EPID 674: Descriptive Statistics

Homework 3 Answer Key

Kelly Bakulski

Last compiled on January 16, 2022

# Load libraries

library(tidyverse)  
library(here)  
library(gtsummary)  
library(flextable)

# Load the dataset

# Load saved NHANES dataset from Homework 2  
load(here(("nhanes\_homework\_dataset.rda")))

# 1. Calculate univariate descriptive statistics on a numeric variable

### Calculate descriptive statistics on the cognition variable. For the cognition variable (CFDDS), calculate the mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum values.

# For the cognition variable (CFDDS), calculate the mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum values.   
nhanes\_homework %>%  
 select(CFDDS) %>%  
 drop\_na() %>%  
 summarise(cog\_mean = mean(CFDDS),  
 cog\_sd = sd(CFDDS),  
 cog\_min = min(CFDDS),  
 cog\_25\_quart = quantile(CFDDS, probs = 0.25),  
 cog\_median = median(CFDDS),  
 cog\_75\_quart = quantile(CFDDS, probs = 0.75),  
 cog\_max = max(CFDDS))

## cog\_mean cog\_sd cog\_min cog\_25\_quart cog\_median cog\_75\_quart cog\_max  
## 1 45.95101 17.23167 0 33 46 58 105

The mean digit symbol substitution score is 45.95 rows and the standard deviation is 17.23 columns. The digit symbol substitution score minimum is 0, the 25th percentile is 33, the median is 46, the 75th percentile is 34, and the maximum is 105.

Interpret the findings: The variable CFDDS appears to have a normal distribution. The mean value is very similar to the median value. The mean and median values are approximately halfway between the 25th and 75th quartile values and they are approximately halfway between the minimum and median values.

# 2. Calculate univariate descriptive statistics for a factor variable

### For the educational attainment variable (education), calculate the number of participants in each category and the percent of the total sample in each category.

nhanes\_homework %>%   
 count(education) %>%  
 mutate(percent = n / sum(n) \* 100)

## education n percent  
## 1 Less than high school 491 26.6702879  
## 2 High school or GED 440 23.9000543  
## 3 More than high school 907 49.2667029  
## 4 <NA> 3 0.1629549

The number of participants with less than high school education is 491, which represents 26.71% of participants. The number of participants with high school or GED level of education is 440, which represents 23.94% of participants. The number of participants with greater than high school education is 907, which represents 49.35% of participants.

Interpret the findings: The educational attainment category with the highest proportion of participants in this dataset is the greater than high school education category.

# 3. Filter dataset and calculate univariate descriptive statistics

### Filter to participants who have mild cognitive impairment. Calculate the number of participants in each educational attainment category and the percent of the total sample in each category.

nhanes\_homework %>%   
 filter(MCI == "Mild Cognitive Impairment") %>%  
 count(education) %>%  
 mutate(percent = n / sum(n) \* 100)

## education n percent  
## 1 Less than high school 208 51.612903  
## 2 High school or GED 99 24.565757  
## 3 More than high school 95 23.573201  
## 4 <NA> 1 0.248139

The number of participants with mild cognitive impairment and less than high school education is 208, which represents 51.74% of participants. The number of participants with mild cognitive impairment and a high school or GED level of education is 99, which represents 24.63% of participants. The number of participants with mild cognitive impairment and greater than high school education is 95, which represents 23.63% of participants.

Interpret the findings: The proportion of people in the lowest educational attainment category among those with mild cognitive impairment is 51.74% of participants, which is higher than the proportion of people in the lowest educational attainment category in the entire study population (51.74% of participants.)

# 4. Calculate bivariate descriptive statistics

### Use the tbl\_summary() function to calculate bivariate descriptive statistics on your dataset, split by mild cognitive impairment status (variable name: MCI). Include the following variables as rows in your table (variable names: sex, RIDAGEYR, race\_eth, INDFMPIR, education, LBXCOT, LBXBPB, LBXBCD).

nhanes\_homework %>%  
 select(sex,  
 RIDAGEYR,  
 race\_eth,  
 INDFMPIR,  
 education,  
 LBXCOT,  
 LBXBPB,  
 LBXBCD,  
 MCI) %>%  
 tbl\_summary(by = MCI,  
 label = list(race\_eth ~ "Race/Ethnicity",  
 education ~ "Educational Attainment"))

## 249 observations missing `MCI` have been removed. To include these observations, use `forcats::fct\_explicit\_na()` on `MCI` column before passing to `tbl\_summary()`.

## Table printed with {flextable}, not {gt}. Learn why at  
## http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
## To suppress this message, include `message = FALSE` in the code chunk header.

| Characteristic | No Impairment, N = 1,1891 | Mild Cognitive Impairment, N = 4031 |
| --- | --- | --- |
| sex |  |  |
| Male | 547 (46%) | 215 (53%) |
| Female | 642 (54%) | 188 (47%) |
| Age in years at screening | 68 (63, 74) | 72 (66, 80) |
| Race/Ethnicity |  |  |
| Non-Hispanic White | 675 (57%) | 130 (32%) |
| Mexican American | 105 (8.8%) | 71 (18%) |
| Non-Hispanic Black | 206 (17%) | 123 (31%) |
| Other Hispanic | 79 (6.6%) | 58 (14%) |
| Other Race | 124 (10%) | 21 (5.2%) |
| Ratio of family income to poverty | 2.68 (1.43, 4.77) | 1.36 (0.86, 2.27) |
| Unknown | 88 | 35 |
| Educational Attainment |  |  |
| Less than high school | 165 (14%) | 208 (52%) |
| High school or GED | 288 (24%) | 99 (25%) |
| More than high school | 736 (62%) | 95 (24%) |
| Unknown | 0 | 1 |
| Cotinine, Serum (ng/mL) | 0 (0, 0) | 0 (0, 1) |
| Unknown | 42 | 18 |
| Blood lead (ug/dL) | 1.31 (0.95, 1.99) | 1.44 (1.01, 2.39) |
| Unknown | 628 | 191 |
| Blood cadmium (ug/L) | 0.34 (0.23, 0.56) | 0.40 (0.25, 0.72) |
| Unknown | 628 | 191 |
| 1n (%); Median (IQR) | | |

Interpret the findings: The median age in years among those with mild cognitive impairment is 72, which is higher than the median age in years among those without impairment (68 years).

The median blood lead level in ug/dL among those with mild cognitive impairment is 1.435, which is higher than the median age in years among those without impairment (1.31 ug/dL).