LING 609: HW 1

Due date: Thursday, October 1st

September 24, 2015

1 Problem 1 (1 point): R coding

Baayen (2008), exercise 1:

The data set spanishMeta contains metadata about fifteen texts sampled from three Spanish authors. Each line in this file provides information on a single text. Provide code that does the following things, as well as the answer provided by your code where relevant. **Note**: in order to access the data set, you must install languageR on your machine.

a) Display the data frame in R. Extract the column names, and the number of rows.

b) Calculate how many different texts are available in spanishMeta for each author. Calculate the mean publication date of the texts sampled for each author.

c) Sort the rows in spanishMeta by year of birth (YearOfBirth) and by the number of words sampled from the texts (Nwords).

d) Extract the vector of publication dates from spanishMeta. Sort this vector. Consult the help page for `sort()`, and sort the vector in reverse numerical order.

e) Extract from spanishMeta all rows with texts that were published before 1980.

2 Problem 2 (3 points): Plotting data

Download [http://people.umass.edu/bwdillon/LING609-2012/judgments.txt](http://people.umass.edu/bwdillon/LING609-2012/judgments.txt), which is the data from a 7-point acceptability survey. The conditions in the experiment were the following:

(1) a. [-WH,SC] Mary heard the sneaky burglar clumsily attempt to open the door.
b. \([-WH,DP]\) Mary heard the sneaky burglar’s clumsy attempt to open the door.

c. \([+WH,SC]\) What did Mary hear the sneaky burglar clumsily attempt to open?

d. \([+WH,DP]\) What did Mary hear the sneaky burglar’s clumsy attempt to open?

Judgments were collected from 12 speakers on a 1 (bad) to 7 (good) scale. There were 6 observations per condition, per speaker. Load this data into R using `read.table()`; consult the help page for `read.table()` if you are unsure of the syntax. Produce a histogram of the ratings for each of the four conditions. Plot a red vertical line on each plot indicating where the median value is. Be sure to appropriately label the axes and title each graph appropriately.

Inspect the distribution of judgments in all four conditions. What appears to be happening? Describe how the judgments differ across conditions, both at the level of the median (or mean) of the condition, as well as the distribution of judgments across the scale.

3 Problem 3 (2 points): Probability theory

from Levy (forthcoming), exercise 2.3:

a) You obtain infinitely many copies of the text *Alice in wonderland* and decide to play a word game with it. You cut apart each page of each copy into individual letters, throw all the letters in a bag, shake the bag, and draw three letters at random from the bag. Suppose that in *Alice in Wonderland*, 12.6\% of the letters in the book are e, 9.9\% are t, and 8.2\% are a. What is the probability that you will be able to spell *tea*?

b) Why did the problem specify that you obtained infinitely many copies of the text? Suppose that you obtained only one copy of the text. Would you have enough information to compute the probability of being able to spell *tea*? If not, what additional piece of information would you need?

4 Problem 4 (4 points): Confidence Intervals

Recall that a 95\% confidence interval is an interval constructed around a sample mean which we are 95\% confident contains the population mean. The 95\%CI is constructed so that, if we were to repeat our experiment again and again, the true
population mean would be contained within the 95%CI around the sample mean on 95% of replications. Evaluate the degree to which this holds for a binomial process with $\pi = 0.5$ by using simulation in R. For your simulated experiments, calculate both conventional binomial proportion CIs using the approximation given in the lecture notes, and Wald-adjusted 95% CIs. Do this for sample size $n = 40$, and for sample size $n = 4000$. In your simulations, how often is the true population mean inside the 95%CI when calculated using the approximate formula for a binomial proportion 95%CI? How often is the true population mean inside the Wald-adjusted CI? What do you conclude about the two different (approximate) methods to calculate 95% CIs for a binomial proportion?