STAT 503 – Statistical Methods for Biology

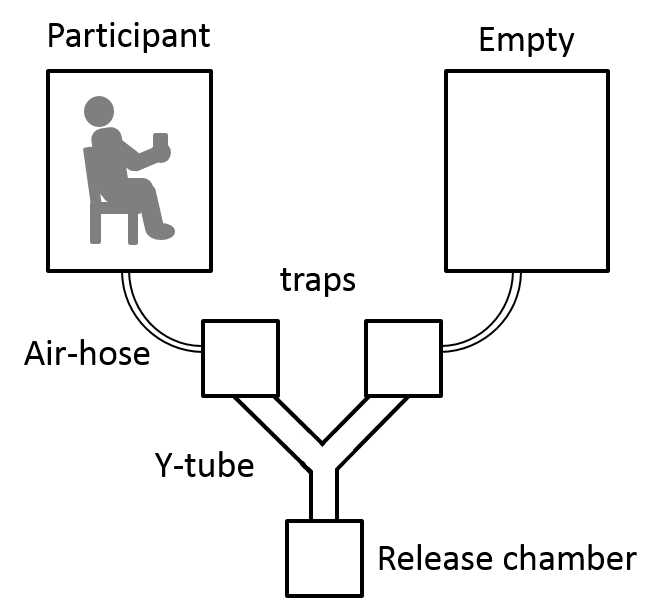
Homework 6

30 Points (34 available). Due at 11:59 PM on Thursday, July 28, 2020

**Please use complete sentences unless the question is marked with an asterisk (\*).**

A study by Lefèvre *et al*. (2010, *PLoSONE* 5: e9546) investigated whether alcohol consumption changed peoples' attractiveness to the mosquito, *Anopheles gambiae*, which is the primary vector for malaria in Africa. A total of 43 human participants were tested, and the attractiveness of each participant was measured twice (see methods below). The first measurement collected a baseline attractiveness for each participant. Then each participant drank a randomly assigned beverage. Twenty-five people drank a liter of beer, and the other 18 drank a liter of water. After 15 minutes, attractiveness was measured a second time. All 43 participants drank only water for 24 hours prior to the baseline measurement.

To measure a person's attractiveness to *A. gambiae*, a vial of 50 mosquitos was released into the Y-tube apparatus shown in Figure 1. Participants were seated in a small tent, which was connected by an air hose to one of two mosquito traps. This allowed the mosquitos to detect and respond to olfactory cues from the participants, without exposing the participants to the mosquitos. The proportion of mosquitos that flew toward the participant was recorded.

Figure 1: Diagram of the Y-tube oldfactometer used by Lefèvre et al. (2010) to study the effects of alcohol consumption on attractiveness to mosquitos. The participant's placement was randomized in each trial.

You can find the Lefèvre data on the website in the file **lefevre\_2010\_mosquitos\_and\_beer.csv**. Each row in the data represents a single participant. The first column is an id code for the person. The second column is a factor that indicates the subject’s treatment-level for drink (beer vs. water). The third column gives the proportion of mosquitos that flew in the participants' direction in the baseline measurement, and the fourth column gives the proportion after drinking. Column five is the change from baseline after drinking (i.e., afterDrink – beforeDrink).

**IMPORTANT:** Make sure you understand how these data are organized before you begin this assignment. The supplemental file found with the assignment can help.

1. In this question, you will conduct two separate analyses, one using the data for water drinkers, and one using data for beer drinkers. Most of the steps below are the same for both drink types, so you will only do them once.
   1. [1 point] Formally state the research question(s) for these analyses. Your question(s) should clearly identify (i) the sample unit, (ii) the response variable, and (iii) the explanatory variable (hint: the explanatory variable here is not the type of beverage!).
      1. **Does alcohol consumption by humans change their attractiveness toward mosquitos?**
   2. \*[1 point] Is the study design for these analyses paired or unpaired?
      1. The study design is unpaired because you are only testing each person once with a single level of alcohol. If you repeated the experiment two days later and had the participants simply switch and record the results, you could have a paired study design.

Paired experiments are good for reducing random variability. For example, certain people might just be less or more attractive to mosquitos which may throw some noise into your study

* 1. \*[1 point] State the significance level you intend to use for these analyses.
     1. α = 0.05
  2. [2 points] Based on your research question, identify your specific (i) goal and (ii) objectives (recall that goals state general, scientific aims, while objectives identify specific statistical aims; see Lecture 5.2, slides 8-11). Assume that we will use both estimation and hypothesis testing. The same goal and objectives will apply to both drink types; you do **not** need to identify separate goals and objectives for water versus beer. Your answer here should match your answer in 1c and your first choice of methods in 1e.
     1. Goal: To identify whether the effect of water vs the effect of alcohol makes the change of mosquito attractiveness into two different distributions
     2. Objective: To find statistically significant difference in the change value between water drinkers and beer drinkers.

* 1. Ideally, we would like to use a -test to analyze these data. However, we should always have a plan to follow if the assumptions for our preferred method cannot be met. Please explain how you plan to proceed if the data fail to meet -test's assumptions, and why you have chosen the backup methods that you have selected. Your plan needs to be reasonable; it does not need to be perfect.
     1. \*[1 point] Explain how estimation will proceed if the data fail to meet the requirements for Student’s *t*.
        1. Calculate confidence intervals

* + 1. \*[1 point] Explain how hypothesis testing will proceed if the data fail to meet the requirements for your preferred method.
       1. Cry
  1. [2 points] You may assume independence and random sampling for the purposes of this homework. Please (i) generate a plot (or a set of plots) to check for normality and to make certain that there are no outliers. (ii) Then state whether or not these requirements have been met. To receive full credit, **your statement must cite specific evidence in your plots** to support your conclusions.

Optionally, you may also present the results from a goodness-of-fit test if you feel that you need them to convince yourself that the assumptions have (or have not) been met. If you do this, please do not paste raw R output into your homework; just report the critical statistics.

Your plot(s) should represent the data for the water group and the beer group separately. This can be done by defining a color aesthetic for drink, or by faceting by drink (see Tutorials 4 and 6 to review plotting methods). The plot(s) that you choose should allow us to check for outliers and to check the assumptions of your chosen analysis (see 1e).

Insert your plot(s) here, with captions.

A screenshot of a cell phone

Description automatically generated

Figure 1. Beer change distribution

A screenshot of a cell phone

Description automatically generated

Figure 2. Water Change distribution

* 1. [2 points] State the null and alternative hypotheses to be tested in these analysis (the same pair of hypotheses will apply to both water and beer - only state them once).
     1. Null Hypothesis: There is no statistical difference between the two populations
     2. Alternate Hypothesis: The distributions are distinct

* 1. [2 points] Please perform the analysis separately for each drink type, and then write a paragraph reporting both sets of results and a biological conclusion. See the Guidelines for reporting statistical results in the Cheat sheets tab of the website for details on what to report. If you would like to, you may present the results in a formatted table. You may also refer to the graph in Question 3.

1. In this question, you will compare the effects of water to the effects of beer. Please use the same value of that you selected in question 1c.
   1. \*[1 point] Is the study design for this analysis paired or unpaired?
      1. Unpaired
   2. [1 points] State a formal research question for this analysis. Be sure to clearly identify (i) the sample unit, (ii) the response variable, and (iii) the explanatory variable.
   3. [2 points] Based on your research question, identify your specific (i) goal and (ii) objectives. Assume that we will use both estimation and hypothesis testing.

* 1. [1.5 points] Again, we would like to use a *t*-test if we can, but we need a backup plan. Please (i) state the particular type of *t*-test that you would like to use. Then identify (ii) a backup method for estimation, and (iii) a backup method for hypothesis testing.
  2. [1 point] Referring back to your previous results in 1f for evidence, identify the method that you will actually use, and justify your selection. You may assume independence and random sampling. If necessary, you may present additional evidence to check any assumptions that cannot be checked using the previous results.
  3. [2 points] State the null and alternative hypotheses to be tested in this analysis.

* 1. [2 points] Please report the results of your analysis and a biological conclusion. Include your point estimate, confidence interval, sample size, test statistic (with degrees of freedom if appropriate), and P-value. You may also refer to the graph in Question 3.

1. [2 points] Present a graph to illustrate the results of your analysis. There are many ways to present these data, and you are free to select whatever method you prefer, as long as your graph clearly illustrates your biological conclusions.
   1. Code to generate these graphs is in part 7

A screenshot of text

Description automatically generated

1. [2 points] Suppose that your friend (who has not taken this class) shows you the results from question 1 using and concludes that the effect of drinking beer on attractiveness to mosquitos is larger than the effect of drinking water. Please explain to your friend (a) why this conclusion is justified in this particular analysis, and (b) why they should not rely on this kind of comparison as a general rule (i.e., why is it better to use the approach in question 2?).
   1. The conclusion is justified; however, the independent t-test (from question 2) is better to use because the comparison of two independent groups would be more accurate.
2. [2 points] Suppose that we would like to conduct an experiment similar to question 2 with *Aedes aegypti*, the primary vector of several diseases, including yellow fever, dengue, chikungunya, and Zika.
   1. The average of the sample standard deviations of the beer and water groups in the *A. gambii* study is 0.15. Taking 0.15 as our assumed population standard deviation, use the power.t.test() function to determine the total number of people that we would we need to include in our study. We would like to achieve a power 0.8 to detect effect sizes 0.1 (that is, 10%) with .
      1. 50
   2. The power.t.test() function assumes a balanced design. Would using an unbalanced design increase or decrease the required sample size?
      1. You would need an increased sample size to prove that one side isn’t just simply being outweighed/offset by the probability that one side has greater values by chance.
3. Human reaction time can be measured by the following procedure (Google "measure reaction time with a ruler" for an illustration):
   1. Hold a ruler vertically by its top end, with zero at the bottom end.
   2. Have the subject place their hand so that their thumb and forefinger are located approximately 1 cm on either side of the ruler, with the bottom of the ruler between their thumb and forefinger.
   3. Instruct the subject to catch the ruler between their thumb and forefinger when it is dropped.
   4. Wait a randomly determined (but small) number of seconds, and then drop the ruler.
   5. Record the number at the top of the subjects' thumb where they caught it. This number is an index of the subject's reaction time (in this usage, an *index* is a variable that correlates closely with the variable you are interested in, but is easier to measure). Optionally, you can then use a gravitational acceleration of 9.8 m/s2 to calculate the subject's actual reaction time.

Suppose that we want to know whether caffeine consumption speeds up reaction time, and we suspect that men and women might be affected differently. Assume that you can recruit a random selection of 50 adults to participate in the study, that all of the participants are willing to drink a caffeinated beverage, and that you can measure each subject a maximum of 2 times. Think about how you would design this study, and then briefly outline the steps that you would take to achieve each goal below (try to identify at least two steps for each goal).

1. [1.5 points] Minimize potential bias.
   1. Repeat the test several times
   2. Make Sure that the dropper of the ruler has no way of providing any downward force. He must simply be dropping it. Some mechanism of ensuring this wouldn’t be too difficult
2. [1.5 points] Maximize precision.
3. Meter-stick with smaller increments delineated on it.
4. Get a baseline measurement without caffeine and make sure no caffeine is drunk within 24 hours of the study
5. Measure Everyone with a certain level of caffeine and then double the dose. Take measurements with both.
6. [1.5 points] Paste the PYTHON code for your analysis in Questions 2-4 here. Please use comments to clearly label each section.
7. import pandas as pd
8. # Note: Change = afterDrink - beforeDrink
9. data\_frame = pd.read\_csv("lefevre\_2010\_moqsuitos\_and\_beer.csv")
10. data\_frame = data\_frame.drop(columns = ["id"])
11. data\_frame
12. # Data frame for just the beer drinkers
13. beer\_df = data\_frame[data\_frame.drink=="beer"]
14. beer\_df
15. # Data Frame for just the water drinkers
16. h2o\_df = data\_frame[data\_frame.drink=="water"]
17. h2o\_df
18. from dataprep.eda import plot
19. # Make a box plot of the data and a line plot of the data - These are the best way to get a good picture of the data
20. box\_plot\_line\_plot = plot(data\_frame, x="drink", y="change")
21. # make figures of each separately
22. plot(h2o\_df, "change")
23. # TIME FOR A T-TEST!
24. import numpy as np
25. from scipy import stats
26. # Numpy Array of the CHANGES of the two groups
27. beer\_change\_array=beer\_df.iloc[:18,4].values
28. beer\_change\_array
29. len(beer\_change\_array)
30. h2o\_change\_array=h2o\_df.iloc[:,4].values
31. h2o\_change\_array
32. len(h2o\_change\_array)
33. # Calculate the variance to get the standard deviation
34. var\_h2o = h2o\_change\_array.var(ddof=1)
35. var\_beer = beer\_change\_array.var(ddof=1)
36. # Standard Deviation
37. std\_dev = np.sqrt((var\_h2o + var\_beer)/2)
38. ## Calculate the t-statistics
39. t = (h2o\_change\_array.mean() - beer\_change\_array.mean())/(std\_dev\*np.sqrt(2/25))
40. t
41. #Degrees of freedom / not the be confused with df for data\_frame
42. df = 2\*18 - 2
43. #p-value after comparison with the t
44. p = stats.t.cdf(t,df=df)
45. p