STAT*2040 DE
Winter 2015

Data Analysis Project

This project (which is more of a major assignment) has a deadline of Wednesday, April 1 at 11:59 pm. You must submit your document as a pdf into the dropbox on courselink by that time.

There are 4 parts to this project:

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 part of a given journal article, answering a few questions, and doing an analysis of data from the article. (5 marks)

Each part is based on information from a published study. Each one of the 4 journal articles is available from the University of Guelph library website. If you are off-campus, then 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, click on E-Journals list, and search for the journal title. You can also search for the article title in Primo (on the library site).

This project 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 project.

You must submit a single pdf document that contains your responses to all 4 sections. Any individual section (not including plots) shouldn't be more than a single double-spaced page. Including plots, your entire submission shouldn't be more than 6 or 7 pages (and can be less).

1 Part I: Crayfish Carapace Lengths

Harlioglu et al. (2012) investigated several characteristics of crayfish in a freshwater lake in Turkey. Twenty-five adult male *Astacus leptodactylus* crayfish were sampled and several variables were recorded. One of the variables was carapace length (mm). The data is contained in the data set s2040DE_W15_crayfish, 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 a boxplot of the carapace lengths and comment on it.

- Plot a normal quantile-quantile plot of the carapace lengths. Comment on the shape of the distribution. Comment on the suitability of the t procedures for this data.

- Use R to calculate a 95% confidence interval for the population mean. Include the output from R in your submission.

- Give an appropriate interpretation of the 95% confidence interval given by R, in the context of the problem. To what population do your conclusions apply? Comment on any biases that might be present.

- If you feel there is an appropriate one-sample hypothesis test here, 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.

Your submission must include the boxplot, the normal QQ plot, and the R output, in addition to your comments and interpretation.

2 Part II: Swimming Speed of Rats

Santori et al. (2014) investigated various characteristics of swimming in different species of semi-aquatic water rat. This part of the project involves analysis of this data.

The study compared various swimming characteristics of 4 species of rat, but here we will look only at *Nectomys rattus* and *Nectomys squamipes*. In one part of the study, the swimming speed of the rats was recorded. Swimming speeds for 14 *N. rattus* and 15 *N. squamipes* rats can be found in the file s2040DE_W15_rats on the courselink site. (The values given in this file are based on the results of the study, but the summary statistics are not exactly the same as those given in the article. You must use the data contained in this file to answer this question.) You will need to import the data into R and do an appropriate statistical analysis. (This will involve a two-sample t procedure of some sort.)

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

- Plot side-by-side box plots of the data (in one plot), and comment on the plot.
- 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. Also, justify your choice of using the pooled-variance t procedure, or the Welch procedure. (Which one 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 swimming speed is the same for both species of rat, and an appropriate interpretation of a relevant confidence interval. Interpretations *must* relate to the problem at hand.
- Comment on what population your conclusions apply to.

Your submission must include the boxplots, normal QQ plots, and the R output, in addition to your comments and interpretation.

3 Response times in truth tellers and 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). We will ignore the rehearsed lying group and compare the unrehearsed lying group to the truth tellers.

In one aspect of the study, the researchers suspected that liars would tend to take longer to respond to questions when compared to truth tellers. Table 1 illustrates the time to respond statistics for yes/no questions.

Truth tellers	$\bar{X}_1 = 638$	$s_1 = 238$	$n_1 = 44$
Unrehearsed liars	$\bar{X}_2 = 881$	$s_2 = 358$	$n_2 = 47$

Table 1: Time to respond (milliseconds) 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.

4 Sudden death in adults

In order to complete this section, you will need to get this paper:

Naneix et al. (2015). Sudden adult death: An autopsy series of 534 cases with gender and control comparison. *Journal of Forensic and Legal Medicine*, 32:10–15

On page 12 of this article, the authors state that "The deaths caused by cardiovascular diseases were more frequent in males than in females ($p < 0.0001$)." In words and symbols, what are the hypotheses of the test that yielded this p-value? In your own words, give a conclusion to the hypothesis test.

Consider the values given in Table 1 of this article. Suppose we wish to test the null hypothesis that male and female victims of sudden death have the same distribution of cause of death. Choose an appropriate test and carry out the test. Give hypotheses, value of the test statistic, p-value and conclusion.

References

Harlioglu et al. (2012). An investigation on the sperm number and reproductive parameters of males in wild caught freshwater crayfish (*Astacus leptodactylus*, eschscholtz). *Animal Biology*, 62:409–418.

Naneix et al. (2015). Sudden adult death: An autopsy series of 534 cases with gender and control comparison. *Journal of Forensic and Legal Medicine*, 32:10–15.

Santori et al. (2014). Swimming performance in semiaquatic and terrestrial Oryzomyine rodents. *Mammalian Biology*, 79:189–194.

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.