STAT*2040: Statistics I

Fall 2017

Data Analysis Project #1

Important information and instructions

- The deadline for this project is **Friday, November 10 at 11:59pm**. Late submissions are not accepted, and will receive a grade of 0.

- There are three questions; you must complete all of them. The breakdown of marks for each question is as follows:
  - Question 1: 5 marks
  - Question 2: 8 marks
  - Question 3: 10 marks

- This project is worth 7.5% of your final grade. You will be assessed on:
  - getting the proper R/RStudio output and plots (note: you must use R or RStudio for this project!)
  - validity of your statistical conclusions and interpretations
  - writing style, including spelling and grammar
  - presentation
- Final report formatting and submission instructions are found at the end of this document. Please read this information carefully! Projects that do not follow the correct format, or that are not submitted properly, will not receive full marks.

1. The following small sample data set represents the price (in British pence) of a 100 gram pack of a brand of instant coffee, from 15 randomly selected shops in and around Milton Keynes (a town about 87 km from London, England) in 1981:

   100, 109, 101, 93, 96, 104, 98, 97, 95, 107, 102, 104, 101, 99, 102

   - Import the data into RStudio, either by entering the data into a spreadsheet first, or entering it directly into a vector in RStudio. **Be sure to enter your values carefully!**
   - Create a stem-and-leaf plot for the data. Comment on the suggested shape of the distribution, and include both your commentary and the stem-and-leaf plot in your final report.
   - Use RStudio to calculate the summary statistics of mean and standard deviation for the data set. Include these values in your final report.
   - Calculate a 95% confidence interval for the true mean price of this brand of instant coffee in Milton Keynes. What assumptions are required for the conclusions from this procedure to be valid? Was it reasonable for you to make these assumptions? Include the confidence interval, an appropriate interpretation, and any comments on assumptions in your final report.
   - The standard deviation you calculated previously is for data sampled in 1981. Let's assume we can take this standard deviation to be the *population* standard deviation for this brand of instant coffee in today's market (ignoring any effect of inflation, etc.). Suppose you wanted to estimate the mean price of the same brand of instant coffee in Canadian dollars, with a margin of error of \$0.05 (i.e. 5 cents) and 95% confidence. How many shops would you have to sample? Explain your procedure. (Hint: a currency converter with daily exchange rates is available at `http://www.xe.com/currencyconverter/`)

2. The length of the forearm (in inches) of 140 randomly selected males from some population of interest are recorded in the file `Forearm_Data.csv`, available on Courselink.

   Import the data into RStudio, and conduct the following analysis:

   - Create a histogram of the data and comment on it (comments can include shape/symmetry, presence of outliers, etc.). Include both the histogram and your comments in your final report.
   - Before carrying out the statistical analysis, it is necessary to ensure the assumptions for the inference procedure have been met. Create a normal quantile-quantile plot (with a normal qq line included) of the data. Comment on the assumptions for the statistical procedure you intend to use, and whether or not they appear to have been satisfied. Include both the plot and your commentary in your final report.
   - Use RStudio to create a 90% confidence interval for μ, the true mean forearm length for this population. Provide this confidence interval, as well as an appropriate interpretation of it, in your final report.
   - If there is a reasonable hypothesis test to conduct, use RStudio to do so. Be sure to state your hypotheses, your results, and any conclusions you can make. Include any relevant RStudio output in your final report. If you do not believe there is a reasonable hypothesis test to conduct, clearly explain why you think this is the case.

3. In the mining of the Alberta oil sands, water is used as part of the extraction process when separating the oil from the sand. The water is recycled, and used in repeated extraction processes until it is saturated and no longer effective. This oil sands process-affected water (OSPW) is then stored in large outdoor ponds, even though it is known to contain compounds that are toxic to various species of wildlife.

   A study published in *Environmental Science & Technology* (see reference below) investigated the health effects of oil sands process-affected water on mallard ducks. Researchers obtained commerical-bred pekin ducks (a subspecies of mallard ducks) to use as their experimental model. At the end of the experiment, a number of health metrics were obtained on the ducks, including information on the corticosterone levels (in nmol/L). The corticosterone levels for ducks exposed to the OSPW can be found in the file `Mallard_Corticosterone_Data.csv` on Courselink. Import the data set into RStudio and conduct the following analysis:

   - Obtain a copy of the article through the University of Guelph Library. Provide a **maximum** five-sentence summary of the study. Don't worry about nitty-gritty details, but instead focus on the big picture of what was done and found.
   - Create a boxplot of the data and comment on it (comments can include shape/symmetry, presence of outliers, etc.). Include both the boxplot and your comments in your final report.
   - Create a normal quantile-quantile plot (with a normal qq line included) of the data. Comment on the assumptions for the statistical procedure you intend to use, and whether or not they appear to have been satisfied based on both the normal qq plot and the study description in the paper. Include both the plot and your commentary in your final report.
   - Create a 95% confidence interval for μ, the population mean corticosterone level using RStudio. Provide this confidence interval, as well as an appropriate interpretation of it, in your final report.
   - The study also reported the corticosterone level for ducks not exposed to OSPW (i.e. "control" ducks). While we will learn a more effective approach in the future, let's assume that the estimated mean corticosterone level for the control ducks can be considered a hypothesized population mean corticosterone level (you should be able to find this value in Table 2 of the paper). Use RStudio to conduct an appropriate hypothesis test on the mean corticosterone level for OSPW exposed ducks. State your null and alternative hypotheses, along with a brief explanation of your choice of alternative hypothesis. Include any relevant RStudio output in your final report, and discuss any conclusions you can make as a result of this hypothesis test.

   REFERENCE:

   Beck, E.M., Smits, J.E.G., & Cassady St. Clair, C. (2014). Health of domestic mallards (*Anas platyrhynchos domestica*) following exposure to oil sands process-affected water. *Environmental Science & Technology*, 48, 8847 - 8854.

   Disclaimer: The data set provided on Courselink is a simulated data set, based on the values obtained from the above referenced paper. While the summary statistics will be close in value to those reported in the paper, they will not be the exact same.

Formatting Details and Submission Instructions

Please read the following formatting information and submission instructions carefully. Projects that do not follow these formatting requirements, and/or that are submitted incorrectly, may not receive full marks.

FORMATTING DETAILS

- You do not need a cover page for your project, but instead you must include your name and student ID# as a header on every page.
- Your project should be written as a short, formal report for each question (i.e. 3 "mini" reports). This includes using full sentences, proper spelling and grammar, and an understandable and logical organization of ideas. This does NOT include a list of bulleted, point form, poorly written answers.
- If you have referenced an external resource (i.e. a textbook, website, etc.) you must provide the reference at the end of each question using the APA style of referencing (an example is here: `http://pitt.libguides.com/c.php?g=12108&p=64730`).
- Your analysis should be written in a word processing software such as Microsoft Word, with double-spaced text and size 12 font.
- Any figures or tables created in R/RStudio should have proper axis labels and titles. Figures/tables should then be copied into your Word document, and appropriately labelled in your final report. For an example of how to label figures/tables properly, refer to the article assigned in Question 3.
- Figures and tables should be scaled to an appropriate size so as to be legible, but still allow you to complete your commentary within the page allotment (see below). You can include text above, below, or around your figures as you see fit.
- Summary statistics and interpretations should include units, where appropriate.
- RStudio output refers to the results you get after you run a command (see lecture notes for an example). Any output included in your final report should be presented in a clear manner, and labelled accordingly.
- You are allotted a MAXIMUM number of pages for each question. Note that you are not required to meet this maximum, and may very well be able to complete each question in fewer pages. The page allotments are:
  - Question 1: maximum of 2 pages
  - Question 2: maximum of 4 pages
  - Question 3: maximum of 4 pages
  - Note: You will need to create separate files for each question (see Submission Instructions below).

SUBMISSION INSTRUCTIONS

- On Friday, November 3 you will receive an email from Crowdmark. **Do not delete this email, or forward it to anyone else**. This is your personalized link for submitting your assignment. If you do not receive it, first check your junk mail folder. If you are sure it is not there, contact me as soon as possible.
- When you are ready to submit your assignment, you can click on the link provided in the email. This will open up a window where you can upload your files.
- You should see a space to upload your files for each question separately. YOU MUST HAVE A SEPARATE FILE FOR EACH QUESTION. The files also must be in PDF format. Word (.docx), text (.txt) or any other such files will not be accepted by the system.
- Upload your files for each question in the appropriate space. **It is your responsibility to ensure the correct files have been uploaded to the correct space**. Answers that have been uploaded to the wrong location will receive a mark of 0.
- Review your assignment to ensure all pages have uploaded correctly. Once you are satisfied, you can submit your assignment.
- You can change and re-upload your assignment files up to the project deadline. After the project deadline, files already uploaded will be "locked in", and you will not be able to change them. Any new projects uploaded after the deadline will be flagged for a 100% late penalty.
- Help for uploading assignments can be found at:
  `https://crowdmark.desk.com/customer/portal/articles/1639407-completing-and-submitting-an-assignment`

Please review the University of Guelph's policies on Academic Misconduct, as mentioned in the course outline and detailed in the University of Guelph Undergraduate Calendar. It is your responsibility to know what constitutes academic misconduct. Students found in violation of any of the University policies on academic integrity will be charged with academic misconduct, and penalized accordingly.