Report_1

Me

23 September 2019

Reading the data into R

The NHANES 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("../nhanes/adult.csv", header=TRUE)
```

The size of this data set is

```r
dim(adult)
```

```
## [1] 20050  1238
```

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 include additional information about these subjects.

Reading the exam.csv file

The following command reads the exam.csv data file into R.

```r
exam = read.csv("../nhanes/exam.csv", header=TRUE)
```

The size of this data set is

```r
dim(exam)
```

```
## [1] 31311  2368
```

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)
```

```
## [1]  3  4  9 10 11 19
```

The following tells us that there is only one row in adult.csv for subject number 4.

```r
dim(adult[adult$SEQN == 4,])
```

```
## [1]    1 1238
```

Variable DMARACER in the adult data set

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:

```
table(adult$DMARACER)
```

```
## 
##     1     2     3     8 
## 13738  5664   640     8
```

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.

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 these subjects. So, if we compute the average weight without accounting for that fact:

```
mean(exam$BMPWT)
```

```
## [1] 5135.447
```

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:

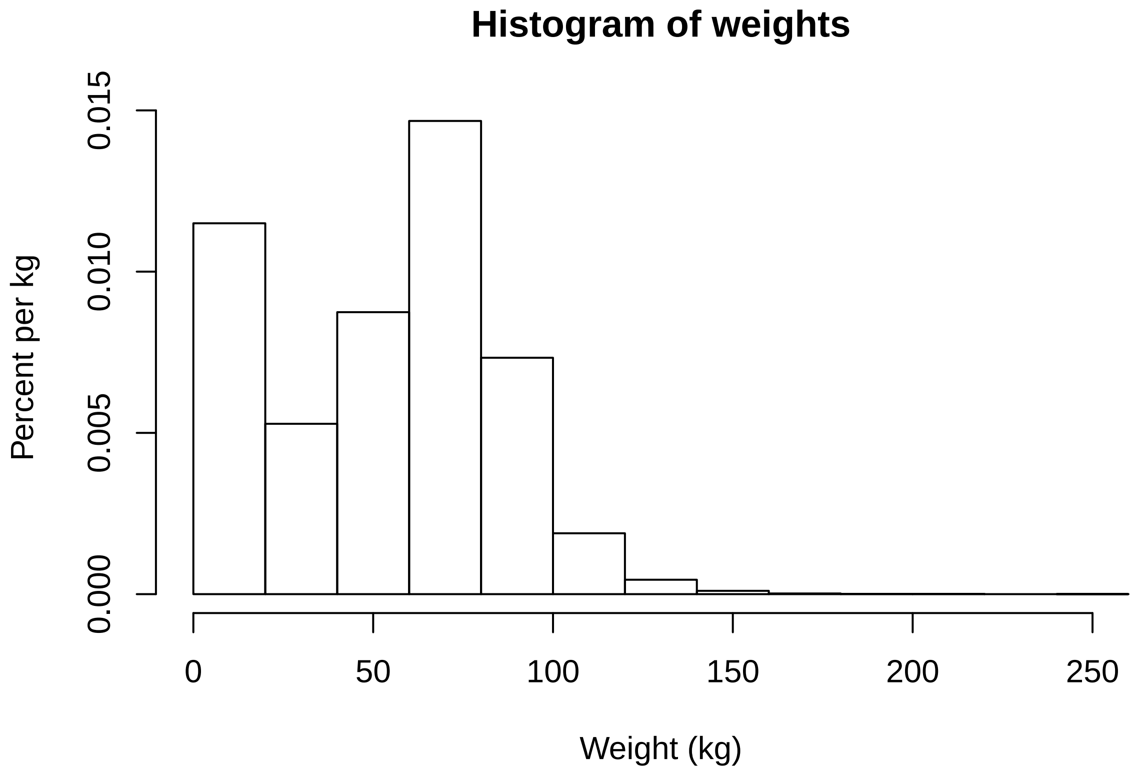
```
e1 = exam[exam$BMPWT != 888888, ]
weights = e1$BMPWT
mean(weights)
```

```
## [1] 54.1257
```

```
sd(weights)
```

```
## [1] 30.30456
```

```
hist(weights, xlab="Weight (kg)", ylab="Percent per kg", probability=T)
```


Histogram of weights
Percent per kg
0.000
0.005
0.010
0.015
0
50
100
150
200
250
Weight (kg)