Math 207
Introduction to Statistics

NHANESIII Data Set: Report I

— Due Friday 27 September 2019 —

In Report 0 you were asked to write your first report using R Markdown. In another exercise, you downloaded the NHANES III data set and code books then read the data into R.

1. **Start Your Report**. Open a new RMarkdown file (File → New File → RMarkdown). Delete everything in the template file then type the following.

```
---
title:  "Report_1"
author: "your name"
date:   "23 September 2019"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
```

2. **Reading the Data into R**.

a) Type the following in your RMarkdown file:

```
## Reading the adult.csv file
The NHANES data consists of four data files:  adult.csv, youth.csv, lab.csv and
exam.csv.  The adult file contains information about subjects who were over
17 years old.  The youth file contains information about other subjects.  The
lab and exam files contain additional data about both adults and youth.

### Reading the adult.csv file
The following command reads the adult.csv data file into R.
```{r}
adult = read.csv("/Users/username/Desktop/math207/adult.csv", header=TRUE)
```
The size of this data set is
```{r}
dim(adult)
```
This means that the file has 20,050 rows and 1238 columns. So, there are
20,050 adults in the data set and 1238 pieces of information about each in
this file. The lab.csv and exam.csv files include additional information
about these subjects.
```

It is a good idea to knit your file frequently to catch errors early.

b) Repeat part a) using the `youth.csv` file. Your new variable should be called `youth`.

c) Repeat part a) using the `lab.csv` file. Your new variable should be called `lab`.

d) Repeat part a) using the `exam.csv` file. Your new variable should be called `exam`.

3. **First Explorations**.

a) Add the following to your RMarkdown file.

````
## First Explorations
The first variable of interest is SEQN.  It stores the anonymous ID
numbers for the subjects.  Each subject has data stored in multiple files.
SEQN allows us to merge this disparate information for participants of
interest. Here are the first few ID numbers.
```{r}
head(adult$SEQN)
```
The following tells us that there is only one row in adult.csv for
subject number 4.
```{r}
dim(adult[adult$SEQN == 4,])
```
````

b) Add the following to your RMarkdown file.

````
### Variable DMARACER in the adult data file
The variable DMARACER gives information about the race of the subjects.
It is a qualitative, unordered variable but is recorded as a number:
1 = white, 2 = black, 3 = other and 8 = Mexican-American of unknown race.
We can confirm the distribution of DMARACER as follows:
```{r}
table(adult$DMARACER)
```
This indicates that there were 13,738 white subjects, 5664 black subjects,
640 classified as other, and 8 classified as Mexican-American of unknown race.
It agrees with page 67 of the ADULT-acc.pdf code book.
````

c) Repeat b) using the variable `DMARETHN`.

d) Repeat b) using the variable `DMAETHNR`.

e) Repeat b) using the variable `HSSEX`.

f) Repeat b) using the variable `HSFSIZER`.

4. **Descriptive Statistics**. Add the following to your RMarkdown file.

```
## Descriptive Statistics
There are two facets to statistics:  descriptive statistics and inferential
statistics.   Descriptive statistics involves organizing, summarizing and
visualizing data.  Inferential statistics involves making predictions.
```

a) Add the following to your RMarkdown file.

```
### Variable BMPWT in the exam data file
Subjects' body measurements were taken in a home exam.  Variable BMPWT
records subjects' weights in kg.  Some subjects did not have their weight
recorded, however.  The BMPWT variable is recorded as 888888 for them.
See page 198 of the exam-acc.pdf code book.  So, if we compute the average
weight without accounting for that fact:
```{r}
mean(exam$BMPWT)
```
we get 5135.477 kg. That is a bit high for human patients. We can extract
the true weights and compute some statistics as follows:
```{r}
e1 = exam[exam$BMPWT != 888888, ]
weights = e1$BMPWT
mean(weights)
median(weights)
sd(weights)
hist(weights, xlab="Weight (kg)", ylab="Percent per kg", col='blue', probability=T)
```
The histogram is not symmetric. That is why the mean and median differ.
```

b) Repeat a) using the variable `BMPHT` from the exam data file. Note: the error code for `BMPHT` is 88888 (not 888888). Also replace

```
weights = e1$BMPWT
```

with the line

```
heights = na.omit(e1$BMPHT)
```

c) Repeat b) using the variable `BMPRECUM` from the exam data file.