact, increased power for some situations)
Restricted permutation of exchangeable units
(exact test, but reduction of power for some situations)
Yes
Permutation of raw data without restriction
(asymptotically exact)

The first step for any permutation test in an ANOVA context is to identify the exchangeable units. These are identifiable by reference to the expected mean squares. The denominator mean square for the *F* ratio of any individual term will indicate which units are exchangeable. For example, the levels of the nested factor are exchangeable for the test of the higher-ranked factor in a nested hierarchy. Similarly, if an interaction term is the denominator of a relevant *F* ratio, then the cells corresponding to that interaction are the exchangeable units for the test. If the denominator mean square is the residual, then individual observations are exchangeable.

The second step is to control for other factors in the model not being tested (which are not already controlled by the choice of exchangeable units). There are two ways of doing this. The first is to restrict permutations within the levels of the other factors. This gives an exact test (i.e., its Type I error is exactly equal to the a priori chosen significance level). The second is to “remove” the effects of factors not being tested by calculating residuals (i.e., subtracting means of levels of factors not under test from each observation). These residuals are then permuted and the relevant *F* ratio under permutation is calculated using them (rather than the original observations), accordingly. Permutation of residuals in this way yields a test that is asymptotically exact (i.e., its Type I error asymptotically approaches the a priori chosen significance level with increases in sample size). Permutation of residuals also generally yields a more powerful test than the test using restricted permutations (M.J. Anderson and C.J.F. ter Braak, unpublished data).

If exact Type I error is of extreme importance, then use an exact test. However, there are several situations where one will prefer to choose an approximate test instead: (*i*) one wishes to increase power, (*ii*) the number of possible permutations is too few with the exact test to give a reasonable *P* value, or (*iii*) the exact test is impossible (e.g., tests of interaction). In these situations, one may choose to use an approximate test such as permutation of residuals. Note that the increase in power does not come at the expense of Type I error, which is maintained for the approximate tests. Also note that permutation of residuals of certain factors in the design never allows one to avoid the issue of the choice of appropriate exchangeable units. These are fixed by the design and cannot be altered by changing the permutational approach.

Should an approximate method be chosen, similar recommendations would apply as for the situation with partial regression. Namely, permutation of residuals under the reduced model is generally to be preferred over permutation of residuals under the full model, as it comes the closest to a conceptually exact test (Anderson and Robinson 2001). Permutation of raw data can only be used if the exchangeable units are the individual observations. It is recommended only in the case of small sample sizes (i.e., as a rule of thumb, if the number of observations used to calculate means for the reduced model is less than 10; Anderson and Legendre 1999).

Discussion

In practice, it is most important that the assumptions underlying the permutation tests used are kept in mind. Since permutation tests are touted as "distribution free," they are often incorrectly construed as having "no assumptions." The general assumption is that errors are iid (see the Background and rationale for permutation tests, above) and, for any test using residuals, an additive linear model is also assumed. At the very least, one must assume that relevant units are exchangeable under some null hypothesis. The examples given here are discussed in the context of a single response variable (e.g., sizes of bivalves, numbers of snails), but the same general principles for permutation apply to tests of multivariate data in complex designs (e.g., Clarke 1993; ter Braak and Šmilauer 1998; Anderson 2001).

The recommendations given in the text are consolidated into flow charts for choosing an appropriate permutation method. These flow charts are provided on the basis of previous empirical results (Gonzalez and Manly 1998; Anderson and Legendre 1999; M.J. Anderson and C.J.F. ter Braak, unpublished data) and theoretical comparisons (Anderson and Robinson 2001), which can be consulted for further details, if required. Figure 5 provides a flow chart for partial regression and Fig. 6 provides a flow chart for analysis of variance. These are general guidelines, applicable to tests of any term in any linear multifactorial model. The logic of the experimental design and the null hypothesis being tested remain paramount in considering what units to permute and how to permute them.

Perhaps the most important consideration in the development of an appropriate permutation test is to determine what units are exchangeable under the null hypothesis. For partial regression, it is useful to write down the full model and then write down what the model would look like if the null hypothesis were true. Assuming the errors under a true null hypothesis are iid, this will indicate what residuals are exchangeable for permutation using the Freedman and Lane

Table 1. List of some computer software packages and their uses for permutation tests.

Computer package	PC or Macintosh	ANOVA or regression	Methods implemented[a]	Reference
CANOCO[b]	PC	ANOVA, regression	D, R, F, S, E	ter Braak and Šmilauer 1998
MULTIV[b]	Macintosh	ANOVA	D, S	Pillar and Orlóci 1996
NPMANOVA[b]	PC	ANOVA	D, R, F, S, E	Anderson 2001
PATN[b]	PC	ANOVA	D	Available from the Commonwealth Scientific and Industrial Research Organisation, Division of Wildlife and Ecology, Australia
PRIMER[b]	PC	ANOVA	D, S, E	Clarke 1993
The R Package[b]	Macintosh	ANOVA, regression	D, R, S	Legendre and Vaudor 1991
RT	PC	ANOVA, regression	D, F, S	Manly 1997
StatXact	PC	ANOVA, regression	D, S	Available from Cytel Software, Cambridge, Mass., U.S.A.
Various Fortran routines	PC or Macintosh	ANOVA, regression	D, S, E	Edgington 1995

[a] D, raw data permutation; R, permutation of residuals under the reduced model; F, permutation of residuals under the full model; S, restricted permutations; E, possible to exchange units other than observations.

[b] Permutation tests in these packages were primarily intended for use with multivariate data, but may be used for univariate applications as well.

(1983) method. Although one has the option in partial regression of using permutation of raw data or the method of ter Braak (1992), the method of Freedman and Lane (1983) is to be preferred, unless sample sizes are very small ($n < 10$), in which case, permuting raw data is preferred, provided there are no outliers in the covariables (Anderson and Legendre 1999).

In the case of analysis of variance, for the exact test, one must consider (*i*) what units to permute and (*ii*) if permutations should be restricted (for an exact test) or if residuals should be permuted (for an asymptotically exact test). To determine *i*, consider the *F* ratio for the particular term being tested. What is the term identified by the denominator mean square? The categories of this factor indicate the units that are exchangeable under the null hypothesis (M.J. Anderson and C.J.F. ter Braak, unpublished data). For example, if the denominator mean square is the residual, then observations themselves can be permuted. If the denominator mean square is an interaction term $A \times B$, then *ab* cells need to be permuted as units. To determine *ii*, recall that the total sum of squares remains constant across all permutations. For an exact test, permutations must be restricted to occur within categories of all other factors in the model not being tested, if their variability is not already controlled by the choice of units in *i*. For an approximate test, one can permute residuals rather than restricting permutations. Note, however, that this does not alter the fact that the correct exchangeable units must still be used.

When an exact test cannot be done (e.g., tests of interaction) or there are too few possible permutations to give enough power with an exact test, then permutation of residuals under the reduced model (in the manner of Freedman and Lane (1983)) is highly recommended. Permuting residuals in ANOVA corresponds to permuting units after subtracting means corresponding to categories of particular factors (Still and White 1981). Use unrestricted permutation of raw data if sample sizes are very small. As a general rule of thumb, if the means (averages) needed to calculate residuals for the Freedman and Lane (1983) method are obtained by fewer than 10 observations, then use unrestricted permutation of raw data (Gonzalez and Manly 1998; Anderson and Legendre 1999).

The recommendations given here provide the basis for constructing exact permutation tests, where possible, or for choosing an optimal approximate permutation test. More detailed statistical information and empirical and theoretical comparisons of these methods can be found elsewhere (Anderson and Legendre 1999; Anderson and Robinson 2001; M.J. Anderson and C.J.F. ter Braak, unpublished data).

A note on computer software

It is always problematic to mention much in the way of available computer software, for as soon as such information is printed, it is out of date. However, Table 1 contains a list, which is by no means intended to be exhaustive, of the available software packages known to me that are capable of performing various permutation procedures referred to in the text. I have not included in this list any of the major statistical software packages, such as SAS, SPSS, S-PLUS, or Minitab, which can be used to perform various permutation methods by programming such procedures using their built-in languages. For programming various methods in FORTRAN, I used the GGPER routine from the International Mathematical and Statistical Library (IMSL) and I have found the books by Edgington (1995) and Manly (1997) to be particularly helpful.

Acknowledgements

I thank P. Legendre, J. Robinson, and C.J.F. ter Braak for their work and assistance in finding appropriate permutation methods for complex models and designs. Earlier versions of this manuscript benefited from comments made by A.J. Underwood, M.G. Chapman, and B. Gillanders, however any mistakes herein are strictly my own.

References

Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. Aust. Ecol. **26**: 32–46.

Anderson, M.J., and Legendre, P. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. J. Statist. Comput. Simul. **62**: 271–303.

Anderson, M.J., and Robinson, J. 2001. Permutation tests for linear models. Austral. N.Z. J. Statist. **43**: 75–88.

Boik, R.J. 1987. The Fisher–Pitman permutation test: a non-robust alternative to the normal theory F-test when variances are heterogeneous. Br. J. Math. Statist. Psychol. **40**: 26–42.

Brown, B.M., and Maritz, J.S. 1982. Distribution-free methods in regression. Austral. J. Statist. **24**: 318–331.

Clarke, K.R. 1993. Non-parametric multivariate analysis of changes in community structure. Aust. J. Ecol. **18**: 117–143.

Edgington, E.S. 1995. Randomization tests. 3rd ed. Marcel-Dekker, New York.

Fisher, R.A. 1935. Design of experiments. Oliver and Boyd, Edinburgh.

Fisher, R.A. 1936. The coefficient of racial likeness and the future of craniometry. J. R. Anthropol. Inst. G.B. Irel. **66**: 57–63.

Fisher, R.A. 1955. Statistical methods and scientific induction. J. Roy. Statist. Soc., Ser. B, **17**: 69–78.

Freedman, D., and Lane, D. 1983. A nonstochastic interpretation of reported significance levels. J. Busin. Econom. Statist. **1**: 292–298.

Gaston, K.J., and McArdle, B.H. 1994. The temporal variability of animal abundances: measures, methods and patterns. Philos. Trans. R. Soc. Lond. B, Biol. Sci. No. 345. pp. 335–358.

Gonzalez, L., and Manly, B.F.J. 1998. Analysis of variance by randomization with small data sets. Environmetrics, **9**: 53–65.

Good, P.I. 2000. Permutation tests: a practical guide to resampling methods for testing hypotheses. 2nd ed. Springer-Verlag, Berlin.

Hayes, A.F. 1996. Permutation test is not distribution free. Psychol. Methods, **1**: 184–198.

Hoeffding, W. 1952. The large-sample power of tests based on permutations of the observations. Ann. Math. Statist. **23**: 169–192.

Hope, A.C. 1968. A simplified Monte Carlo significance test procedure. J. Roy. Statist. Soc. Ser. B, **30**: 582–598.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. **54**: 187–211.

Kempthorne, O. 1955. The randomization theory of experimental inference. J. Am. Stat. Assoc. **50**: 946–967.

Kempthorne, O. 1966. Some aspects of experimental inference. J. Am. Stat. Assoc. **61**: 11–34.

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

Kennedy, P.E., and Cade, B.S. 1996. Randomization tests for multiple regression. Comm. Statist. Simulation Comput. **25**: 923–936.

Legendre, P., and Vaudor, A. 1991. The R package—multidimensional analysis, spatial analysis. Départment de sciences biologiques, Université de Montréal, Montréal.

Manly, B.F.J. 1997. Randomization, bootstrap and Monte Carlo methods in biology. 2nd ed. Chapman and Hall, London.

Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. 1996. Applied linear statistical models. 4th ed. Irwin, Chicago.

Neyman, J. 1923. On the application of probability theory to agricultural experiments: principles. [In Polish with German summary.] Roczniki Nauk Rolniczch, **10**: 1–51.

Pillar, V.D.P., and Orlóci, L. 1996. On randomization testing in vegetation science: multifactor comparison of relevé groups. J. Veg. Sci. **7**: 585–592.

Pitman, E.J.G. 1937*a*. Significance tests which may be applied to samples from any populations. J. Roy. Statist. Soc. Ser. B, **4**: 119–130.

Pitman, E.J.G. 1937*b*. Significance tests which may be applied to samples from any populations II. The correlation coefficient test. J. Roy. Statist. Soc. Ser. B, **4**: 225–232.

Pitman, E.J.G. 1937*c*. Significance tests which may be applied to samples from any populations III. The analysis of variance test. Biometrika, **29**: 322–335.

Romano, J.P. 1988. Bootstrap and randomization tests of some nonparametric hypotheses. Ann. Statist. **17**: 141–159.

Scheffé, H. 1943. Statistical inference in the non-parametric case. Ann. Math. Statist. **14**: 305–332.

Scheffé, H. 1959. The analysis of variance. John Wiley & Sons, New York.

Sokal, R.R., and Rohlf, F.J. 1981. Biometry. 2nd ed. W.H. Freeman and Co., New York.

Still, A.W., and White, A.P. 1981. The approximate randomization test as an alternative to the F test in analysis of variance. Br. J. Math. Statist. Psychol. **34**: 243–252.

ter Braak, C.J.F. 1992. Permutation versus bootstrap significance tests in multiple regression and ANOVA. *In* Bootstrapping and related techniques. *Edited by* K.-H. Jöckel, G. Rothe, and W. Sendler. Springer-Verlag, Berlin. pp. 79–86.

ter Braak, C.J.F., and Šmilauer, P. 1998. CANOCO reference manual and user's guide to CANOCO for Windows: software for canonical community ordination (version 4). Microcomputer Power, Ithaca, N.Y.

Winer, B.J., Brown, D.R., and Michels, K.M. 1991. Statistical principles in experimental design. 3rd ed. McGraw-Hill, New York.