STAT*2040
Winter 2020

Data Analysis Assignment #1

This assignment has a deadline of Sunday March 8 at 11:59 pm. You may submit the work individually or in groups of 2 or 3. There will be very specific submission guidelines to follow, but your submission will be 3 pdf documents (one for each part) submitted and graded using Crowdmark.

There are 3 parts to this assignment. In this assignment you will:

1. Use R to create a histogram, boxplot and normal quantile-quantile plot for a sample data set, then choose an appropriate transformation and plot a histogram, boxplot and normal quantile-quantile plot of your transformed data values.

2. Use R's `t.test` command to calculate a confidence interval for a population mean, then give a proper interpretation of the interval.

3. Read parts of two journal articles and interpret two of the values given in the articles.

I've put up an "Intro to R" document on Courselink, which may help with some of the basic commands.

We've set up drop-in help for R in SSC 1303 every week on Monday 12:30-2:20 and Tuesday through Friday from 1:30 to 3:20 . If you run into snags, feel free to drop in there to ask for help.

The journal articles are 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 Find > E-Journals, and search for the journal title. You can also search for the article title directly in Omni (on the library site).

This assignment is worth 6% 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.

1 Part I: Distribution of bat pass counts (10 marks total)

Appel et al. (2017) investigated night time bat activity for several species of bat. In one part of the study, the authors investigated a possible relationship between moonlight intensity and bat activity. They looked at several species, but here we'll look only at the results for *Pteronotus parnellii.* Bat activity (a count of the number of passes) was recorded by automatic sensors on a number of nights.

The counts for *P. parnellii* are given in the file s2040_W20_batcounts (data estimated from Figure 2A in the article).

For this part of the assignment:

- Plot a histogram, boxplot, and normal quantile-quantile plot of the bat passes. (Include the line in your normal qq plot.) Properly label the axes (you can use the default labelling for the normal QQ plot). Put all 3 plots on one page. Briefly comment on the shape of the distribution.

- Many statistical procedures require the assumption of a normally distributed population, and often this assumption is clearly violated (the data is definitely not normally distributed). But sometimes we can find a transformation of the data that results in data that is approximately normal. Common transformations include the log ($log(x)$), or various power transformations such as the reciprocal ($\frac{1}{x}$), square root ($\sqrt{x}$), or square (x^2). We might consider more exotic options (e.g. $x^{3.17}$ or $\sqrt{log(x+10)}$), but we usually stick with the more common ones when they work reasonably well. Sometimes the nature of the data suggests a certain transformation, and sometimes we use mathematical techniques to suggest one, but often we simply wing it a little, and try a few different ones and see if we can find one that results in our assumptions being satisfied (roughly, at least).

  For this part of the assignment, find a transformation that results in transformed bat counts that are approximately normally distributed. Try a few different ones – I think you can find a good one! Plot a histogram, boxplot, and normal quantile-quantile plot of the transformed values. Properly label your plots. Put all 3 plots on one page.

2 Bone mineral density in Korean judo participants (10 marks total)

Kim et al. (2013) investigated possible differences in bone mineral density between boys who participate in judo and boys that do not. The main point of the study was a comparison between the two groups, but here we'll look only at the boys that participated in judo. As part of the study, the researchers measured the bone mineral density (g/cm^2) in the right femur of the 30 judo-participating boys. The bone densities are given in the file s2040_W20_bonedensity (the data is simulated data based on information from the study). Here we will use R to calculate a confidence interval for the true mean, using the t procedure.

- Plot a normal quantile-quantile of the bone mineral densities. Comment on whether the normality assumption of the t procedure appears to be satisfied here.

- Using the `t.test` procedure in R, obtain a 95% confidence interval for μ. Include the output of the procedure in your submission.

- Give an appropriate interpretation of the 95% confidence interval in the output. Your interpretation *must* relate to the problem at hand. (Don't phone it in. A statement like "we are 95% sure that μ lies between 8.1 and 12.2" is not good. Your interpretation should reflect the true practical meaning of what the interval tells us.)

  NB: One of the trickiest things in statistics is thinking about the sampling design, and what sort of biases might be present, and how that might influence the results. Let's put those troubles aside for now, and assume that the 30 boys can be thought of as a random sample of Korean high school boys who practice judo.

3 Interpreting a standard error and a confidence interval (10 marks total)

Answer the following questions clearly and concisely. Each response should be a sentence or two. You'll need to look up two journal articles and read parts of them to answer the questions.

a) In many journals, when authors 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*. Rivas et al. (2014) compared the 2D:4D ratio (google it!) of people with congenital adrenal hyperplasia to a control group. We're not going to look at the details of the results of their comparison here, as we haven't yet reached that content in our course, but we are going to look at one statistic from the article. At the start of their results section (on page 560), the authors report "0.9806 ± 0.0026".

   What is the meaning of the value 0.0026 here? Clearly and concisely state what the value 0.0026 represents. It's a standard error of some nature, but don't use the term "standard error" when describing it. Your interpretation must relate to the problem at hand. In other words, your interpretation must relate to the variable and the population under study.

b) Xu et al. (2014) investigated differences in estradiol concentrations in US and Chinese men. In this paper, look at the first paragraph in the results section (page 566), where the authors report a confidence interval of (41.9, 44.8). Give an interpretation of that confidence interval, *in the context of the problem at hand.* In other words, your interpretation must relate to the variable, and the population under study. It is nowhere near sufficient to state something like "we are 95% confident that μ lies in the interval."

References

Appel et al. (2017). Aerial insectivorous bat activity in relation to moonlight intensity. *Mammalian Biology*, 85:37–46.

Kim et al. (2013). Beneficial effects of judo training on bone mineral density of high-school boys in korea. *Biology of Sport*, 30:295–299.

Rivas et al. (2014). New studies of second and fourth digit ratio as a morphogenetic trait in subjects with congenital adrenal hyperplasia. *American Journal of Human Biology*, 26:559–561.

Xu et al. (2014). Estradiol concentrations in young healthy US versus Chinese men. *American Journal of Human Biology*, 26:565–569.