Midterm

Due Thursday 11 May by 11:59 PM

This midterm is open note, open internet, open everything *except* open communication. You will be demonstrating your ability to complete the required tasks on your own with the materials available to you.

Set up

Create a *private* GitHub repository for all the data and code associated with this midterm. Name this repo "ENVS-193DS_midterm_lastname-firstname" (fill in with your own name). Organize your files using the **here** package. Commit and push changes as you are working with at least 5 commits, and make sure that your final changes are pushed before submitting. Add me (GitHub username: an-bui) as a collaborator *before you submit*.

Problem 1. Test choice, assumptions, and communication

You're the manager of a reserve that burned in a major wildfire a year ago. Plants have regrown in the burned areas, but things still don't look quite right. You wonder if the phosphorus content (expressed in parts per million, ppm) has something to do with how plants are (or are not) returning to burned areas, and want to know if the mean soil phosphorus concentration in burned areas is equal to the mean soil phosphorus concentration in unburned areas.

a. State the null and alternative hypotheses in statistical and biological terms for this question.
b. What kind of statistical test would you use to test the null hypothesis?
c. What are the assumptions you would have to meet to use this test?
d. If your data did not meet the assumptions of this test, what could you do?
e. You collect random, independent samples from burned and unburned areas in the reserve (n = 34 for both areas), enter your data, and wrangle it. You then check your assumptions: the variances are equal between burned and unburned areas, and the data are normally distributed. You run your statistical test and get the following output in R:

```
data:  p_ppm by treatment

t = 2.5144, df = 66, p-value = 0.01437

alternative hypothesis: true difference in means between group burned and group
unburned is not equal to 0

95 percent confidence interval:

 0.08846159 0.77057443

sample estimates:

  mean in group burned mean in group unburned

             1.2617051           0.8321871
```

You also get the following output from **cohen.d()**:

```
Cohen's d

d estimate: 0.6098373 (medium)
95 percent confidence interval:
       lower  upper
0.1144717 1.1052030
```

> This is very exciting! You want to report your results to the other managers at the reserve in writing. In 2-3 sentences, communicate your results including your: null hypothesis, the statistical test, sample size, α, test statistic, and p-value.

After reporting your results, one of the other managers reveals that there is soil phosphorus data from the burned areas from before the fire. You now want to know if the mean soil phosphorus content from before the fire is different from the mean soil phosphorus content after the fire, only using data from the burned areas.

f. State the null and alternative hypotheses in biological terms for this question.
g. What kind of statistical test would you use to test the null hypothesis?

Problem 2. Reproducible code

Read Fitch and Vaidya, "Roads pose a significant barrier to bee movement, mediated by road size, traffic and bee identity", *Journal of Applied Ecology* (link) **and** download the data and code from the Open Research section.

Understanding research context

a. In 1-2 sentences, summarize
    i. Pathways by which roads might affect plant pollination
    ii. What pigment is a proxy for in this study
    iii. How flower position relative to road, bike path, and pedestrian path influences pigment deposition
    iv. What the authors recommend to allow pollinators to cross roads

Reproducible code

In this exercise, you are strongly encouraged to copy/paste the code provided by the authors on the OSF (Open Science Framework) portal. However, you will need to annotate - and in some cases edit - the code to demonstrate your understanding of how it works. Insert spaces, line breaks, etc. as you see fit.

b. Recreate Figure 1 *without significance codes and icons*

Problem 3. Using new data

Choose one of the two following questions:

1. Is there a difference in median coyote (determined using fragment analysis) scat length between grassland and shrubland sites in the Chihuahuan Desert? (link to data)
2. Is there a difference in median snout-to-vent length between tiger whiptails and common side-blotched lizards near Gateway airport in Mesa, Arizona? (link to data)

After you have chosen a question:

Explore the data

a. Create an exploratory data visualization. Write an accompanying caption with the correct formatting and all relevant information for a reader to understand your visualization.
b. Create a visualization of the missing data in your data set.

Use the appropriate statistical test

c. Choose a statistical test that is appropriate for answering your question. In 1-2 sentences, describe why your test is appropriate for your question and your data.
d. State your null and alternative hypothesis in biological terms.
e. Run your test with any assumption checks as needed.

Communicate about statistical results

f. In 2-3 sentences, describe your results including your: null hypothesis, the statistical test, sample size, α, test statistic, and p-value.

Problem 4. Data visualization

Communicating clearly about data is a core component of data science, especially within environmental science. However, learning how to use the data communication tools means doing things badly first. Using the North Temperate Lakes Ice Cover data set from **{lterdatasampler}** (link) with year as the x-axis and ice duration as the y-axis, create two plots:

a. An ugly plot, with custom (i.e. your own)
    i. Colors for each lake
    ii. X- and y-axis labels, and a title (not meaningful)
    iii. One of the built in themes (link to reference)
    iv. 20 different font, color, width, line type, or other adjustments to arguments in the **theme()** function for the 1) axes, 2) legends, 3) plot, 4) panel, and/or 5) strips (with a comment accompanying each argument describing what it controls in the output)
b. A finalized plot clearly communicating that ice duration (in days) has declined through time at Lakes Mendota and Monona, with
    i. Colors for each lake (different ones from part a)
    ii. X- and y-axis labels, and a title (meaningful)
    iii. One of the built in themes (a more clean-looking one than the one from part a)
    iv. The legend within the panel area
    v. A caption with all relevant information for a reader to understand your visualization

Checklist

Your midterm should include

- ☐ Your name, the date, a title
- ☐ Written responses to problem 1 parts a-g
- ☐ Written responses to problem 2 part a
- ☐ Annotated code and output for problem 2 part b
- ☐ Annotated code, output, and written responses for problem 3 (parts as appropriate)
- ☐ Annotated code, output, and written responses for problem 4 (parts as appropriate)

- [ ] All code *without messages or warnings*

Additionally, it should be

- [ ] Submitted as a single PDF
- [ ] Rendered/knitted from a Quarto Markdown/RMarkdown file
- [ ] On GitHub in a private repo with the source code and associated data, with at least 5 separate commits/pushes to the repo and an-bui as a collaborator