NAME:

STAT 503 – Statistical Methods for Biology

Homework 1

24 Points (24 available). Due at 11:59 PM on Wednesday, Sept 2, 2020

1. In your own words, please briefly explain the difference between:
   1. [2 points] A statistical population and a biological population.
      1. A statistical population is the entirety of data points / individuals that you want to draw a conclusion about. For instance, if I measured the height of 200 males in order to find the average height of males, I am trying to draw conclusions about height on the entire male population.
      2. The biological definition of “population” refers to a group of individuals of the same species, that reside in a proximity that allows them to intermate with each other.
   2. [2 points] A statistical population and a statistical sample.
      1. As in my example above, while I would like to draw conclusions about the entirety of a statistical population (all males on the planet), that is simply unfeasible. I will take 200 individuals that I think make up a representative sampling, and that will be what I use to draw conclusions from that I will then generalize to the entire statistical population.
   3. [2 points] A statistical sample and a biological sample.
      1. A statistical sample refers to the group of all of the individuals of a statistical population that you are observing / measuring / analyzing, while a biological sample refers an individual unit.
      2. Example: If I am trying to find bacteriophages in dirt. Each scoop of dirt that I examine would be a “biological sample”, but when I analyze the entire experiment statistically to say, “what proportion of dirt samples on earth have bacteriophages?”, all of the dirt samples that I analyzed would therefore be my statistical sample. The statistical population that I would be drawing conclusions about would be “all the dirt on earth” … which is impossible to analyze the entirety of …
   4. [2 points] A parameter and a statistic
      1. Essentially the same, except on different scales.
         1. Statistic refers to the variable value of the sample while a parameter refers to the theoretical variable value for the entire population.
2. Whether reading the literature or analyzing your own data, it is important to be able identify the sample units, variables, and their roles in a scientific problem. For each of the following scenarios, please identify (i) the sample unit, (ii) the response variable, and (iii) the explanatory variable. Complete sentences are not required. If there is no explanatory variable write “none” for (iii).
3. [1.5 points] Deposits of a protein known as amyloid- (A) in the brain are closely associated with Alzheimer's disease, but can only be observed in an autopsy. This means that they cannot be used to diagnose Alzheimer's in living patients. Hoping to develop a diagnostic test that would allow earlier treatment of the disease, DeMattos et al. (2002, *Science* 295:2264-2267) injected 49 mice with a monoclonal antibody which causes A that is present in the brian to be released into the blood, where it can be detected. Then they compared the amount of A that accumulated in the blood over a 24-hour period after injection to the A load in the mouse’s brain.
   1. Sample Unit: Each Mouse that is injected
   2. Response Variable: Amyloid-Beta level that is measured in blood
   3. Explanatory Variable: “Amount of” or “Presence of” Amyloid-Beta deposits in the brain of the mouse
4. [1.5 points] Blind forms of the cavefish *Astyanax mexicanus* spend much less time sleeping than their non-blind, surface-living conspecifics. Duboué et al. (2011, *Current Biology* 21:671-676) measured the number of minutes that each of 23 blind cavefish spent sleeping, out of a 24-hour day.
   1. Sample Unit: Each blind cavefish
   2. Response Variable: Number of minutes spent sleeping out of a 24-hour day
   3. Explanatory Variable: None
5. [1.5 points] Ten people volunteered for a study to investigate the relationship between sleep deprivation and cortisol levels (cortisol is a stress hormone). Half of the subjects were randomly assigned to sleep for 8 hours, and the other half slept for 4 hours. In each volunteer, blood samples were collected at 6:00 am on two successive days, before the subject ate breakfast, and the cortisol level on the first day was subtracted from the level on the second day. The change in cortisol levels over the 24-hour period was compared between the two groups.
   1. Sample Unit: Each volunteer
   2. Response Variable: Cortisol level in blood
   3. Explanatory Variable: Number of hours slept
6. [1.5 points] Glucocorticoid hormones are secreted by the adrenal glands in response to acute (short-term) stress and are eventually excreted in urine and feces. Wildlife biologists have taken advantage of this fact to develop sensitive, non-invasive methods to evaluate stress in wild animals. For instance, Creel et al. (2002, *Conservation Biology* 16: 809-814) collected 125 fecal specimens from 34 radio-collared elk (*Cervus elephas*) in Yellowstone National Park, and compared the glucocorticoid levels in these specimens to snowmobile activity in the elk’s location during the preceding day (in elk, it takes 12-24 hours for newly secreted glucocorticoids to appear in feces). Snowmobile activity was classified into four levels: 0, 1-300, 301-500, or > 500 snowmobile visits per day.
   1. Sample Unit: Each fecal specimen
   2. Response Variable: Glucocorticoid hormone levels in the fecal specimen
   3. Explanatory Variable: Level of stress induced or number of snowmobile visits
7. [2 points] For each of the four studies described in Question 2, please indicate whether the study can show a statistical association, a causal relationship, both, or neither, and then briefly explain your answer in terms of the methodology that I have described. You can refer to the studies as (a), (b), (c), and (d).
   1. Statistical Association because you are simply measuring levels and are not causing any changes to an explanatory variable.
   2. Neither. No Association or correlation between variables is being measured. Only observations of a single variable.
   3. Observations are being made; however, a mechanism is unknown so you cannot infer causation. Only a statistical association can be shown.
   4. Only Statistical Association because the molecular mechanism is unknown.
8. Suppose that to estimate the density of garlic mustard (an invasive herb) in a forest, we randomly select a number of 1 m2 quadrats (patches of ground) and count the number of garlic mustard plants in each quadrat. How will the differences in methodology described below affect the (i) accuracy and (ii) precision of our estimates? Remember that accuracy and precision refer to the results that we *would* see if we were to repeat the same study many times.
   1. [1 point] We increase the number of samples from 30 to 50.
      1. Accuracy will likely not change as long as your additional 20 sample units are actually from the same population
      2. Precision Increases, because uncertainty decreases
   2. [1 point] Instead of placing quadrats randomly on a map, we randomly select quadrats within a distance of 50 m from a trail that runs through the forest, to save time. Locations farther than 50 m from a trail are not sampled.
      1. Your population might not vary, but it is all possible that the garlic mustard plants within 50m of the trail are of a specific subpopulation.
         1. This would mean that you accuracy would go down but says nothing about the precision.
9. [1 point] One of our main goals in science is to produce repeatable results. If someone independently replicates our study, then they should be able to reproduce our finding. Please explain how this idea is related to the concepts of sampling error, uncertainty, and precision (hint: how does sampling error affect reproducibility?).
   1. By being able to reproduce a study, you are confirming that there was likely low sampling error in the selection of your sample units.
10. [1 point] Lectures 1.6 and 1.7 discuss randomization in two different contexts: random selection of sample units from the population, and random assignment of treatments to the individual sample units within a study. Please explain how these two forms of randomization differ from each other. What goals do they serve?
    1. Random Selection of Sample Units from the population:
       1. This is to make sure that your statistical sample is representative of the statistical population that you are looking to make conclusions about. By selecting sample units randomly and in an unbiased manner, you have the highest chance of getting a sample that you can accurately draw conclusions from.
    2. Random Assignment of Treatments to the Individual Sample Units Within the study.
       1. Eliminates confounding variables that could cause a change in the response / dependent variable
11. [4 points] This exercise will begin building your coding skills by asking you to translate and debug an unfamiliar script. The code below has 11 statements, labeled A-K. It also contains several errors. Please **copy and paste** the code into a new script in VS Code (do **not** retype it yourself – you might fix a bug without noticing!). Then use the resources that are available to you to fill in the table on the next page. In the first column, either indicate that the statement is okay (it runs without an error), or briefly explain what the mistake was and how you fixed it. In the second column, briefly explain what the statement does (a *statement* includes everything up to the next labeled line of code; e.g., statement D includes two lines).
12. # Stat503 - HW #1
13. # Written in Python 3.7.7
14. # Alexandru Ivan
15. # STID: 0027908411
16. # %%
17. # Import the pandas library
18. import pandas as pd
19. # %%
20. # Seeds Data Array (How many seeds germinated)
21. n\_seeds\_germinated = [11, 9, 20, 11, 23, 16, 16, 25, 24, 27, 27, 30]
22. # %%
23. # Temperature Data Array
24. soil\_temp\_celcius = [15.8, 15.5, 16.5, 15.4, 16.2, 16.3, 16.6, 17.2, 17.5, 17.4, 18.3, 21.0]
25. # %%
26. # Make a dataframe of the two arrays above
27. dataframe = pd.DataFrame(list(zip(n\_seeds\_germinated, soil\_temp\_celcius)),
28. columns =['Number\_Germinated', 'Soil\_Temperature\_Celcius'])
29. dataframe
30. dataframe.mean()
31. # %%
32. # Import graphing tool of your choice, there's like 10 good ones
33. import matplotlib.pyplot as plt
34. # %%
35. # A histogram of number of seeds germinated at various temperatures
36. plt.hist(n\_seeds\_germinated, bins = 5)
37. plt.xlabel("Number of Seeds that Germinated")
38. plt.ylabel("Frequency")
39. # %%
40. # A histogram of how many samples were tested at various bins of temperatures
41. plt.hist(soil\_temp\_celcius, data=dataframe, bins=5)
42. plt.xlabel("Temperature In Celcius")
43. plt.ylabel("Number of Samples")
44. # %%
45. # Scatterplot of the different samples (number germinated in each sample) plotted against the temperature
46. plt.scatter(x=dataframe.Soil\_Temperature\_Celcius, y=dataframe.Number\_Germinated)
47. plt.xlabel("Temperature In Celcius")
48. plt.ylabel("Number of Seeds that Germinated")