STAT*2040
Winter 2020

Data Analysis Assignment #2

This assignment has a deadline of Wednesday April 8 at 11:59 pm. You must submit one pdf document for each part of this assignment (4 pdfs in total). Submissions must be made to Crowdmark, using the personalized link that will be sent to your email address.

There are 4 parts to this assignment:

1. Data analysis and write-up of conclusions for a one-sample problem. (10 marks)

2. Data analysis and write-up of conclusions for a two-sample problem. (10 marks)

3. Data analysis and write-up of conclusions for another two-sample problem. (5 marks)

4. Reading parts of a few journal articles, and giving interpretations of the results of the statistical inference procedures used in the articles. (10 marks)

Each part is based on information from a published study, some with a U of G connection. The journal articles are available from the University of Guelph library website. For accessing off-campus, you must use the off-campus sign on (top right of the page) before proceeding to the journal article. One way to find the articles is to go to http://www.lib.uoguelph.ca, hover over Find and click on E-Journals, then search for the name of the journal. You can also search for the article title in Omni (on the library site).

This assignment is worth 16% of your final grade. You will be marked on: 1) Getting the proper R output and plots, 2) Validity of your statistical conclusions and interpretations, 3) Writing style (grammar and clear concise language count!), 4) Presentation. Note that you *must* use R to complete this assignmen. My "Intro to R" document is available on the Courselink site.

You will have to do some thinking in this assignment. I am not going to tell you exactly what to do, and I would be negligent in my duties as a professor if I were to do so. You are most welcome to ask me questions, and post questions or comments on the discussion board (but refrain from posting specific answers or code that could simply be copied). If you're holding up your end of the bargain, and giving these questions an honest go, then I'm very willing to help when you have questions or concerns. I am not always looking for one specific method of analysis – for some of these questions, there is more than one path to perfect marks.

1 Part I: Birth weight of African elephants born in captivity

Dale (2010) investigated various characteristics of newborn African elephants born in captivity. Here we will look at the birth weight of 27 male African elephants. The file s2040_W20_birthweight contains the birth weight (kg) of these elephants. You must import this data into R to carry out the analysis.

For your write-up to be complete, you must:

- Plot a boxplot of the weights.
- Plot a normal quantile-quantile plot of the weights. Include the appropriate line on your QQ plot.
- Comment on the shape of the distribution. (You may refer to both the boxplot and the normal qq plot when commenting on the shape.) Comment on whether the normality assumption of the *t* procedures appears to be justified.
- Suppose we decide to use the *t* procedures to analyze this data. Use R to calculate a 95% confidence interval for the population mean weight. Include the output from R in your submission.
- Assume that the sample can be thought of as a random sample of male African elephants born in captivity. Give an appropriate interpretation of the 95% confidence interval given by R, in the context of the problem.
- If you feel there is an appropriate hypothesis test to carry out on this data (for just the data in the data set, and not any other data from the paper), then carry it out and properly interpret the results. If you do not feel there is a natural hypothesis test of interest in this situation, then say so and justify your position.
- Read the "Methods and Procedures" section of the paper. Comment on the sampling design used in this study, and how that might impact our statistical inference procedures and the conclusions and interpretations we draw from them.

Your submission must include the boxplot, the normal QQ plot, and the R output, in addition to your comments and interpretation. Your submission for this part should only be two pages, but can be three pages if you feel that is necessary.

2 Part II: Jumping distance of a type of amphibious fish

Brunt et al. (2016) investigated various jumping characteristics of *Kryptolebias marmoratus*, an amphibious fish. The experiment compared jumping characteristics of fish kept in water (Control), fish that had recently spent days out of water (Air), and fish that had spend days out of water and then recovered in water (Recovery). Here we will compare the total distance travelled (cm) by the fish in their two jumping bouts (see the paper's experiment protocol section for full details). (For the purposes of this question we'll ignore the air-treated group and compare the recovery group to the control group.) The data is contained in the data set s2040_W20_fishjump, which can be found on the Courselink site. You must import this data set into R to carry out the analysis.

For your write-up to be complete, you must:

- Plot side-by-side box plots of the data (in one plot). Label the plot appropriately.
- Plot normal quantile-quantile plots for the two groups separately.
- In a single paragraph, comment on the appropriateness of the two-sample t procedures in this setting (i.e. are the assumptions of the procedure satisfied?). You should make reference to the plots. Also, justify your choice of using the pooled-variance t procedure, or the Welch procedure. (Which procedure did you choose, and why.)
- Give the R output for your choice of procedure.
- Interpret the results, including commenting on the results of the test of the null hypothesis that the true mean total jumping distance is the same for both groups, and an appropriate interpretation of a relevant confidence interval. Interpretations *must* relate to the problem at hand.

Your submission must include the boxplots, normal QQ plots, and the R output, in addition to your comments and interpretation. Your submission for this part should only be two pages, but can be three pages if you feel that is necessary.

3 Response times in rehearsed and unrehearsed liars

Walczyk et al. (2013) investigated possible differences between truth tellers and liars when questioned about a mock crime. Participants in a psychology experiment were randomly assigned to a truth telling group, an unrehearsed lying group, or a rehearsed lying group (where the individuals were allowed to see the questions and think about their responses in advance). Here we will ignore the rehearsed lying group and compare the truth tellers with the unrehearsed lying group.

In one aspect of the study, the researchers suspected that liars would tend to to have wordier responses to questions than truth tellers. Table 1 illustrates the summary statistics for a "wordiness" score for one of the question types the researchers used.

Unrehearsed liars	$\bar{X}_1 = 3.03$	$s_1 = 2.25$	$n_1 = 47$
Truth tellers	$\bar{X}_2 = 1.53$	$s_2 = 0.77$	$n_2 = 44$

Table 1: Wordiness of responses for individuals questioned about a mock crime.

Choose an appropriate t procedure to analyze this data, and justify your choice of procedure. Construct a 95% confidence interval for $\mu_1 - \mu_2$ and give a proper interpretation of the interval. Carry out an appropriate hypothesis test (give appropriate hypotheses in words and symbols, test statistic, p-value and conclusion). Interpret the results in the context of the problem at hand. Your submission for this part should be a single page.

4 Interpreting some values in journal articles

Answer the following questions clearly and concisely. Each response should be a single sentence, but you can use two sentences if you feel it is necessary. Your submission for this part should be a single page.

a) In many journals, when they report a result such as 16.8 ± 1.4, the 1.4 is the *standard error* of the statistic and not the *margin of error*. Look again at the Dale (2010) paper, this time at Table 3. The table contains the entry "660.67 ± 5.8". What is the meaning of the value 5.8 here? (Clearly and concisely state what the value 5.8 represents, in the context of this problem. It's a standard error of some nature, but don't use the term "standard error" when describing it.)

b) Let's look again at Brunt et al. (2016). In the first paragraph of the results section (page 3205), they discuss the results of a test regarding lactate concentration that resulted in a p-value of 0.041. Give the null hypothesis that resulted in this p-value, and an appropriate conclusion in the context of the problem.

c) In Table 1 of Bondo et al. (2016) (in the Season subcategory), the authors report a confidence interval of (20.7, 31.8). Give a proper interpretation of this confidence interval.

d) In Table 3 of Bondo et al. (2016), the authors report the results of tests on the equality of population proportions. The authors use a different method than one we've discussed in this course, but the general idea is similar – they test the null hypothesis that there is no difference in the population proportions. (The authors use the *odds ratio*, which equals 1 if the two proportions are equal.) In table 3, the authors compare racoons in different seasons, and report a p-value of 0.289. Give a conclusion to the hypothesis test that resulted in this p-value. Use the terminology of our course, and do not refer to the odds ratio.

References

Bondo et al. (2016). Epidemiology of Salmonella on the paws and in the faeces of free-ranging raccoons (*Procyon Lotor*) in Southern Ontario, Canada. *Zoonoses and Public Health*, 63:303–310.

Brunt et al. (2016). Amphibious fish jump better on land after acclimation to a terrestrial environment. *Journal of Experimental Biology*, 219:3204–3207.

Dale, R. (2010). Birth statistics for African (*Loxodonta africana*) and Asian (*Elephas maximus*) elephants in human care: History and implications for elephant welfare. *Zoo Biology*, 29:87–103.

Walczyk et al. (2013). Eye movements and other cognitive cues to rehearsed and unrehearsed deception when interrogated about a mock crime. *Applied Psychology in Criminal Justice*, 29(1):1–22.