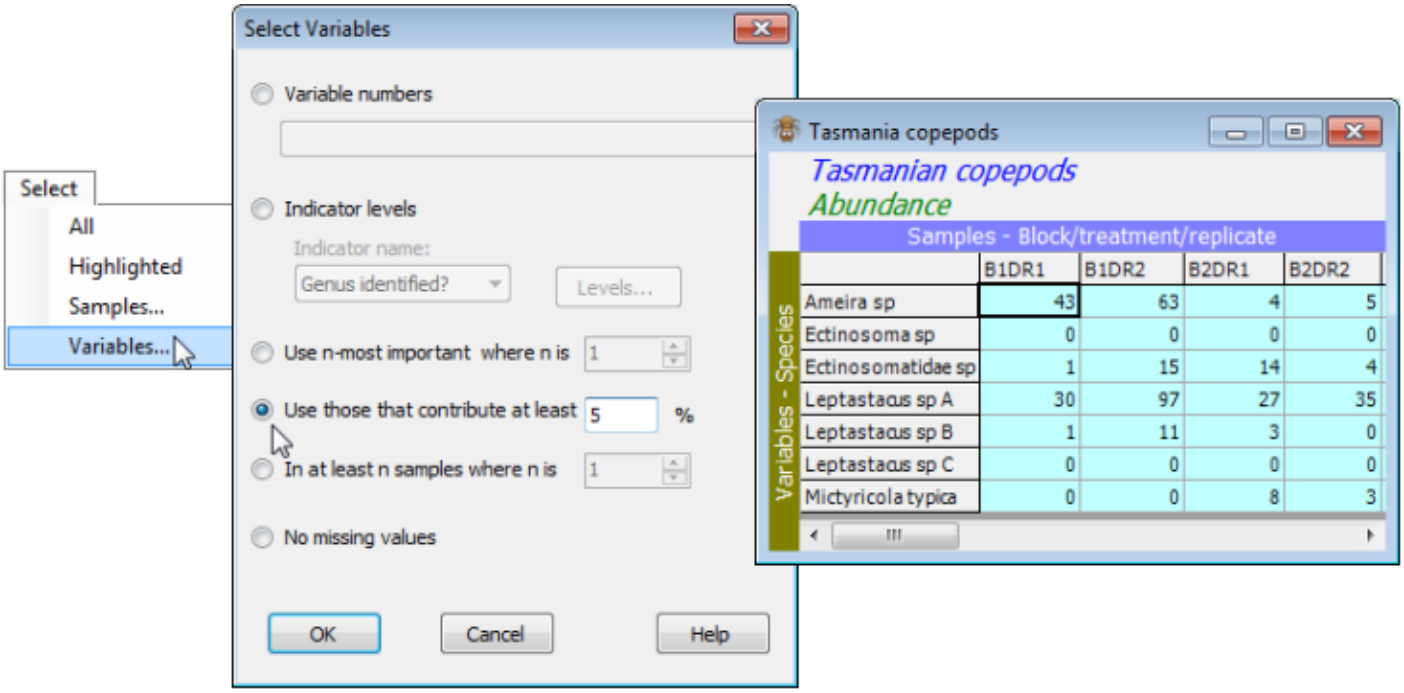
Selecting by ‘most important’

There are, however, three other selection methods under **Select>Variables** that are specific to selecting species (or other taxon-type) variables, in which matrix entries are positive ‘amounts’ of that species (counts, biomass, area cover etc.). The idea of the first two options is to be able to drop species which are not a substantial component of the overall counts (or biomass, area cover etc.) in any sample. The third option, an addition to PRIMER 7, is to drop species which occur in fewer than a specified number of samples, e.g. **Select>Variables>**(•In at least *n* samples where *n* is 2) would drop species which were only seen on one occasion. (It is important to note, however, that removing low abundance or rare species in this way is not required for most of the methods in PRIMER, based on Bray-Curtis similarities for example, and should be done only where there is good reason, e.g. when using a resemblance coefficient which is sensitive to rare species – such as chi-squared distance or Gower, Section 5). The option to **Select>Variables>**(•Use those that contribute at least 5 %) applied to the copepod counts in Tasmania copepods would drop species which, for every sample, account for <5% of its total abundance, leaving only 7 of the original 17 species in the selected sheet. Alternatively, the number of species to retain can be specified, rather than the %, but the principle is the same. Taking **Select>Variables>**(•Use *n*-most important where *n* is 7) generates the same set of species, naturally. If *n* is larger, say 10, then to be retained, the threshold percentage that a species must contribute somewhere will drop – in fact a threshold of around 3% will leave 10 species. If *n* is smaller, say 5, then a higher percentage cut-off is needed (10% in fact). The algorithm simply varies the cut-off percentage until the matrix retains only the exact number of species *n* requested. This means of selecting ‘important’ species (rather than by taking their total abundance across all samples and selecting the top *n*-ranked of those) is preferable because it retains species which are important in impoverished sites, with low total abundance.



The point is re-iterated that **Select>Variables** will operate in combination with **Select>Samples** (unlike repeated **Select>Samples** or **Select>Variables** operations on their own), to ensure the behaviour that would be expected. That is, if a sample selection is in operation then the ‘most important’ 10 species – or the species which occur in at least 2 samples – are determined only with regard to that selection, not using all the samples.

Close the Tasmania ws – there is no need to resave it, since when met in a later section it will not be for a subset of either the samples or species.

Revision #6
Created 21 May 2024 21:55:52 by Arden
Updated 14 January 2025 22:42:09 by Abby Miller