Assignment 1

grienne

February 26, 2019

Packages

```
# specify the packages of interest
packages = c("epiR", "tidyverse", "survival", "readr", "tableone",
    "purrr", "lattice")

# use this function to check if each package is on the local
# machine if a package is installed, it will be loaded if any
# are not, the missing package(s) will be installed and
# loaded
package.check <- lapply(packages, FUN = function(x) {
    if (!require(x, character.only = TRUE)) {
        install.packages(x, dependencies = TRUE)
        library(x, character.only = TRUE)
    }
})
```

Loading Data

Loading Data in R can be intensive, but luckily RStudio makes this fairly easy!

The link below is a website that will guide you through that process in RStudio.

Loading Data: [link]https://support.rstudio.com/hc/en-us/articles/218611977-Importing-Data-with-RStudio

When you load a data file, I encourage you to rename it to something clear and short! The code below does 2 things: 1. I take the loaded dataset rename it to dat1 'dat1 <- academic_dataset_arc.csv' *Note: You can rename it however you want* 2. I take a quick look at the dataset to make sure it looks okay 'head(dat1)'

```
dat1 <- read_csv("academic_dataset_aric.csv")
head(dat1)
names(dat1) <- tolower(names(dat1))

# I removed the enddate and dated11 column as it lengthened
# the output significantly
drop.cols <- c("c7_enddate", "dated11")
dat2 <- dat1 %>% select(-one_of(drop.cols))

dat3 <- dat2 %>% select(v1age01, bmi01, gender, racegrp, hyptmd01,
    diabts03, hdlsiu02, ldlsiu02, tchsiu01, trgsiu01, fast0802,
    cigt01, prvchd05, center)
```

Question 1 - Descriptive Statistics

The code below generates descriptive statistics for the dataset we created above, including means, medians, range, and modes.

Also identifies number of NA observations.

```r
summary(dat3)
```

Question 1a - Race Removal

```r
# removed races A & I
dat4 <- dat3 %>% filter(racegrp != "A" & racegrp != "I")

# turning Racegrp into a factor (i.e. making it categorical)
dat4$racegrp <- as.factor(dat4$racegrp)

# We have removed A&I so when we do summary of dat and just
# check the Racegrp there should only be Black and White
summary(dat4$racegrp)

# Making Center a Factor
dat4$center <- as.factor(dat4$center)

# Summarize data by Center This produces descriptive
# statistics by Center. Including Race Counts
dat4 %>% split(.$center) %>% map(summary)

# remove D & B black participants


# Code Tip: Sometimes you have to effect change on an
# existing dataset. However, if you aren't sure the code will
# work, I will often create a subset of it
dat_r <- dat4[!(dat4$center == "B" & dat4$racegrp == "B"), ]
dat_r <- dat4[!(dat4$center == "D" & dat4$racegrp == "B"), ]

# Removed 6 blacks from Center B & 13 from center D I
# reassign dat_r to dat4 once I have verified I am okay with
# the outcome

dat4 <- dat_r

summary(dat4)

# Histogram is made using the below command.  Tips To make a
# histogram dat3$ and select the variable of interest If you
# want to make a histogram by category the first variable is
# the continuous variable, the second is the categorical
# variable
histogram(~dat4$hdlsiu02 | dat4$racegrp)
```

Question 1 - Race Center

Here I am creating the race center variable

```r
# Create Race Center
dat5 <- dat4 %>% mutate(race.center = ifelse(racegrp == "W" &
    center == "A", 1, ifelse(racegrp == "B" & center == "A",
    2, ifelse(racegrp == "W" & center == "B", 3, ifelse(racegrp ==
```

```
        "W" & center == "D", 4, ifelse(racegrp == "B" & center ==
        "C", 5, NA))))))

# Factor Turning multiple variables into factors/categories
varstofactor <- c("gender", "racegrp", "cigt01", "prvchd05",
    "diabts03")
dat5[varstofactor] <- lapply(dat5[varstofactor], factor)

# Summary I am summarizing dat 5, then summarizing specific
# specific variables from data 5
dat5$race.center <- as.factor(dat5$race.center)
summary(dat5)
summary(dat5$hyptmd01)
summary(dat5$race.center)
```

Question 2 - A: Table One

Goal is to create a a table one with selected variables

```
# I want to quick select the variables from data 5 to create
# my data table.  I use the code, then copy paste the output
# to create vars
dput(names(dat5))


vars <- c("v1age01", "bmi01", "gender", "racegrp", "hyptmd01",
    "diabts03", "hdlsiu02", "ldlsiu02", "tchsiu01", "trgsiu01",
    "fast0802", "cigt01", "prvchd05", "center", "race.center")

#-------------------

# table one is being created by using data 5 and creating a
# table one that incorporates the variables selected above.
tableoneA <- CreateTableOne(vars = vars, data = dat5)

tableoneA
```

Table as Above stratified by Diabetes Status

```
tableoneB <- CreateTableOne(vars = vars, strata = c("diabts03"),
    data = dat5)

tableoneB
```

Question 4 - Complete Case Count

Here I count the complete cases then I sum the cases that aren't complete.

```
ok <- complete.cases(dat4)

sum(!ok)
```