Lab 9. Invertebrate Community Diversity in Wetlands II: Data Analysis

Notes:

- Please brings a laptop for today's lab. A loaner from the library will work, we need to use a web browser.

Objectives:

1. Quantify biodiversity in an ecological system with standardized indices.
2. Statistically analyze and visually display our data to test for an association of wetland characteristics with biodiversity.
3. Interpret the results of our analyses to make inferences about local wetlands.

KEY WORDS: Shannon Diversity Index (H'); Shannon Equitability measure (E); barplot; Kruskal-Wallis test; ANOVA (Analysis of Variance)

Contents

1 Background – Analyzing Our Data

Last lab period, we collected and quantified invertebrates from several local wetlands that varied in size and amount of adjacent forest. We also have data for the same sites from last fall. For each sample, we identified the invertebrates and calculated diversity indices. This week, we will use the diversity indices to determine if invertebrate diversity varies signficantly among the sites or between seasons. The fall samples were collected in early November, when invertebrates are winding down their activity in preparation for winter. Thus, we might expect spring samples to have higher biodiversity as invertebrates hatch out and grow under conditions of plentiful food. To do so, we will employ two methods of statistical analysis: ANOVA and the Kruskal-Wallis test. Both tests address whether our measures of diversity vary significantly based on our explanatory variables (site, season, land use, wetland size). To refresh your memory, our hypothesis and the specific predictions are:

Hypothesis: Natural features of wetlands influence aquatic invertebrate community diversity and species richness.
Prediction 1: Invertebrate biodiversity measures will increase with total wetland area.
Prediction 2: Invertebrate biodiversity measures will increase with amount of adjacent forest (correlating with land use).
Prediction 3: Invertebrate biodiversity measures will be higher for spring samples compared to fall.

2 Importing Data and Calculating Diversity Estimates

Before we can begin any analysis, we need to import the file of raw data, with the counts of each species in each sample. While you also calculated the diversity indices last week, here we're going to see how to use R to calculate them, creating a new dataset object with all the diversity measures (S, H', and E) that we will use as our response

variables in the analysis.

YOU SHOULD REFER TO THE HANDOUT (OR YOUR SCRIPT) FROM THE LAB 7 HANDOUT FOR ANY STEPS YOU DO NOT REMEMBER HOW TO COMPLETE.

1. To start, download the datafile from the lab website (named `lab9dataspring2024.csv`). This file contains all of the data collected by all lab groups for all samples in both fall and spring as a CSV (comma-separated values) text file. Make note of where you save the file!

2. Log into RStudio at https://rstudio.oberlin.edu/. Upload that CSV file into RStudio (hint: Files pane, click Upload).

3. Load the tidyverse package so we can use its functions:

```
library(tidyverse)
```

4. Now, assign the file to a new object so we can use it in R (see below):

```
inverts <- read.csv("lab9dataspring2024.csv") # assign the file to the object data
```

Now check out the data, does everything look as it should?

```
inverts # print the object inverts to the screen
```

```
str(inverts) # show the structure of the object inverts
```

Remember that in this study we are not examining individual species that are found in the wetland samples. Rather, we are interested in the overall diversity of the community—not which taxa are present but how many and in what relative abundances? That means that when we calculate our diversity measures, we'll have one value of each diversity estimate for each replicate from each site.

Calculating species richness (*S*)

5. Now we need to create a new dataset to which we'll add the diversity measures for each sample (identified by `sampleID`). We'll use the group_by and summarize functions to calculate species richness. Notice that we'll create the new object `diversity` to save the results then we'll print the `diversity` object to the screen so we can make sure it worked correctly.

```
diversity <- inverts %>%
    group_by(sampleID, site, team, lab, hectares, landuse, season) %>%
    summarize(S = n_distinct(taxon)) %>%
    ungroup() # calculate species richness for each sampleID
```

```
diversity # print the new object to the screen so we can be sure it worked
```

> Remember: Not all alphanumeric characters work in R when you copy and paste code from outside of R. Type it in, do not copy and paste. This takes little time but will save you from confusing errors!

In the above code, the group_by function tells R to do the summarize step for each different value of the variables listed. It also means all of those variables will appear in our new diversity dataset. The summarize function says to create the new column 'S' and then calculate it by counting up how many different (distinct) taxa are listed for each wetland site.

Calculating the Shannon Diversity Index (*H'*)

6. Our next step is slightly complicated so be sure that you check your work! We're going to write code that will do all the work of calculating H' for each sample. *Note: The Shannon Diversity Index, H', is typically represented by H-prime but R will not accept a variable with a quotation mark in its name so we'll just use H instead.*

Remember from last week's lab that to get H' we need to calculate $H' = -\sum_{i=1}^{S} p_i \ln p_i$ for each sample.

7. First, we'll create a dataset of the p_i values for each sample using the functions `xtabs()` and `prop.table()`. The `xtabs` command reformats our dataset into a table where columns represent the different species and rows represent the different samples (sampleID), thus each table entry is the number of individuals of that species observed in that sample:

```
xtabs(count.per.ml ~ sampleID + taxon, data = inverts) # see what xtabs function does
```

The `prop.table()` command then converts the values from the xtabs object into proportions within each row, where the rows represent different samples (sampleID). Lastly, we'll use the `round()` function to show 4 decimal places for easier reading.

```
xtabs(count.per.ml ~ sampleID + taxon, data = inverts) %>%
    prop.table(margin = 1) %>% round(digits = 4) # see what the prop.table function does
```

8. Let's execute those commands in one go and save the results as an object called "data.pi":

```
data.pi <- xtabs(count.per.ml ~ sampleID + taxon, data = inverts) %>%
    prop.table(margin = 1) %>% round(digits = 4)
head(data.pi) # make sure everything looks good!
```

9. Next, we need to take those p_i values and calculate the product $p_i \ln p_i$ for each of them, saving this in a new dataset called "data.pilnpi":

```
data.pilnpi <- data.pi*log(data.pi)
```

10. We're ready to sum up the values for each sample and multiply them by -1 to complete the calculation of H' for each sample, saving them in a new data frame called "H":

```
H <- data.frame(sampleID = rownames(data.pilnpi), H = round(-1*rowSums(data.pilnpi,
    na.rm = T), digits=3), row.names = NULL) # create new dataset with the H' values
H # check to see what the results look like!
```

11. Note that this new data frame, `H`, has the same `sampleID` column as the `diversity` object we made earlier to store species richness calculations. This means that we can easily merge these two objects so we have both S and H' in one object:

```
diversity <- merge(diversity, H, by = "sampleID")
diversity # check the result!
```

Calculating the Shannon Equitability Index (E)

12. Now we just have one more thing to calculate: species evenness using the Shannon Equitability Index, E. Remember, this measure combines the information from S and H', like so: $E = \frac{e^{H'}}{S}$

This one is pretty straightforward to calculate using our `diversity` object to create a new column for E:

```
diversity <- diversity %>%
    mutate(E = round(exp(H)/S, digits=3)) # make a new column in diversity that calculates
    E for each sample
diversity # check the result!
```

Now we are ready to move on to visualizing and then analyzing our diversity dataset!

3 Visualizing the Data

Remember that the first step in data analysis is to make some pictures of the data—see what patterns are in your data. This helps you to interpret your subsequent statistical results, since they should reflect what you can see in the data.

1. You may remember that we have replicate samples for each site (each row of the diversity dataset is a replicate from one site). Just looking at the dataset, you can see any variation among sites or orther variables.

   ```
   diversity
   ```

2. Let's make a bar plot to get a visual sense of the variation between replicates. With `ggplot()`, we can use `geom_col` to make a barplot. We have to specify the *x* and *y* variables plus we'll color the bars by season, which corresponds to fall or spring. Let's check out our species richness (*S*) values:

   ```
   diversity %>% ggplot() +
       geom_col(mapping = aes(x = sampleID, y = S, fill = season)) +
       coord_flip() +
       ylab("Species richness (S)") +
       ggtitle("Species richness of all samples")
   ```

   Each bar represents one water sample and color indicates season. You can see any variation among sites, replicates, or lab days. Here's another way to show the same data, using points instead of bars:

   ```
   diversity %>% ggplot() +
       geom_jitter(mapping = aes(x = site, y = S, col = season), show.legend=T) +
       theme_bw()
   ```

   Ask yourself: Are there any patterns? Does it look like species richness varies between wetland sites?

   These plots are useful to help us see the total variation in our data but it doesn't really show us whether diversity seems to relate to any of our explanatory variables. That is, this barplot does not help address our hypothesis. Instead, let's create a box-and-whisker plot (or boxplot). With the box plot, we can graph the data by groups, so that we show the minimum data value, the first quartile (the 25th percentile of the data), the median (the 50th percentile or middle of the data), the 3rd quartile (the 75th percentile), and the maximum data value. This can give us a better sense of whether the groups differ from each other on average.

3. For the first boxplot, we'll color the boxes by site. Then we'll make a second boxplot that differentiates site and also season.

   ```
   diversity %>% ggplot() +
       geom_boxplot(mapping = aes(x = site, y = S, fill = site), show.legend = F) +
       ylab("Species richness (S)") +
       ggtitle("Boxplot of species richness for each wetland site") +
       theme_bw()
   ```

   And here's the second boxplot, including season:

   ```
   diversity %>% ggplot() +
       geom_boxplot(mapping = aes(x = season, y = S)) +
       ylab("Species richness (S)") +
       ggtitle("Boxplot of species richness for each season by wetland site") +
       facet_wrap(~ site) +
       theme_bw()
   ```

The line in the middle of the box is the median for that group. The boxed range represents the middle 50% of the data for that group, where the bottom hinge is the 25th quartile and the upper hinge is the 75th quartile. The whiskers span from the lowest observed value to the highest observed value for that group.

Ask yourself: Does this graph show the same patterns you saw in the bar or dot plots? What's the benefit of this box plot versus the bar or dot plot?

4. Now make plots for the other two response variables, H' and E. **Ask yourself: What patterns do see for those variables?**

5. We can also look at a summary of our data in table form. We'll ask R to calculate the mean, standard deviation (sd) and sample size (n) for each diversity measure for each sample, calculating a mean for each season. If we wanted to average by a different variable, like land use, we'd have to change this code slightly.

```
diversity %>%
    group_by(site, season) %>%
    summarize(meanS = mean(S), stdevS = sd(S), meanH = mean(H), stdevH = sd(H),
    meanE = mean(E), stdevE = sd(E), n = length(S))
```

With this function, we are able to specify all the functions we want to run (i.e., mean, standard deviation, and sample size) and what grouping variables to use. You could also include other response or explanatory variables *and* save the output as an object that we could access later (we'll want it for some of the analyses):

```
div.summary <- diversity %>%
    group_by(site, season, landuse) %>%
    summarize(meanS = mean(S), stdevS = sd(S), meanH = mean(H), stdevH = sd(H),
    meanE = mean(E), stdevE = sd(E), n = length(S))

View(div.summary)
```

The table should show the same patterns that you were seeing in your plots earlier.

6. You can also export the summary table as a file that you can then open in Excel or Google Sheets to format it nicely for display. Here's the code to export the object as a CSV file:

```
write_csv(div.summary, file = "diversitysummary.csv", quote = "none")
```

After running that code, you need to find the file you just created in the `Files` pane, click the box next to the file name, and then click `More` → `Export...` Give it a name, click Download, and choose where to save the file on the computer.

4 Analyzing the Data

Now that we've checked out our data graphically, we can run an analysis to get statistics to place some probabilities on any patterns in our data. Remember from Lab 7 Adaptive Morphology that we determine what kind of plot and analysis to use based on the types of variables we have, as in this table.

Explanatory Variable	**Response Variable**	**Graph to use**	**Analysis to use**	**Values to report**
Continuous	Continuous	Scatterplot (`geom_point`)	Linear regression (`lm`)	Slope, y-intercept, *P*-value
Categorical	Continuous	Boxplot (`geom_boxplot`)	ANOVA and Kruskal-Wallis	*P*-value for explanatory variable

Overview of Analysis Methods

Last time, in Lab 7, we ran linear regressions because we had continuous response and explanatory variables. We can do that again if we are using continuous explanatory variables (hectares is continuous, the area of the pond, but lab day, season, and land use are categorical). Refer back to Lab 7 if you need a refresher on how to run a regression and how to understand the results. In other cases, when we have continuous response variables and **categorical** explanatory variables we'll run an ANOVA (Analysis of Variance)—this test checks whether two (or more) categories differ in their average responses. The statistical null hypothesis for an ANOVA is that the different groups (for example, fall or spring season) have the same mean for the response variable. The underlying logic is similar to that of the chi-square test we ran in a previous lab exercise: we calculate the overall average and variance of the dataset, lumping all individuals together as a single group (this is our "expected" value) and we compare that mean to the means and variances we calculate **within** each group (our "observed" values). The bigger the difference between the observed and expected values, the more likely that the groups have different means for the response variable. ANOVA is what's known as a "parametric" test; that is, the method assumes that the data we are analyzing come from a known probability distribution, in this case the normal distribution. When data are normally distributed, it means that we can accurately describe the variation of the sample population using things like the arithmetic average and the sample variance (an estimate of how spread out the data are).

However, not all data meet the assumptions of normality. When data are not normal (nor well-described by other familiar distributions), we can use different tests, known as non-parametric tests. The Kruskal-Wallis test is a non-parametric test that is kind of like an ANOVA but for data that are not normal. Instead of calculating the mean of S, the Kruskal-Wallis test will order our data values and assign them a rank (e.g., 1 for the smallest value, 2 for the next smallest value, etc). Then, you calculate the average rank and the sum of ranks for each treatment group and the total dataset. The test statistic is then calculated from the sum of squared deviations of the groups relative to the total dataset. The Kruskal-Wallis test statistic is somewhat like a chi-square test statistic (remember observed minus expected squared?) and is in fact chi-square distributed. In the end, you get a *P*-value for the probability that the *ranks* differ between the groups.

Which test should we believe? Well, it depends... ANOVA is actually quite robust to deviations from normality, at least certain types of deviations. The Kruskal-Wallis test doesn't assume the data to be normally distributed, but is less powerful than ANOVA and can be misleading under certain conditions—caution may be warranted if groups do not have 5 or more observations each (this is also true for a chi-square test). In our data, we might be worried about the assumption of normality but we also have relatively few observations, so neither test is a perfect fit for this data (as happens all too often in statistical analysis!). Because this is not a statistics course, we're not going to go into detail about checking normality (it's not simple to interpret) or why small numbers of observations can cause problems. Instead, we're going to run both tests and see if they give us similar results, which is what we expect to happen most of the time. Then it's up to you to decide how to interpret them!

Side note: If you have more questions or are interested in learning more, you should consider taking STAT 113! In the meantime, there are some good online resources to explore—see `http://www.biostathandbook.com/index.html` *—and your instructors are happy to discuss further.*

Analyzing *S* with ANOVA

1. When we run the ANOVA, we use the full dataset with all the samples and we include `site` as an explanatory variable. We could also use other explanatory variables at the same time but to start we'll just use site. The function to run an ANOVA in R is `aov()` and as arguments you need to supply a formula (of the $y \sim x$ variety) and the name of the dataset containing variables x and y. We'll run the function `anova` on the output of `aov()` as that will give us a nice summary of the results. Here's the code for an ANOVA model of *S* with site as the explanatory variable:

```
S.site.aov <- aov(S ~ site, data = diversity) %>% anova()
```

And now to print the summary:

```
S.site.aov # print the summary results to the screen
```

You should get results that have the same layout as below but different numbers (the example was run on a different data set from the one you are using):

```
Analysis of Variance Table

Response:  S
             Df  Sum Sq  Mean Sq  F value  Pr(>F)
site          4   5.280   2.6402   0.8956  0.4279
Residuals    27  47.167   2.9479
```

How do we read and interpret this table of information?

- In this table, the `Pr(>F)` is the *P*-value, we have one for each explanatory variable we included in the model.
- The **F-value** is the test statistic, analogous to the χ^2 value of the chi-square test.
- In this example, the first row of the table tells us about the effect of `site` on species richness. For these results, the probability that sites do not differ, on average, in species richness is $P = 0.4279$, which is much higher than 0.05. Thus, we would conclude that these different wetlands do not show significant differences in species richness.
- If you are reporting these results, it would be typical to provide the entire ANOVA table or the F-value, degrees of freedom (Df), and the *P*-value (written as $F_{4,27} = 0.8956$ and $P = 0.4279$ for the effect of site). For example, in this case we could say "there was no significant difference in species richness among sites ($F_{4,27} = 0.8956, P = 0.4279$)".

Ask yourself: What do the ANOVA results tell us about our data, hypothesis, and prediction?

2. Now, if we want to include a second explanatory variable, we change the `anova` code a little bit to include a second variable. Let's try running the anova with both site and land use.

```
S.site.land.aov <- aov(S ~ landuse/site, data = diversity) %>% anova()
```

Notice how `site` is included in the model—in the form a/b—this is read as "b is nested within a". That means that each site `site` has a value of species richness for only a single `landuse`. As a result, the effect of `site` is not independent of the effect of `landuse`, `site` is nested in `landuse`. If we used season instead, that would not be nested and we would write `S ~ site + season` for that part of the code.

And now to print the summary:

```
S.site.land.aov # print the summary results to the screen
```

Referring to the instructions about the first model, how do you think we interpret these results? Keep in mind that the effect of different sites is captured by the `landuse:site` line in the output.

Analyzing *S* with Kruskal-Wallis

3. Now for the Kruskal-Wallis test! For this one, we can only include a single response variable *and* a single explanatory variable. We write the Kruskal-Wallis model similar to that of an ANOVA (i.e., $y \sim x$), only including site and we also provide the dataset name:

```
S.site.kruskal <- kruskal.test(S ~site, data = diversity)
```

```
S.site.kruskal # print the test results to the screen
```

which should print out something like (run on a different dataset from the one you are using):

```
    Kruskal-Wallis rank sum test
data:     S.mean by site
Kruskal-Wallis chi-squared = 1.1429, df = 4, p-value = 0.5647
```

Similar to the ANOVA results, you would typically report the test statistic (1.1429 here), the degrees of freedom (df), and the *P*-value for this test.

Ask yourself: Does the result from the Kruskal-Wallis test agree with that from the ANOVA?

It's Your Turn – Analyze *H'* and *E*

4. Now that you've completed the analysis for *S*, go back and do the same analyses but with the other response variables (H' and E).

   Ask yourself: What do the results tell us about our data, hypothesis, and predictions?

5 Export Your Results

1. Remember, to export graphs, you need to save the file to the folder in RStudio, then go to the `File` panel, click the box to select the graph, and then click `More` → `Export` to download the file to your computer.

2. You may want to export your summary table too (you can then open it in another program, like Excel, to make a table that looks nice):

   ```
   write_csv(div.summary, file = "diversity.table.csv", quote = "none")
   ```

   You can also export the output of the statistical tests this way:

   ```
   write_csv(data.frame(S.site.aov), file="S.site.anova.out.csv", quote = "none")
   ```

   ```
   write_csv(data.frame(S.site.kruskal[]), file="S.site.kruskal.out.csv", quote="none")
   ```

6 Pre-lab exercise

Consider the data that we collected last week in answering the following. You may choose to refer to your lab 8 handout too. This exercise is due by the start of your lab period.

1. What are our explanatory variables?
2. What are our response variables?
3. What is our null hypothesis?
4. If the results of our statistical tests yield $P < 0.05$, what does that tell us about our prediction(s)? About our hypothesis?
5. What does it tell us about our hypothesis/predictions if our *P*-value is greater than 0.05?

7 Post-Lab Assignment

This exercise is due by the start of next week's lab.

To answer the questions below, you will need to do some additional analyses beyond what was described in this handout. We showed you how to analyze the effect of site, answering the question 'Do wetland sites differ in diversity?' Now, you will need to go back and do the same analyses, for S, H', an E as response variables but then including either landuse or season as explanatory variables.

Combine informative, nicely formatted graphs and tables into a document with the relevant P-values next to the graphs. Include typed answers to the following questions and turn them in with your figures next week. Support your answers by referring specifically to your figures and/or the results of your tests of significance, where appropriate. We only did part of these analyses in lab, since we went through using wetland site as an explanatory variable. You will need to run additional analyses to check for the effects of adjacent forest (landuse) or seasonality (season).

A single group of up to 4 people may turn in 1 group assignment, assuming all worked together to complete the assignment.

1. Consider the analyses described here for wetland site. Do we have any evidence that these 5 wetlands differ in diversity, as represented by species richness (S), species diversity (H'), or species evenness (E)? Explain and comment on each of these metrics and how they vary among sites.

2. Describe the additional analysis you chose to do for landuse or season (considering all of S, H', and E). Did you find any evidence that the diversity metrics differed for this additional explanatory variable? Explain, being sure to comment on the output for each diversity measure.

3. After considering your results, think about the hypothesis and predictions we were testing. They don't really tell us about how or why wetland size, adjacent forest, or season would affect aquatic invertebrate diversity. Based in part on your observations and data from last week, propose a new hypothesis (with a mechanism!) and one or two predictions that might explain the variation in diversity we have observed for the community of aquatic invertebrates.

4. Other than wetland size, adjacent forest, or season, are there other comparisons you might make with these data? (Consider other variables that are in the dataset but that we did not use in our analyses.)

5. For these analyses, we only examined planktonic invertebrates that occurred in the water column sampled all in one specific week at two times of the year. How might this particular method of sampling influence the results we obtained?

6. What additional data could you collect to compare biodiversity among wetlands? To answer this question, consider other aspects of the sites that you might want to evaluate.