This assignment has 3 questions, each with multiple parts. Choose two questions, and answer all parts of both. You're welcome to do all 3, but then tell me which two you want me to grade.

Please put your name only on the back of the last page and make sure all pages are securely attached. I prefer to grade anonymously.

1. The Ab.csv file contains the abundance of the tropical tree *Alseis blackiana* in two habitats (OldHigh and OldSlope). The data are the count of stems in 1 ha plots on Barro Colorado Island, Panama. There are 8 plots in the OldHigh habitat and 12 in the OldSlope.

   For most parts of this question, we will assume that the counts follow a Negative Binomial distribution. We will parameterize the Negative Binomial distribution in terms of the mean, μ, and θ, the overdispersion parameter. In the parameterization we will use, the variance is $\mu + \theta\mu^2$.

   See the end of this question for information and computing hints. In particular, note that the dnbinom() function expresses overdispersion as size, for which the variance of the data is $\mu + \mu^2/size$.

   (a) Ignore habitat (i.e., combine all 20 plots into one group) and find the maximum likelihood estimates (mle's) of the mean and θ, the overdispersion parameter.

   (b) Fit a model that allows habitat-specific means and a shared overdispersion. Find the mle's of the difference in means (as OldHigh - OldSlope) and the overdispersion.

   (c) Use results from this model to construct a 95% normal approximation confidence interval for the difference in means.

   (d) And use results from this model (and perhaps other models) to construct a likelihood ratio test of the null hypothesis that the means are equal. Report the value of your test statistic and a p-value.

   (e) Fit a model (or models) that allows habitat-specific means and habitat-specific overdispersion. Find the mle's of the two means and the two overdispersion parameters.

   (f) Use results from this model (and perhaps other models) to construct a likelihood ratio test of the null hypothesis that the two habitats have the same overdispersion. Report the value of your test statistic and a p-value.

   (g) (for a few extra credit points): calculate a 95% profile likelihood confidence interval for the difference in means.
   Note: this will require writing a function that enables finding the mle of one group's mean abundance and the shared overdispersion given a specified difference in means.

   Notes: 1) A function that calculates the log likelihood of (mean, size) given data is:

```
NBlnl <- function(beta, data) {
  mu <- beta[1]
  size <- beta[2]
  sum(dnbinom(data, size, mu=mu, log=T))
  }
```

2) One simple way to include the difference in means as a parameter is to define an indicator variable that has the value of 0 for observations in one group and the value of 1 for observations in the other. data now has two columns. count is the stem count, X is that indicator variable. When X = 0, $\mu = beta[1]$. When X = 1, $\mu = beta[1] + beta[2]$, so beta[2] is the difference.

A function that calculates the log likelihood of a different means, shared overdispersion model is:

```
NBlnl2 <- function(beta, data) {
  mu <- beta[1] + data$X*beta[2]
  size <- beta[3]
  sum(dnbinom(data$count, size, mu=mu, log=T))
  }
```

3) Remember the dnbinom() function defines overdispersion as size, not theta. If you don't remember how to convert between them, look at the variances.

2. We have discussed the model-based approach to the analysis of species composition data. There is a connection between it and the older dissimilarity-based methods. The connection is "model-based dissimilarity", also called "deviance dissimilarity". The idea is that the change in deviance provides a dissimilarity between two samples. This problem explores that connection.

   When counts, Y, have a Poisson distribution with mean λ, the probability mass function, i.e. P[Y = y]= $\frac{e^{-\lambda}\lambda^y}{y!}$. Consider the number of individuals of one species in two sites, Y_1 and Y_2. Both are assumed follow a Poisson distribution, $Y_1 \sim \text{Poisson}(\lambda_1)$ and $Y_2 \sim \text{Poisson}(\lambda_2)$. It is possible to test the hypothesis that $\lambda_1 = \lambda_2$, even when there is only a single observation at each site. The mle of λ when $\lambda_1 = \lambda_2 = \lambda$ is $\hat{\lambda} = (Y_1 + Y_2)/2$.

   (a) Derive the expression for the likelihood ratio test statistic testing $H_0 : \lambda_1 = \lambda_2$ against the alternative $H_a : \lambda_1 \neq \lambda_2$.

   You have counts, Y_{ij}, for all species (j) at each site (i). There are k species all together. All counts are assumed independent Poisson with mean λ_{ij}. The hypothesis is now $\lambda_{1j} = \lambda_{2j}$ for all species. In words, each species has a different mean abundance, but that is the same value at both sites.

   (b) Derive the test statistic that extends the test in question 2a to the hypothesis about all species.

   λ is the expected **count**. Species composition may also be conceived as a proportion, i.e. $\lambda = \pi s$, where s is the total count at a site and π is the proportion of a species. The total count is assumed to be a known value for each site. You want to test the hypothesis that $\pi_1 = \pi_2$. The mle of π when $\pi_1 = \pi_2 = \pi$ is $\hat{\pi} = (Y_1 + Y_2)/(s_1 + s_2)$.

   (c) Return to considering a single species. Derive the expression for the likelihood ratio test statistic testing $H_0 : \pi_1 = \pi_2$ against the alternative $H_a : \pi_1 \neq \pi_2$.

   (d) And extend that test statistic to the all species situation.

   (e) For a few extra credit points:

      i. Derive the mle of π under the hypothesis that $\pi_1 = \pi_2 = \pi$.

ii. The test statistic in question 2d is closely related to a well-known species diversity measure. What is the name of that measure?

3. The data in newfish.csv come from a study of fish communities in three wildlife refuges in Missouri. At each site(=refuge) data were collected in two habitats (main channel and chute). Fish were collected using 5 different gears = sampling techniques (EF: electrofishing, FN: Fyke nets, LH: large hoop nets, SH: small hoop nets, and SN: seine). Fyke nets were used in the chute habitat at site O, so there are no observations there. The intent was to get 4 samples for each combination of site, habitat, and gear, but there are fewer for some combinations. The data values are the count of each fish species in that sample. The version of newfish.csv that I'm providing you omits species found in less than 10% of the samples.

   The newfishenv.csv file has the information on refuge/habitat/gear for each sample. The data are sorted in the same order in both files (newfish.csv and newfishenv.csv).

   Gear is known to have a very large influence on the types of fish one collects. The biologists collecting the data expect differences between sites (refuges) and between habitats. They are very interested in the interaction between site and habitat. The biologists have asked you to analyze the data and provide them answers to the following questions:

   (a) Is there an interaction between site and habitat, i.e., is the difference between the two habitats similar (or not) at the three sites?

   (b) If there is an interaction, which species are significant contributors to that interaction?

   Things that are already known about the data:

   - There is huge overdispersion for many species, so you should assume a negative binomial distribution.
   - These sorts of questions can be analyzed using the manyglm() function in the mvabund library.
   - One combination of site/habitat/gear, has no data, so you must not include site:habitat:gear (the 3 way interaction) in your model. Two-way interactions are fine.

   Still, there are many possible ways to answer the biologist's questions.

   Your answer to this question should be in the form of a partial scientific paper. That is:

   - A statistical methods section that describes the statistical approach you used and how you made any necessary decisions. This should be sufficiently detailed that a reader could recreate your results from the data file and your description of the analysis.
   - A results section that includes appropriate tables and/or figures. This should answer the biologist's questions and may include additional interesting things you discovered.

   It is not necessary to repeat information provided here.