STAT 503 – Statistical Methods for Biology

Homework 4

30 Points (32 available). Due at 11:59 PM on Tuesday, July 14, 2020

**Please use complete sentences unless the question is marked with an asterisk (\*). Round answers to 3 decimal places unless otherwise specified.**

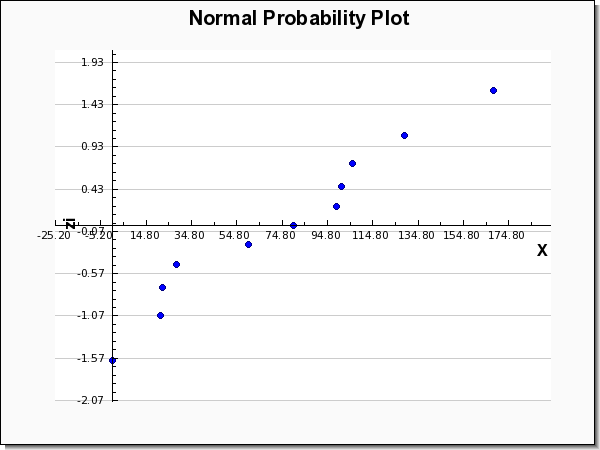
1. In Homework 2, you explored a data set from Norton et al. (2011; *Journal of Neuroscience* 31:13796-13807) on differences in display time between wild type zebrafish (*Danio rerio*) and individuals that possessed a specific mutation in the *Fgfr1a* gene. The data showed a clear difference between the two genotypes; in a boxplot, the display times for approximately 75% of the wild type fish fell below the 25th percentile for the *Spd* mutants.

Although the graph and summary statistics from Homework 2 provide useful descriptions of the data, they only represent the first step in our analysis. Typically, we would also want to estimate the location of each group’s distribution, as well as the difference between the locations (this difference is known as the *effect size*, and we will see how to estimate it in later homeworks). We might also want to test a hypothesis about the difference between the groups.

In this question you will walk through the process of estimating , the true mean display time for wild type fish (the subscript “wt” is for *wild type*). You may use R as a calculator, and to find the quantiles of the distribution in question 1e, but please show your work in these calculations. In addition, for steps a – d, please explain in your own words what information the calculation step contributes to the estimation of the confidence interval.

* 1. [2 points] Calculate the point estimate of the mean for the wild type fish and explain what it contributes to the interval estimate.
     1. This point estimate made from the sample mean will be used to represent the population mean, is estimated from the sample data. It will be used as an estimate for the peak of the distribution curve - which should be the center point for the confidence interval estimate.
     2. (0 + 21 + 22 + 28 + 60 + 80 + 99 + 101 + 106 + 129 + 168) / 11 =
        1. 74 seconds
  2. [3 points] Calculate the standard error of the mean for the wild type fish, and explain what it contributes to the interval estimate.
     1. Standard Error of the Mean is generally referred to as the standard deviation. It tells us how far the average distance from the sample mean is when comparing the sample mean to each individual data point.
     2. S = [1/(n) \* ∑ (xi – x(average))] = [ 1/10 \* [(0 – 67.83) ^2 + (21 – 67.83)^2 … + (168 – 67.83)^2 ] ^0.5
        1. 52.5 / (11)^0.5 = 15.83 seconds
  3. [2 points] Identify the correct number of degrees of freedom for the distribution, , and explain what this number represents (it may be helpful to review the notes about the sample variance).
     1. Because there is only 1 sample with n = 11 fish for the wild type distribution. There are n-1 degrees of freedom = 11 -1 = 10 degrees of freedom. This tells us that if we know 10 of the values, we know the 11th value given that we know variance or standard deviation of the mean.
  4. [2 points] Find the quantiles of the Student's distribution with degrees of freedom, , that correspond to the 95% confidence interval. You can do this in R if you want to, or you can use the -table found on the website. What do these quantiles contribute to the interval estimate?
     1. The value of t for the 95% interval for 10 degrees of freedom is 2.23 (From the table)
     2. This means that the limits of the 95% interval are as follows:
        1. Lower (2.5% Quantile): 74 - 2.23\*15.83 = 38. 7 seconds
        2. Upper (97.5% Quantile): 74 + 2.23\*15.83 = 109.3 seconds
  5. [2 points] Calculate the 95% confidence interval for .
     1. (38.7, 109.3)

1. [3 points] Please explain how the interval calculated in question 1e should be interpreted (hint: imagine we repeat the study many times…).
   1. If you were to repeat this study many times, each sample would have a 95% confidence interval. If you look at all of the 95% CI of the samples, you would expect roughly 5% of them to not include the true population mean.
2. [3 points] List the conditions that must be satisfied for the confidence interval that you calculated in question 1 to be valid?
   1. The sample should be of a subpopulation that is roughly representative of the population.
      1. The data must be randomly sampled and unbiased
3. [3 points] Present a quantile plot to check the normality of the data for question 1 (remember that you only want the data for the wild type fish; filter() will be helpful). How confident are you that the data are normally distributed? Please briefly explain your reasoning.
   1. Q – Q Plot



1. \*[3 points] Use R to calculate the 95% Student's confidence interval for each of the two genotypes (Wt and *Spd*), and then present your results as a table containing the following columns: sample size (*n*) per group, group mean (mean), standard error of the mean (SE), lower 95% confidence limit (lower), and the upper 95% confidence limit (upper). You are welcome to reuse code from previous homeworks, but this is not required. Note that you can use your result here to double check your answer in question 1.
2. In this exercise, you will repeat the confidence interval calculation for both groups, using bootstrap resampling instead of the Student's distribution. This is easiest to code as a two-part process: first, we will write a new function that returns the selected quantiles of the bootstrap distribution for the mean. Then, we will use group\_by() and do() to run the analysis.
   1. \*[1 point] Present a table similar to the table from question 5, showing the bootstrap standard error and bootstrap 95% confidence intervals for the mean display time in each genotype. Please use 2000 resamples. You will need to modify the function to get the mean and sample size for each group.
   2. [1 point] If you rerun the function several times, the answers will change. Why?
   3. [1 point] Compare your bootstrap confidence interval to the interval that you calculated using the Student's t distribution in question 5. How do the two sets of intervals differ? Why is this difference is occurring ?
   4. [1 point] Suppose that we had sampled 500 fish from each genotype. How would the intervals compared in question 6c change? Your answer should address how the individual intervals would change, how the comparison would change, and why each of these changes occurs.
3. *Transitive inference* is the form of logical reasoning that allows us to conclude that, if and , then must be . Recently, Tibbetts *et al.* (2019, *Biology Letters* 15: 20190015) used the following setup to determine whether paper wasps (*Polistes dominula*) can use this kind of logic:
   1. An individual wasp was placed in the middle of a tube that had colored plates at either end. The wasp was then trained to prefer one color over another (it received a minor shock if it walked toward the "wrong" color, but not if it walked toward the "right" color).
   2. In this manner, each wasp was trained over a 4-day period to select color A over color B, B over C, C over D, and D over E.
   3. On day 5, the wasp was tested with a color combination that it had never previously seen: A versus E. No shocks were administered during the testing phase of the study.
   4. A total of 23 wasps were tested, and for each wasp, the authors recorded the number of correct choices out of 10 A vs E trials.

The data for these tests may be found in tibbetts\_et\_al\_2019\_wasp\_ti.csv, on the website. The number of correct trials per wasp is in the Correct column. We would like to model these data using a binomial distribution.

* 1. \*[1 point] Let represent the number of correct choices made by the *i*th wasp. What is the value of the parameter ?
     1. The parameter n stands for the number of trials that were run for each wasp = 10 trails
  2. [1 point] For the binomial model to be valid, what assumption(s) do we need to make about the individual wasps' behavior and probabilities of success?
     1. We have to make assumptions that all wasps are equally capable, trained in identical manners and are given the same 10 trials.
     2. For fun, we can also make sure none of the wasps see the trials ahead of time while other wasps are running the test because we can’t be sure that they don’t comprehend. They may be smarter than we think
  3. [2 points] Use this data and the method of maximum likelihood to estimate a 95% confidence interval for the probability that wasps will pick the right color. You may either code this yourself or use the **bbmle** package (or you can do both and compare the results!).
     1. Binomial Distribution is the same as a normal distribution except instead of continuous values, it is used for discrete values. If there are enough discrete values, it may look very similar to a normal distribution

Write 1-2 sentence to formally report your point estimate and confidence interval. Based on this result, would you conclude that wasps use transitive inference? Why, or why not?

**5.729821105468706 6.391304347826087 7.052787590183468**

**This is the direct output from my code.**

**This represents the lower limit of the 95% CI, the point estimate, and the upper limit of the 95% CI, respectively**

**From this, I conclude that wasps can use transitive inference, because overall they guess correctly substantially over 50% of the time. However, this is based on only a single species of wasp, with only 23 as the sample size and only 10 runs each. Larger numbers would namke this a lot more certain. I would also try with 3, 4, 5, 6, 7 different values to compare and see if they can succeed with simpler and more difficult problems of transitive inference. I could see wasps doing better if they only compared A, B, and C, instead of A through E**

1. \*[1 point] Paste your code for questions 4-7 here. Please make sure your code is organized and the code for each question is clearly labeled.

BELOW IS THE CODE FOR #7

1. # Stat503 - HW #4
2. # Written in Python 3.7
3. # Alexandru Ivan
4. # STID: 0027908411
5. # import some packages for analyzing dataset
6. import numpy as np
7. import scipy.stats
8. # A function that finds that lower limit of the 95% CI, the point estimate, and the uppower limit of the 95% CI
9. def mean\_confidence\_interval(data, confidence=0.95):
10. a = np.array(data)
11. n = len(a)
12. m, se = np.mean(a), scipy.stats.sem(a)
13. h = se \* scipy.stats.t.ppf((1 + confidence) / 2., n-1)
14. return m-h, m, m+h
15. # Import the pandas library and seaborn for beautiful printing of plots
16. import pandas as pd
17. import matplotlib.pyplot as plt
18. import seaborn as sns
19. # Load up the .csv file that we are using
20. data\_frame = pd.read\_csv("tibbetts\_et\_al\_2019\_wasp\_ti.csv")
21. data\_frame.head(23)
22. # Make the beautiful binomial plot... kinda. I actually made a histogram that has so many bins that it just looks like a binomial plot
23. # Hope this is ... acceptable
24. sns.distplot(data\_frame["Correct"], hist=True, kde=False)
25. plt.show()
26. # Return and print the results of the function at the top functioning on the wasp correct dataset
27. lower, mid, upper = mean\_confidence\_interval(data = data\_frame["Correct"])
28. print(lower, mid, upper)