Exercise 3: Packages and Phylogenetic Methods

The power of R comes from the approximately 2,700 R packages that are freely available. In your free time check out these resources to find R packages that will be helpful in your own research:

CRAN Task View: http://cran.r-project.org/web/views/
Rseek: http://www.rseek.org
Rgraph gallery: http://gallery.r-enthusiasts.com/thumbs.php

Part A Installing and Loading Packages

1) Packages can be installed in RStudio by clicking on the packages tab
   a) Click on Install packages
   b) Type the name of the package in the new window
2) Once a package has been installed you will still need to load it if you want to use it in an analysis. To load a package simply add a line to the top of your script like this:
   ```
   require(package_name)
   ```
3) For today's exercise install and load the package "ape"

Part B Intro to phylogenetic methods in R: This exercise will introduce you to a lot of different things that you can do with phylogenies and species level data. However, the phylogenetic methods available in R are vast. If you have a specific idea or question come talk to me and I can help you find what you are looking for. Your goal today is to simulate a tree and discrete data and then test two models using a likelihood ratio test and determine if there is sufficient signal to recover the correct answer.

1. When you are doing simulations it is always nice to set a seed this insures that your results can be replicated
   ```
   set.seed(21)
   ```
2. Now lets simply create a phylogenetic tree. There are many functions that do this we will use a simple one from the package APE called rcoal. Lets also plot it and add a scale bar.
   ```
   tree <- rcoal(100)
   plot(tree, show.tip.label=F)
   add.scale.bar()
   ```
3. You should notice that this is a very shallow tree. Much more shallow than the trees that most of you will be dealing with in your research. R stores trees in such a way that we can fix this quite easily. A pylogenetic tree is an object of type list. One of the elements of the list is the branch lengths. They are stored as a vector in the element "edge.length". So lets just multiply that vector by a scalar (100) to make your tree have a depth that is a bit more common:

   First check out the structure of a tree
   ```
   str(tree)
   ```
   Now lets stretch our tree out
   ```
   tree$edge.length <- tree$edge.length*100
   ```
   Now lets replot and make sure it worked
   ```
   plot(tree, show.tip.label=F)
   add.scale.bar()
   ```

4. Next let's go ahead and simulate a discrete trait. To do this I chose the function rTraitDisc. We will specify a tree, a transition matrix, and a root state and the funct ion will stochastically evolve character data across the tree. The output of the function will be a vector representing the character state of the extant species.

```
tips<-rTraitDisc(tree,model=matrix(c(0,.04,.1,0),2),root.value=1)
```

This is a probability matrix. It describes the chance of transitioning from one state to another in one unit of time.

	A	B
A	0	0.8
B	0.4	0

This matrix described our evolutionary model in this case an all rates different (ARD) model if the two values are the same then it is an all rates equal model (ARE)

5. Lets visualize our tip data. First create a vector of colors and then plot symbols at the tip of tree to represent the character states:

```
co <- c("blue","yellow")
tiplabels(pch = 22, cex = .5, bg = co[as.numeric(tips)])
```

6. Now for the fun part: we are going to generate a maximum likelihood estimate of the probability matrix based on the tip data and the tree. The function that does this will also calculate the likelihood of our model. We will do this once with the rate of change into and out of state 1 and 2 equal and another time with a model that allows for different rates. The function that we will use for this is called "ace" and it is also from APE.

```
ans.are <- ace(tips, tree,model=matrix(c(0,1,1,0),2),type = "d")
ans.ard <- ace(tips,tree,model=matrix(c(0,1,2,0),2),type = "d")
```

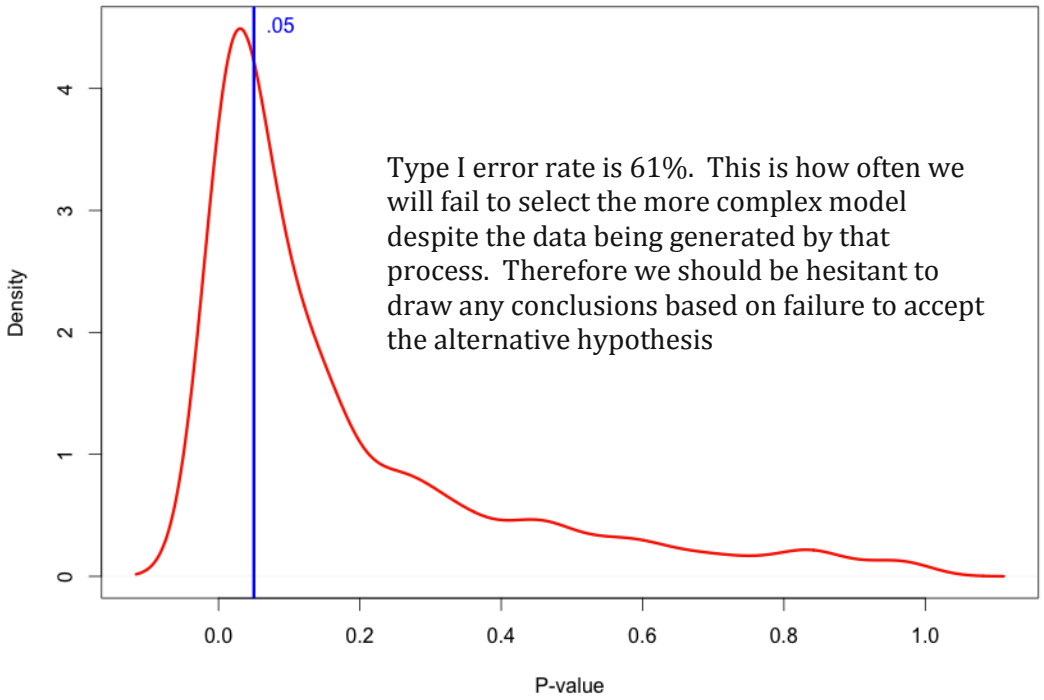
7. Now you can compare the likelihood of the two models using a likelihood ratio test. This will tell us if one of the models fits the data significantly better. First calculate the likelihood ratio then compare it to the chi-square distribution with 1 degree of freedom to get a p-value.

```
LR <- -2*ans.are$loglik + 2*ans.ard$loglik
pval <- 1-pchisq(LR,1)
```

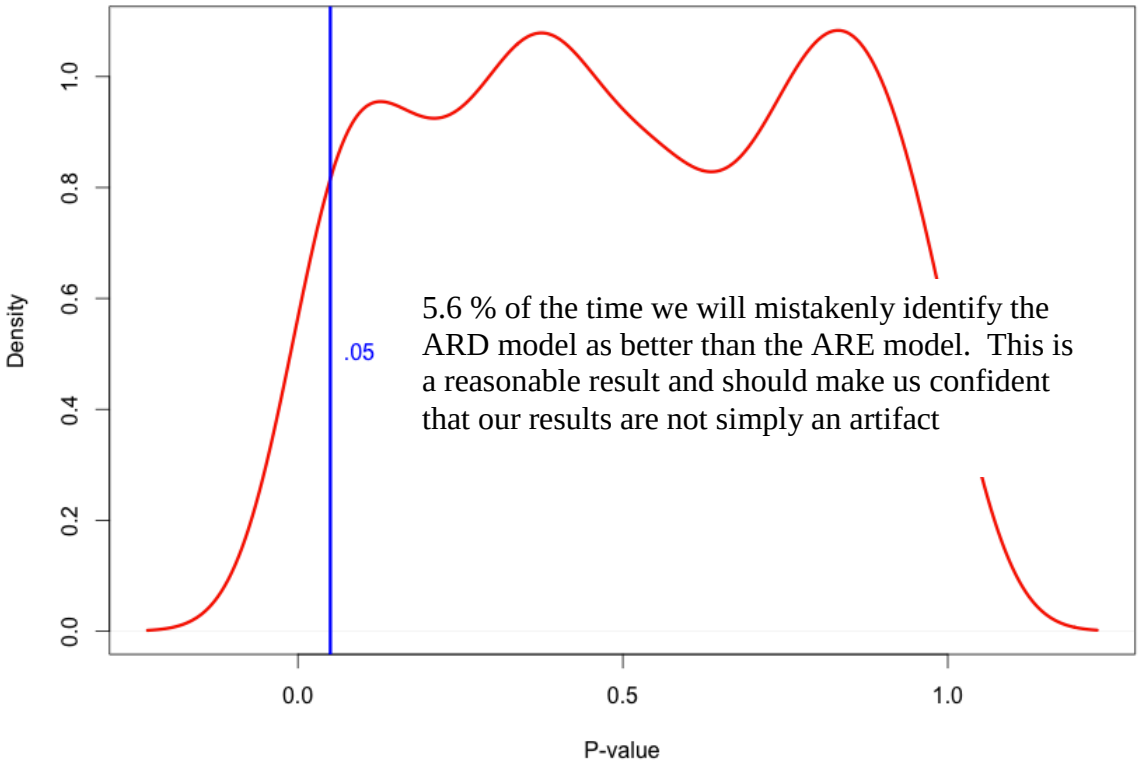
8. You should get a p-value of .041. This means that if the null hypothesis (ARE model) is true we would get a test statistic at least this large 4.1% of the time. Most people would say that this result is sufficient to reject the null hypothesis in favor of the alternative hypothesis (ARD model).
9. If I was going to publish results like this I would want to see how often we can recover this result with the types of trees and rates present in our data. Let's go ahead and calculate both the type I and II error rates. Do this by taking out the seed and enclosing all of your code in a loop.

```
p.val.results <- vector()       #place this line before your code
for(i in 1:100){                #this starts the loop

##ALL YOUR EXISTING CODE

p.val.results[i]<-pval          #Record the result of the current loop
}                               #ends the loop
length(which(p.val.results<.05))/100  #How often we can detect
                                      #that the ARD model is better
```

To estimate the type one error rate I conducted the above loop for 1000 generations:



Type II error rate: to do this I want to simulate the data under an ARE model and see how often we pick the ARD model. Here is the graph that I generated when I did this:
Type II error rate 5.6%



The issue of Type one and two error rates in this type of test (and other model choice tests) is an area of ongoing discussion and was recently a topic of a thread on the rsigphylo discussion list. Lots of the people that have developed these functions chimed in and it was quite interesting here is a link to it: https://stat.ethz.ch/pipermail/r-sig-phylo/2012-August/002202.html

The best quote from the discussion was definitely Joe Felsenstein saying that "*my phylogeny book has a simple, elegant, and clear explanation -- which I wrote in a hurry while excited that I finally understood this, and which turns out to make no sense whatsoever and should be firmly ignored by all*"

Why should you care? To me this is really one of the most basic evolutionary questions that you can ask about species level variation. Does (insert your trait of interest) have its current distribution due to a higher rate of loss or gain in comparison to the alternatives? If you are interested in these types of analyses you might want to look at a couple of recent articles that attempt to draw conclusions based on these types of tests.

Kimball, Rebecca T., Colette M. St Mary, and Edward L. Braun. "A Macroevolutionary Perspective on Multiple Sexual Traits in the Phasianidae (Galliformes)." *International journal of evolutionary biology* 2011 (2011).

Anacker, Brian L., et al. "Origins and consequences of serpentine endemism in the California flora." *Evolution* 65.2 (2011): 365-376.