Homework 4 – Intro to Probability and Statistics

Your name here

Instructions:

Due: 05/07 at 11:59PM.

What am I expecting? An R Markdown with the answers.

Have fun!

Revisiting the Gay Marriage Experiment

In this exercise, we revisit the gay marriage study we analyzed in the previous problem set. In May 2015, three scholars reported several irregularities in the data set used to produce the results in the study: link here. They found that the gay marriage experimental data were statistically indistinguishable from data in the Cooperative Campaign Analysis Project (CCAP), which interviewed voters throughout the 2012 US presidential campaign. The scholars suggested that the CCAP survey data—and not the original data alleged to have been collected in the experiment—were used to produce the results reported in the gay marriage study. The release of a report on these irregularities ultimately led to the retraction of the original article. In this exercise, we will use several measurement strategies to reproduce the irregularities observed in the gay marriage data set. To do so, we will use two CSV data files: a reshaped version of the original data set in which every observation corresponds to a unique respondent, `gayreshaped.csv`, and the 2012 CCAP data set alleged to have been used as the basis for the gay marriage study results, `ccap2012.csv`. Note that the feeling thermometer measures how warmly respondents feel towards gay couples on a 0–100 scale.

Gay Marriage Reshaped Data

Name	Description
`study`	Which study the data set is from (1 = study 1, 2 = study 2)
`treatment`	Five possible treatment assignment options
`therm1`	Survey thermometer rating of feeling towards gay couples in wave 1 (0–100)
`therm2`	Survey thermometer rating of feeling towards gay couples in wave 2 (0–100)
`therm3`	Survey thermometer rating of feeling towards gay couples in wave 3 (0–100)
`therm4`	Survey thermometer rating of feeling towards gay couples in wave 4 (0–100)

CCAP Survey Data

Name	Description
`caseid`	Unique respondent ID
`gaytherm`	Survey thermometer rating of feeling towards gay couples (0–100)

Question 1

In the gay marriage study, researchers used seven waves of a survey to assess how lasting the persuasion effects were over time. One irregularity the scholars found is that responses across survey waves in the control group (where no canvassing occurred) had unusually high correlation over time. What is the correlation between respondents' feeling thermometer ratings in waves 1 and 2 for the control group in study 1? To

handle missing data, we should set the use argument of the `cor()` function to "complete.obs" so that the correlation is computed using only observations that have no missing data. Provide a brief substantive interpretation of the results.

Question 2

Repeat the previous question using study 2 and comparing all waves within the control group. Note that the `cor()` function can take a single data frame with multiple variables. To handle missing data in this case, we can set the use argument to `"pairwise.complete.obs"`. This means that the `cor()` function uses all observations that have no missing values for a given pair of waves even if some of them have missing values in other waves. Briefly interpret the results.

Question 3

Most surveys find at least some outliers or individuals whose responses are substantially different from the rest of the data. In addition, some respondents may change their responses erratically over time. Create a scatter plot to visualize the relationships between wave 1 and each of the subsequent waves in study 2. Use only the control group. Interpret the results.

Question 4

The researchers found that the data of the gay marriage study appeared unusually similar to the 2012 CCAP data set even though they were supposed to be samples of completely different respondents. We use the data contained in `ccap2012.csv` and `gayreshaped.csv` to compare the two samples. Create a histogram of the 2012 CCAP feeling thermometer, the wave-1 feeling thermometer from study 1, and the wave-1 feeling thermometer from study 2. There are a large number of missing values in the CCAP data. Consider how the missing data might have been recoded in the gay marriage study. To facilitate the comparison across histograms, use the breaks argument in the histogram.

Question 5

A more direct way to compare the distributions of two samples is through a quantile–quantile plot. Use this visualization method to conduct the same comparison as in the previous question. Briefly interpret the plots.