

Computer-intensive methods

Marie-Josée Fortin, Geoffrey M. Jacquez & Bill Shipley

Volume 1, pp 399–402

in

Encyclopedia of Environmetrics
(ISBN 0471 899976)

Edited by

Abdel H. El-Shaarawi and Walter W. Piegorsch

© John Wiley & Sons, Ltd, Chichester, 2002

Computer-intensive methods

With recent advances in the power of personal computers, reference distributions of statistics are now often generated using computer-intensive methods. Several terms are erroneously used interchangeably when referring to such computer-intensive statistical methods, including ‘resampling’, ‘Monte Carlo’, ‘permutation’, ‘randomization’ and ‘bootstrap’. These techniques are not interchangeable but are fundamentally different in terms of their statistical mechanics. The generic terms ‘resampling’ and ‘computer-intensive methods’ refer to all methods in which the observed data are used to generate a reference distribution by means of randomization [9]. This reference distribution is then used to assess the significance of a statistic calculated from the observed (not randomized) data. Significance is evaluated under the assumption that the statistic computed using the observed data is sampled from the reference distribution generated with a randomization technique.

Resampling methods have fewer assumptions than traditional statistics. For example, resampling does not require the data to follow a known analytical distribution. Hence, an attractive feature of resampling is that statistical significance is evaluated based on empirical distributions generated from the observed sample. This ‘distribution-free’ alternative to parametric statistics is quite appealing to, say, ecologists faced with small data sets that do not meet the assumed parametric distribution. While computer-intensive methods may involve fewer assumptions [2], this does not mean no assumptions. Hence, caution must be exercised because the randomization procedures often assume data independence, and this assumption is invalid when the data are spatially or temporally autocorrelated [1, 3, 4, 6–8]. The presence of autocorrelation necessitates the use of restricted randomization to account for the spatial or temporal structure of the data (see below and [7]). In addition, the particular choice of a computer-intensive method in effect determines the statistical null and alternative hypotheses. Moreover, test implementation itself can have subtle and often unintended impacts on the hypotheses tested. Depending on the test chosen, and its subsequent

implementation, environmetricians may or may not correctly test the hypothesis they originally intended to analyze.

Randomization Test

The terms ‘permutation test’ and ‘randomization test’ are often used interchangeably. Formally, Fisher [5] used the former term (*see Permutation tests*) to refer to a method that performed population-level inference, while randomization test was applied to methods for sample-based inference. The term permutation test, or **exact randomization test** is also used when the test provides the exact significance levels by *exhaustive* computation of all possible rearrangements (permutations) of the data. In practice, even with current powerful computers, permutation tests can only be performed when the number of observations (n) is small, because the number of permutations increases as the factorial of the sample size ($n!$). For example, a sample size of 7 implies 5040 possible permutations, but with only three more observations a sample size of 10 implies 3 628 800 permutations!

When the number of observations precludes an exact test, an ‘approximate randomization’ test is used instead. This is a randomization test that generates a subset of the possible permutations. Because only a subsample of all possible permutations is calculated, many authors recommend that 10 000 or more randomizations be used when constructing the reference distribution [2, 7]. The reference distribution of any statistic is obtained using a five-step randomization procedure [8] that repeatedly reallocates the value of the observations over the sample, and then recalculates the statistic to generate the null reference distribution:

1. define the ecological question or hypothesis;
2. choose a statistic (already existing or design a new one);
3. compute a statistic T^* for the observed data;
4. generate the null reference distribution by:
 - rearranging the order of the observed data over the entire sample by shuffling them randomly (i.e. the values of the response variable are shuffled over all the samples, where each sample keeps its spatial, temporal or phylogenetic identity);

2 Computer-intensive methods

- compute the statistic for the data so randomized; and
 - repeat this step k times (e.g. 10 000 or more times according to the number of observations analyzed); and
5. compare the observed statistic to the reference distribution.

Statistical inference in this framework is based on a comparison of the T^* statistic to the reference distribution. The statistical decision of accepting or rejecting the null statistical hypothesis, in the final step, has a corresponding scientific interpretation. The number of randomizations k determines the precision of significance level that can be resolved by the test. For example, a reference distribution based on an approximated randomization of $k = 100$ values can resolve a precision of $P \leq 0.01$, 1000 values of $P \leq 0.001$, and so on.

With a small sample size, randomization tests can produce an odd reference distribution. For example, at sample size $n = 5$, only 120 permutations are possible. Furthermore, depending on the distribution of the observed values, the reference distribution could be skewed or bimodal.

Randomization tests have several advantages, including flexibility and relative ease of implementation. They support significance testing without distributional assumptions (e.g. normality) and complex designs for which parametric tests do not exist. For comparable statistics, randomization tests are as powerful as parametric tests when the number of randomizations is large.

Restricted Randomization Tests

There are always two sides to a coin. The flexibility of randomization tests is offset by the relatively high level of understanding required to access this flexibility. For example, randomization tests assume that the data are independent so that rearrangements (i.e. reordering, shuffling) of the samples are equally likely, and therefore *exchangeable* [6]. Exchangeability does not imply, however, independence. Such an assumption of independence is violated when data show an inherent structure, such as spatial and temporal autocorrelation.

What are the implications of blindly using randomization tests with spatially or temporally autocorrelated data? The statistical null hypothesis

under randomization tests implied by random shuffling is that the observations are spatially independent. Translated into practical terms, spatial independence means that, under the test's null hypothesis, no spatial pattern is present. Under such a condition, no spatial autocorrelation must exist at the variable, or species, level to be able to shuffle the data independently. Randomization tests based on complete random shuffling are therefore not appropriate for evaluating ecological or other hypotheses for which spatial autocorrelation is present. Generally, these kinds of spatial interaction must be incorporated into the null hypothesis because they form the fabric against which specific ecological hypotheses are evaluated (*see Spatial analysis in ecology*). To account for spatial autocorrelation 'restricted randomization' procedures have been developed [7] as a proposed alternative to randomization tests per se. For example, when it is natural to partition the spatial field into several regions one may randomize within but not across regions. This procedure assumes that an observation from one region is equally likely to have occurred at any other location within that region, but that the observation has zero probability of occurring at locations in other regions. Another type of restricted randomization procedure is to keep the spatial or temporal structure of the data fixed and then slide the 'map' of one variable over the other. For spatial data, this is achieved using a two-dimensional torus constructed by connecting the map margins and then sliding one variable map over the other. This procedure maintains the spatial interactions, and therefore the spatial autocorrelation at the variable or species level. Finally, another way to ensure that the spatial structure is preserved is to use restricted randomizations with the same degree of spatial autocorrelation as the observed one.

Other Computer-intensive Methods

Monte Carlo is a general term that refers to any test that uses a **random number generator**. Applications include risk assessment and the propagation of uncertainty (*see Risk assessment, management and uncertainties*). Specifically, Monte Carlo simulations draw random samples from a modeled population with probability distributions estimated from a real population, and test whether the observed sample was randomly drawn from that real population. As

such, the resampling is accomplished in the process of generating samples from the modeled population. This framework supports inference at the population level, with the caveat that inference is premised on the appropriateness of the model (*see Simulation and Monte Carlo methods*).

Bootstrap resampling is another resampling technique that assumes that the observations are independent and performs sampling with replacement from an infinite population, where each observation has the same probability of being chosen each time [8]. Hence, a bootstrap sample from an original sample of size n consists of randomly choosing one of the n values and repeating this k times, putting the chosen value back into the pool each time. The desired statistic is calculated from this bootstrap sample and the entire procedure is repeated k times. This ensures that each bootstrap sample is independent. Unlike randomization tests, statistical inference for the bootstrap applies to the population. The common use of bootstrap methods is to provide estimates of standard errors, approximate confidence intervals for unconventional statistics and approximate probability estimates relative to some null hypotheses. Bootstrap hypothesis tests are often inferior to tests based on parametric or permutation methods when such equivalent methods exist [4].

Finally, **jackknife resampling** is a technique comparable to bootstrapping. However, in the jackknife technique a reference distribution is enumerated by repeatedly leaving out one observation at a time and then recalculating the test statistic. The jackknife technique is mostly used to quantify the bias and **standard error** of an estimate [4].

Final Remarks

Randomization tests open a wide range of possibilities for testing specific hypotheses with complex, dependent environmental data [7]. However, without clear understanding of the type of randomization test employed, an environmentalist may or may not be testing the hypothesis he or she intended. Selection of an inappropriate test may result in wrong statistical conclusions and incorrect scientific interpretations. The above discussion

and classification of randomization procedures illustrate that randomization tests are particularly prone to mis-specification of the null hypothesis, primarily because the null hypothesis is embedded in the randomization procedure and is not self-evident. For example, a complete randomization in which all observations are equiprobable corresponds to a statistical null hypothesis of *Complete Spatial Randomness* (*see Ripley's K function*). In many instances such a null hypothesis is ridiculous, indicating that restricted randomization should be used instead. However, care must be exercised to ensure that the null hypothesis implied under restricted randomization is environmentally tenable. A firm understanding of the null and alternative hypotheses of the chosen randomization test is required in order to ensure that the biological, ecological, or environmental questions under study are correctly addressed.

References

- [1] Cressie, N.A.C. (1993). *Statistics for Spatial Data*, Wiley, New York.
- [2] Crowley, P.H. (1992). Resampling methods for computation-intensive data analysis in ecology and evolution, *Annual Review of Ecological Systems* **23**, 405–447.
- [3] Edgington, E.S. (1995). *Randomization Tests*, 3rd Edition, Marcel Dekker, New York.
- [4] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.
- [5] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.
- [6] Good, P. (1993). *Permutation Tests: a Practical Guide to Resampling Methods for Hypothesis Testing*, Springer-Verlag, New York.
- [7] Manly, B.F.J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd Edition, Chapman & Hall, London.
- [8] Noreen, E. (1989). *Computer-intensive Methods for Testing Hypotheses*, Wiley, New York.
- [9] Simon, J. (1995). *Resampling: the New Statistics*, Resampling Stats, Arlington.

(*See also Statistical computing in environmental science*)

MARIE-JOSÉE FORTIN, GEOFFREY M. JACQUEZ
& BILL SHIPLEY