Photo credit: Omar de Kok-Mercado

L2: Data Analysis with R (20 points)

Instructor: Nolan Bett
TAs: Wenhuan Xu & Abdulateef Gafar

DUE DATE: Friday, January 27th @ 9pm

OVERVIEW OF LAB:

In this lab, you will continue to build your R skills as you refamiliarize yourself with fundamental statistical concepts.

The datasets for this lab come from agricultural landscapes in Iowa. In the first part of the lab, you will examine production costs in conventional agricultural systems. In the second part of the lab, you will examine data from an innovative experiment in Iowa designed to reduce the environmental impacts and improve biodiversity and ecosystem function in monocrop agricultural systems. In the experiment, farmers planted strips of native prairie plants (similar to hedgerows but with prairie grasses and flowers) in the middle of monoculture row crop fields in Iowa. Scientists then assessed how these strips affected biodiversity and ecosystem services, and the trade-offs between enhanced ecosystem function versus the reduction in crop yields. While working on this lab, think about a farmer's decision-making process, the tradeoffs between various services, and how the valuations of ecosystem services might differ among various stakeholders.

The lab is based on the following study: (an interesting read!)

> Schulte, L. A., Niemi, J., Helmers, M. J., Liebman, M., Arbuckle, J. G., James, D. E., . . . Witte, C. (2017). Prairie strips improve biodiversity and the delivery of multiple

ecosystem services from corn-soybean croplands. Proceedings of the National Academy of Sciences of the United States of America, 114(42), 11247-11252. doi:10.1073/pnas.1620229114

LEARNING OBJECTIVES:

By the end of this exercise, you should be:

- increasingly comfortable with basic R skills (e.g., loading packages; importing, summarizing, and cleaning data; making histograms and basic plots)
- able to conduct and interpret a t-Test in R to determine if there are significant differences between two groups
- able to conduct and interpret an ANOVA analysis in R to determine if there are significant differences between three or more groups
- able to conduct and interpret a Regression analysis in R to determine the relationship between two continuous variables

DELIVERABLES:

Please upload to Canvas two documents:

1. Your r-script (including additional code you wrote with comments)
2. A Word or PDF document with answers to questions 1-8 below

Note that the first few questions are worth fewer points, and the later questions are worth more. For the later questions in particular, you want to clearly explain the results of your analysis, supporting your statements with quantitative results (e.g., means, p-values, etc). If you include graphs or plots, these should have axis labels, captions, and figure numbers, and you should refer to the figure numbers in your text.

Grading will be based on: completeness, correct answers, and readability, so please format and organize your answers well before you submit.

Q1 (1 point): Compare the mean and standard deviation of nitrogen costs, assuming a) that the NA values are real (i.e., remove them from the analysis) versus (b) that they should be 0$. Also, take a look at the dataset to see where the NAs/zeros exist. Which do you think is correct (i.e., do you think we should change the NAs to 0?)? Why or why not?

Q2 (1 point): Depending on what type of data you are using and how the data was collected, sometimes you might want to replace missing values with the mean value of the variable. For example, if there were missing values in commodity prices, most likely the prices wouldn't have dropped to $0. What code would you write to replace the NA values with the mean price in the Price.Mg column? Do this in R, then paste the code below.

Q3 (1 point): What is the mean of the new 'Total_Cost' variable? What is the mean and standard deviation for Annual_Revenue?

Q4 (1 point): What is the mean Seed Cost for a) Soybean and b) Corn, as shown in the T.test output? Are the means statistically different? How do you know?

Q5 (2 point): What is the relationship between the cost of labour and the commodity price? Is the relationship significant? How do you know? Export the graph above and include it in your answer sheet. Make sure to include a caption and label your axes.

Q6 (4 points): Use your coding knowledge to create two subset datasets: one for soybean and one for corn. Then conduct linear regressions for each dataset to examine the effect of herbicide costs on average yield. Is there a linear relationship for either crop? Why might this be? Is the relationship statistically significant? Include your well-formatted graphs and support your answer with your statistical results.

Q7 (5 points): Examine each of the four other ecosystem service variables to assess how they vary by crop type (row crop vs 10% prairie strips vs 20% prairie strips). Before you do the analysis, start with a hypothesis of the differences that you expect to see, compare the different treatments with a boxplot, then run your ANOVA. Which ecosystem services are significantly different among crop types? Report and interpret the results of your analysis. As you do so, draw out the larger patterns in your results, and reflect on what the results suggest about the use of prairie strips in intensive ag systems. (see note at end on reporting results)

Q8 (5 points): Run a post-hoc test for each of the five ecosystem services. Are all of the differences significant? What do the results suggest about the synergies or trade-offs between ecosystem services? Report and interpret the results of your analysis (remember to include your code), setting it into larger context and reflecting on the implications of your findings. Support your findings with p-values and graphs as needed.

Note: Good ways to report your scientific results

In reporting your results, it is often most effective to describe the main trend in your result, while including enough specific information that the reader can understand and interpret the details. For example, to report the results of a Pollinator Abundance ANOVA, you might write: "Pollinator abundance was significantly different between the three crop types (p=0.003). Fields with 20% prairie strips had the highest pollinator abundance (mean = X), followed by fields with 10% strips (mean = X) and conventional fields without prairie strips (mean = X). This funding suggests that…"