STAT 503 – Statistical Methods for Biology

Homework 4

24 Points (28 available). Due before midnight on Monday, October 19, 2020

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

1. [1.5 points] Please identify a parametric distribution that we might reasonably use to model the response variable in each of the following scenarios, and briefly explain your reasons for selecting the chosen distribution. The goal here is to pick a distribution based on the characteristics of the variable to be modeled, before the data have been collected.
   1. A super-resolution optical microscope will be used to capture an image of the surface of a cell membrane. Then a grid will be laid over the image to divide it into 200 200 nm quadrats, and the number of visible cell-surface proteins will be counted in several randomly selected quadrats.
      1. **I would choose a normal distribution as the number of proteins on the cell surface membrane at any given point is random. Although some factors such as dimerization/polymerization may cause clumping together of proteins.**
      2. **But you could probably also use a Poisson distribution as technically these are “counts” of receptors.**
   2. Blood glucose concentrations will be measured in 30 healthy subjects after a 12 hour fast, in mg/dL.
      1. **I would choose a lognormal distribution as there are many factors at play in determining blood glucose level; however, it is still a biological process and therefore should be normally distributed to some extent.**
   3. Researchers plan to collect a random sample of 500 medical histories for patients at a hospital who underwent the same surgical procedure and then determine the number of those patients that had postoperative surgical site infections.
      1. **It sounds like just a normal distribution if it is based on existing data from that same hospital.**
2. [1.5 points] Most of you have heard that the average oral temperature reading for a healthy human adult is 98.6F (37C). This value is based on a paper published in 1868 by Carl Wunderlich, who also reported a standard deviation of 0.6F. Dr. Wunderlich's results have been questioned (Mackowiak et al. 1992, *Journal of the American Medical Association* 268:1578-1580), but we will accept them for the purposes of this question.
   1. \* What range of temperatures contains 95% of the most common temperature values for healthy adults?
      1. **(Mean – 2sd , mean +2sd) = (97.4 F, 99.8 F)**
   2. \* Please use either the pnorm() or qnorm() function in R to answer this question and report your code in question 8. Suppose that we collect oral body temperatures on 500 apparently healthy people, and 11 of them have temperatures > 99.8F. Should we be surprised by this result? Why, or why not?
      1. **11/500 is basically the same as the upper limit estimate of the 95% confidence interval according to the function.**

Norton et. al (2011, *Journal of Neuroscience* 31:13796-13807) examined the behavioral effects of a mutation in the *Fgfr1a* gene in zebrafish (*Danio rerio*; *Fgfr1a* stands for "fibroblast growth factor receptor 1a"). Their data describe the amount of time in seconds that each of 21 fish spent on aggressive territorial displays when they were presented with a mirror. There are two genotypes: 11 wild type (wt) fish, and 10 fish with the so-called Spd mutation. The full data are on the website in norton\_et\_al\_2011\_zebrafish.csv. The times for the Spd fish are:

(96, 97, 100, 127, 128, 156, 162, 170, 190, 195) seconds

1. This question will walk you through the process of estimating the mean display time for the Spd mutants in the Norton dataset, using Student’s *t* . **The objective in this question is to understand how the different parts of a confidence interval are put together, and why they are put together that way**. Please show your work for the calculations (you may use R as a calculator, and to find the quantiles of the distribution in question 3e). In addition, please explain in your own words what each value represents and contributes to the confidence interval. To do this, think about (1) what would cause the term to get bigger or smaller, and (2) how a change in the term would affect the confidence interval.
2. [2 points] Calculate the point estimate of the mean for the Spd group, , 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 and 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.
     2. (96 + 97 + 100 + 127 + 128 + 156 + 162 + 170 + 190 + 195)/10 = 142.1
3. [2 points] Calculate the standard error of the mean for the Spd group 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 \* [(96– 142.1) ^2 + (97 – 142.1)^2 … + (195 – 142.1)^2 ] ^0.5
4. [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 from section 2 in the lecture notes).
   * 1. Because there is only 1 sample with n = 10 fish for the spd distribution. There are n-1 degrees of freedom = 10 -1 = 9 degrees of freedom. This tells us that if we know 9 of the values, we know the 10th value given that we know variance or standard deviation of the mean.
5. [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 using the qt() function. For the sake of practice, I encourage you to also try it using 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 9 degrees of freedom is 2.262 (From the table)
     2. This means that the limits of the 95% interval are as follows:
        1. Lower (2.5% Quantile): 142.1 - 2.23\*(standard error from b)
        2. Upper (97.5% Quantile): 142.1 + 2.23\*(standard error from b)
6. [4 points] Calculate the 95% confidence interval for , and then explain the interpretation of this interval. What does it tell us?

Range = (142.1 - 2.23\*(standard error from b), 142.1 + 2.23\*(standard error from b))

This is where 95% of our values are within the distribution.

1. [2 points] In general, what conditions (i.e., assumptions) must be satisfied for a confidence interval like the one that you calculated in question 3 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
   2. The distribution should be roughly normally distributed -> Variance is independent of the mean
2. \*[2 points] Use Python 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).
   1. Function Code recorded in question 8.
3. In this exercise, you will repeat the confidence interval calculation for both groups, but this time, you will use 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 summarize() to run the analysis.

The following code defines the function for the first step. Copy it and paste it into your script, and then run it in the console to define the function. DO NOT MAKE ANY CHANGES TO IT, but you should take the time to try to understand how it works.

bootstrap\_mean <- function(

data, # a data frame

variable, # the name of a variable, in quotes

probs, # a vector of probabilities for the quantiles that you want

nboot # the number of bootstrap resamples to use

){

# extract the column named in the "variable" argument

# the double-square-bracket notation returns a vector

# instead of a one-column data frame

x <- data[[variable]]

# resample the variable nboot times with replacement

xBoot <- replicate(nboot, sample(x, replace = TRUE))

# get the mean for each bootstrap sample

xBarBoot <- colMeans(xBoot)

# find the standard error

se <- sd(xBarBoot)

# find the quantiles

qs <- quantile(xBarBoot, probs = probs)

# format the results so each quantile is a column in a data frame

# t() is the transpose function; it forces res to be row instead of a column

res <- c(SE = se, qs)

as.data.frame(t(res))

}

If I wanted to run this function on the entire dataset called myData, I would call:

bootstrap\_mean(myData, variable="myResponse", probs=c(0.25, 0.75), nboot=100)

The arguments define the dataset, the name of the variable that I want to estimate the mean for (in quotes!), the probability levels for the confidence interval, and the number of bootstrap samples to be used.

To separate data by the groups, you can use group\_by() and do(), like this:

myData %>% group\_by(myExplanatoryVariable) %>% do(

bootstrap\_mean(., variable="myResponse", probs=c(0.25, 0.75), nboot=100)

)

Notice that I have replaced the first argument with a period (.). This tells do() where to place the grouped data that it gets from group\_by(). To attach these results to a table wit other summary statistics generated by summarize(), I can use bind\_cols() (for “bind columns”):

groupedData <- myData %>% group\_by(myExplanatoryVariable)

mySum1 <- groupedData %>% summarize(n = n())

mySum2 <- groupedData %>% do(bootstrap\_mean(., “myResponse”) # default args

mySummary <- bind\_cols(muSum1, mySum2)

Use the bootstrap\_mean() function to answer the following questions:

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 time in each group. Use 2000 bootstrap samples, and round the results to 2 decimal places.
2. [1 point] Rerun the function several times, and the results (you do not need to report them). What happens to the standard error and confidence interval? Why?
   1. It varies because the sample data that you generate via bootstrapping is randomly chosen from the subset of existing values.
3. [1 point] Compare your bootstrap confidence interval to the interval that you calculated using the Student's t distribution in question 5. How are they different, and why this difference is occurring?

Same as part b

1. [1 point] Should we trust this bootstrap confidence interval? Why, or why not?
   1. Yeah, the answers aren’t too far from each other as to make them mutually exclusive. Bootstrap resampling is very good when you are confident in your small set of data.
2. [1 point] What would happen to the confidence interval in 6a if we had data on 500 fish instead of 21 fish? Would the larger sample size change your answer in 6d?
   1. Increasing the sample size decreases the width of confidence intervals, because it decreases the standard error as there are many values that will be closer to the mean.
3. See Lecture 4.6 for assistance on this question. In the Oppliger *et al.* (1996) dataset, 7 out of 35 birds in the control group were found to be infected with avian malaria.
   1. [1 point] Use the method of maximum likelihood to find a point estimate and 95% confidence interval for the baseline prevalence of malaria in this population (*prevalence* is the proportion of a population that has a disease). You may either code this yourself or use the **bbmle** package (or you can do both and compare the results!). Write 1-2 sentences reporting your results.
   2. [1 point] What assumptions do we need to meet for his interval to be valid (hint: what distribution are you using to model the population?)?
   3. [1 point] If we increase the sample size but the ratio of sick to total birds remains 1:5, the confidence interval will get narrower. Why does this happen?
4. \*[1 point] Paste your code here. Please make sure your code is organized and the code for each question is clearly labeled.
   1. QUESTION 2
5. from scipy.stats import norm
6. estimate = norm.cdf(98.8, 98.6, 0.6)
7. prob\_of\_temp\_above\_998 = 1 - estimate
8. sample\_temperature\_above\_998 = 11/500
9. print(prob\_of\_temp\_above\_998, sample\_temp\_above\_998)

Question 5

1. # A function that finds that lower limit of the 95% CI, the point estimate, and the uppower limit of the 95% CI
2. def mean\_confidence\_interval(data, confidence=0.95):
3. a = np.array(data)
4. n = len(a)
5. m, se = np.mean(a), scipy.stats.sem(a)
6. h = se \* scipy.stats.t.ppf((1 + confidence) / 2., n-1)
7. return m-h, m, m+h