
Excerpt (pp. 17-26) from:

Legendre, P. & L. Legendre. 1998. Numerical ecology, 2nd English edition. Elsevier Science BV, Amsterdam.

1.2 Statistical testing by permutation

Statistic

The role of a statistical test is to decide whether some *parameter* of the reference population may take a value assumed by hypothesis, given the fact that the corresponding statistic, whose value is estimated from a sample of objects, may have a somewhat different value. A *statistic* is any quantity that may be calculated from the data and is of interest for the analysis (examples below); in tests of significance, a statistic is called *test statistic* or *test criterion*. The assumed value of the parameter corresponding to the statistic in the reference population is given by the statistical null hypothesis (written H_0), which translates the biological null hypothesis into numerical terms; it often negates the existence of the phenomenon that the scientists hope to evidence. The reasoning behind statistical testing directly derives from the scientific method; it allows the confrontation of experimental or observational findings to intellectual constructs that are called hypotheses.

Testing is the central step of inferential statistics. It allows one to generalize the conclusions of statistical estimation to some reference population from which the observations have been drawn and that they are supposed to represent. Within that context, the problem of multiple testing is too often ignored (Box. 1.3). Another legitimate section of statistical analysis, called descriptive statistics, does not rely on testing. The methods of clustering and ordination described in Chapters 8 and 9, for instance, are descriptive multidimensional statistical methods. The interpretation methods described in Chapters 10 and 11 may be used in either descriptive or inferential mode.

1 — Classical tests of significance

Null hypothesis

Consider, for example, a correlation coefficient (which is the statistic of interest in correlation analysis) computed between two variables (Chapter 4). When inference to the statistical population is sought, the null hypothesis is often that the value of the correlation parameter (ρ , rho) in the statistical population is zero; the null hypothesis may also be that ρ has some value other than zero, given by the ecological hypothesis. To judge of the validity of the null hypothesis, the only information available is an *estimate* of the correlation coefficient, r , obtained from a sample of objects drawn from the statistical population. (Whether the observations adequately represent the

Multiple testing

Box 1.3

When several tests of significance are carried out simultaneously, the probability of a type I error becomes larger than the nominal value α . For example, when analysing a correlation matrix involving 5 variables, 10 tests of significance are carried out simultaneously. For randomly generated data, there is a probability $p = 0.40$ of rejecting the null hypothesis at least once over 10 tests, at the nominal $\alpha = 0.05$ level; this can easily be computed from the binomial distribution. So, when conducting multiple tests, one should perform a global test of significance in order to determine whether there is any significant value at all in the set.

The first approach is Fisher's method for combining the probabilities p_i obtained from k independent tests of significance. The value $-2\sum \ln(p_i)$ is distributed as χ^2 with $2k$ degrees of freedom if the null hypothesis is true in all k tests (Fisher, 1954; Sokal & Rohlf, 1995).

Another approach is the Bonferroni correction for k independent tests: replace the significance level, say $\alpha = 0.05$, by an adjusted level $\alpha' = \alpha/k$, and compare probabilities p_i to α' . This is equivalent to adjusting individual p-values p_i to $p'_i = kp_i$ and comparing p'_i to the unadjusted significance level α . While appropriate to test the null hypothesis for the whole set of simultaneous hypotheses (i.e. reject H_0 for the whole set of k hypotheses if the smallest unadjusted p-value in the set is less than or equal to α/k), the Bonferroni method is overly conservative and often leads to rejecting too few individual hypotheses in the set k .

Several alternatives have been proposed in the literature; see Wright (1992) for a review. For non-independent tests, Holm's procedure (1979) is nearly as simple to carry out as the Bonferroni adjustment and it is much more powerful, leading to rejecting the null hypothesis more often. It is computed as follows. (1) Order the p-values from left to right so that $p_1 \leq p_2 \leq \dots \leq p_i \leq \dots \leq p_k$. (2) Compute adjusted probability values $p'_i = (k - i + 1)p_i$; adjusted probabilities may be larger than 1. (3) Proceeding from left to right, if an adjusted p-value in the ordered series is smaller than the one occurring at its left, make the smallest equal to the largest one. (4) Compare each adjusted p'_i to the unadjusted α significance level and make the statistical decision. The procedure could be formulated in terms of successive corrections to the α significance level, instead of adjustments to individual probabilities.

An even more powerful solution is that of Hochberg (1988) which has the desired overall ("experimentwise") error rate α only for independent tests (Wright, 1992). Only step (3) differs from Holm's procedure: proceeding this time from right to left, if an adjusted p-value in the ordered series is smaller than the one at its left, make the largest equal to the smallest one. Because the adjusted probabilities form a nondecreasing series, both of these procedures present the properties (1) that a hypothesis in the ordered series cannot be rejected unless all previous hypotheses in the series have also been rejected and (2) that equal p-values receive equal adjusted p-values. Hochberg's method presents the further characteristic that no adjusted p-value can be larger than the largest unadjusted p-value or exceed 1. More complex and powerful procedures are explained by Wright (1992).

For some applications, special procedures have been developed to test a whole set of statistics. An example is the test for the correlation matrix \mathbf{R} (eq. 4.14, end of Section 4.2).

statistical population is another question, for which the readers are referred to the literature on sampling design.) We know, of course, that a sample is quite unlikely to produce a parameter estimate which is exactly equal to the true value of the parameter in the statistical population. A statistical test tries to answer the following question: given a hypothesis stating, for example, that $\rho = 0$ in the statistical population and the fact that the estimated correlation is, say, $r = 0.2$, is it justified to conclude that the difference between 0.2 and 0.0 is due to sampling error?

Pivotal
statistic

The choice of the statistic to be tested depends on the problem at hand. For instance, in order to find whether two samples may have been drawn from the same statistical population or from populations with equal means, one would choose a statistic measuring the difference between the two sample means ($\bar{x}_1 - \bar{x}_2$) or, preferably, a *pivotal* form like the usual t statistic used in such tests; a pivotal statistic has a distribution under the null hypothesis which remains the same for any value of the measured effect (here, $\bar{x}_1 - \bar{x}_2$). In the same way, the slope of a regression line is described by the slope parameter of the linear regression equation, which is assumed, under the null hypothesis, to be either zero or some other value suggested by ecological theory. The test statistic describes the difference between the observed and hypothesized value of slope; the pivotal form of this difference is a t or F statistic.

Alternative
hypothesis

Another aspect of a statistical test is the alternative hypothesis (H_1), which is also imposed by the ecological problem at hand. H_1 is the opposite of H_0 , but there may be several statements that represent some opposite of H_0 . In correlation analysis for instance, if one is satisfied to determine that the correlation coefficient in the reference population (ρ) is significantly different from zero in either the positive or the negative direction, meaning that *some* linear relationship exists between two variables, then a *two-tailed* alternative hypothesis is stated about the value of the parameter in the statistical population: $\rho \neq 0$. On the contrary, if the ecological phenomenon underlying the hypothesis imposes that a relationship, if present, should have a given sign, one formulates a *one-tailed* hypothesis. For instance, studies on the effects of acid rain are motivated by the general paradigm that acid rain, which lowers the pH, has a negative effect on terrestrial and aquatic ecosystems. In a study of the correlation between pH and diversity, one would formulate the following hypothesis H_1 : pH and diversity are positively correlated (i.e. low pH is associated with low diversity; $H_1: \rho > 0$). Other situations would call for a different alternative hypothesis, symbolized by $H_1: \rho < 0$.

The expressions *one-tailed* and *two-tailed* refer to the fact that, in a two-tailed test, one would look in both tails of the reference statistical distribution for values as extreme as, or more extreme than the reference value of the statistic (i.e. the one computed from the actual data). In a correlation study for instance, where the reference distribution (t) for the test statistic is symmetric about zero, the probability of the null hypothesis in a two-tailed test is given by the proportion of values in the t distribution which are, *in absolute value*, as large as, or larger than the *absolute value* of the reference statistic. In a one-tailed test, one would look only in the tail corresponding to the sign given by the alternative hypothesis; for instance, for the proportion of values

in the t distribution which are as large as or larger than the *signed value* of the reference t statistic, for a test in the right-hand tail ($H_1: \rho > 0$).

In standard statistical tests, the *test statistic* computed from the data is referred to one of the usual statistical distributions printed in books or computed by some appropriate computer software; the best-known are the z , t , F and χ^2 distributions. This, however, can only be done if certain assumptions are met by the data, depending on the test. The most commonly encountered are the assumptions of normality of the variable(s) in the reference population, homoscedasticity (Box 1.4) and independence of the observations (Box 1.1). Refer to Siegel (1956, Chapter 2), Siegel & Castellan (1988, Chapter 2), or Snedecor & Cochran (1967, Chapter 1), for concise yet clear classical exposés of the concepts related to statistical testing.

2 — *Permutation tests*

Randomi-
zation

The method of *permutation*, also called *randomization*, is a very general approach to testing statistical hypotheses. Following Manly (1997), permutation and randomization are considered synonymous in the present book, although *permutation* may also be considered to be the technique by which the principle of *randomization* is applied to data during permutation tests. Other points of view are found in the literature. For instance, Edgington (1995) considers that a randomization test is a permutation test based on randomization. A different although related meaning of *randomization* refers to the random assignment of replicates to treatments in experimental designs.

Permutation testing can be traced back to at least Fisher (1935, Chapter 3). Instead of comparing the actual value of a test statistic to a standard statistical distribution, the reference distribution is generated from the data themselves, as described below; other randomization methods are mentioned at the end of the present Section. Permutation provides an efficient approach to testing when the data do not conform to the distributional assumptions of the statistical method one wants to use (e.g. normality). Permutation testing is applicable to very small samples, like nonparametric tests. It *does not* resolve problems of independence of the observations, however. Nor does the method solve distributional problems that are linked to the hypothesis subjected to a test*. Permutation remains the method of choice to test novel or other statistics whose distributions are poorly known. Furthermore, results of permutation tests are valid even with observations that are not a random sample of some statistical population; this point is further discussed in Subsection 4. Edgington (1995) and Manly (1997)

* For instance, when studying the differences among sample means (two groups: t -test; several groups: F test of ANOVA), the classical Behrens-Fisher problem (Robinson, 1982) reminds us that two null hypotheses are tested simultaneously by these methods, i.e. equality of the means and equality of the variances. Testing the t or F statistics by permutations does not change the dual aspect of the null hypothesis; in particular, it does not allow one to unambiguously test the equality of the means without checking first the equality of the variances using another, more specific test (two groups: F ratio; several groups: Bartlett's test of equality of variances).

have written excellent introductory books about the method. A short account is given by Sokal & Rohlf (1995) who prefer to use the expression “randomization test”. Permutation tests are used in several Chapters of the present book.

The speed of modern computers would allow users to perform any statistical test using the permutation method. The chief advantage is that one does not have to worry about distributional assumptions of classical testing procedures; the disadvantage is the amount of computer time required to actually perform a large number of permutations, each one being followed by recomputation of the test statistic. This disadvantage vanishes as faster computers come on the market. As an example, let us consider the situation where the significance of a correlation coefficient between two variables, x_1 and x_2 , is to be tested.

Hypotheses

- H_0 : The correlation between the variables in the reference population is zero ($\rho = 0$).
- For a two-tailed test, $H_1: \rho \neq 0$.
- Or for a one-tailed test, either $H_1: \rho > 0$, or $H_1: \rho < 0$, depending on the ecological hypothesis.

Test statistic

- Compute the Pearson correlation coefficient r . Calculate the pivotal statistic $t = \sqrt{n-2} [r / \sqrt{1-r^2}]$ (eq. 4.13; n is the number of observations) and use it as the reference value in the remainder of the test.

In this specific case, the permutation test results would be the same using either r or t as the test statistic, because t is a monotonic function of r for any constant value of n ; r and t are “equivalent statistics for permutation tests”, *sensu* Edgington (1995). This is not always the case. When testing a partial regression coefficient in multiple regression, for example, the test should not be based on the distribution of permuted partial regression coefficients because they are not monotonic to the corresponding partial t statistics. The partial t should be preferred because it is pivotal and, hence, it is expected to produce correct type I error.

Considering a pair of equivalent test statistics, one could choose the statistic which is the simplest to compute if calculation time would otherwise be longer in an appreciable way. This is not the case in the present example: calculating t involves a single extra line in the computer program compared to r . So the test is conducted using the usual t statistic.

Distribution of the test statistic

The argument invoked to construct a null distribution for the statistic is that, if the null hypothesis is true, all possible pairings of the two variables are equally likely to occur.

The pairing found in the observed data is just one of the possible, equally likely pairings, so that the value of the test statistic for the unpermuted data should be typical, i.e. located in the central part of the permutation distribution.

- It is always the null hypothesis which is subjected to testing. Under H_0 , the rows of \mathbf{x}_1 are seen as “exchangeable” with one another if the rows of \mathbf{x}_2 are fixed, or conversely. The observed pairing of \mathbf{x}_1 and \mathbf{x}_2 values is due to chance alone; accordingly, any value of \mathbf{x}_1 could have been paired with any value of \mathbf{x}_2 .
- A realization of H_0 is obtained by permuting at random the values of \mathbf{x}_1 while holding the values of \mathbf{x}_2 fixed, or the opposite (which would produce, likewise, a random pairing of values). Recompute the value of the correlation coefficient and the associated t statistic for the randomly paired vectors \mathbf{x}_1 and \mathbf{x}_2 , obtaining a value t^* .
- Repeat this operation a large number of times (say, 999 times). The different permutations produce a set of values t^* obtained under H_0 .
- Add to these the reference value of the t statistic, computed for the unpermuted vectors. Since H_0 is being tested, this value is considered to be one that could be obtained under H_0 and, consequently, it should be added to the reference distribution (Hope, 1968; Edgington, 1995; Manly, 1997). Together, the unpermuted and permuted values form an estimate of the sampling distribution of t under H_0 , to be used in the next step.

Statistical decision

- As in any other statistical test, the decision is made by comparing the reference value of the test statistic (t) to the reference distribution obtained under H_0 . If the reference value of t is typical of the values obtained under the null hypothesis (which states that there is no relationship between \mathbf{x}_1 and \mathbf{x}_2), H_0 cannot be rejected; if it is unusual, being too extreme to be considered a likely result under H_0 , H_0 is rejected and the alternative hypothesis is considered to be a more likely explanation of the data.

Significance level • The significance level of a statistic is the proportion of values that are as extreme as, or more extreme than the test statistic in the reference distribution, which is either obtained by permutations or found in a table of the appropriate statistical distribution. The level of significance should be regarded as “the strength of evidence against the null hypothesis” (Manly, 1997).

3 — Numerical example

Let us consider the following case of two variables observed over 10 objects:

\mathbf{x}_1	-2.31	1.06	0.76	1.38	-0.26	1.29	-1.31	0.41	-0.67	-0.58
\mathbf{x}_2	-1.08	1.03	0.90	0.24	-0.24	0.76	-0.57	-0.05	-1.28	1.04

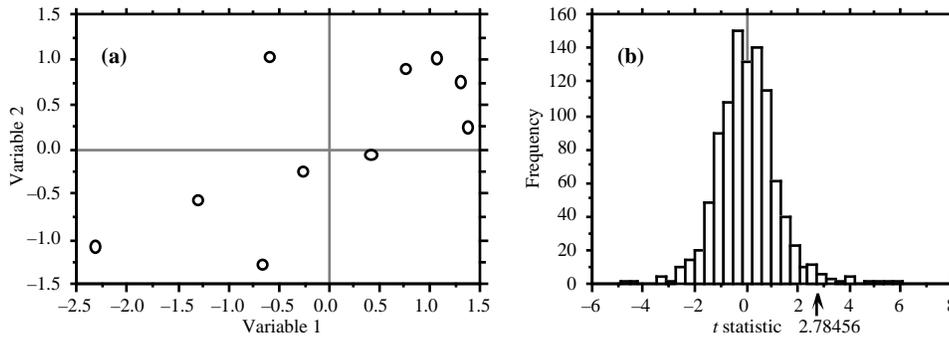


Figure 1.6 (a) Positions of the 10 points of the numerical example with respect to variables x_1 and x_2 . (b) Frequency histogram of the (1+999) permutation results (t statistic for correlation coefficient); the reference value obtained for the points in (a), $t = 2.78456$, is also shown.

These values were drawn at random from a positively correlated bivariate normal distribution, as shown in Fig. 1.6a. Consequently, they would be suitable for parametric testing. So, it is interesting to compare the results of a permutation test to the usual parametric t -test of the correlation coefficient. The statistics and associated probabilities for this pair of variables, for $\nu = (n - 2) = 8$ degrees of freedom, are:

$r = 0.70156, t = 2.78456, n = 10$;
 prob (one-tailed) = 0.0119, prob (two-tailed) = 0.0238.

There are $10! = 3.6288 \times 10^6$ possible permutations of the 10 values of variable x_1 (or x_2). Here, 999 of these permutations were generated using a random permutation algorithm; they represent a random sample of the 3.6288×10^6 possible permutations. The computed values for the test statistic (t) between permuted x_1 and fixed x_2 have the distribution shown in Fig. 1.6b; the reference value, $t = 2.78456$, has been added to this distribution. The permutation results are summarized in the following table, where “ $|t|$ ” is the (absolute) reference value of the t statistic ($|t| = 2.78456$) and “ t^* ” is a value obtained after permutation. The absolute value of the reference t is used in the table to make it a general example, because there are cases where t is negative.

	$t^* < - t $	$t^* = - t $	$- t < t^* < t $	$t^* = t $	$t^* > t $
Statistic t	8	0	974	1 [†]	17

[†] This count corresponds to the reference t value added to the permutation results.

For a one-tailed test (in the right-hand tail in this case, since $H_1: \rho > 0$), one counts how many values in the permutational distribution of the statistic are equal to, or larger than, the reference value ($t^* \geq t$, there are $1 + 17 = 18$ such values in this case). This is

the only one-tailed hypothesis worth considering, because the objects are known in this case to have been drawn from a positively correlated distribution. A one-tailed test in the left-hand tail ($H_1: \rho < 0$) would be based on how many values in the permutational distribution are equal to, or smaller than, the reference value ($t^* \leq t$, which are $8 + 0 + 974 + 1 = 983$ in the example). For a two-tailed test, one counts all values that are as extreme as, or more extreme than the reference value *in both tails of the distribution* ($|t^*| \geq |t|$, which are $8 + 0 + 1 + 17 = 26$ in the example).

Probabilities associated with these distributions are computed as follows, for a one-tailed and a two-tailed test (results using the t statistic would be the same):

One-tailed test [$H_0: \rho = 0$; $H_1: \rho > 0$]:
 $\text{prob}(t^* \geq 2.78456) = (1 + 17)/1000 = 0.018$

Two-tailed test [$H_0: \rho = 0$; $H_1: \rho \neq 0$]:
 $\text{prob}(|t^*| \geq 2.78456) = (8 + 0 + 1 + 17)/1000 = 0.026$

Note how similar the permutation results are to the results obtained from the classical test, which referred to a table of Student t distributions. The observed difference is partly due to the small number of pairs of points ($n = 10$) sampled at random from the bivariate normal distribution, with the consequence that the data set does not quite conform to the hypothesis of normality. It is also due, to a certain extent, to the use of only 999 permutations, sampled at random among the $10!$ possible permutations.

4 — *Remarks on permutation tests*

In permutation tests, the reference distribution against which the statistic is tested is obtained by randomly permuting the data under study, without reference to any statistical population. The test is valid as long as the reference distribution has been generated by a procedure related to a null hypothesis that makes sense for the problem at hand, irrespective of whether or not the data set is representative of a larger statistical population. This is the reason why the data do not have to be a random sample from some larger statistical population. The only information the permutation test provides is whether the pattern observed in the data is likely, or not, to have arisen by chance. For this reason, one may think that permutation tests are not as “good” or “interesting” as classical tests of significance because they might not allow one to infer conclusions that apply to a statistical population.

A more pragmatic view is that the conclusions of permutation tests may be generalized to a reference population if the data set is a random sample of that population. Otherwise, they allow one to draw conclusions only about the particular data set, measuring to what extent the value of the statistic is “usual” or “unusual” with respect to the null hypothesis implemented in the permutation procedure. Edgington (1995) and Manly (1997) further argue that data sets are very often not drawn at random from statistical populations, but simply consist of observations which happen

to be available for study. The generalization of results, in classical as well as permutation tests, depends on the degree to which the data were actually drawn at random, or are equivalent to a sample drawn at random, from a reference population.

Complete permutation test For small data sets, one can compute all possible permutations in a systematic way and obtain the complete permutation distribution of the statistic; an *exact* or *complete permutation test* is obtained. For large data sets, only a sample of all possible permutations may be computed because there are too many. When designing a *sampled permutation test*, it is important to make sure that one is using a *uniform random generation algorithm*, capable of producing all possible permutations with equal probabilities (Furnas, 1984). Computer programs use procedures that produce random permutations of the data; these in turn call the 'Random' function of computer languages. Such a procedure is described in Section 5.8 of Manly's book (1997). Random permutation subprograms are also available in subroutine libraries.

The case of the correlation coefficient has shown how the null hypothesis guided the choice of an appropriate permutation procedure, capable of generating realizations of this null hypothesis. A permutation test for the difference between the means of two groups would involve random permutations of the objects between the two groups instead of random permutations of one variable with respect to the other. The way of permuting the data depends on the null hypothesis to be tested.

Some tests may be reformulated in terms of some other tests. For example, the *t*-test of equality of means is equivalent to a test of the correlation between the vector of observed values and a vector assigning the observations to group 1 or 2. The same value of *t* and probability (classical or permutational) are obtained using both methods.

Restricted permutations Simple statistical tests such as those of correlation coefficients or differences between group means may be carried out by permuting the original data, as in the example above. Problems involving complex relationships among variables may require permuting the residuals of some *model* instead of the raw data; *model-based permutation* is discussed in Subsection 11.3.2. The effect of a nominal covariable may be controlled for by *restricted permutations*, limited to the objects within the groups defined by the covariable. This method is discussed in detail by Manly (1997). Applications are found in Brown & Maritz (1982; restrictions within replicated values in a multiple regression) and in Sokal *et al.* (1987; Mantel test), for instance.

In sampled permutation tests, adding the reference value of the statistic to the distribution has the effect that it becomes impossible for the test to produce no value "as extreme as, or more extreme than the reference value", as the standard expression goes. This way of computing the probability is biased, but it has the merit of being statistically valid (Edgington, 1995, Section 3.5). The precision of the probability estimate is the inverse of the number of permutations performed; for instance, after $(999 + 1)$ permutations, the precision of the probability statement is 0.001.

How many permutations?

The number of permutations one should perform is always a trade-off between precision and computer time. The more permutations the better, since probability estimates are subject to error due to sampling the population of possible permutations (except in the rare cases of complete permutation tests), but it may be tiresome to wait for the permutation results when studying large data sets. In the case of the Mantel test (Section 10.5), Jackson & Somers (1989) recommend to compute 10000 to 100000 permutations in order to ensure the stability of the probability estimates. The following recommendation can be made. In exploratory analyses, 500 to 1000 permutations may be sufficient as a first contact with the problem. If the computed probability is close to the preselected significance level, run more permutations. In any case, use more permutations (e.g. 10000) for final, published results.

Interestingly, tables of critical values in nonparametric statistical tests for small sample sizes are based on permutations. The authors of these tables have computed how many cases can be found, in the complete permutation distribution, that are as extreme as, or more extreme than the computed value of the statistic. Hence, probability statements obtained from small-sample nonparametric tests are exact probabilities (Siegel, 1956).

Monte Carlo

Named after the famous casino of the principality of Monaco, Monte Carlo methods use random numbers to study either real data sets or the behaviour of statistical methods through simulations. Permutation tests are Monte Carlo methods because they use random numbers to randomly permute data. Other such methods are based on computer-intensive resampling. Among these are the jackknife (Tukey 1958; Sokal & Rohlf, 1995) and the bootstrap (Efron, 1979; Efron & Tibshirani, 1993; Manly, 1997). In these methods, the values used in each iteration to compute a statistic

Jackknife
Bootstrap

form a subsample of the original data. In the jackknife, each subsample leaves out one of the original observations. In the bootstrap, each subsample is obtained by resampling the original sample *with replacement*; the justification is that resampling the original sample approximates a resampling of the original population.

As an exercise, readers are invited to figure out how to perform a permutation test for the difference between the means of two groups of objects on which a single variable has been measured, using the t statistic; this would be equivalent to a t -test. A solution is given by Edgington (1995). Other types of permutation tests are discussed in Sections 7.3, 8.9, 10.2, 10.3, 10.5, 10.6, 11.3, 12.6, 13.1 and 13.3.