Descriptive Analysis of a Multilevel Data Set

Katherine Wolf

February 20, 2020

Report

n = 4000 participants, for whom 40 were missing age, 112 were missing race, 77 were misisng education, and 580 were missing data on income.

```
knitr::opts_chunk$set(echo = FALSE,
                      warning = FALSE,
                      message = FALSE)


# need to call the libraries every time you begin a new R session
library(dplyr)
library(ggplot2)
library(nnet)
library(tidyverse)
library(tableone)
library(xtable)
library(knitr)
library(tableone)
library(kableExtra)

# read in data - suppose the file dataset.csv contains continuous variables var1 and var2, and a binary
nyses_raw_data <- read_csv("NYSES data for class.csv")

# make working data file
nyses_to_edit <- nyses_raw_data

# id
nyses_to_edit$id <- nyses_to_edit$QKEY2

# community district
nyses_to_edit$`Community district` <- nyses_to_edit$cd

# neighborhood median income
nyses_to_edit$`Neighborhood median income ($)` <- nyses_to_edit$medinc

# neighborhood percent below poverty
```

Table 1: Descriptive statistics for participants in NYSES cohort, stratified by binge drinking.

	Level	No binge drinking	Binge drinking
n		3562	438
Neighborhood median income ($) (mean (SD))		40139.3 (14740.2)	42332.7 (16865.5)
Neighborhood poverty (%) (mean (SD))		20.8 (10.6)	20.6 (11.3)
Borough (%)	Bronx	550 (15.4)	58 (13.2)
	Brooklyn	1079 (30.3)	119 (27.2)
	Manhattan	707 (19.8)	139 (31.7)
	Queens	1036 (29.1)	98 (22.4)
	Staten Island	190 (5.3)	24 (5.5)
Age (years) (%)	18-24	284 (8.0)	66 (15.1)
	25-34	540 (15.2)	145 (33.1)
	35-44	711 (20.0)	104 (23.7)
	45-54	751 (21.1)	57 (13.0)
	55-64	570 (16.0)	42 (9.6)
	65+	667 (18.7)	23 (5.3)
	NA	39 (1.1)	1 (0.2)
Race/ethnicity (%)	White	1382 (38.8)	234 (53.4)
	African American	968 (27.2)	87 (19.9)
	Asian	144 (4.0)	20 (4.6)
	Hispanic/Latinx	877 (24.6)	81 (18.5)
	Other	85 (2.4)	10 (2.3)
	NA	106 (3.0)	6 (1.4)
Education (%)	Less than high school	467 (13.1)	41 (9.4)
	High school/GED	836 (23.5)	87 (19.9)
	Some college	781 (21.9)	98 (22.4)
	College graduate	758 (21.3)	125 (28.5)
	Graduate work	647 (18.2)	83 (18.9)
	NA	73 (2.0)	4 (0.9)
Income ($) (%)	>= 40,000	1460 (41.0)	145 (33.1)
	40,001 to 80,000	938 (26.3)	155 (35.4)
	> 80,000	612 (17.2)	110 (25.1)
	NA	552 (15.5)	28 (6.4)

```r
nyses_to_edit$`Neighborhood poverty (%)` <- nyses_to_edit$pbelowpv

# borough
borough_labels <-
  c("Bronx",
    "Brooklyn",
    "Manhattan",
    "Queens",
    "Staten Island") # (value = order)

nyses_to_edit$Borough <-
  factor(nyses_to_edit$boro,
         labels = borough_labels)

# age
age_labels <-
  c("18-24",
    "25-34",
    "35-44",
    "45-54",
    "55-64",
    "65+") # (value = order)

nyses_to_edit$`Age (years)` <-
  factor(nyses_to_edit$agecat,
         labels = age_labels)

# race/ethnicity
race_labels <-
  c("White",
    "African American",
    "Asian",
    "Hispanic/Latinx",
    "Other") # (value = order)

nyses_to_edit$`Race/ethnicity` <-
  factor(nyses_to_edit$racecat, labels = race_labels)

# education
ed_labels <-
  c("Less than high school",
    "High school/GED",
    "Some college",
    "College graduate",
    "Graduate work") # (value = order)

nyses_to_edit$`Education` <-
  factor(nyses_to_edit$edcat, labels = ed_labels)

# income
income_labels <-
  c(">= 40,000",
    "40,001 to 80,000",
    "> 80,000") # (value = order)

nyses_to_edit$`Income ($)` <-
  factor(nyses_to_edit$inc3cat, labels = income_labels)

# binge drinking
binge_labels <-
  c("No",
    "Yes") # (value = order)

nyses_to_edit$`Binge drinking` <-
  factor(nyses_to_edit$binge,
         labels = binge_labels)

nyses_analyze <-
  nyses_to_edit %>%
  select(id,
         `Community district`,
         `Neighborhood median income ($)`,
         `Neighborhood poverty (%)`,
         Borough,
         `Age (years)`,
         `Race/ethnicity`,
         Education,
         `Income ($)`,
         `Binge drinking`)


# create a list of variables for the table
# (not including the stratification variable)
table_one_variables <- c("Neighborhood median income ($)",
                         "Neighborhood poverty (%)",
                         "Borough",
                         "Age (years)",
                         "Race/ethnicity",
                         "Education",
                         "Income ($)")

# create a list of which ones are categorical (factor)
factor_variables <- c("Borough",
                      "Age (years)",
                      "Race/ethnicity",
                      "Education",
                      "Income ($)")

table_1 <- CreateTableOne(vars = table_one_variables,
                          factorVars = factor_variables,
                          strata = "Binge drinking",
                          data = nyses_analyze,
                          test = FALSE,
                          includeNA = TRUE)

save(table_1,
     file = "table_1.rdata")

# print(table.1) # Standard output

# Creates a formatted table, using kable from the knitr package
# Would want to clean this up for publication purposes:
hi <- kable(print(table_1,
                  showAllLevels = TRUE,
                  printToggle = FALSE,
                  noSpaces = TRUE,
                  catDigits = 1,
                  contDigits = 1),
            col.names = c("Level", "No binge drinking", "Binge drinking"),
            caption=paste("Descriptive statistics for participants in",
                          "NYSES cohort, stratified by binge drinking."))


# hi <- kable(print(table_1,
#                     showAllLevels = TRUE,
#                     printToggle = FALSE,
#                     noSpaces = TRUE,
#                     catDigits=1,
#                     contDigits=1),
#       caption=paste("Descriptive statistics for participants in",
#                      "NYSES cohort, stratified by binge drinking."))

hi


# check for missing data (nas)
na_counts <-
  map(nyses_to_edit,  # cycles through all variables
      function(x) sum(is.na(x)))  # sums all "T" values from nas
na_counts

range(nyses_to_edit$pbelowpv)
ggplot(data = nyses_to_edit, aes(x = binge, y = pbelowpv)) +
  geom_boxplot()

range(nyses_to_edit$cd)
ggplot(data = nyses_to_edit, aes(x = binge, y = cd)) +
  geom_boxplot()



range(nyses_to_edit$medinc)
ggplot(data = nyses_to_edit, aes(x = binge, y = medinc)) +
  geom_boxplot()



# # load packages used in this assignment
# # only need to install packages once
# install.packages("dplyr")
# install.packages("ggplot2")
# install.packages("nnet")

# need to call the libraries every time you begin a new R session
library(dplyr)
library(ggplot2)
library(nnet)

# read in data - suppose the file dataset.csv contains continuous variables var1 and var2, and a binary
df <- read.csv("c:/dataset.csv")

# create a categorical variable from a continuous variable
df$catvar1 <- df$var1
df$catvar1 <- ifelse(df$catvar1<=500,0,ifelse(df$catvar1>500,1,NA))

# describe variables
summary(df$var1)
table(df$var1)
hist(df$var1)

with(df, table(var1,var2))
with(df, table(var1,var2, exclude=NULL))

df %>%
  group_by(catvar1) %>% summarise(mean_outcome = mean(outcome))


# bivariable relations

# to do a Pearson's chi-squared test
x2 <- chisq.test(df$var1, df$var2)

# to see the results of the test
x2

# to see the table of observed
x2$observed

# to see the percents by row
prop.table(x2$observed, 1)

# to see the percents by column
prop.table(x2$observed, 2)

# calculate correlation between variables
cor.test(df$var1, df$var2)

# summarize relationship between variables in a plot with a lowess line
ggplot(df) + geom_point(aes(x=var1, y=outcome))  +
  geom_line(aes(x=var1, y=predict(loess(outcome~var1)))) + theme_bw()
```