DS5110 HW 4 - Due March 1

Kylie Ariel Bemis

2/19/2019

Content Note: Problem 5 includes references to statistics about suicide. Please contact the instructor if you have difficulty completing the assignment for personal reasons due to this.

Instructions

Create a directory with the following structure:

- `hw4-your-name/hw4-your-name.Rmd`
- `hw4-your-name/hw4-your-name.pdf`

where `hw4-your-name.Rmd` is an R Markdown file that compiles to create `hw4-your-name.pdf`.

Do not include data in the directory. Compress the directory as `.zip`.

Your solution should include all of the code necessary to answer the problems. All of your code should run (assuming the data is available). All plots should be generated using `ggplot2`. Missing values and overplotting should be handled appropriately. Axes should be labeled clearly and accurately.

To submit your solution, create a new private post of type "Note" on Piazza, select "Individual Student(s) / Instructor(s)" and type "Instructors", select the folder "hw4", go to Insert->Insert file in the Rich Text Editor, upload your `.zip` homework solution. Title your note "[hw4 solutions] your name" and post the private note to Piazza. **Be sure to post it only to instructors**

Part A

Problems 1–2 use the US Department of Education's Civil Rights Data Collection from Homework 2. It is available at https://www2.ed.gov/about/offices/list/ocr/docs/crdc-2015-16.html. Use the `read_csv()` function to import the dataset into R, handling missing data appropriately. Again, treat all reserve codes as missing.

Problem 1

For each school, calculate the proportions of enrolled students of each race (among all enrolled students at each school). Then create side-by-side boxplots showing the distributions of these proportions for each race.

Problem 2

For each school, calculate the proportions of students of each race enrolled in a Calculus class (*among students enrolled in calculus classes at each school*). (It may be helpful to filter out schools without data for Calculus class enrollment.) Then create side-by-side boxplots showing the distributions of these proportions for each race.

Comment on any similarities or differences between these distributions and the ones plotted in Problem 1.

Part B

Problems 3–5 use data collected from the Virginia Transgender Health Initiative Study (THIS). It is available via the Inter-university Consortium for Political and Social Research (ICPSR), of which Northeastern University is a member, at http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/31721 or via a proxy link at http://www.icpsr.umich.edu.ezproxy.neu.edu/icpsrweb/ICPSR/studies/31721 (if you are not on a campus internet connection). You will need to create a free MyData account as well as login with your myNEU credentials to gain access to the public version of the dataset.

Download the R data (`.rda`) version of the dataset and load it into R using the `load()` function.

Problem 3

We would like to investigate how certain questions break down among trans women, trans men, and non-binary participants. However, the survey sometimes uses outdated terminology and includes many gender-nonforming and questioning participants who are difficult to categorize this way, as well as erroneously and confusingly including "transgender" as a distinct gender category.

Transform the data to include only 3 gender categories for trans men, trans women, and non-binary participants. Use the following definitions when transforming the dataset: (1) trans women are women who were assigned-male-at-birth; (2) trans men are men who were assigned-female-at-birth; (3) combine the "Genderqueer" and "Androgynous" categories to create a single "Non-binary" category. Filter the dataset to include only participants in these categories.

Create a bar plot showing the number of participants of each of the above genders.

Then create bar plots showing the proportion of participants who have ever been homeless, faceted for each of the above genders. (Do not include missing data in the plot.)

According to statistics from a 2003 study (https://ourworldindata.org/homelessness), roughly 6.2% of the general U.S. population has ever been homeless. How does that compare to the participants of this survey?

Problem 4

Using the full dataset again, transform the dataset to have a column for `race` indicating the race of the participant. Include only the racial demographics with publicly available data (i.e., African American, Caucasian, Hispanic/Latinx, and Native American).

(Participants with two or more races may create multiple rows: this is fine for now. Do NOT use the pre-calculated 'RACE' column in the dataset, which does not properly disaggregate multiracial participants.)

Then create bar plots showing the proportions of participants who have ever been homeless, faceted for African American, Caucasian, Hispanic/Latinx, and Native American demographics. (Do not include missing data in the plot.)

How do these numbers compared to the statistic of 6.2% of the U.S. general population experiencing homelessness in their lifetime?

Problem 5

One of the findings reported in the 2015 U.S. Transgender Survey (http://www.ustranssurvey.org) was that a staggering 40% of the respondents reported attempting suicide in their lifetime, nearly nine times the attempted suicide rate of the general U.S. population (4.6%).

Using the full dataset, calculate the *total* proportion of participants who have *attempted* suicide in the Virginia THIS survey. Is it higher or lower than the national average for trans people? Is it higher or lower than the national average for the general population?

We would like to know if having a birth family who is supportive of one's gender identity and expression reduces the risk of suicide. Using the full dataset, filter the dataset to remove participants who answered "Not applicable to me" to the question about familial support, and then create bar plots showing the proportions of participants who have *thought* about killing themselves for each level of familial support. (Do not include missing data in the plot.)

What do you notice?