Biol 206/306 – Advanced Biostatistics
Lab 8 – Mantel and Randomization Tests
Fall 2016

By Philip J. Bergmann

0. Laboratory Objectives
1. Learn what a randomization test is and how it relates to a permutation test
2. Learn how to calculate distance matrices in R
3. Learn how the Mantel test works and how to do it in R
4. Learn how the partial Mantel test works and how to do it in R
5. Learn when to use a randomization test
6. Do a randomization two-factor ANOVA in R

1. Randomization and Permutation Tests
Randomization and permutation tests are considered non-parametric tests because they avoid using a standard test statistic distribution, such as an F or t distribution. Instead, they rely on the generation of an empirical null distribution derived from resampling the original data to calculate a p-value. These tests consist of three basic steps:
1. Use a dataset to calculate a test statistic
2. Randomize the dataset s times, calculating the same test statistic from each randomized dataset – these s test statistics form the null distribution
3. Calculate the p-value by compaing the actual test statistic to the null distribution

Randomization and permutation tests work exactly the same way, but a permutation test samples all possible permutations of the data, while a randomization test samples only a subset of the possible permutations. A permutation test allows for the estimation of the most precise p-value, but is impractical when the dataset is large and there are a very large number of possible permutations. The more permutations that one samples, the more precisely the p-value can be estimated. In a randomization/permutation test, the p-value is calculated as the number of null test statistics that are more extreme than the empirical test statistic, divided by the number of permutations, s. If you are conducting a two-tailed test, then this p-value should be multiplied by two. You can determine the precision with which the p-value is estimated by pretending that there is only one null test statistic greater than the empirical test-statistic, and considering how p changes as you manipulate s. ***Do this for the following number of permutations. Pretend you are doing a one-tailed test.***

| s	10	20	100	1,000	10,000
p_{min}					

2. The Mantel Test
The Mantel test is a special type of randomization test that is very useful because it allows one to test for an association between two matrices. Typically, it is used to compare matrices that contain pairwise distance data, similarity data, or dissimilarity (similar to distance) data. In all of these cases, the data matrix is square and symmetrical, and the diagonal does not contain useful

information. This is because when an observation is compared to itself, it is identical to itself, and therefore, has a distance/dissimilarity of zero or a similarity of one.

Testing for association between two matrices presents two problems: the data in the cells are not independent, and you are comparing two bivariate objects. The randomization aspect of the approach deals with the non-independence of distances. To deal with how to compare two bivariate objects, the Mantel test essentially straightens out the lower triangle of each matrix into a vector and then calculates the correlation coefficient between the two vectors. This is the empirical test statistic. One of the matrices is then permuted s times and a correlation is calculated for each random replicate to generate the null distribution.

We will be analyzing a lizard locomotion dataset using the Mantel test. The data are actually two datasets – one is a set of measurements of body and limb shape for sixteen species of lizard, and the other is a set of locomotor variables for the same species. The morphological data include measurements of snout-vent length, body width and height, head length and width, and front and hind limb length. The locomotor variables include maximal and average velocity, stride and step duration and length, the angle the body is bent during running, and angles of protraction and retraction of the hind limb. One PCA was run on the morphological data, and a second was run on the locomotor data to reduce the datasets. The data supplied to you include the species mean values for each of the first four PCs of each PCA. You are interested in whether species differences in morphology are associated with species differences in locomotor variables. One might expect this to be the case because one would expect differently proportioned species to move differently. Below is a figure plotting morphological and kinematic PC-1s and PC-2s (from *Bergmann and Irschick. 2010. Alternate pathways of body shape evolution translate into common patterns of locomotor evolution in two clades of lizards. Evolution 64: 1569-1582.*). Note that the points in the figure represent individuals, not species. Open squares are skink lizards, all other symbols are Phyrnosomatine lizards. On the following page, there is a phylogeny showing the relationships of the species, with branch lengths. Skinks belong to the genera *Lerista* and *Ctenotus*, and Phrynosomatines are everything else.
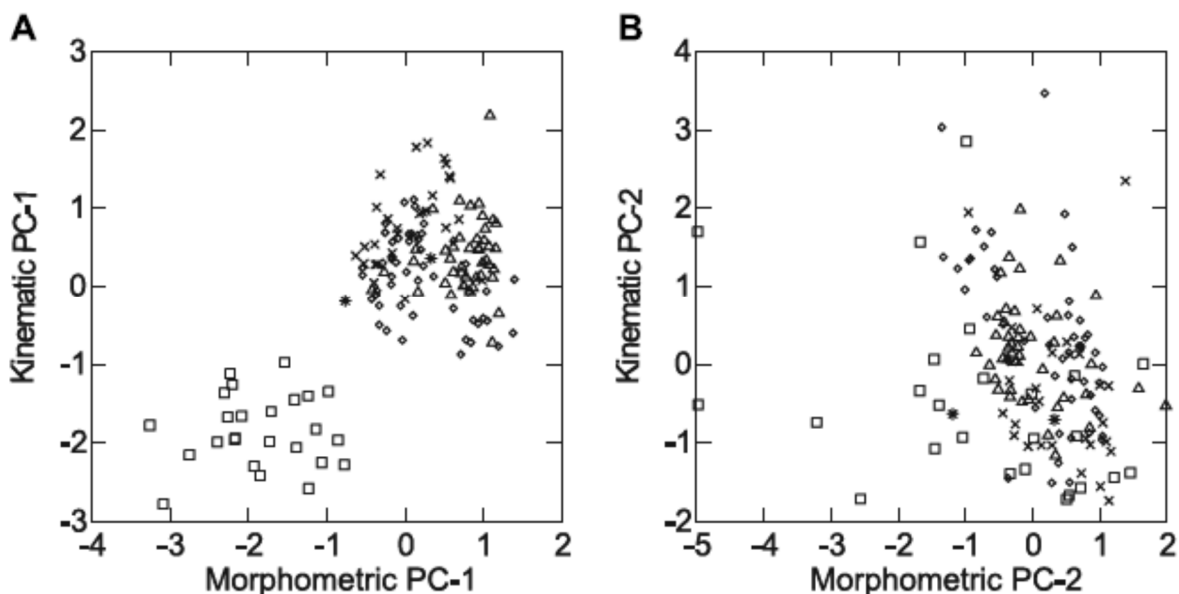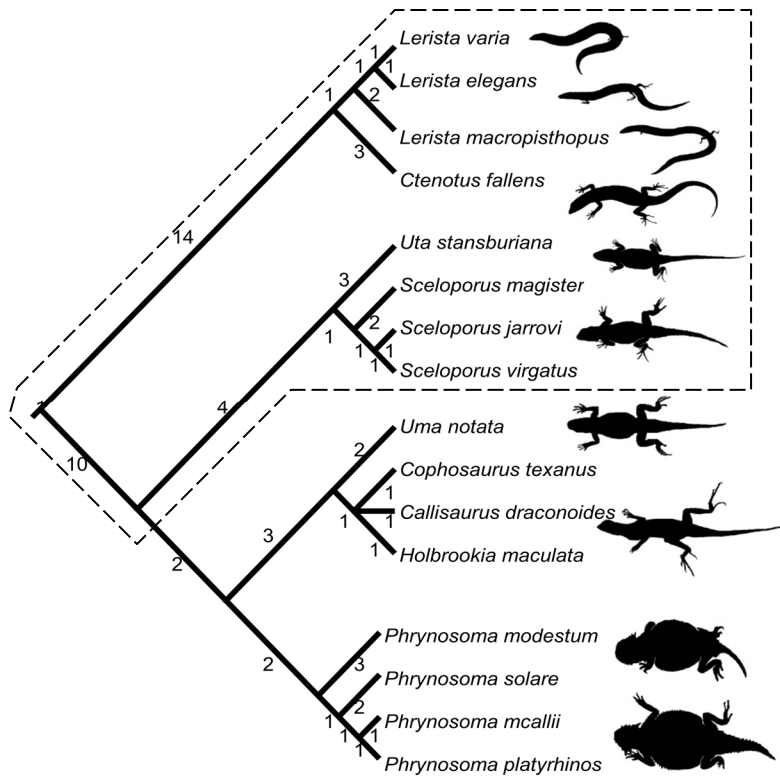
Ignore the phylogeny for now, but note that the skinks tend to have very elongate bodies, while the Phrynosomatines tend to be much stockier.

Download the dataset, open it in Excel, examine it, and save it as a text file. Then import it into R, saving it to an object named "PC_data". Note that in the current format, the data consist of four-dimensional coordinates (four PCs from each PCA) for species in morphospace and locomotor space. To analyze them using a Mantel test, they need to be converted to distances. You can do this using the following function:

```
> dist(x, method=c("euclidean", "maximum", "manhattan",
"binary"), diag=FALSE, upper=FALSE)
```

In this function, **x** is a data frame of *p* variables for which you want to calculate *p*-dimensional distances, **method** specifies what type of distance you want to calculate, **diag** specifies whether or not to include the diagonal in the matrix, and **upper** specifies whether or not to include the upper triangle of the distance matrix. The default is to include only the lower triangle, which is what is desired. ***Use the dist function to calculate pairwise Euclidean distances among species based on the four morphological PCs and then on the four locomotor PCs. Do this separately, assigning the resulting morphological distance matrix to an object "morph_dist", and the locomotor distance matrix to an object "loco_dist". Examine the matrices. How big are they (how many rows and columns)?***

Now you are ready to do a Mantel test for association of the two matrices. ***You will need the package "vegan" to do the Mantel test. Install it if you do not yet have it installed, and then load it. Use the "mantel" function to do your test of matrix association:***
```
> mantel(xdist, ydist, method=c("pearson", "spearman",
"kendall"), permutations=999, strata)
```

Here, **xdist** and **ydist** are dis/similarity/distance matrices to be tested for association, **method** specifies the type of correlation to use, **permutations** is the number of randomizations of the dataset, and **strata** is optional and can include a factor that specifies groups within which

randomization is done. Pearson is the typically used method, and it is the first matrix, **xdist**, that is randomized.

What is the null hypothesis of the Mantel test you have conducted?

Do the mantel test on your two distance matrices with 10, 100, 1000, 10000, and 100000 permutations. Provide the results in the table below. How do the r and p values change with the number of permutations?

Permutations	r	p
10		
100		
1000		
10000		
100000		

How many permutations would you use if you could do the test only once? Explain.

What are your biological conclusions from the Mantel test that you did?

3. The Partial Mantel Test
The partial Mantel test is a useful extension of the regular Mantel test that allows you to account for a confounding dis/similarity/distance matrix. It tests for an association between two such matrices while taking the effects of a third matrix into account. The third matrix is a nuisance variable, but note that it must also be in the form of a matrix of the same size as the other two matrices. Like the regular Mantel test, the general approach can be explained as straightening all three matrices, **X**, **Y**, and **Z**, into vectors, doing a regression between vectored matrices **Y** and **Z**, and then doing a Mantel test on matrix **X** and the matrix of residuals from the **YZ** regression.

In the lizard body shape and locomotion data, an important nuisance variable is the evolutionary relationships among the species. The species have known evolutionary relationships, shown using the phylogeny above, and it is important to take these into account. The species can also be divided into two major clades that are only distantly related: the Scincidae (including *Lerista* and *Ctenotus*), and the Phrynosomatinae (including everything else). You already have your morphology and locomotion distance matrices. What you need now is a phylogenetic distance matrix. We will do this in two different ways to help show you different approaches to making

such matrices. The first approach is to simply consider the two main clades (Scincidae and Phrynosomatinae). If two species belong to the same clade, they get a zero, if they belong to different clades, they get a one. ***Why is this not the best approach to incorporating the evolutionary relationships among species into the analysis?***

Briefly describe one other scenario in which making a matrix of zeros and ones for use in a partial Mantel test would be useful (i.e., phylogenetic relationships shouldn't be part of your scenario).

The second approach, which more fully uses the phylogenetic information available to you, is to sum the branch lengths connecting each pair of species to one another. The branch lengths are listed on the phylogeny and are measures of dissimilarity. For example, the phylogenetic distance between *Sceloporus virgatus* and *Phrynosoma modestum* on the phylogeny above is: $1+1+1+4+2+2+3=14$.

Assignment: (2 points)
In the matrix below, and only for the species in the dashed box on the phylogeny, put the same/different clade (0/1) distances in the upper triangle, and the actual phylogenetic distances in the lower triangle.

	L.varia	L.eleg.	L.macro.	Ctenotus	Uta	S.virg.	S.jarr.	S.mag.
L.varia								
L.eleg.								
L.macro.								
Ctenotus								
Uta								
S.virg.								
S.jarr.								
S.mag.								

Note that when producing a matrix like this the rownames and column names have to match those in your other matrices exactly. Download the Excel spreadsheet that contains both the same/different clade and phylogenetic distance matrix, and save each as a tab-delimited text file. Import them in R as objects "sd_dist" and "phylo_dist". View the matrices to ensure that they are imported correctly. Notice that both triangles are filled in and the diagonal is composed of zeros. Producing such a matrix is made easier if you notice that (1) the

phylogenetic distance is always twice the distance from one of the species to the common ancestor, and (2) the phylogenetic distance from any species in one clade to any species in a second clade will be the same. Finally, the reason that the hand-made distance matrices do not precisely match those produced by the **dist** function is because those produced by **dist** are objects of class dist, while the ones you are making in Excel are being imported as objects of class data.frame. A data frame class object cannot handle a lower triangle matrix – leading to an error stating that the first line does not have the right number of entries.

The following function can be used to do a partial Mantel test:

```
> mantel.partial(xdist, ydist, zdist, method="pearson", permutations=999, strata)
```

The function is almost the same as the **mantel** function, but has an added argument, zdist. Note that residuals are calculated from a regression run on ydist and zdist, so it makes the most sense to use zdist as the nuisance matrix, and ydist as the matrix most likely to be influenced by it. In our example, morphology tends to have a stronger phylogenetic influence than locomotion, so xdist should refer to the **loco_dist** object.

Assignment: (3 points)
Conduct a partial Mantel test on the morphology and locomotion distances with sd_dist as the nuisance matrix. Then repeat with phylo_dist as the nuisance matrix. Do both analyses with 10,000 permutations of the data. Report the statistical results for both tests below. Do the conclusions differ? What is your biological interpretation of these analysis results?

Thinking more generally, what could doing both tests (with phylogenetic distance represented in the two different ways) tell you about how phylogeny influences an association between body shape and locomotion?

4. Randomization Two-Factor ANOVA

Although we started off learning the Mantel test, randomization is a general approach that can be applied to virtually any test available. For example, one could do a randomization two-sample t-test, ANOVA, Shapiro-Wilks test, or any other test that you might want to use. The strength of the randomization approach is that it relaxes test assumptions about distribution. This is because you no longer base your conclusions on a standard test distribution, but instead create a null distribution that is unique to your dataset.

Consider this with a hypothetical example. You need to do a single-factor ANOVA, but you find that both your response variable and the residuals are not normally distributed. Your Shapiro-Wilks test for normality returns a highly significant p-value (say, $p=0.0000001$), so the assumption of normality is strongly violated. What are your options in this case? You could ignore your test of normality and do the ANOVA anyway, but any conclusion that you come to with this approach would be highly suspect and it is strictly incorrect. You could do a Kruskal-Wallis test instead – the non-parametric version of the single-factor ANOVA. This is a perfectly fine alternative, but has the disadvantage of not being very powerful (your response variable gets transformed into a rank variable, so you lose a lot of information). You could also do a randomization ANOVA. This combines the power of a regular ANOVA with relaxed distributional assumptions, like a Kruskal-Wallis test. In many ways, a randomization test is a great way to go because it always is a custom test for your particular dataset. The costs are that it needs to be implemented (many software packages do not do it), and doing a lot of randomizations and repeating the analysis for each one takes computational time (as you saw with the Mantel test done with 100,000 permutations). With fast computers, the computational time is not much of an issue (does it matter that it takes 10s instead of 0.01s to do the analysis?). This leaves implementation. Dr. Bergmann has implemented a randomization fully crossed two-factor ANOVA as an R function, which is what we will be using today, but keep in mind that any test that we have done this semester could be done using a randomization approach. For this reason, we will also take a look at the code for Dr. Bergmann's function.

Download the text file "randomized_twofactor_anova.txt". Then load it into R using the "File>Source R code…" menu item. The function name is "twofac_anova", and it should appear in your list of objects in the R workspace. Type in the function name in R (or open the text file) to view the code.

Note that the code is documented: any line that starts with "#" is a documentation line that describes what the code beneath it does. It is wise to document code that you write so that anyone using it knows how it works. This can be useful to you as well, if you revisit some function that you wrote a year or two ago.

The documentation for some lines has been replaced with "AA", "BB", and "CC". What do the four lines under "# AA" do?

What do the five lines under "# BB" do?

What do the four lines under "# CC" do?

We will do a randomization two-factor ANOVA on the mosquito dataset that we analyzed in lab 3 (the ANOVA lab). Take the following steps:

1. *Import the lab 3 mosquito dataset into R.*
2. *Do a two-factor ANOVA on hatch rate, with larval density and larval species as two fixed factors. Be sure to include the interaction term. Use the "aov" function and assign the output of aov (not summary(aov()) to an object.*
3. *Use a Shapiro-Wilks test on the response variable to test for normality.*
4. *Do another Shapiro-Wilks test to test for normality of the residuals.*

What are the results of the Shapiro-Wilks tests? Complete the table below and give a verbal conclusion.

Variable	W	p
Hatch Rate		
Residuals		

<u>Assignment: (5 points)</u>
Continue the analysis by taking the summary of your aov object to get an ANOVA table. Then use the "twofac_anova" function to do the randomization anova. Do 10,000 randomizations. What did you type to use the "twofac_anova" function?

Explain why the F-statistics for the regular two-factor ANOVA and the randomization ANOVA are identical (the df, SS, and MS are also identical, so are not shown in the randomization output)?

Complete the table below, providing results for both the regular ANOVA and the randomization ANOVA.

Effect	df	MS	F	p_{ANOVA}	$p_{rand(ANOVA)}$
Density					
Larva					
Density:Larva					
Residual					

Do your conclusions change from doing the randomization ANOVA (compared to the regular ANOVA)?

What are your biological conclusions from this analysis?