9.59 Lab in Psycholinguistics: Problem Set #3

The goal of this problem set is to get you comfortable doing some exploratory data analysis and significance testing in R.

In the first pset, you used the reaction times data set. Look at pset 1 to figure out how to load the RT's into R. You may want to re-use some of your code from that problem set to answer the questions below. As before, RT refers to RTlexdec.

1. Plot a histogram to look at the distribution of RT's in the data set. Make sure the number of bins is such that you can clearly see the distribution. Do RT's appear to be normally distributed? How many 'peaks' are there in the distribution?

2. Now use facetting to make separate histograms for young and old subjects. What does this reveal about the first histogram? Do the young person data look normally distributed?

3. IMPORTANT: For all remaining questions, use only the young subject data. That is, throw out the old subjects. Now (as in pset 1) calculate z-scores for each of the data points. Remember that the z-score is $(x - \mu)/sd$.

    a. If the data were normally distributed, what percentage of the data would you expect to have a z-score greater than 1.96? Less than -1.96?

    b. What percentage of the data actually has a z-score above 1.96? What percentage is actually below -1.96? If one of these things is very different from what you expect (hint, hint), why might that be the case?

    c. What percentage of words, if the data were normally distributed, would have a z-score higher than 3? Look at the words that DO have a z-score higher than 3. Why do you think they do?

4. Compare the median and mean of young people RT's. How close are they? Now compare the median and mean of the column NounFrequency. What's going on here? (i.e., why the big difference in one case?). It may be helpful to plot the distribution.

5. We previously looked at the mean RT for words that start with 'p' vs all other words. Do a two-sided (i.e., default) t-test using `t.test()` on the p-word RT's vs the other RT's to see if the difference is significant. Report the t-value and the p value, and say whether it is significant at 95%. What can we conclude based on this?

6. Using ggplot2, create a boxplot comparing noun RT's to verb RT's. Do the outliers generally appear above or below? Why? Run the function `fivenum()` on the data and compare to the boxplot values.

7. Create a bar plot with error bars showing 95% confidence intervals for the mean noun RT vs. the mean verb RT.

8. Make a boxplot (one box for each letter) showing the mean RT for each initial letter of the word. I.e., one box for a, one for b, one for c, etc.

9. Compare RTs for words that start with 2 consonants to the RTs of all other words (i.e. any word that starts with something other than 2 consonants). Using an appropriate test of your choice, give a reasonable discussion of whether that difference is significant.