Math 207
Introduction to Statistics

First Experiments with the NHANESIII Data Set

For much of this course we will explore the NHANESIII data set in order to illustrate the concepts that we cover. This is a large survey of health statistics of U.S. residents. The data set website is:

https://wwwn.cdc.gov/nchs/nhanes/nhanes3/datafiles.aspx

The data consists of four large files. On the CDC website these are in a format developed for the SAS statistical software package. I had these converted to csv (comma-separated-value) text files then put them in a zip file which is on our course website.

1. **Download the data set**.
   - Create a folder called `math207` on your desktop.
   - Download the `NHANESIII.zip` file from Canvas.
   - Move the `NHANESIII.zip` file into your `math207` directory.
   - Open a terminal and type the following:
     ```
     cd Desktop
     cd math207
     unzip NHANESIII.zip
     ```

The `math207` folder should contain the following files:

Data File	Contents
adult.csv	survey information about adults subjects
youth.csv	survey information about youth subjects
lab.csv	blood and urine test data
exam.csv	large collection of medical exam data
ADULT-acc.pdf	information file for the adult.csv data file
YOUTH-acc.pdf	information file for the youth.csv data file
lab-acc.pdf	information file for the lab.csv data file
exam-acc.pdf	information file for the exam.csv data file

2. **Experimenting with the Adult Survey Data**. Open RStudio and the ADULT-acc.pdf code book (statisticians' term for a document that describes what is in a data set). Try the following in RStudio. Explain to yourself what `R` thinks it is doing.

```
# First, load the data <- This is a comment, you don't have to type it.
# USE YOUR USERNAME not the word "username".
adult = read.csv("/Users/username/Desktop/math207/adult.csv", header=TRUE)
class(adult)

# This will tell you the number of rows (20,050) and columns (1238).
dim(adult)

# Here are the 1238 (cryptic) variable names.  They are desribed briefly
# on pages 30 -- 66 of the code book.  Pages 66 -- 391 give some basic
# statistics for each variable.  The rest of the document gives more
# detailed descriptions.
names(adult)

# The first variable of interest is SEQN.  It stores the anonymous ID
# numbers for participants.  Each subject has data stored in multiple files.
# SEQN allows us to collect such disparate information for participants.
head(adult$SEQN)
```

```
# The following tells us that there is only one row in adult.csv for
# subject number 4.
dim(adult[adult$SEQN == 4,])

# Confirm the range stated on page 66 of the code book.
range(adult$SEQN)

# Confirm the race distribution stated on page 67 of the code book.
table(adult$DMARACER)
```

3. Continue using the adult survey data set in the following.

a) Confirm the sex distribution stated on page 68 of the code book.

b) Confirm the family size distribution using `HSFSIZER` (page 70).

c) Make a histogram of the age distribution.

```
hist(adult$HSAGEIR, breaks=seq(15,95,by=5), xlab="Age",
     ylab="Percent per Year", probability=TRUE)
```

d) Confirm the census region distribution using `DMPCREGN` (page 71) then make a barplot.

```
barplot(table(adult$DMPCREGN), names.arg=c("Northeast", "Midwest", "South", "West"))
```

e) Were the days of the week of the survey exams fairly uniform? Use variable `MXPTIDW`. If not, can you suggest a reason?

f) What percent of respondents worked for the federal government? Use variable HFD12 (page 124).

g) What percent of respondents worked for the either federal, state or local government? Again, use variable HFD12 (page 124).

h) Did more respondents own cats or fish? Use variables HFE8B and HFE8D.

i) Is it possible that more respondents owned either a bird or fish (or both) than owned cats? Use variables HFE8B and HFE8D.

4. The variables on pages 131 – 132 give information about smoking habits in subjects' homes. Let's construct a barplot of the number of cigarettes a day smoked by person 1 in subjects' homes. Do the following in RStudio.

```
cigs1  = adult$HFF3A
cigs1b = cigs1[cigs1 < 800]  # What are responses 888 and 999?
table(cigs1b)                # What does the value 777 mean?

cigs1b[cigs1b == 777] <- 0   # Look at the table below to see what this does.
table(cigs1b)

barplot(table(cigs1b))       # Now make the barplot.
```

5. **Experimenting with the Exam Data**. Open the `exam-acc.pdf` code book. In RStudio, load the exam data set.

```
exam = read.csv("/Users/username/Desktop/math207/exam.csv", header=TRUE)

# It's a data frame too.
class(exam)

# How big is it?
dim(exam)

# Subject ID numbers are stored in both files.
head(adult$SEQN)
head(exam$SEQN)

# Here is survey information for subject 11.
adult[adult$SEQN == 11, ]

# Here is exam information for that subject.
exam[exam$SEQN == 11, ]



# How tall (in cm) was subject 11? See page 199 of the code book
exam[exam$SEQN==11, ]$BMPHT

# How much did that subject weigh (in kg)?  See page 198.
exam[exam$SEQN==11, ]$BMPWT

# Let's examine the variation of weight with height.
# Try the following then explain what goes wrong.
plot(exam$BMPHT, exam$BMPWT, xlab="height (cm)", ylab="weight (kg)")

# We need to extract the blank values.
w1 = exam[(exam$BMPWT != 888888) & (exam$BMPHT != 88888), ]$BMPWT
h1 = exam[(exam$BMPWT != 888888) & (exam$BMPHT != 88888), ]$BMPHT

plot(h1, w1, xlab="height (cm)", ylab="weight (kg)")
```

Does weight increase with height? Is the variation linear? That is, does a straight line fit the data well?