LING82100: homework 1

(Due 2/26)

1 Arithmetic

Give R expressions that compute the following arithmetic quantities:

1. $\frac{1}{3} + \frac{1}{4}$
2. $2^{10} + 1$
3. $1127 \ln(1 + \frac{f}{700})$ (the conversion from Hz to mel) where $f = 440$
4. $\frac{-b+\sqrt{b^2-4ac}}{2a}$ (one solution for a quadratic equation) where $a = 2$, $b = 4$, and $c = -4$

What to turn in

Provide a list of the R expressions you used to compute these values.

Hints

- In R, the natural logarithm $\ln x$ is computed by the `log` function.
- It may be easier to assign the values (like $f = 440$) to variables, for instance by typing `f <- 440` and then using `f` (instead of `440`) in a second expression.
- If you're stuck or otherwise feeling uncomfortable with R, do the first (free) lesson of the "Introduction to R" course on DataCamp.[1]

2 Categorical data

The *comma-separated values* (CSV) file `NYC.csv`[2] contains data from William Labov's 1966 department store study. In this study, Labov investigated *r*-lessness in the English of New York City residents. Labov went to three department stores: Sak's 5th Ave., Macy's Herald Square, and S. Klein's (now defunct, but was located on the east side of Union Square). At each store, Labov elicited instances of the phrase "fourth floor" (e.g., by asking where the men's shoes were) from employees. Labov then coded whether or not the words "fourth" and "floor" had audible post-vocalic *r*'s. Load the CSV file into R. Then, using the `table`, and/or `xtabs` functions, answer the following questions:

[1] `https://www.datacamp.com/courses/free-introduction-to-r`
[2] `http://wellformedness.com/courses/LING82100/Data/NYC.csv`

- How many times did employees at the three department stores use *r* in the word "fourth" in the emphatic condition?
- What percentage of the time did employees at S. Klein's use *r* in the word "floor"?

2.1 What to turn in

Provide a list of the R commands you used, and the answers you obtained.

2.2 Stretch goal

Provide some additional summaries of this data. For example, which department stores have the highest and lowest rates of *r*-lessness? How is *r*-lessness related to emphasis? Does *r*-lessness vary depending on whether the word is "fourth" or "floor'? Give simple statistical summaries (like percentages) to back up your claims.

2.3 Hints

- URLs are (or may be) case-sensitive.
- You can either download the CSV file to your working directory, or, if you're connected to the internet, you can actually pass the URL as a string to R's `read.csv` command.
- Baayen §1.4 provides some relevant examples.

3 Ratio data

The *tab-separated values* (TSV) file `VOT.tsv`[3] contains simulated *voice onset time* data for a series of word-initial coronal stops in English and Spanish (Casillas et al. 2015). Load the TSV file into R and compute:

- Sample quartiles for VOT (NB: the 2nd quartile, AKA the 50% percentile, is the median)
- The mean of Spanish speakers' VOTs
- The (sample) standard of English speakers' VOTs

3.1 What to turn in

Provide a list of the R commands you used, and the answers you obtained.

[3] `http://wellformedness.com/courses/LING82100/Data/VOT.tsv`

3.2 Stretch goal

Instead of using built-in R functions, do this manually, e.g., for sample standard deviation, use the formula

$$s = \sqrt{\frac{\sum_i^N (x_i - \bar{x})^2}{N-1}}$$

3.3 Hints

Instead of using `read.csv`, you can use the `read.table` function. By default, it assumes a tab separator. However, you may need to tell it that there is a header row (with the names of the variables). Type `?read.table` into R to read the relevant documentation.

References

Casillas, J. V., Y. Diaz, and M. Simonet. 2015. Acoustics of Spanish and English coronal stops. In *ICPhS 18*. Glasgow.

Labov, William. 1966. *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.