606-01 - Lab 1_Introduction to data

Shyam BV

September 3, 2016

Contents


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

1. How many cases are there in this data set?

```
nrow(cdc)
```

```
## [1] 20000
```

Part 2: How many variables?

```
ncol(cdc)
```

```
## [1] 9
```

Part 3: 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 ...
```

2. Create a numerical summary for `height` and `age`, and compute the interquartile range for each.

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

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

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

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

Part 2: Compute the relative frequency distribution for `gender` and `exerany`.

```
table(cdc$gender)/nrow(cdc)
```

```
##
##       m       f
## 0.47845 0.52155
```

```
table(cdc$exerany)/nrow(cdc)
```

```
##
##      0      1
## 0.2543 0.7457
```

Part 3: How many males are in the sample?

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

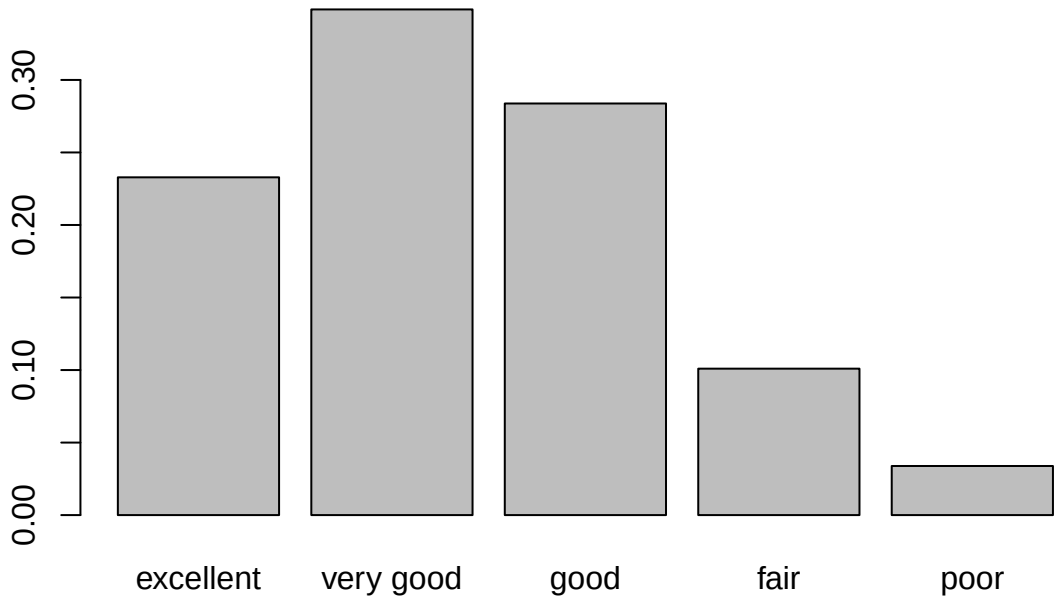
```
##
##     m     f
##  9569 10431
```

Part 4: What proportion of the sample reports being in excellent health?

```
table(cdc$genhlth)/nrow(cdc)
```

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

```
barplot(table(cdc$genhlth)/nrow(cdc))
```

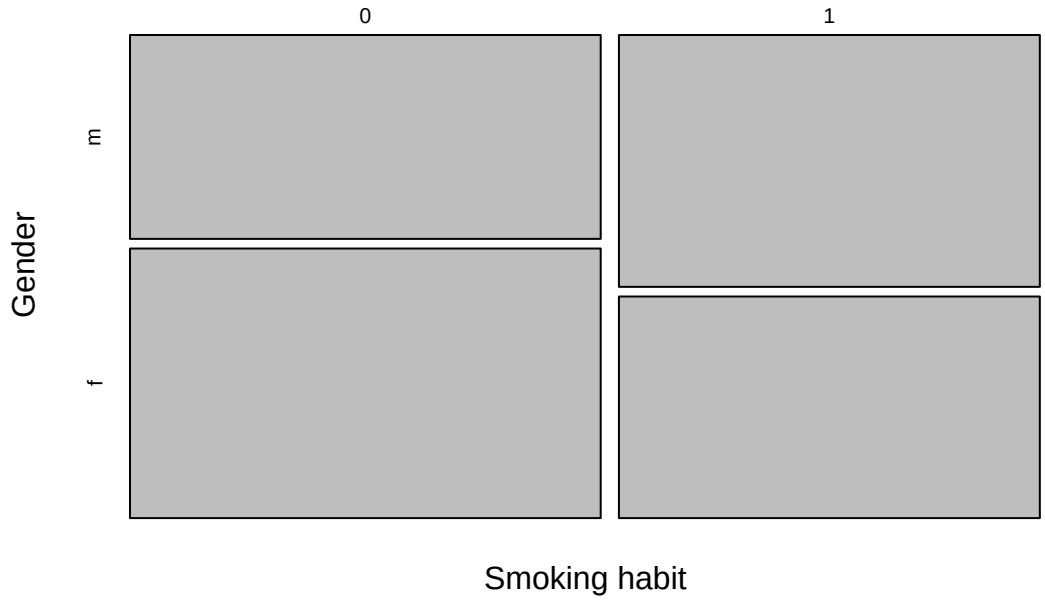


3. What does the mosaic plot reveal about smoking habits and gender?

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

```
##
##        m    f
##   0 4547 6012
##   1 5022 4419
```

```
mosaicplot(table(cdc$smoke100,cdc$gender),main="Smoking habits and gender",xlab="Smoking habit",ylab="G
```



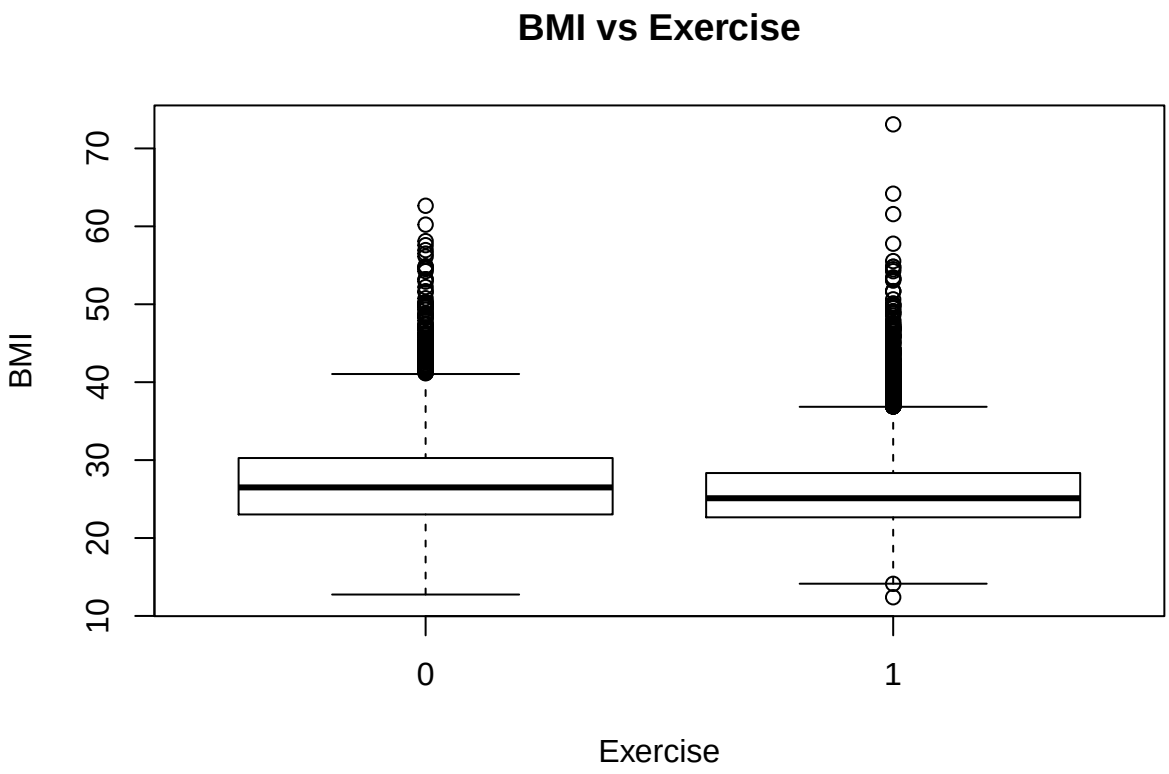
4. Create a new object called `under23_and_smoke` that contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime. Write the command you used to create the new object as the answer to this exercise.

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

```
##       genhlth exerany hlthplan smoke100 height weight wtdesire age gender
## 13  excellent       1        0        1     66    185      220  21      m
## 37  very good       1        0        1     70    160      140  18      f
## 96  excellent       1        1        1     74    175      200  22      m
## 180      good       1        1        1     64    190      140  20      f
## 182 very good       1        1        1     62     92       92  21      f
## 240 very good       1        0        1     64    125      115  22      f
```

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$exerany,main="BMI vs Exercise",ylab="BMI",xlab="Exercise")
```

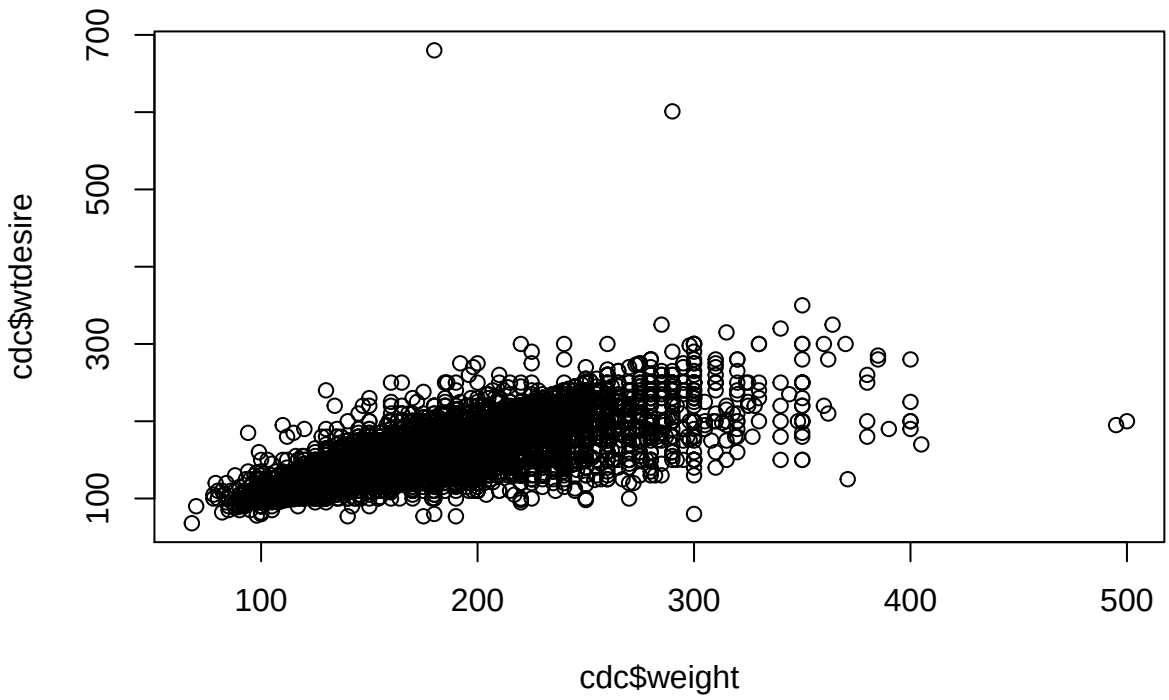


From the above plot and data, it shows that the people who excercise under normal BMI (18.5 - 24.9). People who dont excercise have a higher BMI than normal.

On your own

1. Make a scatterplot of weight versus desired weight. Describe the relationship between these two variables.

```
plot(cdc$weight,cdc$wtdesire)
```



There is a strong corelation between weight vs desired weight. Both are linear.

2. Let's consider a new variable: the difference between desired weight (`wtdesire`) and current weight (`weight`). Create this new variable by subtracting the two columns in the data frame and assigning them to a new object called `wdiff`.

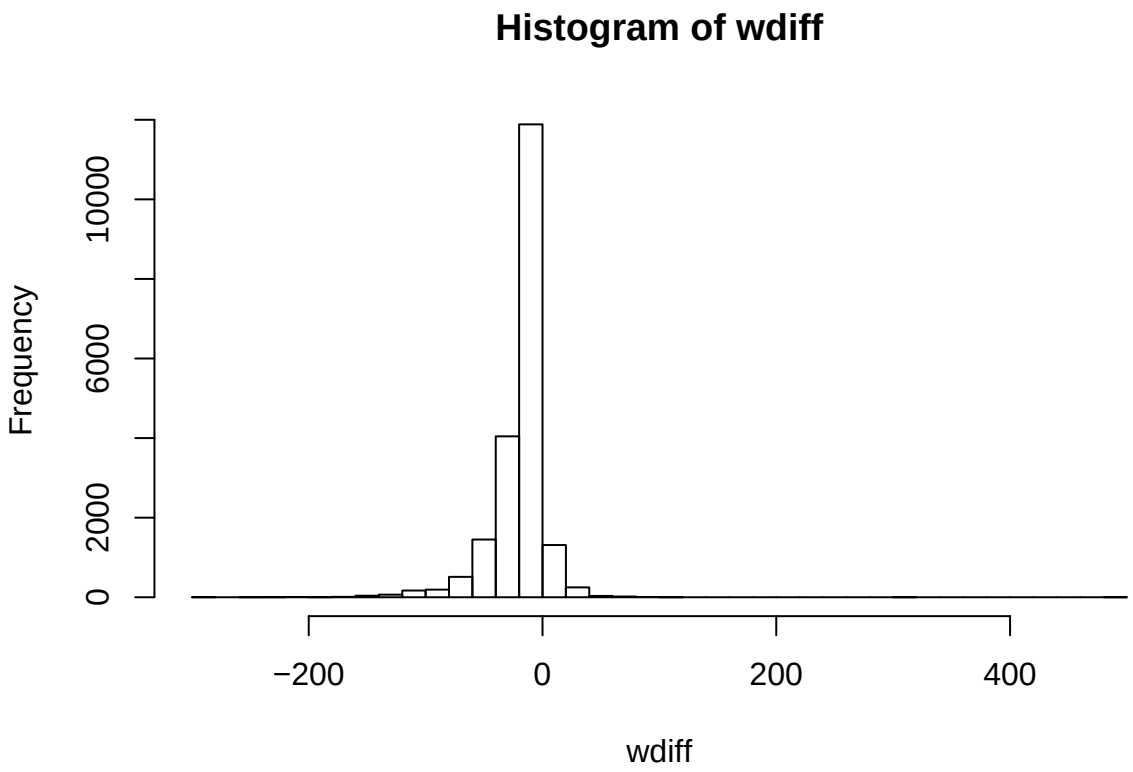
```
wdiff <- cdc$wtdesire - cdc$weight
```

3. What type of data is `wdiff`? If an observation `wdiff` is 0, what does this mean about the person's weight and desired weight. What if `wdiff` is positive or negative?

Answer: If the `widff` is 0, then the desired weight and current weight is equal. So he is on ideal weight. if possitive, then the weight of the person is less. He still need to gain this much weight to be in ideal range. if negative, then the weight of the person is more. He need to reduce this much weight to be in ideal range.

4. Describe the distribution of `wdiff` in terms of its center, shape, and spread, including any plots you use. What does this tell us about how people feel about their current weight?

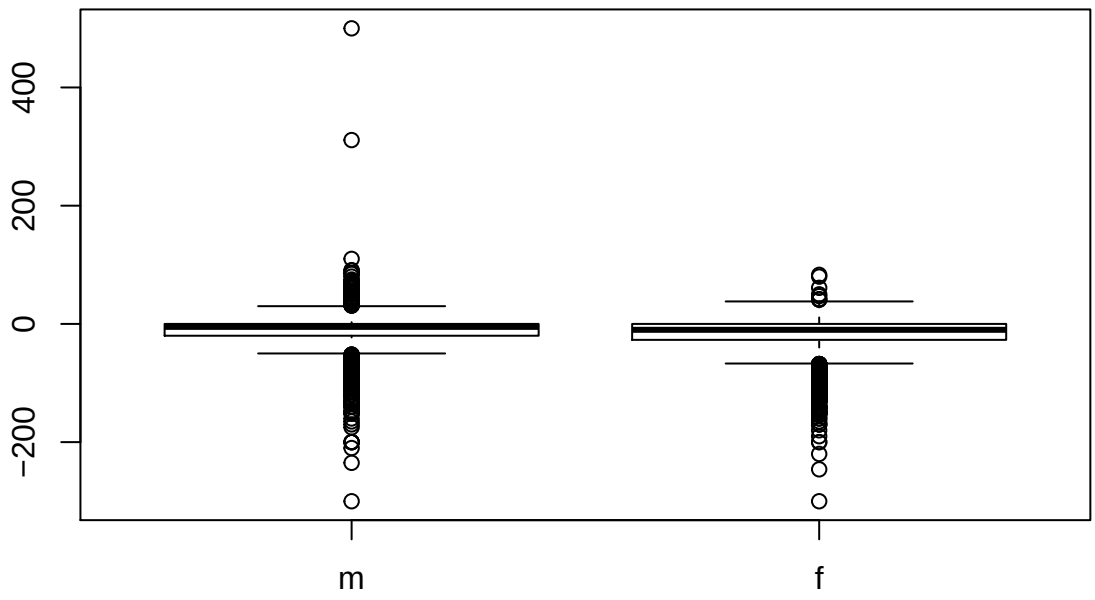
```
hist(wdiff,breaks=50)
```



Above chart shows that the center is around 0 and the shape is a bell curve. Also it is not widely spread. This means that of the people are near their ideal weight.

5. Using numerical summaries and a side-by-side box plot, determine if men tend to view their weight differently than women.

```
cdcc_withdiff <- data.frame(cdc,wdiff)
plot(cdcc_withdiff$gender,cdcc_withdiff$wdiff)
```



The median is almost equal for men and women.

6. Now it's time to get creative. Find the mean and standard deviation of weight and determine what proportion of the weights are within one standard deviation of the mean.

```
cdc_meanweight <- mean(cdc$weight)
cdc_sdweight <- sd(cdc$weight)
nrow(subset(cdc,cdc$weight > (cdc_meanweight - cdc_sdweight) & (cdc_meanweight - cdc_sdweight)))/nrow(c
```

```
## [1] 0.86085
```

Around 86% fall under one Standard deviation.