Problem Set 2

Stats 506, F19

Due: Saturday October 12 by 9 am

Instructions

- Submit the assignment by the due date via canvas. Assignments may be submitted up to 72 hours late for a 5 point reduction.
- All files read, sourced, or referred to within scripts should be assumed to be in the same working directory (`./`).
- Your code should be clearly written and it should be possible to assess it by reading it. Use appropriate variable names and comments. Your style will be graded using the style rubric (./StyleRubric.html) [10 points].
- Some of these exercises may require you to use commands or techniques that were not covered in class or in the course notes. You can use the web as needed to identify appropriate approaches. Part of the purpose of these exercises is for you to learn to be resourceful and self sufficient. Questions are welcome at all times, but please make an attempt to locate relevant information yourself first.

Questions

Question 1 [35 points]

In this question, you will answers several questions using the 2015 RECS survey data available here (https://www.eia.gov/consumption/residential/data/2015/index.php?view=microdata). For each question, write a sentence or two that directly answers the question(s) and, when indicated, produce a well formatted table or graph as indicated. For all statistics presented as point estimates, also include 95% confidence intervals. Please use ‘dplyr’ and ‘tidyr’ for all data manipulations, ‘stringr’ manipulating strings, and ‘ggplot2’ for plotting. For this question, do not use any other non-base packages outside of the “tidyverse”.

a. [5 pts] What is the national average home temperature at night in winter, among homes that use space heating?
b. [10 pts] Create a table showing the proportion of homes using each level of “main space heating fuel” within each unique combination of census division and census (2010) urban type among homes that use space heating.
c. [10 pts] Create a plot comparing, by census division and urban type, the average winter home temperatures at night, during the day with someone home, and during the day with no one home (when applicable).

d. [10 pts] Among homes that use space heating, what is the (national) median difference between the daytime (with someone home) and nighttime temperatures for each level of “main heating equipment household behavior”? You can exclude those for whom the latter variable is not applicable. Create a nicely formatted table or graph to display your results. **Hint: The weighted median is the value for which the cumulative weight first exceeds half the total weight.**

Question 2 [35 points]

In this question, you will return to data from mouse-tracking experiments of the type encountered in problem set one. In particular, please install the R package `mousetrap` and refer to the data set `mousetrap::KH2017_raw` for all parts of this question. Use “dplyr” and “tidyr” for data manipulations and the “lme4” package for part d (you may also use “lmerTest” if you prefer).

a. [5 pts] Create a file `ps2_q2_funcs.R` containing only the (possibly corrected) functions you wrote for question 3, problem set 1. Use a `source()` call to make these functions available in your solution script.

b. [10 pts] Load the data and examine the columns `xpos_get_response`, `ypos_get_response` and `timestamp_get_response`. Write a function to extract the x, y, and t components of a trajectory into a numeric vector. Apply this function to the data to represent the trajectories in a numeric format.
Hint: Either use “list” columns for each component or create separate data frames for each trial and then bind into a single longer-format dataset.

c. [10 pts] Use your functions from problem set 1 to compute curvature measures for each trajectory. Then, create a “tidy” data frame that meets the following criteria:
   - one row per subject / trial,
   - filter to trials in which the subject chose the correct response,
   - variables identifying the subject, trial, “Condition”, Exemplar", and curvature measures.

d. [10 pts] Use the R package `lme4` to fit linear mixed models exploring how each curvature measure differs by condition. You should fit one model per curvature measure with “Condition” as the only fixed effect. To account for the repeated measures in the data, include a random intercept for each subject and additional random intercepts for each “Exemplar”. Use a log transformed curvature measure as the response in each model. For which curvature measure(s) does condition have the largest (relative) effect? Create a nicely formatted table or plot to justify your answer.