Week2-Lab2

Amol Gote

9/7/2019

source("http://www.openintro.org/stat/data/cdc.R")

View the names of the variables

names(cdc)

## [1] "genhlth" "exerany" "hlthplan" "smoke100" "height" "weight"   
## [7] "wtdesire" "age" "gender"

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

**Number of cases from the cdc dataset = 20000**

nrow(cdc)

## [1] 20000

**Number of Variables = 9**

ncol(cdc)

## [1] 9

For each variable, identify its data type. genhlth - Categorical

exerany - Categorical

hlthplan - Categorical

smoke100 - Categorical

height - Discrete

weight - Discrete

wtdesire - Discrete

age - Discrete

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?

**Numerical Summary 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

**Numerical summary 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

Interquartile Range = Q3 - Q1 Interquatile range for height: **6**

70.00 - 64.00

## [1] 6

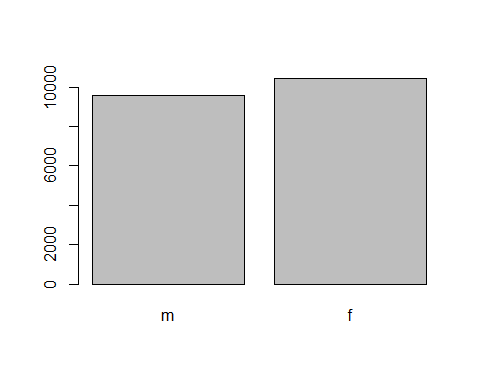
Interquatile range for age : **26**

57.00 -31.00

## [1] 26

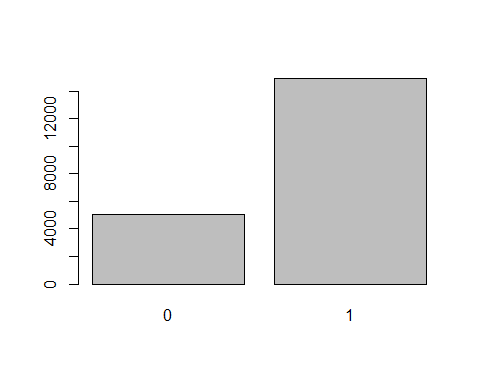
**Relative frequency distribution for gender**

barplot(table(cdc$gender))



Relative frequency distribution for exerany

barplot(table(cdc$exerany))



Number of males in the sample: **9569**

table(cdc$gender)

##   
## m f   
## 9569 10431

Proportion of the sample which report being in excellent health : **0.23** or **23%**

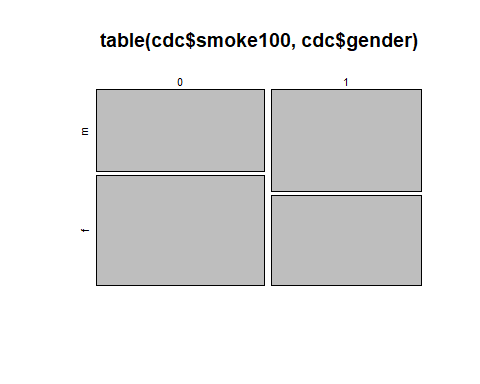
table(cdc$genhlth)/nrow(cdc)

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

**Excercise 3**

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

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



Mosaic plot reveals that there are more females who have **not** smoked 100 cigarettes in thier lifetime. Males who have smoked 100 cigrattes outnumber females.

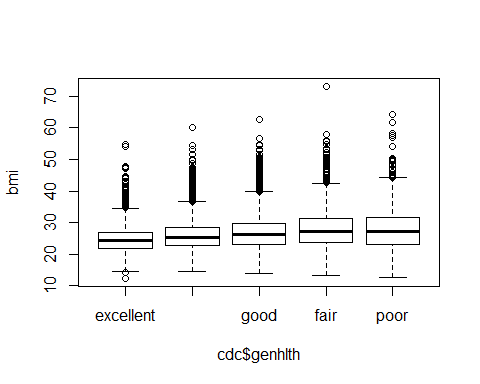
**Excercise 4** Create a new object called under23\_and\_smoke that contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime. Write the command you used to create the new object as the answer to this exercise.

under23\_and\_smoke <- subset(cdc, age < 23 & smoke100 == 1)

**Excercise 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)



summary(bmi)

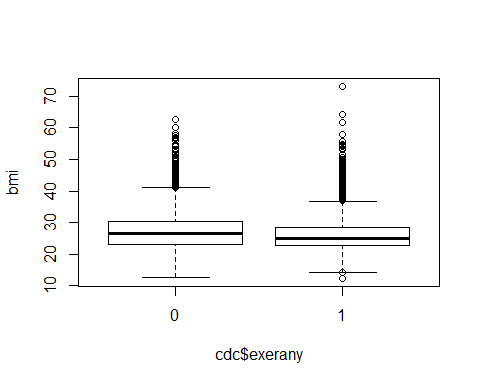
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.40 22.71 25.60 26.31 28.89 73.09

Box plot indicates

1. The median BMI across all genhlth categories increases from excellent to poor
2. Difference between upper and lower whisker keeps on increasing bewtween excellent to poor.
3. IQR increases from excellent to poor.

Another Categorical variable that might be linked to BMI is **exerany**. Reason for the same is that peope who excercise tend have to have good BMI. Below box plot does prove that theory as well.

boxplot(bmi ~ cdc$exerany)



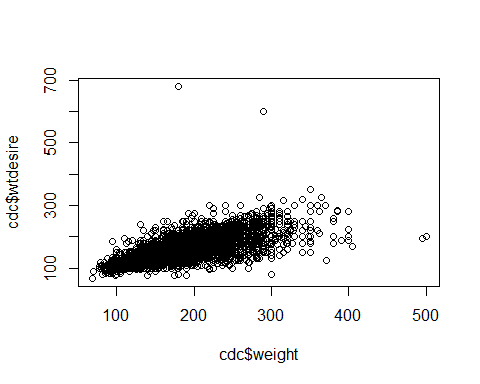
Observations from the box plot.

1. Median of BMI of people who have excercised in last 1 month is lower that that who have not excercised in last 1 month.
2. Upper Whiskers for people who have execrcised seems to ne lower than those who have not in last 1 month.

**Activities 1–6 of Lab 2**

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

plot(cdc$weight, cdc$wtdesire)



1. Weight and desired weight appear to be linearly positively associated with each other.
2. Majority of the population between 100 to 200 pounds seem to have weight and desired weight nearly same.
3. 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

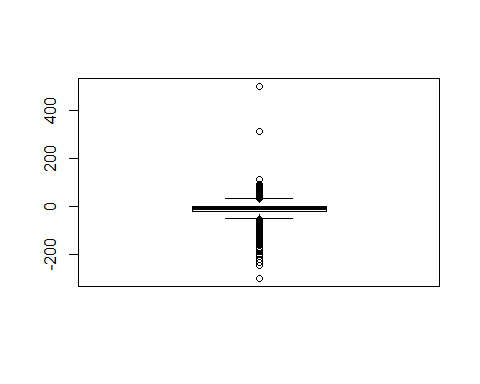
1. What type of data is wdiff? If an observation wdiff is 0, what does this mean about the person’s weight and desired weight. What if wdiff is positive or negative?
2. wdiff is numerical (integer) discrete data type.

typeof(wdiff)

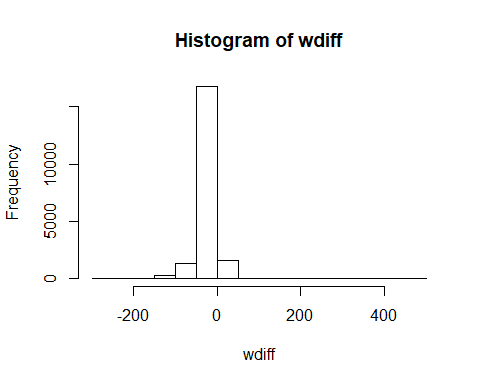
## [1] "integer"

1. if wdiff = 0 then weight and desired weight are equal and the person has a perfect BMI.
2. wdiff is positive that means that person is overweight and has to loose weight to achieve near perfect BMI and if wdiff is negative then person is underweight and has to gain weight to achieve near perfect BMI.
3. 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?

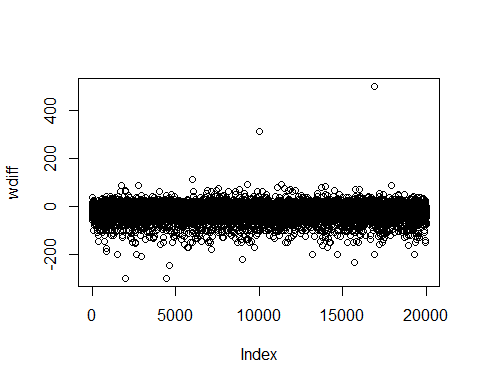
boxplot(wdiff)



hist(wdiff, breaks=20)



plot(wdiff)



summary(wdiff)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -300.00 -21.00 -10.00 -14.59 0.00 500.00

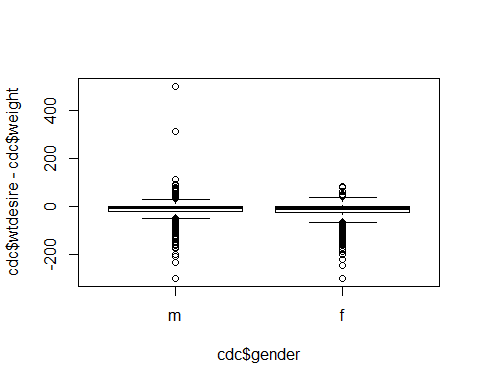
Average/mean of wdiff is -10 and median is -14.59, both are negative which indicates most of the people (males/female) need to loose weight around 10 to 15 pounds.

Histogram - It is right skewed slightly and unimodal with single peak. This again proves the theory that people need to loose weight and there are far fewer who need to gain weight.

IQR = 0.00 - (-21) = 21. So looking at box plot there are many outliers but most outliers fall in category of loosing weight and less in the category of gaining weight.

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

boxplot(cdc$wtdesire-cdc$weight ~ cdc$gender)



genderWeightDifference <- data.frame(wdiff, cdc$gender)  
summary(subset(genderWeightDifference, cdc$gender == "m"))

## wdiff cdc.gender  
## Min. :-300.00 m:9569   
## 1st Qu.: -20.00 f: 0   
## Median : -5.00   
## Mean : -10.71   
## 3rd Qu.: 0.00   
## Max. : 500.00

summary(subset(genderWeightDifference, cdc$gender == "f"))

## wdiff cdc.gender  
## Min. :-300.00 m: 0   
## 1st Qu.: -27.00 f:10431   
## Median : -10.00   
## Mean : -18.15   
## 3rd Qu.: 0.00   
## Max. : 83.00

Median and mean for men -5.00 and -10.71 Median and mean for women -10.00 and -18.15

Based on median indicator more women (-10) need to loose few more pounds than men (-5).

IQR MEN = 0.00 - (-20) = 20 IQR WOMEN = 0.00 - (-27) = 27 Based on above IQR number, women have large range than men.

Based on the box plot looking at outliers it is evident that more men need to gain weight than women.

1. 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.

mean\_wt = mean(cdc$weight)  
sd\_wt = sd(cdc$weight)  
mean\_wt

## [1] 169.683

sd\_wt

## [1] 40.08097

cdc\_one\_standard\_deviation = subset(cdc, cdc$weight < (mean\_wt +sd\_wt) & cdc$weight > (mean\_wt - sd\_wt))  
nrow(cdc\_one\_standard\_deviation)

## [1] 14152

nrow(cdc)

## [1] 20000

(nrow(cdc\_one\_standard\_deviation)/ nrow(cdc) ) \* 100

## [1] 70.76

70.76% of the population weights fall within one standard deviation of the mean weight.