Hao-1

Bruce Hao

September 2, 2016

Setup lab.

```
library(IS606)
source("more/cdc.R")
```

Exercise 1: How many cases are there in this data set? How many variables? For each variable, identify its data type (e.g. categorical, discrete).

```
str(cdc)
```

```
## 'data.frame':    20000 obs. of  9 variables:
##  $ genhlth : Factor w/ 5 levels "excellent","very good",..: 3 3 3 3 2 2 2 2 3 3 ...
##  $ exerany : num  0 0 1 1 0 1 1 0 0 1 ...
##  $ hlthplan: num  1 1 1 1 1 1 1 1 1 1 ...
##  $ smoke100: num  0 1 1 0 0 0 0 0 1 0 ...
##  $ height  : num  70 64 60 66 61 64 71 67 65 70 ...
##  $ weight  : int  175 125 105 132 150 114 194 170 150 180 ...
##  $ wtdesire: int  175 115 105 124 130 114 185 160 130 170 ...
##  $ age     : int  77 33 49 42 55 55 31 45 27 44 ...
##  $ gender  : Factor w/ 2 levels "m","f": 1 2 2 2 2 2 1 1 2 1 ...
```

As shown above, there are 20,000 observation and 9 variables with the following data types:

- genhlth: categorical ordinal
- exerany: categorical
- hlthplan: categorical
- smoke100: categorical
- height: numeric continuous (discrete when captured in whole units)
- weight: numeric continuous (discrete when captured in whole units)
- wtdesire: numeric continuous (discrete when captured in whole units)
- age: numeric continuous (discrete when captured in whole units)
- gender: categorical

Exercise 2: Create a numerical summary for height and age, and compute the interquartile range for each. Compute the relative frequency distribution for gender and exerany. How many males are in the sample? What proportion of the sample reports being in excellent health?

Summary and relative frequency distribution for height:

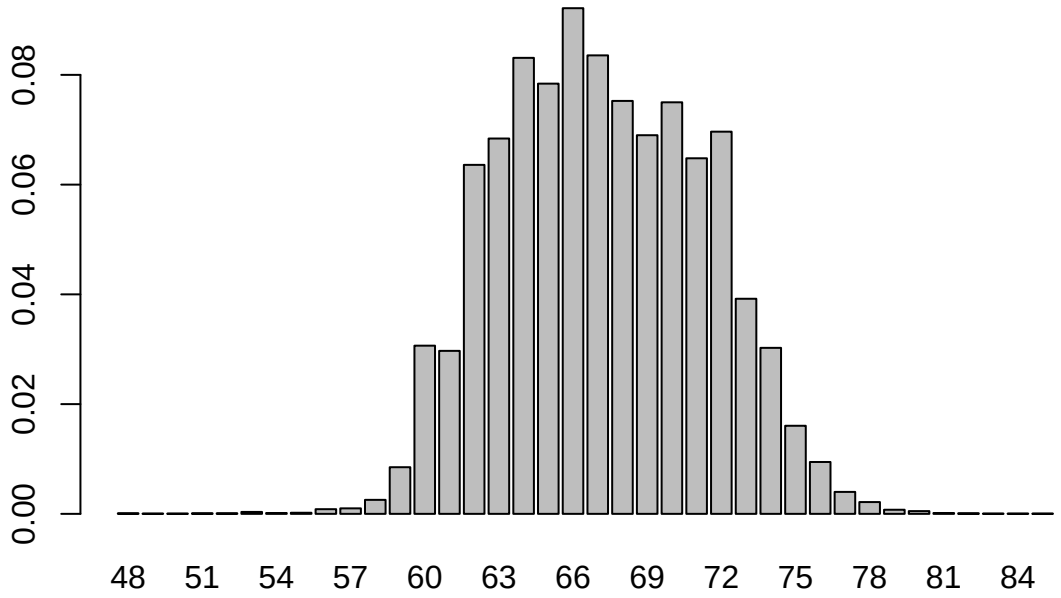
```
summary(cdc$height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   48.00   64.00   67.00   67.18   70.00   93.00
```

```
prop.table(table(cdc$height))
```

```
##
##      48      49      50      51      52      53      54      55      56
## 0.00010 0.00005 0.00005 0.00010 0.00010 0.00035 0.00015 0.00020 0.00085
##      57      58      59      60      61      62      63      64      65
## 0.00100 0.00255 0.00850 0.03065 0.02970 0.06360 0.06840 0.08310 0.07840
##      66      67      68      69      70      71      72      73      74
## 0.09215 0.08355 0.07525 0.06900 0.07500 0.06480 0.06965 0.03920 0.03025
##      75      76      77      78      79      80      81      82      83
## 0.01605 0.00945 0.00400 0.00215 0.00075 0.00050 0.00015 0.00010 0.00005
##      84      93
## 0.00005 0.00005
```

```
barplot(prop.table(table(cdc$height)))
```



Summary relative frequency distribution for age:

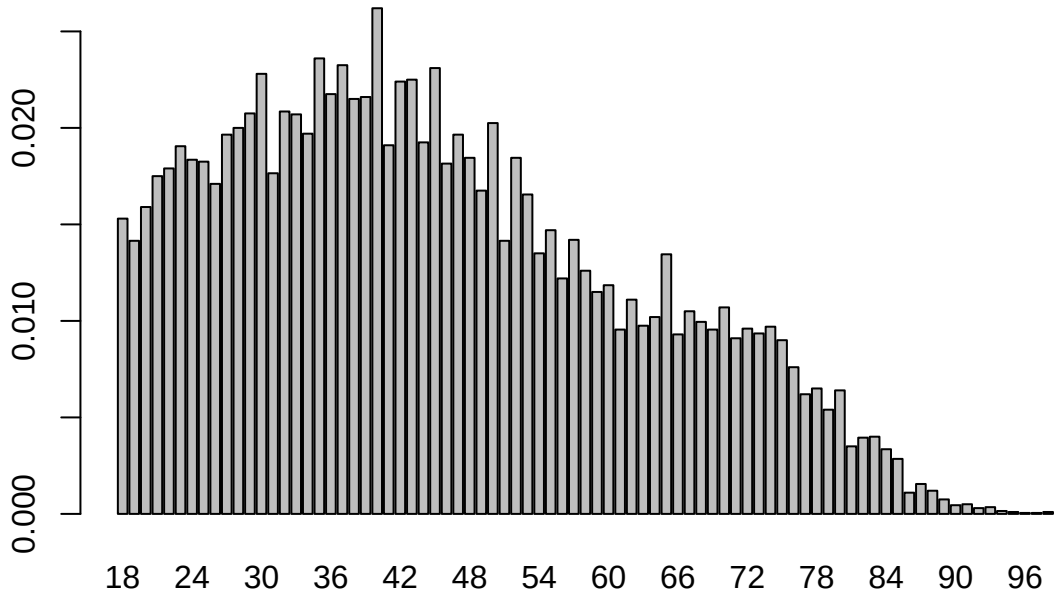
```
summary(cdc$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   31.00   43.00   45.07   57.00   99.00
```

```
prop.table(table(cdc$age))
```

```
##
##      18      19      20      21      22      23      24      25      26
## 0.01530 0.01415 0.01590 0.01750 0.01790 0.01905 0.01835 0.01825 0.01710
##      27      28      29      30      31      32      33      34      35
## 0.01965 0.02000 0.02075 0.02280 0.01765 0.02085 0.02070 0.01970 0.02360
##      36      37      38      39      40      41      42      43      44
## 0.02175 0.02325 0.02150 0.02160 0.02620 0.01910 0.02240 0.02250 0.01925
##      45      46      47      48      49      50      51      52      53
## 0.02310 0.01815 0.01965 0.01845 0.01675 0.02025 0.01415 0.01845 0.01655
##      54      55      56      57      58      59      60      61      62
## 0.01350 0.01470 0.01220 0.01420 0.01260 0.01150 0.01185 0.00955 0.01110
##      63      64      65      66      67      68      69      70      71
## 0.00975 0.01020 0.01345 0.00930 0.01050 0.00995 0.00955 0.01070 0.00910
##      72      73      74      75      76      77      78      79      80
## 0.00960 0.00935 0.00970 0.00900 0.00760 0.00620 0.00650 0.00540 0.00640
##      81      82      83      84      85      86      87      88      89
## 0.00350 0.00395 0.00400 0.00335 0.00285 0.00110 0.00155 0.00120 0.00075
##      90      91      92      93      94      95      96      97      99
## 0.00045 0.00050 0.00030 0.00035 0.00015 0.00010 0.00005 0.00005 0.00010
```

```
barplot(prop.table(table(cdc$age)))
```



```
table(cdc$gender)
```

```
##
##     m     f
##  9569 10431
```

There are 9,569 males in the sample.

```
prop.table(table(cdc$genhlth))
```

```
##
## excellent very good      good      fair      poor
##   0.23285   0.34860   0.28375   0.10095   0.03385
```

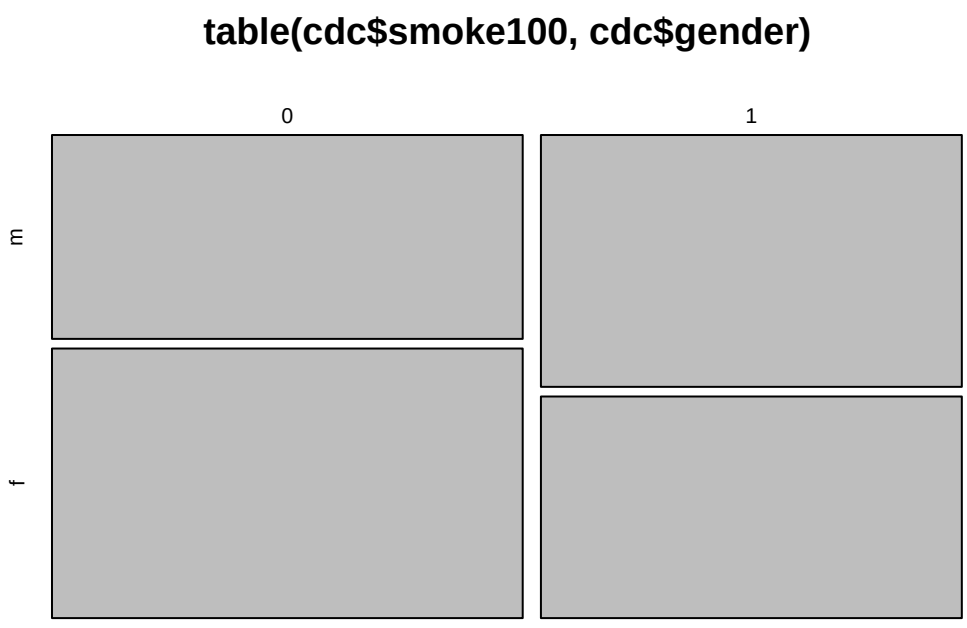
23.3% of the sample reports being in excellent health.

Exercise 3: What does the mosaic plot reveal about smoking habits and gender?

```
prop.table(table(cdc$smoke100, cdc$gender))
```

```
##
##           m       f
##   0 0.22735 0.30060
##   1 0.25110 0.22095
```

```
mosaicplot(table(cdc$smoke100, cdc$gender))
```



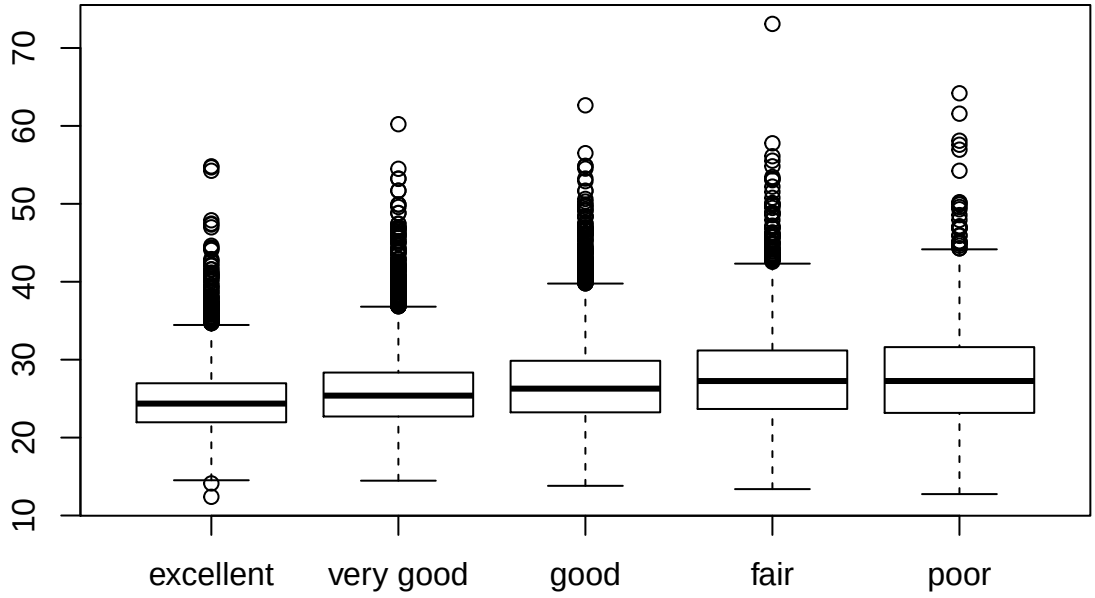
Within the sample, more than half of men reported having smoked (at least a 100 cigarettes); whereas less than half of females reported the same.

Exercise 4: Create a new object called under23_and_smoke that contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime. Write the command you used to create the new object as the answer to this exercise.

```
under23_and_smoke = subset(cdc, age < 23 & smoke100 == 1)
```

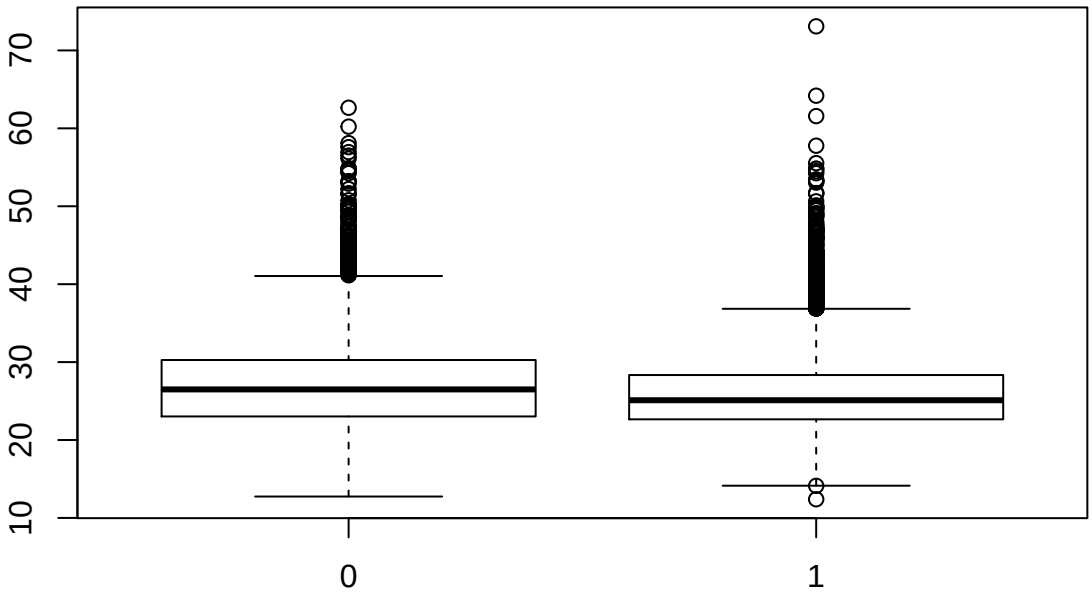
Exercise 5: What does this box plot show? Pick another categorical variable from the data set and see how it relates to BMI. List the variable you chose, why you might think it would have a relationship to BMI, and indicate what the figure seems to suggest.

```
bmi <- (cdc$weight / cdc$height^2) * 703
boxplot(bmi ~ cdc$genhlth)
```



The boxplot shows that BMIs trend upward as health worsens - medians, IQRs and overall distributions increase monotonically as health worsens.

```
boxplot(bmi ~ cdc$exerany)
```



The boxplot above shows that BMIs for those who reported to have exercised over the past month appear to be lower and more tightly distributed than those who did not exercise.