

Problem Set #2

EH6105 - Quantitative Methods

Steven V. Miller

This homework makes use of data available in `{stevedata}` and implies the use of `{tidyverse}` to answer the questions. `{tidyverse}` is not necessary to answer these questions though it will assuredly make the process easier. Load these two libraries to get started answering these questions.

```
library(tidyverse)
library(stevedata)
```

The Usual Daily Drinking Habits of Americans

This homework will refer to the `nesarc_drinkspd` data set that is available in `{stevedata}`. This data set is loosely modified from Wave 1 (2001-2) of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) data set, which, in part, asks a nationally representative survey of 43,093 Americans to say how often they consume alcohol on a usual day in which they drink alcohol. You can find out more information about the data by visiting [this part of the package's website](#), or with the following command.¹

```
?nesarc_drinkspd
```

Here's a little preview of these data.

```
nesarc_drinkspd
```

```
## # A tibble: 43,093 x 8
##   idnum ethrace2a region  age  sex marital  educ s2aq8b
##   <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     1     1     5     4    23     0     6     2    NA
## 2     2     2     5     4    28     1     1     2     1
## 3     3     3     5     3    81     1     3     1    NA
## 4     4     4     5     2    18     0     6     2    NA
## 5     5     5     2     2    36     0     1     4    NA
## 6     6     6     2     2    34     1     1     4     1
## 7     7     7     2     2    19     0     6     3     1
## 8     8     8     1     4    84     1     6     2     1
## 9     9     9     1     3    29     1     6     4     1
## 10    10    10     5     4    18     1     6     2     6
## # ... with 43,083 more rows
```

¹You really should always look at the codebook, and definitely here.

Answer these questions. A successful answer of these question must include the R code you used to help you answer the question.

1. What is the mode, median, and mean respondent age in these data?
2. What is the mode, median, and mean of the `ethrace2a` variable, and why is this a ridiculous question to ask?
3. What is the mode, median, and mean of the `sex` variable? What does the mean tell you for a binary variable like this?
4. Recode the `ethrace2a` variable into a dummy `white` variable that equals 1 if the respondent is white (not Hispanic) and 0 if the respondent is some other race or is Hispanic.
5. Look at the `marital` variable. How many people in the data say they living with someone as married (i.e. “sambo”, as you might call them in Sweden)?
6. Get a histogram of the `s2aq8b` variable and tell me what you see.
7. The `s2aq8b` variable is taken “as is” from the original data set, and communicates the number of drinks of any alcoholic type that the respondent says they consume on a given day drinking alcohol in the past 12 months. There are two problematic possible values in here that no longer make this a count variable, as is. What are they, and why are they problematic?
8. Recode this variable into a more informative `drinks_consumed` variable, given what you know about the variable from the codebook. Explain what you’re doing.