7. Finite factors

- 7.1 Overview - Finite factors
- 7.2 Dichotomy: fixed vs random factors
- 7.3 Not a dichotomy: a progression from fixed to random
- 7.4 Example: environmental impact on molluscs
- 7.5 Broader implications for detecting impact

7.1 Overview - Finite factors

ANOVA is one of the most widely used statistical techniques, providing a partitioning of the measured variation of a random variable in response to one or more factors in complex experimental designs and sampling programmes. A *factor* is a categorical variable that identifies several groups or *levels* that are of special interest to the researcher (e.g., treatment *vs* controls), or that are contributing a potentially important source of variation in the study design (e.g., sites). To make rigorous inferences in multi-factorial ANOVA settings, we need to ascertain, for each and every factor in a given experiment or sampling protocol, whether that factor is *fixed or random*. Classically, the levels of a fixed factor are viewed as being finite, while those of a random factor are viewed as being drawn randomly from an infinite (or, at least, an uncountably large) population of possible levels. The choice of whether any given factor is fixed or random is viewed as a dichotomy. The need to make appropriate decisions about this for every factor in the design before embarking on any statistical analysis is essential for dissimilarity-based PERMANOVA, just as it is for univariate ANOVA. There are important consequences of these choices on the results and the inferences that can be drawn from them.

What would happen if we have a factor that would typically be thought of and treated as random, but the population of possible levels is *finite* ? Well, if we can sample *all* of the levels, we might then just treat the random factor as fixed. However, what if we can't sample all of them, but we *can* sample a *substantial fraction* of them? Anderson *et al.* (2025) describe how the dichotomy of fixed *vs* random can, instead, be viewed as a *progression*, which depends on how much of the population of possible levels of a given factor has actually been sampled (i.e., the sampling fraction).

Finite factors tend to occur at large spatial scales. For example, suppose there is a cluster of 20 islands in a given region, and suppose 10 of these have undergone some intensive restoration of habitat. We may not be able to sample all of the islands, but perhaps we can sample 4 restored and 4 unrestored islands (in each case, out of a possible 10). By treating the factor of 'Islands' as 'finite', and specifying the size of the population and hence identifying the sampling fraction (the sampling fraction here is 4/10), we are able to get much stronger and more powerful inferences regarding the effectiveness of the restoration than we would otherwise obtain if we were to treat the factor of 'Islands' as random.

7.2 Dichotomy: fixed vs random factors

Consider the classical one-way linear ANOVA model, as described in section 6.2 above. Specifically, we have a random variable Y, and we have taken a sample of size n_i from each of $i = 1,\ldots, a$ groups to obtain observed values y_{ij}. Thus, factor A has a groups or levels. We can visualise the ANOVA model graphically as shown in Fig. 7.1.

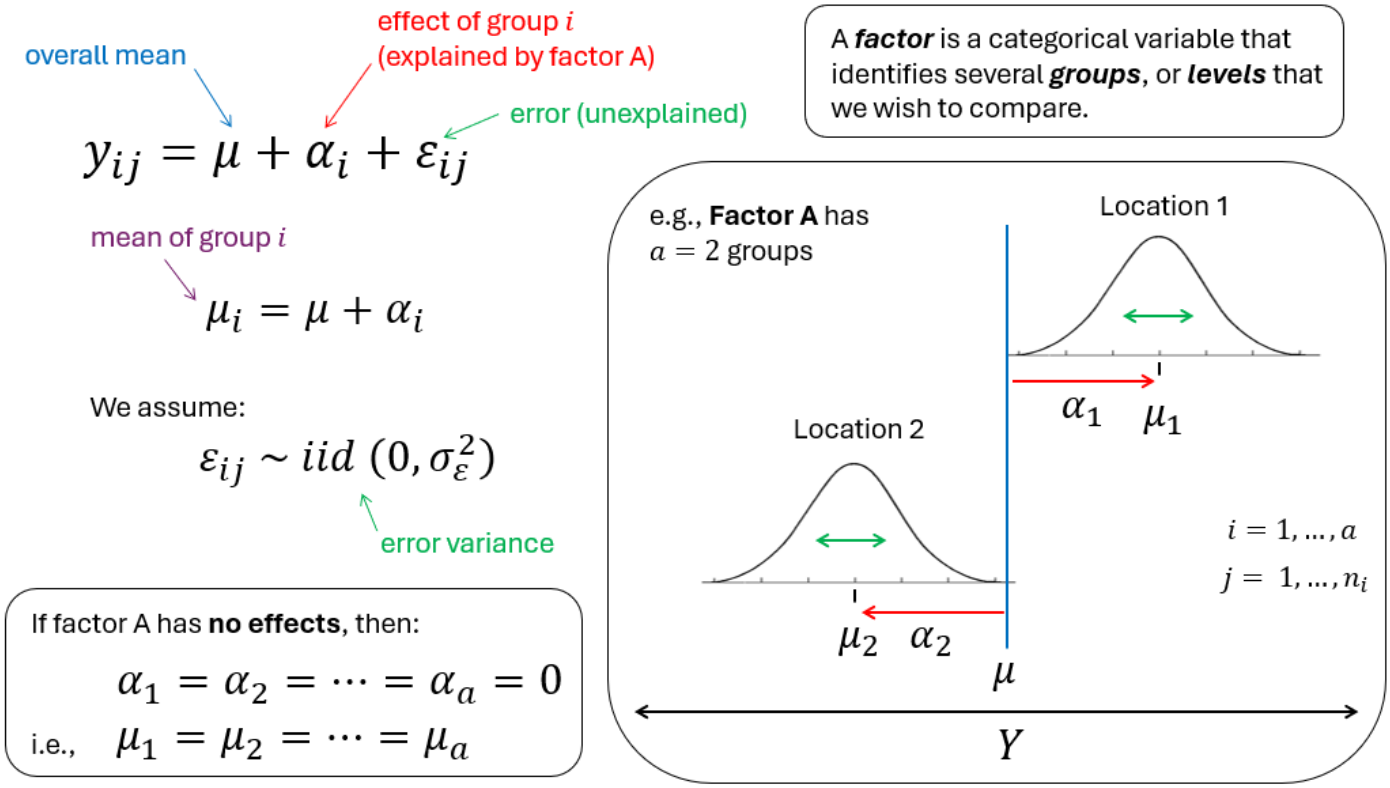


Fig. 7.1. A schematic diagram of the ANOVA linear model.

Specifically, we can imagine (in the univariate case) that Y is represented on a number line going from left to right. To arrive at any particular value y_{ij}, we begin at the *overall mean*, μ. To this we add the *effect* of being in a particular group, α_i. If the effect is non-zero it will shift us (up or down) along the number line a distance of α_i away from the overall mean, μ, to arrive at the position μ_i, which is the mean for group i. To arrive at the value equal to our particular observation y_{ij}, we must then also add the *error*, ε_{ij} associated with that individual observation.

What do we assume about the overall mean, the effects and the errors?

- For the overall mean, we (classically) assume that it is a fixed unknown constant.
- For the errors, we assume that they are each drawn randomly and independently from an (effectively) infinite population distribution that has a mean of zero and a variance of σ_ε^2.

- What we assume about the effects depends on whether factor A is fixed or random.
  - If factor A is *fixed*, then the effects, α_i, are considered to be fixed unknown constants (like μ).
  - If factor A is *random*, then the effects, α_i, are considered each to have been drawn randomly and independently from an (effectively) infinite population distribution that has a mean of zero and a variance of σ_α^2.

The criteria that are typically used in order to decide whether an individual factor is fixed or random are given in the Table below.

Table 7.1. Criteria for identifying a given factor as either fixed or random.

Fixed factor	**Random factor**
The effects, α_i, are fixed unknown constants. There is a finite number of levels, and all of them (or all of them of interest) occur in the study.	The effects, α_i, are a subset of levels drawn from an infinite (or uncountably large) population of possible levels.
If we were to repeat the study, the same levels would be chosen.	If we were to repeat the study, the same levels might not be chosen.
Interest lies in the individual effects, α_i.	Interest lies in the variance component, σ_α^2.
We will want to do pair-wise comparisons for this factor, if the factor is significant.	We are typically not interested in pair-wise comparisons for this factor.
H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_a = 0$	H_0: $\sigma_\alpha^2 = 0$
Inferences are about only the specific levels of the factor included in our study.	Inferences are about the whole population of possible levels for that factor that we could have sampled.

Examples of *fixed factors* could be:

- Habitat: {seagrass, kelp forest, rocky reef}
- Maturity: {adult, juvenile}
- Treatments: {high-dose, low-dose, placebo, control}
- Status: {pristine, restored, disturbed}

In each of the above examples, we can see that the names of the levels have a particular meaning in the context of the experiment or sampling design. These levels are not a random sample of possible levels. In each case, they correspond to something specific that we have chosen *a priori* to investigate. We would definitely like to know, for example, what the effect is of (say) the low-dose treatment compared to that of the high-dose treatment, if any. If the levels of the factor have specific names or labels like this, then this is a direct indication that the factor is fixed; the researcher cares about the levels themselves, and has set up the sampling design or experiment to measure and understand these particular effects. We would (almost certainly) be keen to follow-up any significant F-ratio with subsequent pair-wise comparison tests to examine how the individual levels might differ from one another.

Even if the levels included in the study do not correspond to 'all possible levels' of a given factor, there is still a clear sense that the levels of a fixed factor that are included in the study correspond to all those that are of interest. For example, suppose we design an experiment to investigate the

effect of temperature on the growth of an organism, and have set up a series of (replicated) microcosms at each of three different levels: {20°C, 23°C, and 26°C}. Clearly, we have not sampled all possible temperatures. Nevertheless, these are the only temperatures about which we intend to make inferences, and we would therefore treat 'Temperature' as a fixed effect.

Examples of *random factors* could be:

- Batches: {B_1, B_2, B_3, ...}
- Transects: {T_1, T_2, T_3, ...}
- Sites: {S_1, S_2, S_3, ...}
- Years: {Y_1, Y_2, Y_3, ...}

For random factors, typically the levels are not, individually, of any particular interest. A random factor might be included in a study purely for logistic reasons. For example, suppose you are studying the behavioural response of soldier crabs to human by-standers. It may not be feasible to get sufficient replication of your experimental protocols in the field all at once, simultaneously, so you might have to run your study in batches over a period of days. Thus, 'Batch' (and/or 'Day') then becomes a random factor that contributes a potential source of variation to your results.

In other cases, random factors occur because interest lies precisely in measuring variability at one or more spatial or temporal scales. For example, we might expect variation in the abundance of settlers for organisms that are brooders to be higher than that of broadcast spawners in marine environments at large spatial scales. Settlers of each type of organism (brooders and spawners) can be monitored at a series of sites (separated by, say, hundreds of metres), and the factor of 'Sites' would be a random factor. There may be a very large number of sites that we could have sampled, but we will likely be able to sample only a very tiny fraction of these as a subset (suppose we sample 10 sites). Measuring the abundance of settlers in replicate quadrats at each of the a = 10 sites, we care not at all how the mean may differ between (say) site 3 and site 8; pair-wise comparisons are of no interest. We *do* wish to examine if site-to-site variation is detectable over-and-above residual variation (hence, to test H_0: $\sigma_\alpha^2 = 0$), and, if so, to estimate the size of σ_α^2 separately for brooders and spawners.

Importantly, whenever we set up an experimental/sampling design (e.g., in a'Design' file in PERMANOVA for PRIMER), the choices that are made in this regard (for each factor) will have very important consequences for:

- the assumptions underlying the PERM(ANOVA) model;
- the expected values of mean squares (EMS) for each term in the model;
- the construction of an appropriate F ratio for individual terms in the model;
- the construction of an appropriate permutation algorithm (under exchangeability);
- the hypothesis being tested by the F ratio; and
- the extent and the nature of the statistical inferences.

7.3 Not a dichotomy: a progression from fixed to random

What is meant by a 'finite' factor?

Suppose, for any factor, there are a total of A levels in the population. In some cases, A is absolutely enormous and it may be effectively infinite in the sense of being uncountable (e.g., blades of seagrass in a large seagrass meadow). In other cases, A might well be *finite* (e.g., there might only be a total of A = 10 restored areas).

In any given study, the researcher may sample (i.e., randomly and representatively draw, without replacement) a levels out of the A total possible levels for any given factor. The *sampling fraction* is therefore a/A. The larger the sampling fraction, the more the researcher will know about the system, hence, the greater the potential power in drawing inferences from the study about that factor.

Now, a fixed factor occurs where all possible levels are drawn, so $a=A$ and the sampling fraction is $a/A = 1$. The finiteness of fixed factors is quite clear. For example, the levels 'treatment' and 'control' do not come from a wider population: together they comprise a 'population' of only 2 levels. These are naturally the only two levels of interest in the study and $a = A$ = 2.

On the other hand, a random factor occurs where A is extremely large (effectively infinite), and hence our sampling fraction a/A is very tiny (approaching zero in the limit).

A progression of steps from fixed to random

It is quite easy to conceive, however, of a finite population of levels where A is known, but we cannot sample all possible levels, and $A > a$. For example, suppose I am able to sample a = 4 restored habitats (islands) out of a total of A = 10 restored habitats that occur in a given region of interest. In this case, the sampling fraction is a/A = 4/10 = 2/5. This fraction is neither trivially small (random), yet nor is it precisely equal to 1 (fixed). In this way, more generally, we can see that there is an incremental progression of steps, from fixed to random, that depends on the sampling fraction (Fig. 7.2).

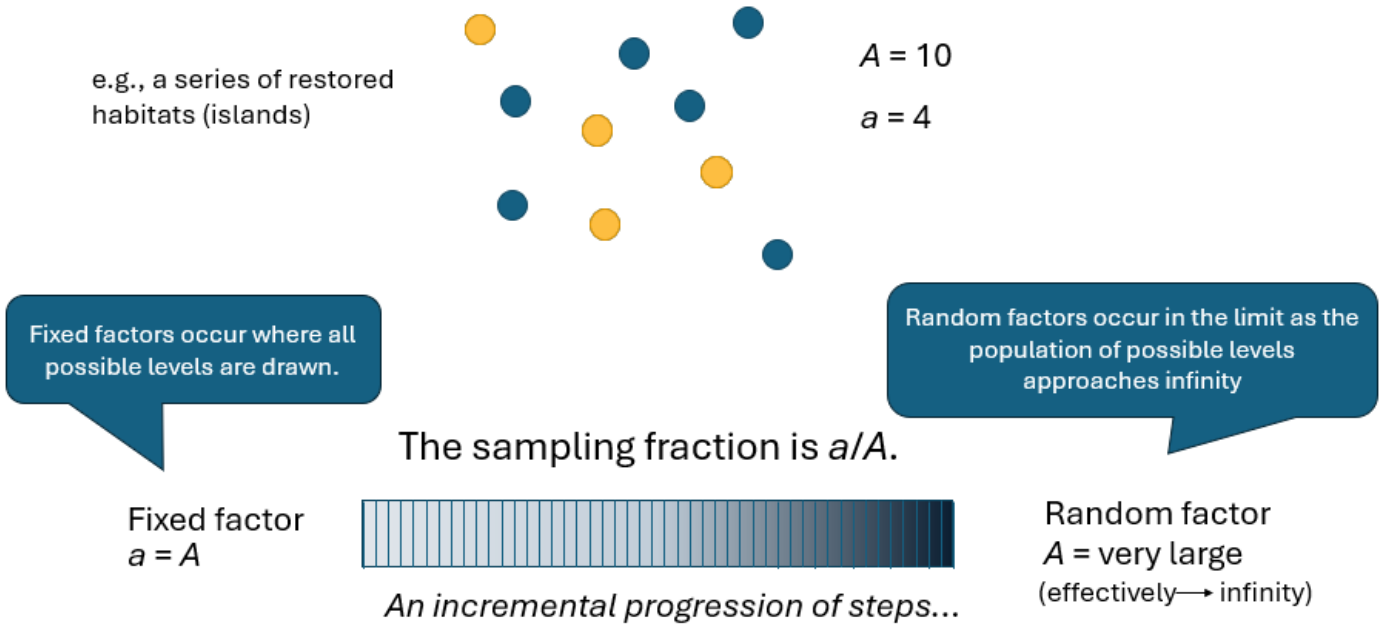


Fig. 7.2. A series of steps in the progression from fixed to random factors.

The finiteness of the population of possible levels will tend to become more apparent (and more important) as the spatial or temporal scale of the factor gets larger. For example, suppose I repeat an experiment on the effects of fish predators at each of three separate bays along a coastline. I may well wish to include the factor of 'Embayment' in my study design. What are these three embayments intended to represent? Are there many such embayments, or only a handful? Have I sampled all of them or a substantial fraction of them? These are important questions to answer so as to ensure we achieve maximum power to test relevant hypotheses in our study.

Statistical derivations of EMS

Cornfield & Tukey (1956) articulated the concept of the experimenter sampling levels of factors from finite *vs* infinite populations, and they showed the resulting outcomes for expectations of mean squares (EMS), and therefore how to construct correct F tests, in two-way and three-way crossed balanced designs for univariate ANOVA cases.

Anderson *et al.* (2025) combined these results with the landmark work by Hartley (1967), Rao (1968) and Hartley *et al.* (1978) for balanced and unbalanced cases, thereby incorporating the sampling fraction from finite populations into the derivation of EMS by 'synthesis' for any general complex ANOVA design. Anderson *et al.* (2025) further extended these results to multivariate dissimilarity-based tests using PERMANOVA. The new PERMANOVA routine in PRIMER 8 fully implements the methodology described by Anderson *et al.* (2025), with correct tests constructed by reference to the EMS *via* 'synthesis'. Specifically, the new PERMANOVA routine in PRIMER 8 permits:

- any individual factor to be specified as 'fixed', 'random' or 'finite' (a new factor type);
- the size(s) of any finite populations (i.e., the total number of levels) to be specified for each finite factor;
- finite factors in asymmetrical designs (e.g., where there may be different numbers of levels in different parts of the study design, such as 1 impact location and multiple controls).

Motivation

Motivation for the development of an option to fit finite factors in PERMANOVA arises especially in the context of ecological studies of environmental impact. We wish generally to permit flexibility in the definitions of factors where the sampling fraction is neither equal to 1, nor infinitely small. Studies of environmental impact will often contrast responses of organisms measured at a purportedly impacted location *vs* one or more 'control' (reference or unimpacted) locations (Underwood (1991) , Underwood (1992)). In such a design, one views the control locations as being a random sample from some larger population of control locations that are (apart from the impact itself) environmentally similar to the impacted location (Underwood (1994) , Glasby (1997)). It is desirable to sample as many control locations as logistics/time/funding will permit, so as to increase both the power of the test and the scope of the inferences (Glasby (1997) , Glasby & Underwood (1998)). In practice, however, the population of possible control locations is likely to be both finite and limiting, particularly at large spatial scales. In such cases, we might consider the (single) impacted location as being 'fixed' (e.g., a single oil spill, a single sewage outfall, a single storm, etc.), while the reference locations can be treated as either random or drawn from a finite population of a specified size.

Next, we shall provide an example of a PERMANOVA analysis involving a finite factor in the context of a study of the potential environmental impact of a sewage outfall on mollusc assemblages inhabiting rocky subtidal habitats on the coast of Italy (Mediterranean Sea).

7.4 Example: environmental impact on molluscs

The study design

We consider here a study examining effects of a sewage outfall for p = 151 mollusc species from subtidal habitats (3-4 m depth) in the Mediterranean Sea along the southwestern coast of Apulia, Italy (Terlizzi *et al.* (2005)). Abundances of each species were obtained from each of n = 9 replicate 20 cm x 20 cm quadrats in each of three random sites (separated by 80-100 m) at the outfall location ('I', putatively impacted), and at each of two control locations ('C1' and'C2'). Overall, there was therefore a total of N = 81 sampling units: 9 replicates within each of 3 sites within each of 3 locations (Fig.7.3).

Important: the two control locations were chosen randomly from a set of 8 possible such locations along that particular coastline, which were separated by at least 2.5 km and which provided comparable environmental conditions (in terms of slope, wave exposure, type of substratum) to those occurring at the outfall.

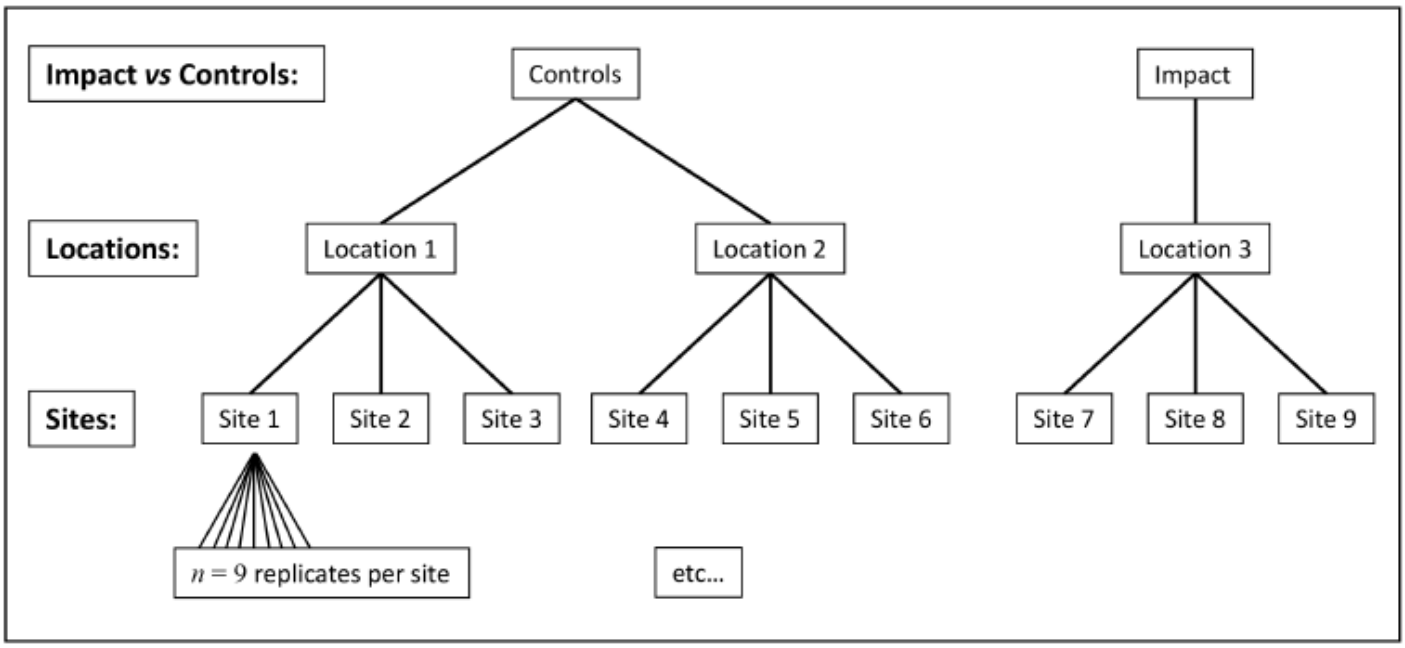


Fig. 7.3 Schematic diagram of the hierarchical design used by Terlizzi et al. (2005) to examine effects of a sewage outfall on subtidal molluscan assemblages.

The ANOVA model for this design has three factors:

- Impact *vs* Controls ('IvC', fixed with a = 2 levels)
- Locations ('L', nested in IvC, with b_1 = 2 controls and b_2 = 1 impact location)

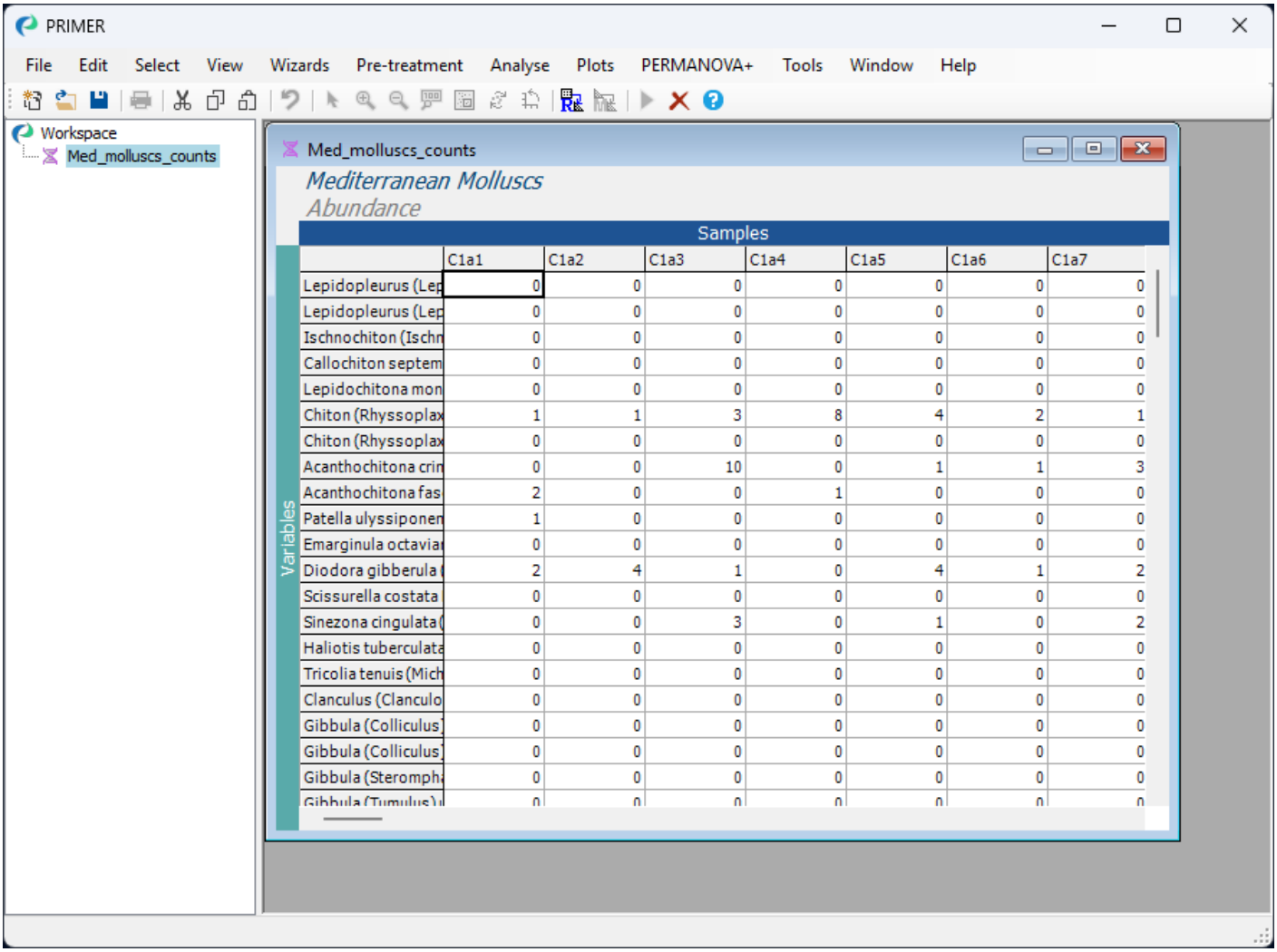
- Sites ('S', random and nested in L with $c_{j(i)}$ = 3 for all $j = 1,\ldots,b_i$ and all $i=1,\ldots, a$.

Is the factor of 'Locations' fixed or random? Well, the b_1 = 2 control locations were chosen randomly from a *finite population*, having a total size of B_1 = 8 possible control locations. This yields a sampling fraction of b_1/B_1 = 1/4 for the controls, whereas there was only one (fixed) impact location, hence B_2 = 1 and b_2/B_2 = 1. When we set up the design file to run this model to assess the response of the whole assemblage (based on the Bray-Curtis resemblance measure), using the PERMANOVA routine in PRIMER 8, we will be able to specify precisely this design, *including* details of the finite nature of the population of control locations from which we have sampled.

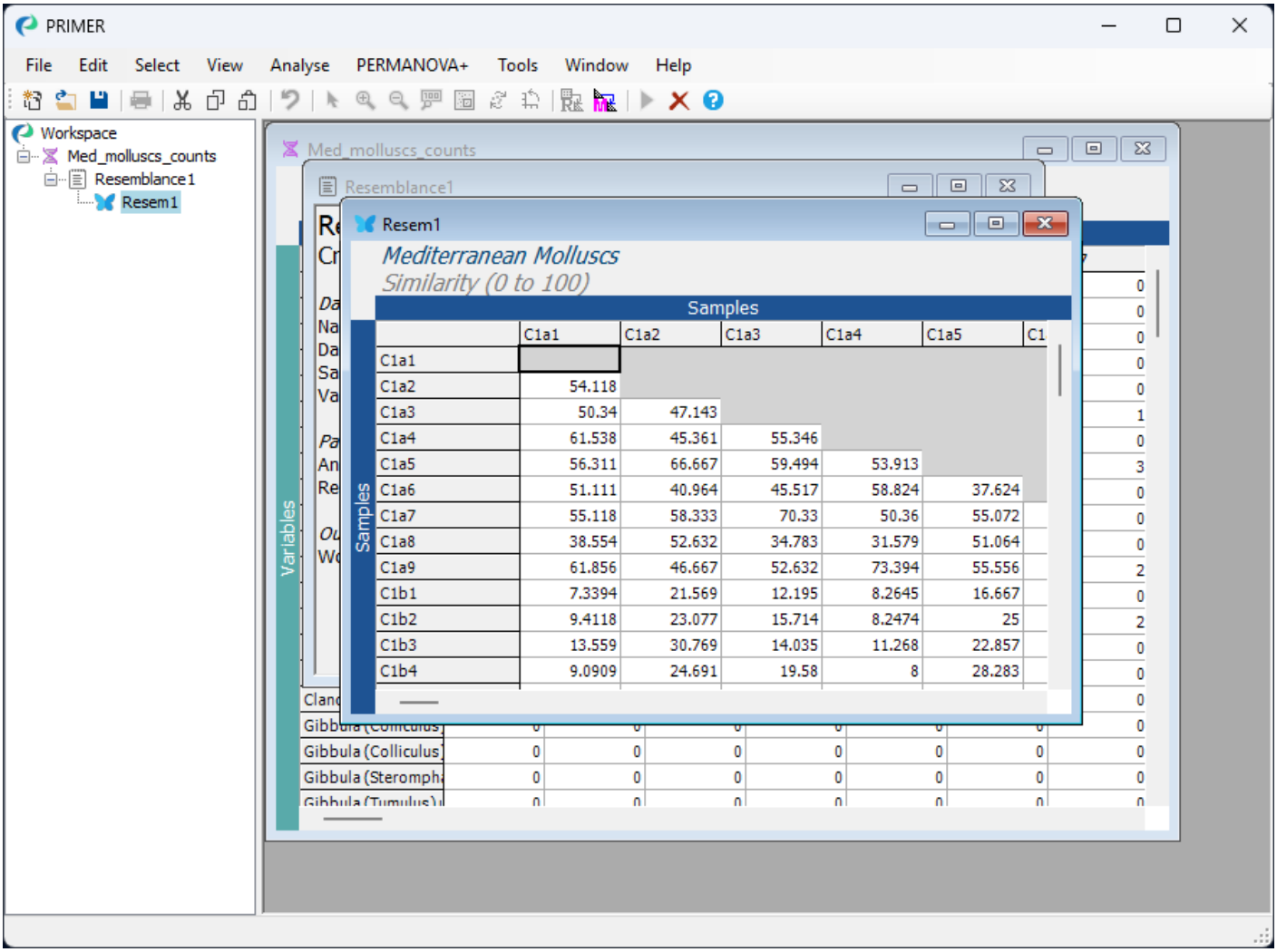
Note that the onus is on us, as researchers, to specify the size(s) of any finite population(s) of levels for each factor as precisely as we can, as part of our design. This means we have to consider carefully just what the population of levels actually is that we are drawing from for any random factor, and, therefore, the genuine spatial extent of the inferences we intend, and will be able to make, from the analysis. In some cases, particularly at large spatial scales, as in this case, a random factor may be *finite*. This shifts it, along the sampling-fraction progression, towards being considered more like (though not entirely as) a fixed factor (see Fig. 7.2). Conceptually, it makes sense that if we know more about the population (because we have sampled a greater fraction of its possible levels), then we will accordingly (generally) have more power to draw specific inferences about that population.

Examine patterns in the multivariate data

1. Start running **PRIMER 8**, then click **File** > **Open...** to open the data file named 'Med_molluscs_counts.pri' (found inside the 'Examples_P8 > Med_molluscs' folder).



2. Get the resemblance matrix among the sampling units based on the Bray-Curtis measure. Click **Analyse** > **Resemblance...** > (Measure > $\bullet$ Bray-Curtis similarity) & (Analyse between > $\bullet$ Samples). The resulting resemblance matrix will be called 'Resem1'.



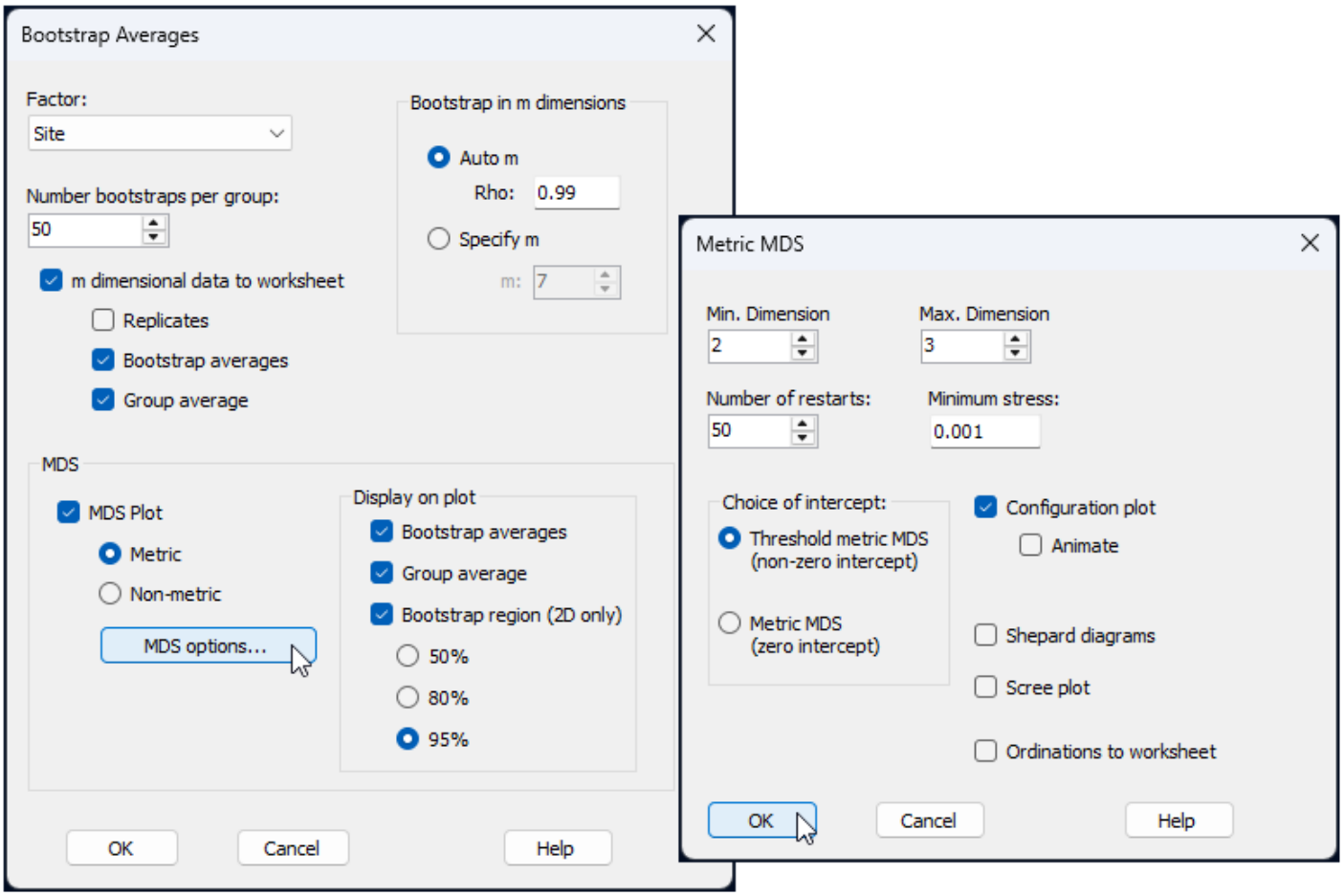
Interest lies in examining variation among sites and locations in this hierarchical design. It would therefore be useful to observe an ordination of resemblances among the *averages* of the 9 sites (based on the identities and relative abundances of mollusc species they contain), and to get a visual sense of the *variability* in those averages across the locations, rather than viewing a (noisy and high-stress) ordination of relationships among all of the replicate quadrats.†

We will therefore examine an ordination of *bootstrap averages* at the scale of Sites for this example. Specifically, within each site, we will bootstrap re-sample (with replacement, in a suitable number‡ of dimensions, m) the n = 9 quadrats a total of n_{boot} = 50 times. We then calculate fifty corresponding bootstrap averages (1 from each bootstrap re-sample) for each site, and plot these all together in a threshold-metric multi-dimensional scaling ordination based on the Bray-Curtis resemblances among them. We will also choose to output the m-dimensional bootstrap averages themselves, just to give us more flexibility for plotting.

3. From the 'Resem1' similarity matrix, click **Analyse** > **Bootstrap Averages...** and choose the following options (leaving the rest as defaults):
   $\hspace{0.5 cm}$> Factor: Site
   $\hspace{0.5 cm}$> Number bootstraps per group: 50
   $\hspace{0.5 cm}$> $\checkmark$ m dimensional data to worksheet
   $\hspace{1.0 cm}$ $\checkmark$ Bootstrap averages
   $\hspace{1.0 cm}$ $\checkmark$ Group average
   $\hspace{0.5 cm}$> $\checkmark$ MDS Plot > $\bullet$ Metric, then click the button

that says 'MDS options...' and pick
$\hspace{1.0 cm}$ Choice of intercept: $\bullet$ Threshold metric MDS (non-zero intercept)

all as shown in the dialog windows below:



The default bootstrap average ordination plot (called 'Graph1' in the Explorer tree after you run the bootstrap averaging) can be tweaked by changing the colours, labels, symbols, etc. (just click **Graph** > **Sample Labels & Symbols...** and click on the 'Key' button). If you want even finer control, you can work from the m-dimensional bootstrap averages themselves, output as 'Data1' here. Specifically, from 'Data1':

- (i) calculate the Euclidean distances among all of these samples (they are the full set of bootstrap averages) by clicking **Analyse** > **Resemblance...** > (Measure > $\bullet$ Euclidean distance) & (Analyse between > $\bullet$ Samples).; and
- (ii) create a threshold metric MDS of this Euclidean distance matrix by clicking **Analyse** > **MDS** > **Metric MDS (mMDS / tmMDS...)**.

Changing the symbols/labels of the output graphic (which would be output in an item labeled ' Graph3' in the Explorer tree, based on the above steps) essentially gives Fig. 7.4, shown below.

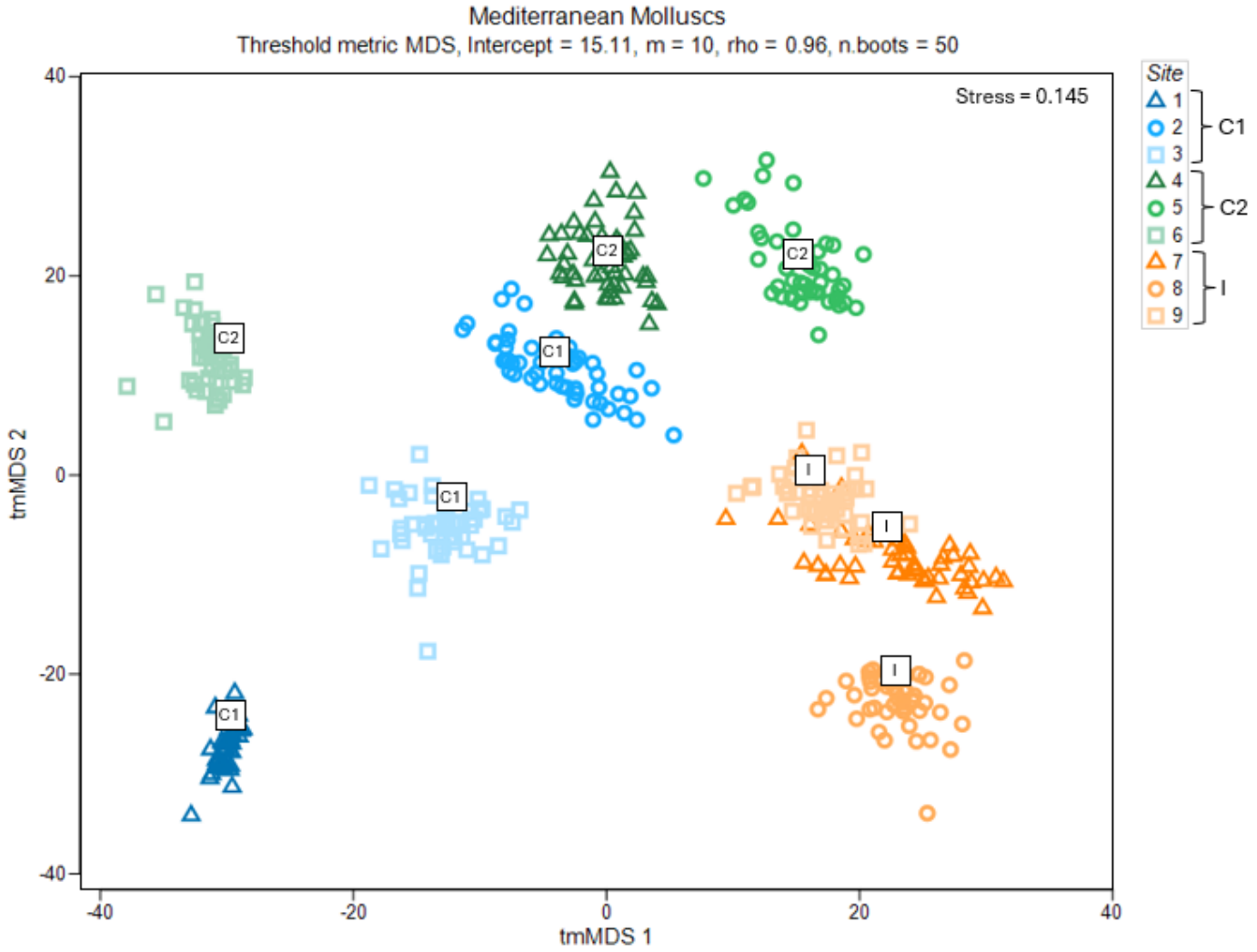


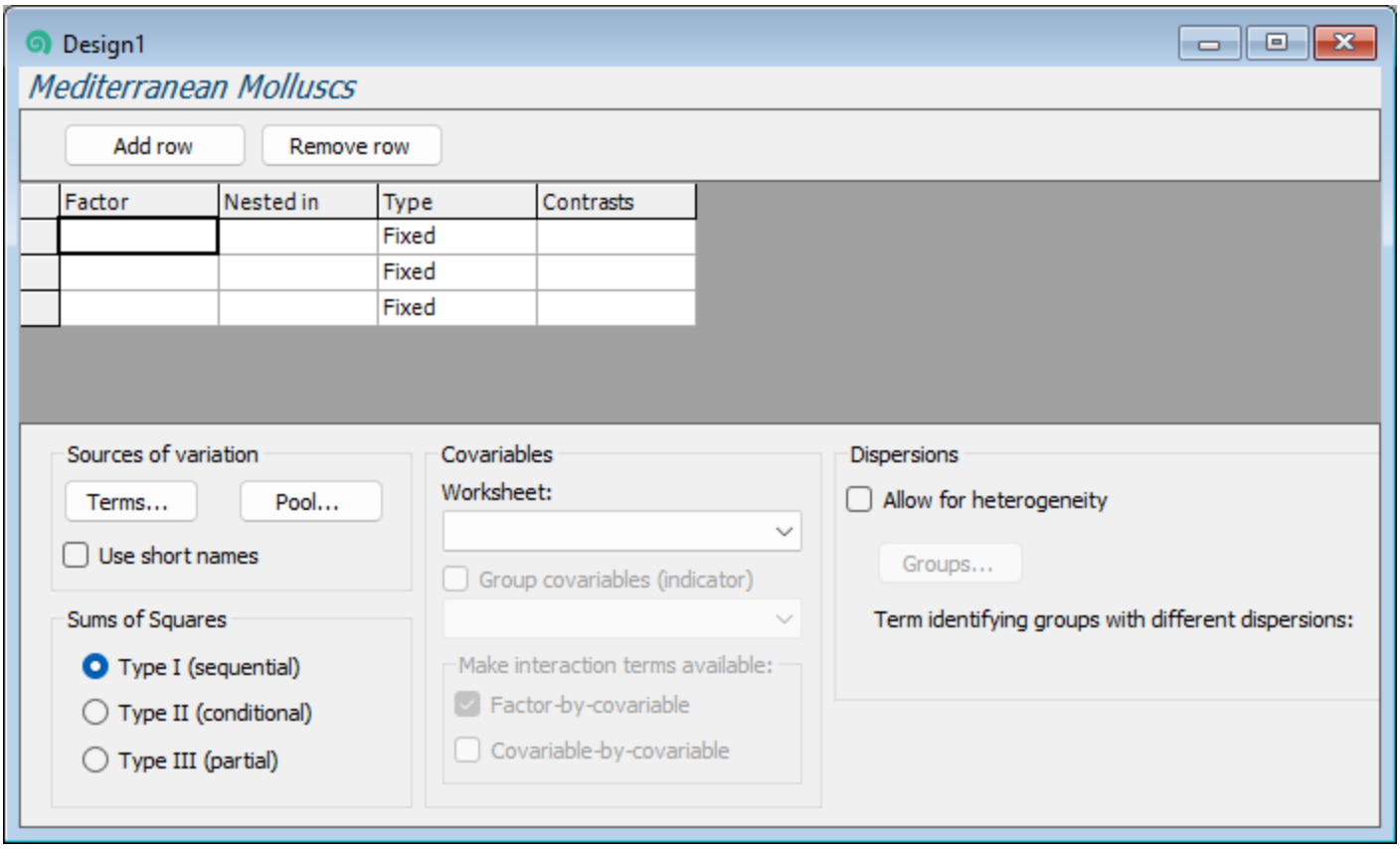
Fig. 7.4 Threshold-metric MDS plot of the averages of three sites from each of three locations in the study design ("C1", "C2" and "I", labeled inside white squares), along with 50 bootstrap averages (coloured symbols specific to each site) for subtidal molluscan assemblages in the Mediterranean, based on Bray-Curtis resemblances.

Average assemblages in sites at the impact location ("I", shown using symbols coloured with orange hues) appear to be quite separate and distinguishable from those occurring at either of the control locations ("C1" in blue hues and "C2" in green hues) (Fig. 7.4). It is not clear, on the face of it, however, whether this difference would be sufficient to be detected as statistically significant, because there is quite a lot of variability among control sites and potentially a difference between the two control locations as well.

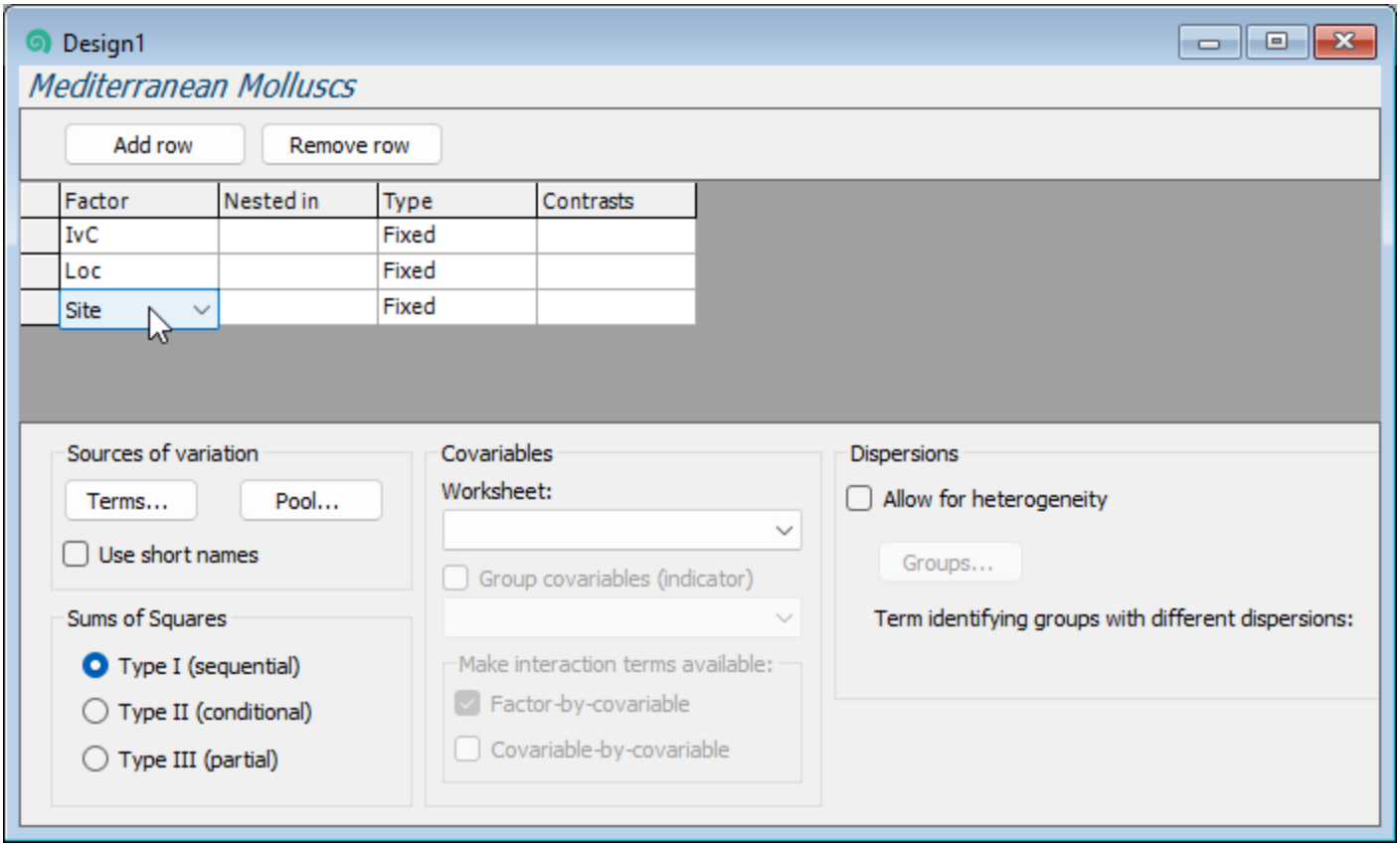
Create the design file

To run a PERMANOVA on these data, we first need to create a design file that matches the study design.

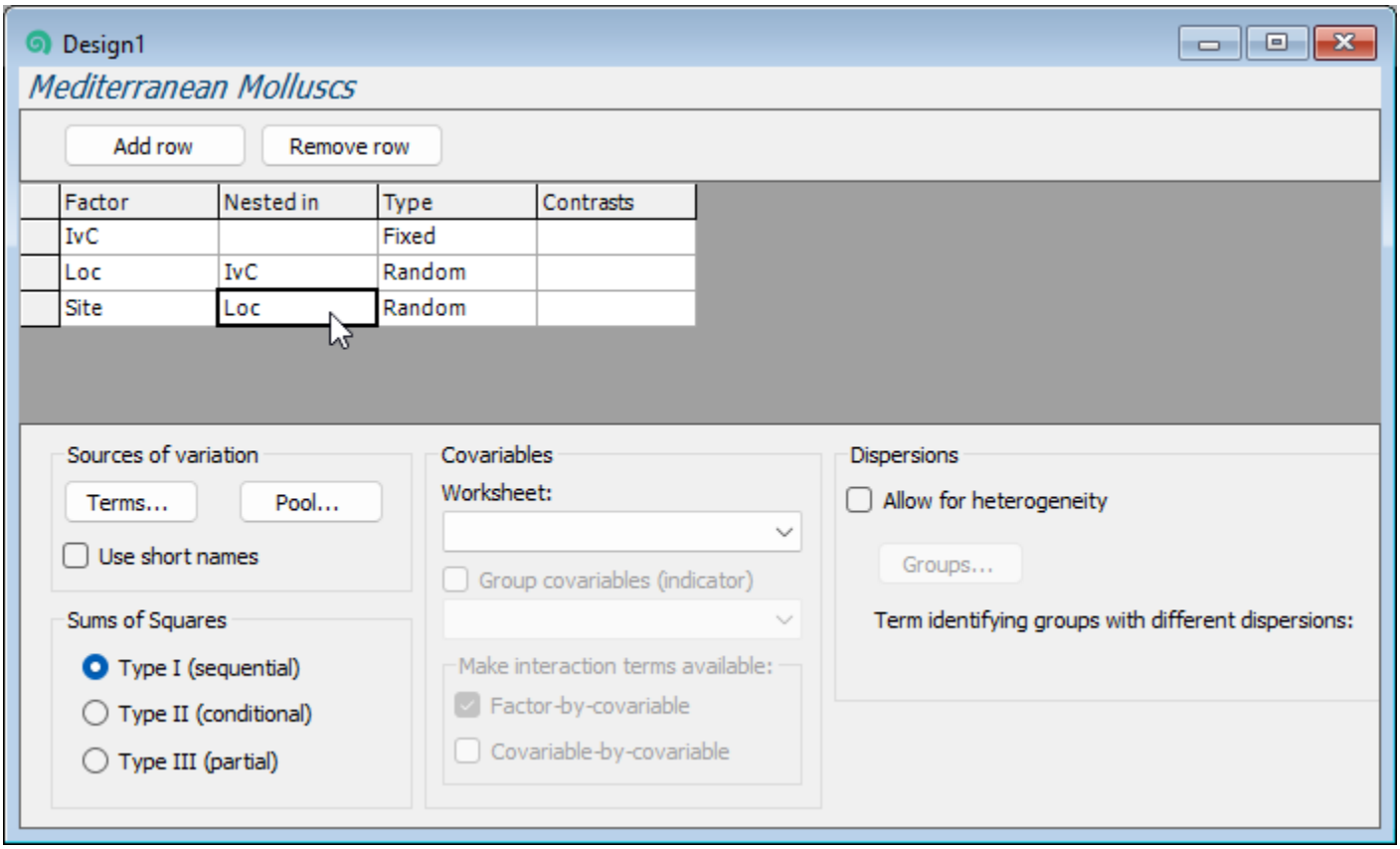
4. **Make as many rows as there are factors** - From the original resemblance matrix ('Resem1'), click **PERMANOVA+** > **Create PERMANOVA Design...**, then click twice on the button to add a row (Add row) so that there are three rows in total.



5. **Choose the name of the factor you want in each row** - In this design file (called 'Design1'), double-click inside each cell of the first column (headed 'Factor') in order to choose the following factors so that they occur sequentially in rows 1, 2, and 3, respectively: row 1 = 'IvC', row 2 = 'Loc' and row 3 = 'Site', like so:

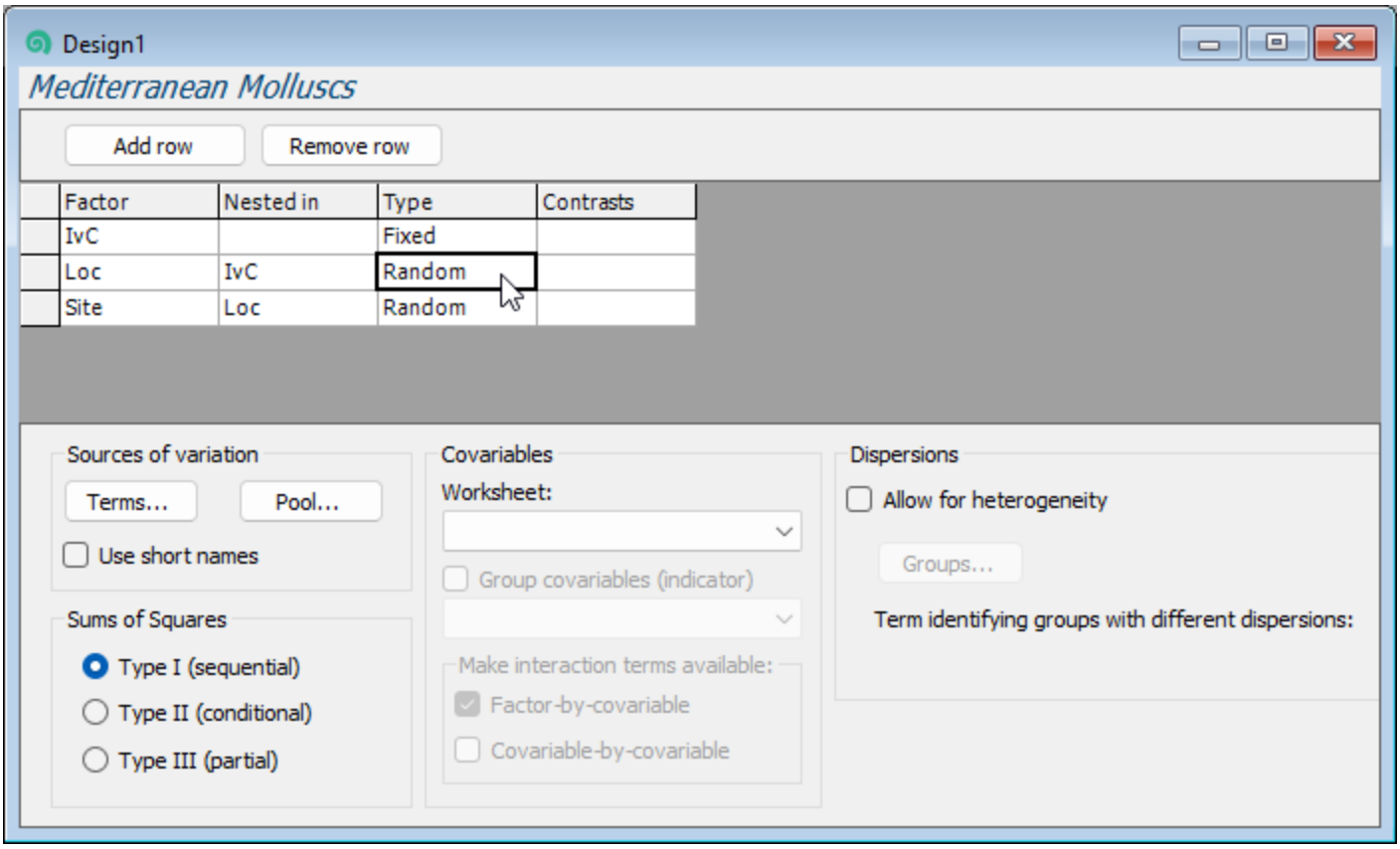


6. **Specify the nested structure** - Next, in column 2, specify the nested relationships among the factors in this hierarchical design. Specifically, the factor 'Loc' in row 2 should be nested in 'IvC' and the factor 'Site' in row 3 should be nested in 'Loc', like so:

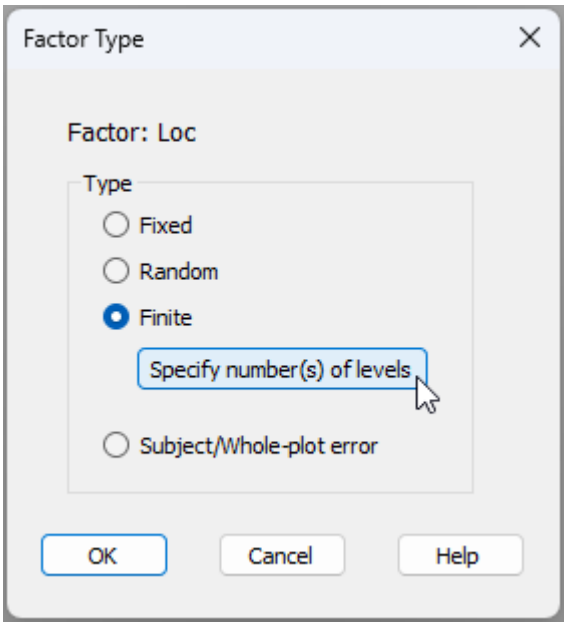


You will notice, after specifying the nested structure of the design, that the 'Type' of factor (column 3) has automatically been changed from 'Fixed' to 'Random' in rows 2 and 3, corresponding to the two nested terms in the model, by default. In our case, we are indeed happy to treat 'Sites' as a random factor, because there is a very large (perhaps uncountable) number of sites that we could have chosen from within each of the locations. We are also (naturally) happy for the factor 'IvC' to remain fixed. There are only two levels of this factor (impact and control), and we are only interested in sampling and comparing these two specific levels (there are no other levels to consider), so the sampling fraction is 1. When it comes to the 'Location' factor, however, we know that it is finite and should be treated as such.

7. **Specify any finite factor(s) and provide relevant population-level details** - To specify that the 'Location' factor is finite and give further details, first click inside the cell in row 2, column 3 of the design file (shown below).



In the 'Factor Type' dialog window, choose Type > $\bullet$ Finite, and click the button to 'Specify number(s) of levels', like so:

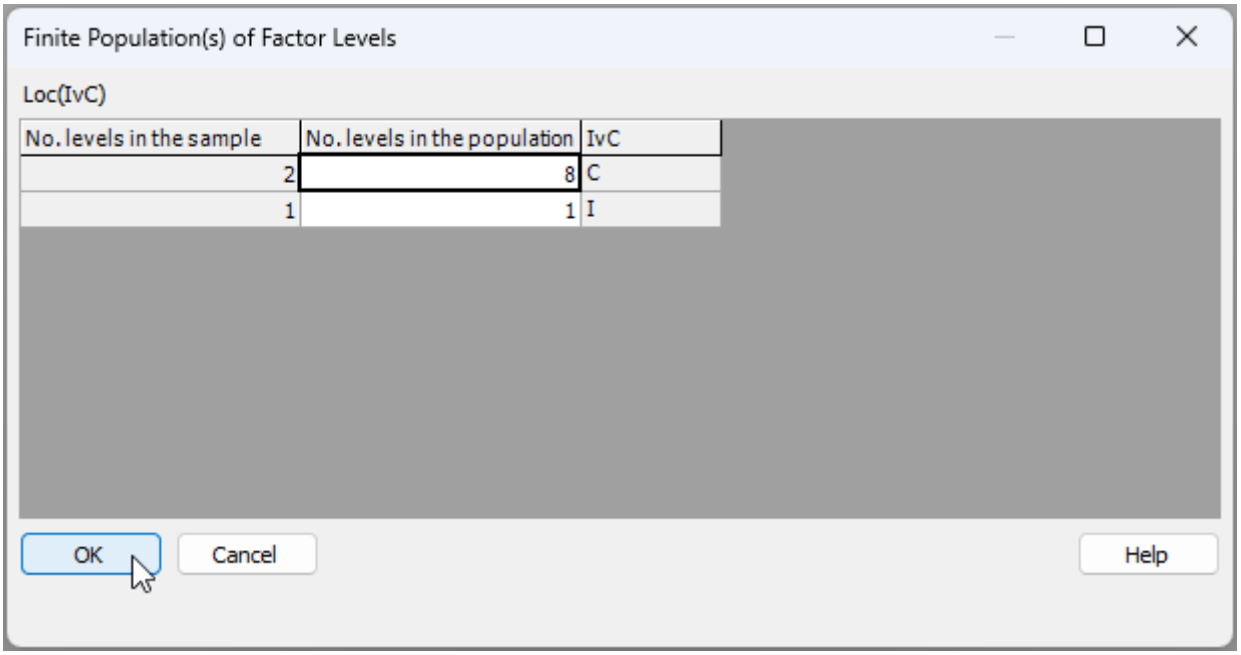


In the resulting dialog window, you will need to articulate *the size of the population of levels from which sampled levels have been drawn*. PRIMER will already assert (in column 1 below) the number of levels of the finite factor have been sampled, given the name of the factor you have provided. But the number of levels in the population (from which sampled levels have been drawn) cannot be gleaned directly from the data and its associated factor information. So, the researcher has to provide it.

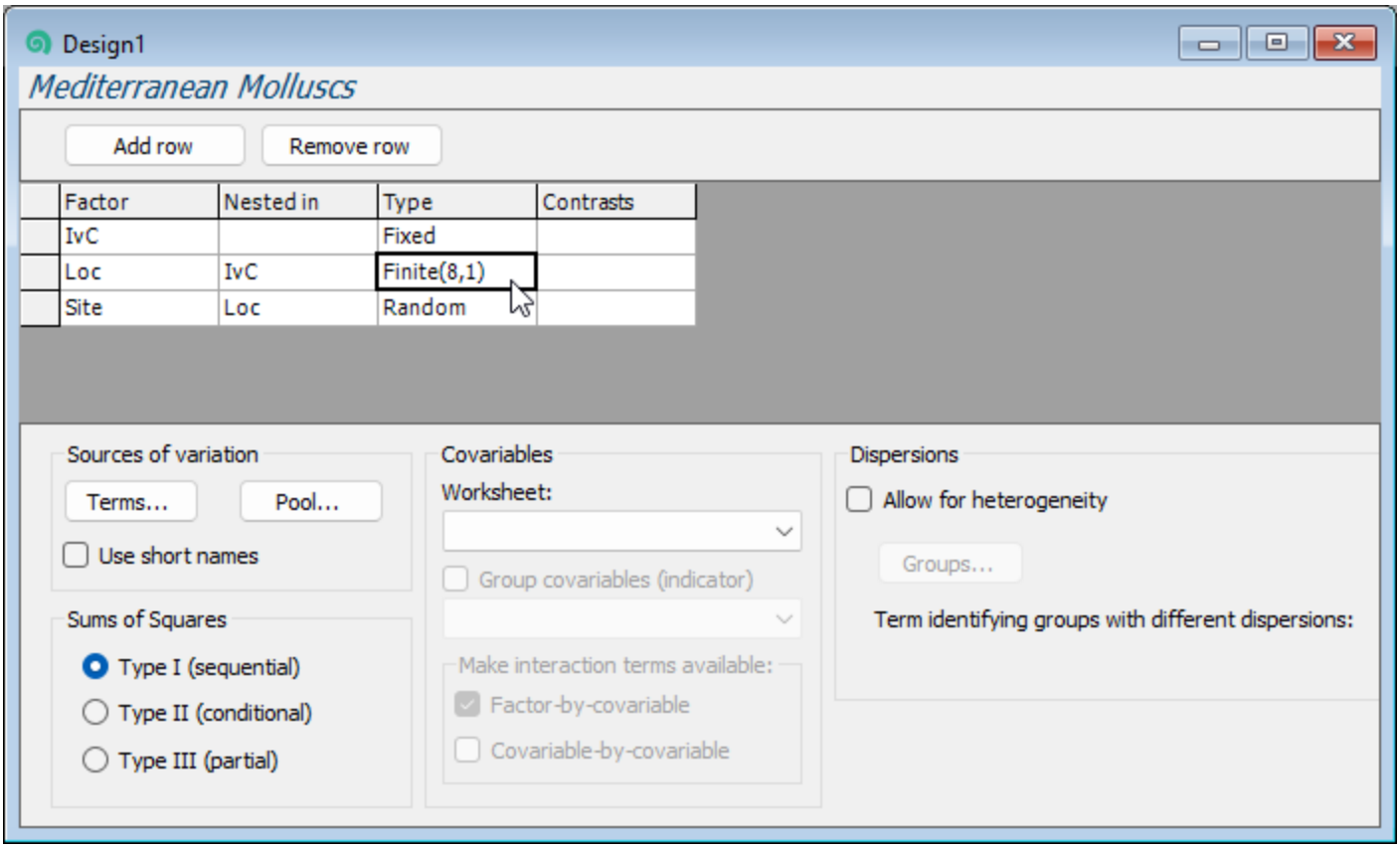
Note also that, in this particular case, the factor of 'Location' is nested in 'IvC'. So we will need to specify the size of the population of levels for each of:

- (i) the control locations (shown in row 1 below); and
- (ii) the impact locations (shown in row 2 below).

We drew b_1 = 2 control locations out of B_1 = 8, so we should enter '8' for the 'No. levels in the population' for the control state ('C'). There is, however, only one impact location in total (B_2 = 1), so we enter '1' for the 'No. levels in the population' for the impact state ('I'), as shown below:



The final completed design file will look like this:

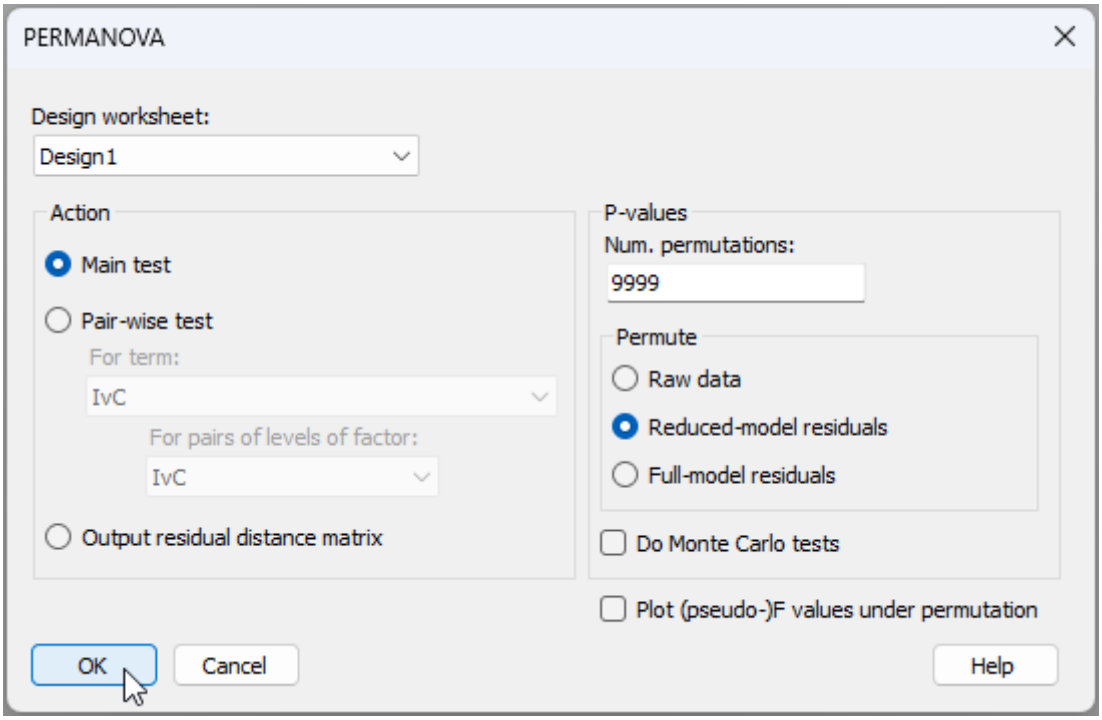


Run the PERMANOVA

When some identifiable fraction of a finite population of possible levels are drawn, the factor can be thought of as somewhere in between fixed and random, and can be analysed explicitly as finite directly within the ANOVA framework. Anderson *et al.* (2025) have provided (and PERMANOVA+ in PRIMER 8 implements directly) the important methodology to derive expectations of mean squares (EMS) for any ANOVA design having any types of factors along the entire graded progression from fixed to random, inclusive. Furthermore, just as for any PERMANOVA model, tests of hypotheses are carefully achieved here under minimal assumptions of exchangeability, using appropriate permutation algorithms for each term in the model. Inclusion of finite factors merely requires, in each case, the explicit specification of the population size from which observed levels are drawn.

Having specified the full such design for the present case-study, we are now ready to embark on the PERMANOVA analysis itself.

8. **Run the analysis** - From the 'Resem1' matrix, click **PERMANOVA+** > **PERMANOVA**. Ensuring that the design worksheet is 'Design1', we can take the defaults for the rest and click '**OK**'.



The resulting output file, 'PERMANOVA1', is shown below.

PERMANOVA1

PERMANOVA
Permutational MANOVA

Resemblance worksheet
Name: Resem1
Data type: Similarity
Selection: All
Resemblance: S17 Bray-Curtis similarity

PERMANOVA design worksheet
Name: Design1
Title: Mediterranean Molluscs

Sums of squares type: Type I (sequential)
Permutation method: Permutation of residuals under a reduced model
Number of permutations: 9999

Factors

Name	Type	Levels
IvC	Fixed	2
Loc	Finite(8,1)	3
Site	Random	9

Excluded terms
No terms excluded

Inestimable terms
No inestimable terms.

PERMANOVA table of results

Source	df	SS	MS	Pseudo-F	P(perm)	Unique perms
IvC	1	30352	30352	3.7936	0.021	839
Loc(IvC)	1	10507	10507	1.4664	0.1972	830
Site(Loc(IvC))	6	42992	7165.3	4.1045	0.0001	9794
Res	72	1.2569E+05	1745.7			
Total	80	2.0954E+05				

Estimates of components of variation

Source	Estimate	Sq.root
S(IvC)	620.85	24.917
V(Loc(IvC))	123.77	11.125
V(Site(Loc(IvC)))	602.17	24.539
V(Res)	1745.7	41.782

Details of the expected mean squares (EMS) for the model

Source	EMS
IvC	1*V(Res) + 9*V(Site(Loc(IvC))) + 6.75*V(Loc(IvC)) + 36*S(IvC)
Loc(IvC)	1*V(Res) + 9*V(Site(Loc(IvC))) + 27*V(Loc(IvC))
Site(Loc(IvC))	1*V(Res) + 9*V(Site(Loc(IvC)))

Construction of Pseudo-F ratio(s) from mean squares

Source	Numerator	Denominator	Num.df	Den.df
IvC	1*IvC	0.75*Site(Loc(IvC)) + 0.25*Loc(IvC)	1	5.46
Loc(IvC)	1*Loc(IvC)	1*Site(Loc(IvC))	1	6
Site(Loc(IvC))	1*Site(Loc(IvC))	1*Res	6	72

These results show a statistically significant impact of the sewage outfall on assemblages of molluscs along this coastline in the Mediterranean (i.e., for the 'IvC' term, we have $F_{1,5.46}$ = 3.79 and P = 0.021). Also, although there is relatively high site-level variation ($F_{6,72}$ = 4.10, P = 0.0001), there is no evidence for significant location-level variation, over and above this, among the control locations ($F_{1,6}$ = 1.47, P = 0.1972). Note that our inferences here

extend to the finite population of 8 locations along the Italian Mediterranean coast from which we have sampled, but not beyond that.

[¶]*There are some rather large abundance values here and, given the relatively large number of replicates per site, these data would be a great candidate for the application of dispersion weighting, as a pre-treatment option. Alternatively, we might think it sensible to apply a fourth-root transformation as a pre-treatment, to downweight the influence of the more abundant species. Here, we have chosen to analyse the data without any transformation or pre-treatment, simply to maintain consistency with the approach taken by Terlizzi et al. (2005) .*

[†]*One could also look at (say) a non-metric MDS ordination among all of the original N = 81 replicates. The lowest stress achieved for the 2-dimensional solution is 0.196, which makes it difficult to interpret finer details with much confidence (due to high variation among replicates). Another alternative is to calculate and then plot distances among the site centroids, without doing any bootstrapping; however, on its own, this won't show the variability in those site centroids. We acknowledge that bootstrap routines to examine variation in centroids in the dissimilarity space (rather than averages of the original variables) would be a desirable addition.*

[‡]*Note that, for this example, the 'appropriate number of dimensions' in which to do the bootstrapping (by default) turns out to be m = 10 metric MDS axes, which achieves a matrix correlation with the original resemblance matrix of ρ = 0.960. This can be seen in the 'Diagnostics' section of the output file called 'Bootstrap Average1', produced by the Bootstrap Average routine when run with these parameters on this dataset.*

7.5 Broader implications for detecting impact

Comparison of results treating 'Locations' as random

Historical wisdom for such a design would have treated 'Locations' as a random factor (Underwood (1992) , Glasby (1997)). It is quite instructive to consider what the results of this analysis might have been had we done this, instead of treating the 'Locations' factor as finite. Below are the two output files:

- treating 'Loc' as a *finite factor* with a sampling fraction of 2/8 (= 1/4) for the controls ('PERMANOVA1')

PERMANOVA1, Loc finite

PERMANOVA table of results

Source	df	SS	MS	Pseudo-F	P(perm)	Unique perms
IvC	1	30352	30352	3.7936	0.021	839
Loc(IvC)	1	10507	10507	1.4664	0.1972	830
Site(Loc(IvC))	6	42992	7165.3	4.1045	0.0001	9794
Res	72	1.2569E+05	1745.7			
Total	80	2.0954E+05				

Estimates of components of variation

Source	Estimate	Sq.root
S(IvC)	620.85	24.917
V(Loc(IvC))	123.77	11.125
V(Site(Loc(IvC)))	602.17	24.539
V(Res)	1745.7	41.782

Details of the expected mean squares (EMS) for the model

Source	EMS
IvC	1*V(Res) + 9*V(Site(Loc(IvC))) + 6.75*V(Loc(IvC)) + 36*S(IvC)
Loc(IvC)	1*V(Res) + 9*V(Site(Loc(IvC))) + 27*V(Loc(IvC))
Site(Loc(IvC))	1*V(Res) + 9*V(Site(Loc(IvC)))

Construction of Pseudo-F ratio(s) from mean squares

Source	Numerator	Denominator	Num. df	Den. df
IvC	1*IvC	0.75*Site(Loc(IvC)) + 0.25*Loc(IvC)	1	5.46
Loc(IvC)	1*Loc(IvC)	1*Site(Loc(IvC))	1	6
Site(Loc(IvC))	1*Site(Loc(IvC))	1*Res	6	72

- treating 'Loc' as a *random factor* ('PERMANOVA2').

PERMANOVA2, Loc random

PERMANOVA table of results

Source	df	SS	MS	Pseudo-F	P(perm)	Unique perms
IvC	1	30352	30352	2.8886	0.3309	3
Loc(IvC)	1	10507	10507	1.4664	0.1997	830
Site(Loc(IvC))	6	42992	7165.3	4.1045	0.0001	9810
Res	72	1.2569E+05	1745.7			
Total	80	2.0954E+05				

Estimates of components of variation

Source	Estimate	Sq.root
S(IvC)	551.23	23.478
V(Loc(IvC))	123.77	11.125
V(Site(Loc(IvC)))	602.17	24.539
V(Res)	1745.7	41.782

Details of the expected mean squares (EMS) for the model

Source	EMS
IvC	1*V(Res) + 9*V(Site(Loc(IvC))) + 27*V(Loc(IvC)) + 36*S(IvC)
Loc(IvC)	1*V(Res) + 9*V(Site(Loc(IvC))) + 27*V(Loc(IvC))
Site(Loc(IvC))	1*V(Res) + 9*V(Site(Loc(IvC)))

Construction of Pseudo-F ratio(s) from mean squares

Source	Numerator	Denominator	Num.df	Den.df
IvC	1*IvC	1*Loc(IvC)	1	1
Loc(IvC)	1*Loc(IvC)	1*Site(Loc(IvC))	1	6
Site(Loc(IvC))	1*Site(Loc(IvC))	1*Res	6	72

There are a number of things that are different between these two outputs (Table 7.2). Look specifically at the EMS for the 'IvC' term and, hence, the construction of the pseudo F ratio from mean squares and associated degrees of freedom for that test. These, in turn, obviously affect the observed value of the pseudo F statistic for the test of 'IvC' and its associated P value as well.

Table 7.2. Key essential differences in the PERMANOVA output when we treat 'Locations' as a finite population with 8 levels versus treating it as a random factor.

Point of difference	**Treat 'Loc' as finite (8 levels)**	**Treat 'Loc' as random**
Number of levels in the population for the factor of 'Location' (among Controls)	$B_1 = 8$	$B_1 = \infty$
Coefficient, K, on the 'Loc' variance component (Loc(IvC)) in the EMS for 'IvC'	$K = 6.75$	$K = 27$
Denominator in the construction of the pseudo F test statistic to test 'IvC'	$0.75 \cdot \text{MS}_ {\text{Sites}} + 0.25 \cdot \text{MS}_ {\text{Loc}}$	$\text{MS}_ {\text{Loc}}$
Denominator degrees of freedom for the test of 'IvC'	$\text{df}_ {\text{denom}} = 5.46$	$\text{df}_ {\text{denom}} = 1$
pseudo F test-statistic for 'IvC'	$F = 3.7936$	$F = 2.8886$
P value for the test of 'IvC'	$P = 0.021$	$P = 0.331$¶

An interesting thing to note is that the denominator that needs to be used for the construction of the pseudo F test-statistic for the test of 'IvC' when we treat 'Loc' as finite has to be constructed as a *linear combination of mean squares*. This is also the essential reason that the denominator

degrees of freedom for the test of 'IvC' in the finite-factor case is a non-integer value (i.e., $df_{denom} = 5.46$).

The implications of all of this are far-reaching, because the test of the 'IvC' term is definitely the most important test for the researcher in this particular study design. The specification of the 'Loc' factor as 'finite' has clearly provided *more power* for this key test of environmental impact in the present case. In general, we can expect that the power will increase whenever there is an increase in the denominator degrees of freedom (all else being equal).

Changes in the size of the inference space

To demonstrate the effect of changing the *size of the inference space* on the analysis, we can posit what the results for this study would look like - specifically for the test of the 'IvC' term in the model - if the number of 'Control' locations in the *population* (i.e., B_1), were larger (Table 7.3). We have already seen what would happen if we consider that this population is infinite (i.e., if we treat 'Locations' as a random factor). The table below looks at a progression of values for B_1, corresponding to a gradation in the sampling fraction ($f = b_1/B_1$) from fixed ($f=1$) to random (where $B_1 = \infty$ so f is effectively zero), showing the concomitant change in the results.

Table 7.3. Effect of a change in the size of the inference space (sampling fraction) on construction of the F statistic and associated denominator degrees of freedom in a PERMANOVA test for the factor 'IvC'.

B_1	**Sampling fraction (f)**	**Denominator for the test of 'IvC'**	**df_{denom}**	**F_{IvC}**
∞ (random)	$\underset{B_1 \to \infty}{\lim} f = 0$	MS_{Loc}	1	2.889
100	$1/50$	$0.67 \cdot \text{MS}_{\text{Sites}} + 0.33 \cdot \text{MS}_{\text{Loc}}$	4.35	3.676
30	$1/15$	$0.69 \cdot \text{MS}_{\text{Sites}} + 0.31 \cdot \text{MS}_{\text{Loc}}$	4.57	3.699
20	$1/10$	$0.70 \cdot \text{MS}_{\text{Sites}} + 0.30 \cdot \text{MS}_{\text{Loc}}$	4.72	3.716
10	$1/5$	$0.73 \cdot \text{MS}_{\text{Sites}} + 0.27 \cdot \text{MS}_{\text{Loc}}$	5.21	3.767

B_1	Sampling fraction (f)	Denominator for the test of 'IvC'	df_{denom}	F_{IvC}
8	$1/4$	$0.75 \cdot \text{MS}_{\text{Sites}} + 0.25 \cdot \text{MS}_{\text{Loc}}$	5.46	3.794
4	$1/2$	$0.83 \cdot \text{MS}_{\text{Sites}} + 0.17 \cdot \text{MS}_{\text{Loc}}$	6.62	3.930
2	1	MS_{Sites}	6	4.235

Note that we are not changing anything about the number of levels *actually sampled*, here: i.e., for all of the lines in Table 7.3 above, we have the same b_1 = 2 sampled control locations.

A word of caution

We are not typically at liberty simply to choose whatever inference space we want. Ethical as well as scientific considerations may well come in to play here. It has to be recognised that the specific pseudo F ratio in every line of Table 7.3 actually corresponds to a test of a different null hypothesis. Each line presents a test with a different breadth of inference: the top line has the broadest scope, while the bottom line has the narrowest. Indeed, treating the control locations as fixed (bottom line in Table 7.3) might well give us more power, but it also severely limits our statistical inferences to just those particular locations in our study and to no others. It is clear that the inferences from the true study design, where we sampled 2 out of 8 control locations, apply to the whole of the Italian Mediterranean coastline that is spanned by those 8 potential locations, no less and no further. Hence, this excellent new tool permitting specification of finite factors, although it affords us greater flexibility and (potentially) power to test the terms of greatest interest, also comes with a special dose of responsibility. We need, as researchers, to articulate carefully the broader population, hence the scale and extent of the inferences that we shall draw from any study.

¶*Note that the p-value here is being limited by the fact that there are only 3 possible permutations of the 3 locations across the 2 groups 'I' and 'C'. We do have the option to obtain a Monte Carlo approximation to the p-value when we run the PERMANOVA, by ticking the box in the PERMANOVA run dialog: ($\checkmark$Do Monte Carlo tests). Assuming that each of the principal coordinate (PCO) axes representing the sample points in the chosen resemblance space (Bray-Curtis in this example) are asymptotically normal, then the PERMANOVA pseudo-F statistic is distributed as a ratio of two linear forms in chi-square under a true null hypothesis, from which we can take a random Monte Carlo draw (the linear forms are supplied by the eigenvalues of the PCO). The Monte Carlo approximate p-value for the test of the 'IvC' term for this example, treating 'Locations' as a random and not a finite factor, is P ~ 0.03.*