Week 4 Problem Set: Estimation and maximum likelihood

**(30 pts)**

**PART I** **- Warm Up**

We will be using a recently published dataset on ancient rings and ribs that may have been used as early forms of money. These data come from the paper “The origins of money: Calculation of similarity indexes demonstrates the earliest development of commodity money in prehistoric Central Europe” by M.H.G. Kuijpers and C. Popa (PLoS ONE, January 20, 2021).

Download the data in the file “rings\_and\_ribs.csv”. (If you haven’t set your working directory, you will need to type in the full path)

> money<-read.csv(“rings\_and\_ribs.csv”,header=T)

(These data are the first tab of the worksheet included in the paper’s Supplement.)

Look at the column names:

> names(money)

The columns we will need to use for this problem set are the column $Weight, which lists the weight of each object in grams, the column $Type, which denotes whether each object was a Rib or a Ring, and the column $Country, which lists the country where the object was found.

We want to use these data to do two things 1: To summarize the data and 2: To make inference about the population from which this sample was drawn.

**Q1.** For now, let’s group all the locations together. Make a histogram of ring weight (COPY AND PASTE BELOW). (1 pt)

**Q2.** What is the FORMULA for the sample mean and standard deviation (in other words, what is the formula you would want to use if you wanted to estimate the population mean [the μ parameter assuming a normal distribution] and the population variance [if we assume a normal distribution] from a sample that represented a random subset of the entire population)? (1 pt)

**Q3.** What is the population about which we are trying to make inference (1 pt)?

**Q4.** Using R, what are the mean and the standard deviation of ring weight? (1 pt)

**Q5.** What is the formula for the standard error of the mean (heretofore s.e.)? (1 pt)

**Q6.** Describe the difference between the s.e. (of the mean) and the s.d. (1 pts)

**Q7.** Finish the sentence: The standard error is the standard deviation of \_\_\_\_\_\_\_\_\_\_\_\_. (1 pt)

**Q8.** Use the MASS package’s ‘fitdistr’ function to fit a normal distribution to the ring weight data. Do you get the roughly same answer as above? (1 pts)

**Q9.** Why might it not be valid to group the different locations when summarizing the data? (1 pts)

**Q10.** Using the R command ‘boxplot’, create a boxplot to compare ring weight across different countries. (Remember, use “?boxplot” to get more information about the parameter inputs to boxplot.) (COPY AND PASTE BELOW) (1 pt)

**Q11.** Make the case (in words and/or mathematically) that average ring weight in Germany is or is not statistically different from the average ring weight in Austria. (2 pts) (As we have not covered this material yet, I’m asking you to make an argument. I’m less interested in the right answer here than I am the basic processing of reasoning through the answer.)

**Q12.** So far, we’ve only been focused on ring weight. Let’s go back and make a histogram of rib weight. Why would testing a hypothesis about average rib weight be harder? (1 pt)

**PART II – Maximum Likelihood**

Assume an experiment in which the number of plants in 16 experimental plots is counted as: 8,0,3,3,4,3,2,3,3,2,4,2,3,1,4,1

We want to model the number of plants in each plot as being distributed according to a Poisson distribution.

**Q1.** Starting with the probability density function for the Poisson, manually derive the maximum likelihood estimate for λ, the parameter for the Poisson distribution. (5 pts)

**Q2.** Write an R function to calculate the negative log-likelihood for this data as described by the Poisson distribution. (5 pts)

**Q3.** Using your function for the negative log-likelihood, calculate the MLE for λ and the 95th percentile confidence interval. What would the 99th percentile confidence interval be? (4 pts) [Hint: The 95th percentile cut-off is (0.95) = 0.5\*qchisq(0.95,df=1)=1.92; in other words, the NLL would have to be >1.92 higher than the minimum to fall outside the 95th percentile confidence interval. Likewise, the 99th percentile cut-off is (0.99) =3.32.]

**Q4.** Plot the likelihood over a range of parameter values and plot the boundaries of the 95th and 99th percentile confidence interval. Remember: The confidence interval is a range of parameter values, it is NOT the likelihood values itself. (COPY AND PASTE PLOT) (3 pts) [Hint: Consider using the command ‘abline’ (in the “base” package) or ‘segments’ (found in the “graphics” package) to draw the boundaries of the confidence intervals.]