Project 1 **Math 222**
Statistical Inference with R **Due Friday, Jan 29**

This project will use .csv files on my website (under the Projects tab). Please download any of those files into your SageMathCloud account so that you can use R with the data. Links to the files are also are included here, if you click on the file names below.

1. *USA Today* and *The Washington Post* are very different newspapers. *USA Today* aims to have a very broad readership, while *The Washington Post* has a reputation for serious journalism. The sentences.csv file contains data on the number of words in each sentence for the lead editorials from January 22, 2007 in both newspapers. Longer sentences could indicate that a newspaper is writing for better educated readers.
   (a) Make side by side box-plots of the two newspaper's sentence lengths. Describe in a paragraph any features and differences of the two distributions that are clear from the image.
   (b) Make two histograms and two normal quantile plots, one for each newspaper. Based on the images, discuss briefly whether the conditions for a two-sample t-test are satisfied.
   (c) Use R to conduct a two-sample t-Test to test whether the sentences in *USA Today* are significantly shorter than those in *The Washington Post*.
   (d) How big is the difference between the two newspapers? Use a confidence interval to state your answer.
   (e) Do you think this is conclusive evidence one-way or the other that *The Washington Post* is a more serious and intellectual newspaper than *USA Today*? What other quantitative variables could we have looked at instead to contrast the two newspapers?
   (f) Were there any outliers in the data? Repeat the two-sample t-Test with the outliers removed. Does this change the conclusions? Does it make more sense to include the outliers or remove them in this example?

2. To study the long-term effects of preschool programs for poor children, the High/Scope Educational Research Foundation has followed two groups of Michigan children since early childhood. One group of 62 attended preschool as three and four year olds. A control group of 61 children from the same area and similar backgrounds did not attend preschool. Over a 10-year period as adults, 38 of the preschool sample and 49 of the control sample needed social services (mainly welfare).
   (a) Was this an experiment or an observational study? Explain how you can tell.
   (b) Identify the variables of interest in the study. What are the explanatory variable and response variables? What other variables did the researchers try to control for? Are there any possible lurking variables that were not controlled?
   (c) Does this study provide significant evidence that children who attend preschool have less need for social services as adults?
   (d) How large is the difference between the proportions of the preschool and non-preschool populations that require social services? Include a margin of error for your estimate.
   (e) Explain what inference techniques you used to answer the previous two parts, and describe whether the conditions for those techniques have been met.