AMPA E4900: Modeling Social Data

Spring 2019

modelingsocialdata.org

Homework 2

The first problem looks at the link between coffee and cancer, the second problem examines an experiment on whether yawning is contagious, and the third problem involves replicating the results of a paper about the Google ngram dataset. Details are in the README.md file for each problem.

Your code and a brief report with your results are to be submitted electronically in one zipped (or tarball-ed) file through the CourseWorks site. All solutions should be contained in the corresponding files provided here. Code should be written in bash / R and should not have complex dependencies on non-standard libraries. Each problem contains an Rmarkdown file which should be rendered as a pdf to be submitted with your solution. All work should be your own and done individually.

1. Coffee and cancer

Note: This is adapted from a discussion by David Spiegelhalter.

In March of 2018 there was a court case in California that hinged on whether a chemical called acrylamide found in coffee causes cancer.

The judge asked the defendants to show that coffee was "safe" in that drinking coffee caused fewer than 1 in 100,000 extra cases of cancer. The defendants were unable to demonstrate this. As a result, companies selling coffee in California must now display a cancer warning alongside it.

Imagine you were the defendants and wanted to set up a randomized experiment to demonstrate that coffee is indeed "safe" and causes fewer than 1 in 100,000 extra cases of cancer. Assuming the lifetime risk of getting cancer is 40%, this means that your experiment would have to tell the difference between 40,000 out of 100,000 non-coffee drinkers getting cancer versus 40,001 out of 100,000 coffee drinkers getting cancer.

a) Assuming you could actually pull of such an experiment, how many participants would you need to reject the null hypothesis that fewer than 1 in 100,000 extra cases of cancer occur among people assigned to drink coffee?

For the sake of the calculation assume you'd like to have the following long-run error rates. If coffee is indeed "safe", you'd like your test to falsely return "unsafe"

at most 1 in 10 times. If, however, coffee is "unsafe", you'd like your test to detect this effect 4 out of 5 times.

You can use one of R's built-in power functions to calculate an answer.

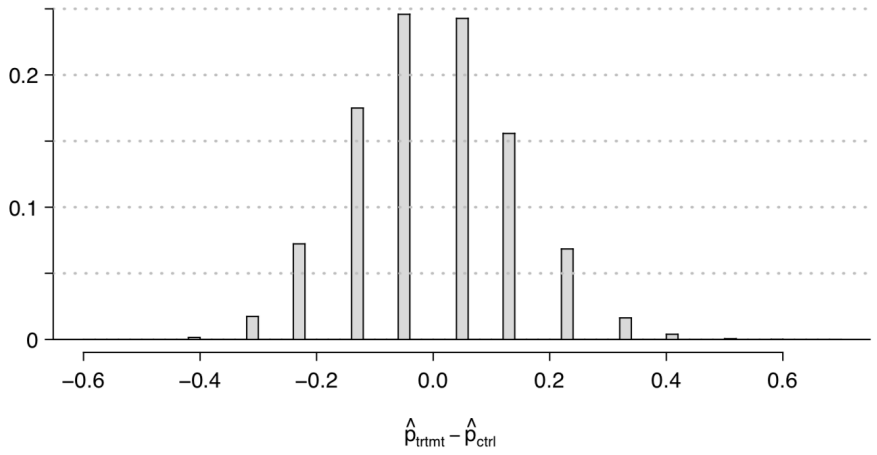
b) Comment on challenges you might face in running this experiment. ### 2. Is Yawning Contagious?

Note: this problem is adapted from Exercise 2.6 in Introductory Statistics with Randomization and Simulation by David Diez, Christopher Barr, and Mine Cetinkaya-Rundel.

An experiment conducted by the MythBusters, a science entertainment TV program on the Discovery Channel, tested if a person can be subconsciously influenced into yawning if another person near them yawns. 50 people were randomly assigned to two groups: 34 to a group where a person near them yawned (treatment) and 16 to a group where there wasn't a person yawning near them (control). The following table shows the results of this experiment.

	Treatment	Control	Total
Yawn	10	4	14
No yawn	24	12	36
Total	34	16	50

A simulation was conducted to understand the distribution of the test statistic under the assumption of independence: having someone yawn near another person has no influence on if the other person will yawn. In order to conduct the simulation, a researcher wrote yawn on 14 index cards and not yawn on 36 index cards to indicate whether or not a person yawned. Then he shuffled the cards and dealt them into two groups of size 34 and 16 for treatment and control, respectively. He counted how many participants in each simulated group yawned in an apparent response to a nearby yawning person, and calculated the difference between the simulated proportions of yawning in the treatment and control groups. This simulation was repeated 10,000 times using software to obtain 10,000 differences that are due to chance alone. The histogram shows the distribution of the simulated differences.



a) Write code that implements this simulation and produces a similar distribution.

b) Calculate the observed difference between the yawning rates in treatment and control.

c) Estimate the p-value by comparing this observed difference to the distribution from the simulation.

d) Do you have sufficient evidence to reject the null hypothesis that yawning isn't contagious? For the sake of completeness, you can assume that you'd like your long-run error rate to be such that if yawning isn't contagious, your test will falsely identify it as such no more than 1 in 20 times.

e) Now simulate repeating this experiment 1,000 times in a world where yawning is actually contagious, and the true probability of yawning in the treatment and control groups is exactly equal to what was found in this one experiment. Calculate the power for this test by measuring how often your test rejects the null that yawning isn't contagious in these 1,000 experiments.

f) Imagine you were going to produce another episode on this topic for Mythbusters. How would you change the experiment based on the answer to the previous question? ### 3. Culturomics

Problem statement

Note: This is Exercise 6 in Chapter 2 of Bit By Bit: Social Research in the Digital Age by Matt Salganik.

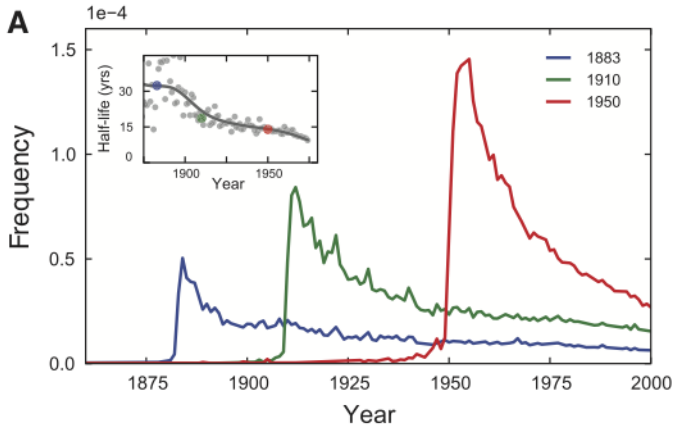
In a widely discussed paper, Michel and colleagues (2011) analyzed the content of more than five million digitized books in an attempt to identify long-term cultural trends. The data that they used has now been released as the Google

NGrams dataset, and so we can use the data to replicate and extend some of their work.

In one of the many results in the paper, Michel and colleagues argued that we are forgetting faster and faster. For a particular year, say "1883," they calculated the proportion of 1-grams published in each year between 1875 and 1975 that were "1883". They reasoned that this proportion is a measure of the interest in events that happened in that year. In their figure 3a, they plotted the usage trajectories for three years: 1883, 1910, and 1950. These three years share a common pattern: little use before that year, then a spike, then decay. Next, to quantify the rate of decay for each year, Michel and colleagues calculated the "half-life" of each year for all years between 1875 and 1975. In their figure 3a (inset), they showed that the half-life of each year is decreasing, and they argued that this means that we are forgetting the past faster and faster. They used Version 1 of the English language corpus, but subsequently Google has released a second version of the corpus. Please read all the parts of the question before you begin coding.

This activity will give you practice writing reusable code, interpreting results, and data wrangling (such as working with awkward files and handling missing data). This activity will also help you get up and running with a rich and interesting dataset.

The full paper can be found here, and this is the original figure 3a that you're going to replicate:



a) Get the raw data from the Google Books NGram Viewer website. In particular, you should use version 2 of the English language corpus, which was released on July 1, 2012. Uncompressed, this file is 1.4GB.

b) Recreate the main part of figure 3a of Michel et al. (2011). To recreate this figure, you will need two files: the one you downloaded in part (a) and the "total counts" file, which you can use to convert the raw counts

into proportions. Note that the total counts file has a structure that may make it a bit hard to read in. Does version 2 of the NGram data produce similar results to those presented in Michel et al. (2011), which are based on version 1 data?

c) Now check your graph against the graph created by the NGram Viewer.

d) Recreate figure 3a (main figure), but change the y-axis to be the raw mention count (not the rate of mentions).

e) Does the difference between (b) and (d) lead you to reevaluate any of the results of Michel et al. (2011). Why or why not?

f) Now, using the proportion of mentions, replicate the inset of figure 3a. That is, for each year between 1875 and 1975, calculate the half-life of that year. The half-life is defined to be the number of years that pass before the proportion of mentions reaches half its peak value. Note that Michel et al. (2011) do something more complicated to estimate the half-life—see section III.6 of the Supporting Online Information—but they claim that both approaches produce similar results. Does version 2 of the NGram data produce similar results to those presented in Michel et al. (2011), which are based on version 1 data? (Hint: Don't be surprised if it doesn't.)

g) Were there any years that were outliers such as years that were forgotten particularly quickly or particularly slowly? Briefly speculate about possible reasons for that pattern and explain how you identified the outliers.

Template

A template has been provided for a solution.

Edit the `01_download_1grams.sh` file to download the `googlebooks-eng-all-1gram-20120701-1.gz` file and the `02_filter_1grams.sh` file to filter the original 1gram file to only lines where the ngram matches a year (output to a file named `year_counts.tsv`).

Then edit the `03_download_totals.sh` file to down the `googlebooks-eng-all-totalcounts-20120701.txt` and file and the `04_reformat_totals.sh` file to reformat the total counts file to a valid csv (output to a file named `total_counts.csv`).

Place the rest of your solution in the `05_final_report.Rmd` file and render it to a pdf file.

Finally, edit the `Makefile` in this directory to execute the full set of scripts that download the data, clean it, and produce this report. This must be turned in with your assignment such that running `make` on the command line produces your final pdf file.