SUNY Geneseo
KnightScholar

Principles of Ecology Sustainability Curriculum

2018

POE 3 data analysis assignment

Suann Yang
SUNY Geneseo, yang@geneseo.edu

Follow this and additional works at: https://knightscholar.geneseo.edu/sustainability-curriculum-principles-of-ecology

This work is licensed under a Creative Commons Attribution-Share Alike 4.0 License.

Recommended Citation
Yang, Suann, "POE 3 data analysis assignment" (2018). *Principles of Ecology*. 3.
https://knightscholar.geneseo.edu/sustainability-curriculum-principles-of-ecology/3

This Open Educational Resource is brought to you for free and open access by the Sustainability Curriculum at KnightScholar. It has been accepted for inclusion in Principles of Ecology by an authorized administrator of KnightScholar. For more information, please contact KnightScholar@geneseo.edu.

Food Sustainability: Data analysis and interpretation assignment

Overview

This assignment is designed to

2. Evaluate ways to alleviate the environmental degradation that can be caused by agriculture, including:
    a. Interpreting statistical results to evaluate a method's effectiveness.
    b. Conducting a two-sample comparison in R.

Students will be analyzing a subset of data, modified from Bender and van der Heijden (2015). In this study, Bender and van der Heijden (2015) ask the question "Does enriching the soil microbiota improve soil quality and crop yield?"

There can be one or two class periods allotted for this assignment. Students will need to bring their laptops with RStudio on both days. The assignment is introduced at the end of the class prior to the two class periods dedicated to it. For homework, students work on their own to create graphs in R of a common data set. This is facilitated by an online assignment (distributed via course management system, or paper handout) to practice the R syntax. During the next class, students will practice (or learn!) to do a two-sample *t*-test in R, and interpret those results. Then, each group divides up the remaining data sets among the students in the group. Two options are possible from here: If there is time, each student will then create graphs, conduct a two-sample *t*-test, and interpret his/her results for his/her data set. Near the end of this class period, the group will compile their results, and discuss whether or not adding soil microbiota is an effective way to improve soil quality and crop yield. Each student submits an individual, written evaluation based on the group discussion for homework that is due at the beginning of the next class period. If there is not an opportunity for students to get started during class on their own data set, then they will write the code for homework (and perhaps try to run it). Then, any troubleshooting will occur in the next class while each person analyzes his/her assigned data set. Compiling and discussing results follow, and students should complete their written evaluation by the end of this class period.

Materials

1. Dataset modified from Bender and van der Heijden (2015) provided as "POE 3 soil biota data.csv"
2. Slides of Bender and van der Heijden (2015) question and experimental design, provided as "POE 3 mini-lecture.pptx"
3. Hartvigsen, G. 2014. *A Primer in Biological Data Analysis and Visualization Using R.* Columbia University Press.
4. R script for visualizing the data and demonstrating how to conduct a *t*-test, provided as "POE 3 graphs and t-test assignment.r"

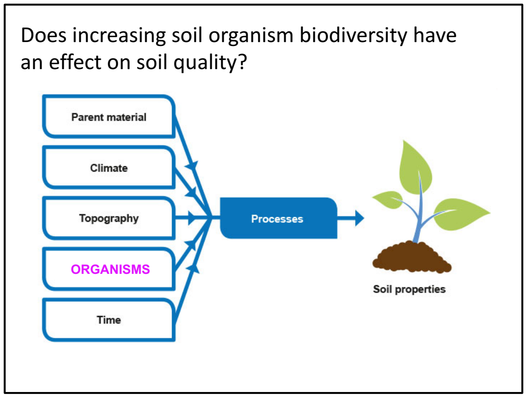
Instructor Notes

(end of last class before beginning the assignment)

Introduction

After exploring the ways in which agriculture can alter decomposition and nutrient cycling, students are introduced to the existence of the variety of sustainable farming practices.

One specific method is then investigated:



Does increasing soil organism biodiversity have an effect on soil quality?

Parent material

Climate

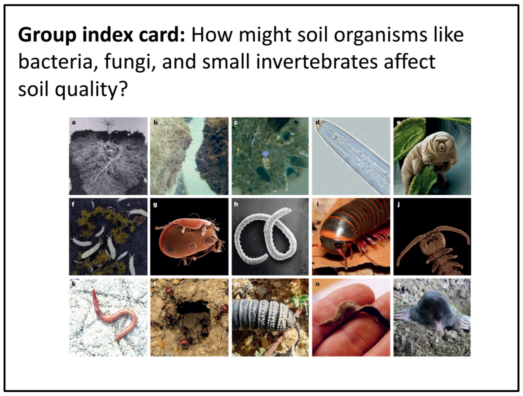
Topography

ORGANISMS

Time

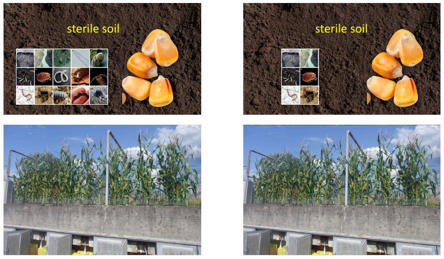
Processes

Soil properties



Group index card: How might soil organisms like bacteria, fungi, and small invertebrates affect soil quality?

The group writing will become the group's hypothesis for the experiment:



1. Collected leachate below the box after watering
2. Analyzed leachate for different forms of nitrogen
3. Harvested corn for biomass measurement
4. Analyzed corn biomass for nitrogen content

In this component of Bender and van der Heijden's (2015) experiment, they added enriched soil microbiota or reduced microbiota to sterilized soil, and then grew corn in it.

The overview for the rest of the assignment is then described to the class, and students begin the Canvas activity.

Online activity for homework, due next class period

The purpose of the activity is to prepare students to decide what kind of graphs to draw and what code to write. They upload PDFs of their graphs to a course management system (or printed and turned in at beginning of the next class).

Create a quiz or tutorial in your course management system (or make a paper handout) as follows:

Preparation steps:
1) Make a working folder (e.g., Desktop//Ecology//R_analyses.)
2) Next, we will create an R Script file.

- Read 1.5 Using Script Files (pp. 19-20) in the Hartvigsen *Primer*.
- Open RStudio. From your toolbar, create an R Script by clicking on File→New File→R Script.

- Save this R Script to your working folder (see prep step 1), with a logical, descriptive file name (<Date>_Nutrient analysis_<your name>.R).

(True/False) We should type our code in the R Script window (top left in RStudio) instead of the command line console window (bottom left in RStudio).

Graphing Step 1: Upload the data set to RStudio to create the dataframe. Read 2.2 READING DATA FROM AN EXCEL SPREADSHEET of the Hartvigsen *Primer,* (pp. 27-30, paying special attention to Box 2.2).

Apply the section you just read to the soil biota data CSV (comma separated values) file. Which is the right way to construct the code?

A) `sbd = read.table(“POE 3 soil biota data.csv”, header = T)`
B) `sbd = read.csv(“POE 3 soil biota data.csv”, header = T)`
C) `read.csv(“POE 3 soil biota data.csv”, header = T)`
D) `sbd = read.csv(POE 3 soil biota data.csv, header = T)`

Graphing Step 2: Draw histograms of each treatment

Histograms of each treatment will allow us to assess the range of variation in biomass that was observed.

2a. We want to obtain each treatment’s histogram of biomass, so we first have to create subsets of the data, one for the enriched plots and one for the reduced plots. Read 3.4 SUBSETTING DATA of the Hartvigsen *Primer* (pp. 40-41).

You will see that the syntax for subsetting looks like this:

    data1 = subset(the name of the dataframe, the name of the column that identifies what we want to subset == “the name of the items we want”)

data1 is the name of the dataframe that will hold the subset that we want R to make
subset() is the function we want R to use to make our subset
all the stuff in the parenthesis specifies what will go into the subset

Apply to the corn biomass data set. Which is the right way to construct the code in the R Script (not the Console window) for extracting just the corn biomass for the enriched treatment?

A) `enriched = subset(sbd, treatment == “enriched”)`
B) `enriched = subset(sbd, treatment = “enriched”)`
C) `enriched = subset(sbd, enriched == “treatment”)`
D) `subset = enriched(sbd, treatment)`

Write the code to subset the corn biomass for the reduced treatment.

2b. Now we are ready to make the histograms. Read 5.2 HISTOGRAMS of the Hartvigsen *Primer* (pp. 69-70).

You will see that the format of the code is like this:
hist(name of the variable we want the histogram of, xlab = “the name of the variable (units if applicable)”, col = “name of a color”, main = “the title of the histogram”)

hist() is the name of the function that we want R to use to draw our histogram
all the stuff in the parentheses describes what data we want R to use, and how the histogram should be formatted

Apply to the enriched treatment’s corn biomass. Which is the right way to format the code for a histogram of the corn biomass of the enriched treatment?
A) `hist(enriched$total_corn_biomass_tonne_per_ha, xlab = “Corn biomass (tonne/ha)”, col = “blue”, main = “Corn biomass for the enriched treatment”, breaks = 3)`
B) `hist(sbd$total_corn_biomass_tonne_per_ha, xlab = “Corn biomass (tonne/ha)”, col = “blue”, main = “Corn biomass for the enriched treatment”)`
C) `hist(total_corn_biomass_tonne_per_ha, xlab = “Corn biomass (tonne/ha)”, col = “blue”, main = “Corn biomass for the enriched treatment”)`

Write the code to draw a histogram of the corn biomass for the reduced treatment. You may want to use a different color.

Graphing Step 3: Draw a summary graph of means

The two histograms are not a very compact way to compare the two treatments to each other. Another way would be to graph the independent variable on the x-axis and the dependent variable on the y-axis. What is the independent variable?
A) enriched
B) reduced
C) treatment
D) total corn biomass

Considering the fact that the independent variable is categorical (discrete), what is the best graph type to use?
A) scatter plot
B) bar plot
C) pie chart
D) it doesn’t matter

3a. To draw a bar graph (bar plot), we must first make a table of mean biomass for each treatment. The format of the code is like this:

table<-tapply(the name of the target column we want to summarize, the name of the column that has the categories, mean)

table is the name of the table we are asking R to make
tapply is the function we want R to use to make our table
all the stuff in the parentheses describes what kind of calculation we want R to do to make the table (in this case, the mean of the values from the target column organized by the category column)

Apply to the data set provided. Which is the right way to construct the code?

A) `table = tapply(sbd$total_corn_biomass_tonne_per_ha, sbd$treatment, mean)`
B) `tapply = table(sbd$treatment, sbd$total_corn_biomass_tonne_per_ha, mean)`
C) `mean = tapply(sbd$treatment, sbd$total_corn_biomass_tonne_per_ha, table)`
D) `sbd = tapply(sbd$total_corn_biomass_tonne_per_ha, table, mean)`

Think about how you want the bar graph to look. Which goes on the x-axis?
A) total_corn_biomass_tonne_per_hectare
B) treatment
C) mean
D) table

3b. Draw the bar graph. Read 5.4 Barplots in the Hartvigsen *Primer* (pp. 72-76).

You will see that the format of the code is like this:

bar = barplot(table, xlab=”write the independent variable’s name”, ylab=”write the dependent variable’s name”, col=c(“black”, “gray50”))

bar is the name of the bar graph we want R to draw
barplot is the function we want R to use to draw our bar graph
all the stuff in the parentheses specifies what R should put in our bar graph
table is what data to plot (the table we created earlier)
xlab is the label for the independent variable
ylab is the label for the dependent variable
col says what colors to use; always put the colors in quotations
c() tells R that we want everything in the parentheses that follows to go together; colors in this case for the two bars

Instructor Notes

Apply to the table you created of the corn biomass. Which is the right way to construct the code? (Note that titles aren’t included here for brevity. To include a title, use the argument main = “An appropriate title” like in the histogram code.

A) `barplot(table, xlab=”Treatment”, ylab=”Corn biomass (tonne/ha)”, col=c(“black”, “gray50”))`

B) `bar = barplot(table, xlab=”Treatment”, ylab=”Corn biomass (tonne/ha)”, col=c(“black”, “gray50”))`

C) `bar = barplot(table, xlab=”Corn biomass (tonne/ha)”, ylab=”Treatment”, col=c(“black”, “gray50”))`

D) `plot(table, xlab=”Treatment”, ylab=”Corn biomass (tonne/ha)”, col=c(“black”, “gray50”))`

You should be saving your work regularly when working in R.
A) True
B) False

Upload your histogram for the enriched treatment.
Upload your histogram for the reduced treatment.
Upload your bar graph of the corn biomass per treatment.
Upload your R script.

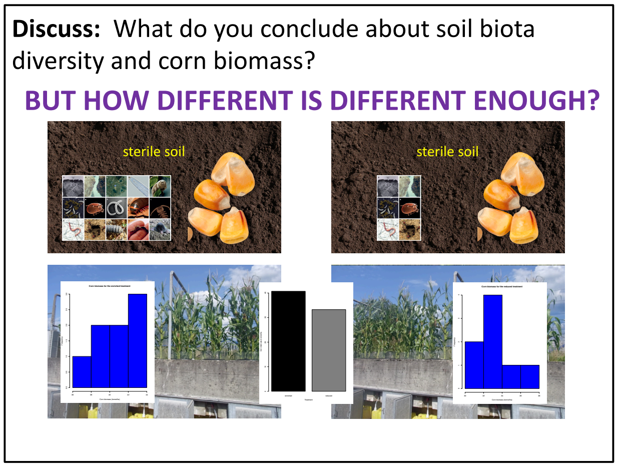
Instructor Notes

(First full class period)

t-test

Students will need to bring their laptops and a copy of Hartvigsen's *Primer.*

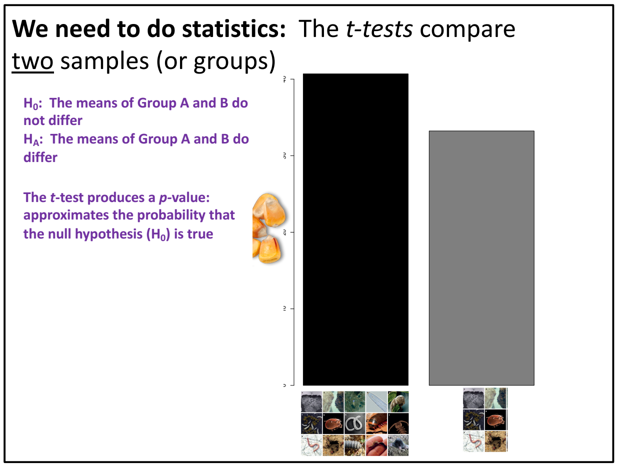
After reviewing the histograms and bar graphs generated by the class, students decide whether or not the treatment affects the total corn biomass.



Discuss: What do you conclude about soil biota diversity and corn biomass?

BUT HOW DIFFERENT IS DIFFERENT ENOUGH?

sterile soil
sterile soil



We need to do statistics: The *t-tests* compare two samples (or groups)

H_0: The means of Group A and B do not differ

H_A: The means of Group A and B do differ

The *t*-test produces a *p*-value: approximates the probability that the null hypothesis (H_0) is true

Demonstrate how to conduct a *t*-test on the corn biomass data:

1. Check the assumptions. (Shapiro-Wilk, H_0 is that the data are normal; *F*-test, H_0 is that the variances are equal).
2. Proceed to the *t*-test.
3. The class interprets the *p*-value together.

(Homework or next class period)

Jigsaw

Each group divides up the remaining dependent variables: Total corn N content, total N leaching, organic N leaching, nitrate leaching. They may want to do Total corn N content together first, then split the remaining 3.

Group members upload their individual work to a Google Document. A template for this document, for one dependent variable, is provided below.

Google Document Template

Total Corn Nitrogen Content Analysis

Replace this text with the name of the person who completes this section (e.g., by Charles Darwin)

Why is this measurement informative for the question we are studying?

Replace this text with your answer.

Graphs

Replace this text with your histograms and bar graph. Make sure your graph axes are labeled appropriately. Include a title and caption (Figure 1a. Blah blah blah. Use Figure 1a, 1b, 1c).

Statistical analysis

Statistical hypotheses

Null hypothesis:

Replace this text with your null hypothesis.

Alternative hypothesis:

Replace this text with your alternative hypothesis. (Add a sentence to explain why you have chosen Enriched>Reduced, Enriched<Reduced, or Enriched≠Reduced).

Is a *t*-test appropriate to conduct?

Assumption	Test	*p*-value
Normal distributions	Shapiro-Wilk	TYPE YOUR P VALUES HERE
Variances are equal	F test	TYPE YOUR P VALUE HERE

Replace this text with your interpretations of the *p*-values above to answer the question.

Fill out the table to report the results from your *t*-test.

t-statistic	df	*p*-value

Interpret the results of your *t*-test statistically.

Replace this text with a few sentences that interprets the *p*-value with respect to your statistical hypotheses.

Interpret the results of your *t*-test biologically.

Replace this text with a few sentences that interprets the *t*-test results with respect to the question we are studying. It may be helpful to review the reason you gave for why this measurement is informative (first question).

Final step: individual writing assignment

Individuals upload their work to the individual assignment space on Canvas:

Using the analyses completed by your group to address the question:

- **How effective is enriching soil biota diversity for soil quality and crop yield? Apply your findings to suggest strategies for reducing the impact of agriculture on nutrient cycling.**

In your typed, double-spaced response, include and refer to appropriate figures that either you or your group mates made (label with figure numbers and captions) and statistical results (include *p*-values and mean values). Propose an ecological explanation for what your found. *Print (and* *staple if applicable) and bring to our next class. Include your name, section, and new group number please.*

External references are not required, but if you use any, please include citations.

FAQs

1. Some of my group did not finish analyzing their data sets. What do I do?
   - You could communicate to find out when each person is able to have the work completed.
   - You could find out if they need help, and either offer assistance or refer your group mate to office hours.
2. I'm one of those people who did not finish analyzing my data set, because I'm not sure what is going on. How can I get help in a timely fashion?
   - You could contact the instructor or TA.
   - You could ask one of your group mates for help.
3. Which graphs should I include in my write-up?
   - You should include the graphs that support the message that you are conveying in your writing, without redundancy. *(Hint: for each dependent variable, the two histograms we drew are equal to their one bar graph)*
4. How do I summarize a statistical result?
   - You want to report a *p*-value along with the biological interpretation.
   - You want to be concise.
   - Good summary: Corn had significantly higher biomass when the soil biota was enriched compared to when soil biota was reduced (t = 8.0732, df = 14, p<0.0001). *(Note, the cutoff for reporting an exact p-value is 0.0001; any smaller than that is written p<0.0001)*
   - Bad summary: Corn biomass differed in reduced vs. enriched treatments. *(Note, this does not communicate what the difference is. Plus, there are no statistical results here.)*
   - Worse summary: Because *p*<0.05, I rejected the null hypothesis. *(Note, this does not communicate anything at all about the experiment)*

Grading rubric

Component	Satisfactory (3)	Progressing (2)	Needs improvement (1)
Structure: organized, few errors			
Statistics: *p*-values reported and interpreted correctly			
Figures: graphs support the argument with redundancy			
Argument: conclusions are			

consistent with the statistical results			
Ecological explanation: mechanisms are biologically correct and relevant			

Instructor Notes

References:

Bender SF, van der Heijden MGA (2015) Soil biota enhance agricultural sustainability by improving crop yield, nutrient uptake and reducing nitrogen leaching losses. Journal of Applied Ecology 52(1): 228-239. https://doi.org/10.1111/1365-2664.12351

Bender SF, van der Heijden MGA (2014) Data from: Soil biota enhance agricultural sustainability by improving crop yield, nutrient uptake and reducing nitrogen leaching losses. Dryad Digital Repository. https://doi.org/10.5061/dryad.p5v4g