DA 460 - Lab 1

Marjorie Blanco

July 08, 2018

# Handout 1 R: Introduction to data

Data Source: US CDC website.

* genhlth: A categorical vector indicating general health, with categories excellent, very good, good, fair, and poor.
* exerany: A categorical vector, 1 if the respondent exercised in the past month and 0 otherwise.
* hlthplan: A categorical vector, 1 if the respondent has some form of health coverage and 0 otherwise.
* smoke100: A categorical vector, 1 if the respondent has smoked at least 100 cigarettes in their entire life and 0 otherwise.
* height: A numerical vector, respondent’s height in inches.
* weight: A numerical vector, respondent’s weight in pounds.
* wtdesire: A numerical vector, respondent’s desired weight in pounds.
* age: A numerical vector, respondent’s age in years.
* gender: A categorical vector, respondent’s gender.

## Getting started

# Load BRFSS data  
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"

### 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).

he dimensions of the data frame is 20000, 9. There are 20000 records and 9 variables in this data set. The variable names are: genhlth, exerany, hlthplan, smoke100, height, weight, wtdesire, age, gender.

Four variables are discrete (height, weight, wtdesire and age) and the remaining five are categorical. One categorical variable, genhlth, is ordinal.

* genhlth: categorical (ordinal)
* exerany: categorical
* hlthplan: categorical
* smoke100: categorical
* height: numerical
* weight: numerical
* wtdesire: numerical
* age: numerical
* gender: categorical

# Look at the first few entries (rows) for our data set  
head(cdc)

## genhlth exerany hlthplan smoke100 height weight wtdesire age gender  
## 1 good 0 1 0 70 175 175 77 m  
## 2 good 0 1 1 64 125 115 33 f  
## 3 good 1 1 1 60 105 105 49 f  
## 4 good 1 1 0 66 132 124 42 f  
## 5 very good 0 1 0 61 150 130 55 f  
## 6 very good 1 1 0 64 114 114 55 f

# Look at the last few entries (rows) for our data set  
tail(cdc)

## genhlth exerany hlthplan smoke100 height weight wtdesire age  
## 19995 good 0 1 1 69 224 224 73  
## 19996 good 1 1 0 66 215 140 23  
## 19997 excellent 0 1 0 73 200 185 35  
## 19998 poor 0 1 0 65 216 150 57  
## 19999 good 1 1 0 67 165 165 81  
## 20000 good 1 1 1 69 170 165 83  
## gender  
## 19995 m  
## 19996 f  
## 19997 m  
## 19998 f  
## 19999 f  
## 20000 m

## Summaries and tables

# summary statistics for weight  
summary(cdc$weight)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 68.0 140.0 165.0 169.7 190.0 500.0

# mean (average) for weight  
mean(cdc$weight)

## [1] 169.683

# variance for weight  
var(cdc$weight)

## [1] 1606.484

# median for weight  
median(cdc$weight)

## [1] 165

# compute the interquartile range for the respondents’ weight  
190 - 140

## [1] 50

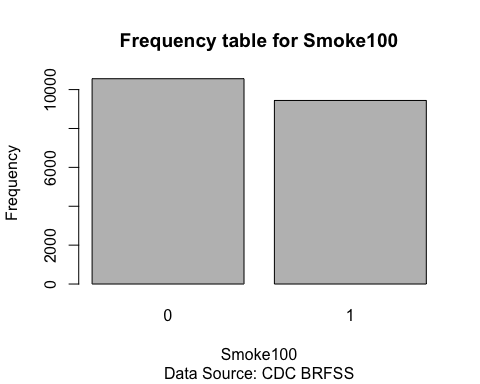
# sample frequency for smoke100  
table(cdc$smoke100)

##   
## 0 1   
## 10559 9441

# relative frequency distribution for smoke100  
table(cdc$smoke100)/20000

##   
## 0 1   
## 0.52795 0.47205

# barplot  
barplot(table(cdc$smoke100), main = "Frequency table for Smoke100",  
 xlab = "Smoke100", ylab = "Frequency", sub="Data Source: CDC BRFSS")  
  
# object, a table, called smoke  
smoke <- table(cdc$smoke100)  
# make a bar plot of the entries  
barplot(smoke, main = "Frequency table for Smoke100",  
 xlab = "Smoke100", ylab = "Frequency", sub="Data Source: CDC BRFSS")



### 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?

Interquartile range for height: between 48 and 93 with a range of 45 Interquartile range for age: between 18 and 99 with a range of 81

There are 9569 males in the sample.

A total of 4657 out out of 20000 reported being in excellent health. This is 23.285% of the sample.

# numerical summary for height and age  
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

# interquartile range for height and age  
max(cdc$height) - min(cdc$height)

## [1] 45

max(cdc$age) - min(cdc$age)

## [1] 81

# frequency distribution for gender and exerany  
table(cdc$gender)

##   
## m f   
## 9569 10431

table(cdc$exerany)

##   
## 0 1   
## 5086 14914

# relative frequency distribution for gender and exerany  
table(cdc$gender) / length(cdc$gender)

##   
## m f   
## 0.47845 0.52155

table(cdc$exerany) / length(cdc$exerany)

##   
## 0 1   
## 0.2543 0.7457

# frequency distribution and relative frequency distribution for genhlth  
table(cdc$genhlth)

##   
## excellent very good good fair poor   
## 4657 6972 5675 2019 677

table(cdc$genhlth) / length(cdc$genhlth)

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

### Exercise 3

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

The mosaic plot shows that more males smoke at least 100 cigarettes than women.

# examine which participants have smoked across each gender  
table(cdc$gender, cdc$smoke100)

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

# mosaicplot for gender and smoking habits  
mosaicplot(table(cdc$gender, cdc$smoke100), main = "Gender and Smoked100",  
 ylab = "Smoked100", xlab = "Gender")



## Interlude: How R thinks about data

# size of the data frame   
dim(cdc)

## [1] 20000 9

# sixth variable of the 567th respondent  
cdc[567,6]

## [1] 160

names(cdc)

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

# weights for the first 10 respondents  
cdc[1:10,6]

## [1] 175 125 105 132 150 114 194 170 150 180

# range 1 through 10  
1:10

## [1] 1 2 3 4 5 6 7 8 9 10

# data for the first 10 respondent  
cdc[1:10,]

## genhlth exerany hlthplan smoke100 height weight wtdesire age gender  
## 1 good 0 1 0 70 175 175 77 m  
## 2 good 0 1 1 64 125 115 33 f  
## 3 good 1 1 1 60 105 105 49 f  
## 4 good 1 1 0 66 132 124 42 f  
## 5 very good 0 1 0 61 150 130 55 f  
## 6 very good 1 1 0 64 114 114 55 f  
## 7 very good 1 1 0 71 194 185 31 m  
## 8 very good 0 1 0 67 170 160 45 m  
## 9 good 0 1 1 65 150 130 27 f  
## 10 good 1 1 0 70 180 170 44 m

# see the weights for all 20,000 respondents  
cdc[,6]  
cdc$weight

# see the weight for the 567th respondent  
cdc$weight[567]

## [1] 160

# see the weight for first 10 respondents  
cdc$weight[1:10]

## [1] 175 125 105 132 150 114 194 170 150 180

## A little more on subsetting

# test condition to check if person was male   
cdc$gender == "m"  
# test condition to check if person was older than 30   
cdc$age > 30

# extract just the data for the men in the sample  
mdata <- subset(cdc, cdc$gender == "m")  
# shows first few rows for mdata  
head(mdata)

## genhlth exerany hlthplan smoke100 height weight wtdesire age gender  
## 1 good 0 1 0 70 175 175 77 m  
## 7 very good 1 1 0 71 194 185 31 m  
## 8 very good 0 1 0 67 170 160 45 m  
## 10 good 1 1 0 70 180 170 44 m  
## 11 excellent 1 1 1 69 186 175 46 m  
## 12 fair 1 1 1 69 168 148 62 m

# data for men over the age of 30.  
m\_and\_over30 <- subset(cdc, gender == "m" & age > 30)  
# data for people who are men or over the age of 30   
m\_or\_over30 <- subset(cdc, gender == "m" | age > 30)

### Exercise 4

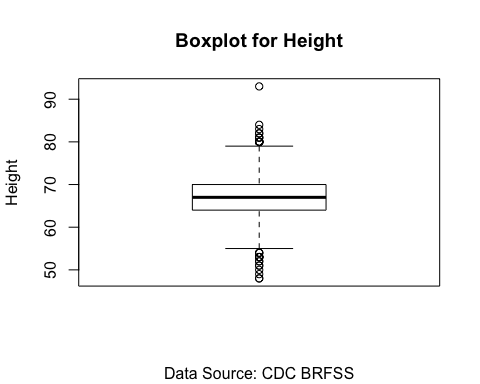
Create a new object called under23\_and\_smoke that contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime. Write the command you used to create the new object as the answer to this exercise.

# contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime  
under23\_and\_smoke <- subset(cdc, age < 23 & smoke100 == 1)  
# shows first few rows for under23\_and\_smoke  
head(under23\_and\_smoke)

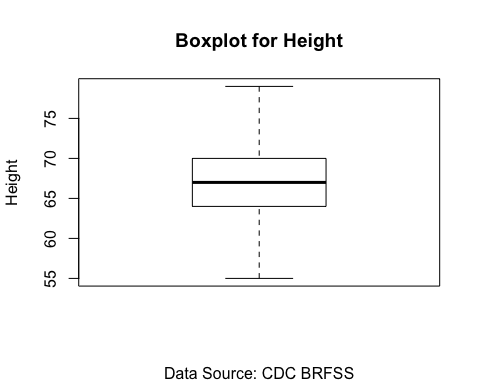
## genhlth exerany hlthplan smoke100 height weight wtdesire age gender  
## 13 excellent 1 0 1 66 185 220 21 m  
## 37 very good 1 0 1 70 160 140 18 f  
## 96 excellent 1 1 1 74 175 200 22 m  
## 180 good 1 1 1 64 190 140 20 f  
## 182 very good 1 1 1 62 92 92 21 f  
## 240 very good 1 0 1 64 125 115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

## Quantitative data

# construct a box plot for a single variable: height  
boxplot(cdc$height, main = "Boxplot for Height", ylab = "Height",  
 sub="Data Source: CDC BRFSS")



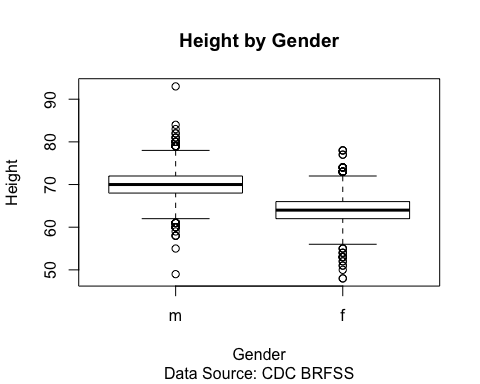
# construct a box plot for a single variable: height  
boxplot(cdc$height, main = "Boxplot for Height", ylab = "Height",  
 outline = FALSE,  
 sub="Data Source: CDC BRFSS")



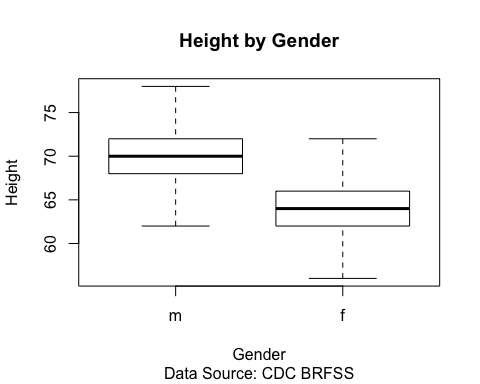
# examining the summary statistics for a single variable: height  
summary(cdc$height)

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

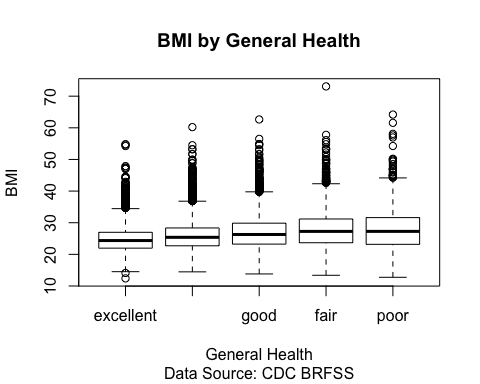
# compare the heights of men and women  
boxplot(cdc$height ~ cdc$gender, main = "Height by Gender",   
 xlab = "Gender", ylab = "Height",  
 sub="Data Source: CDC BRFSS")



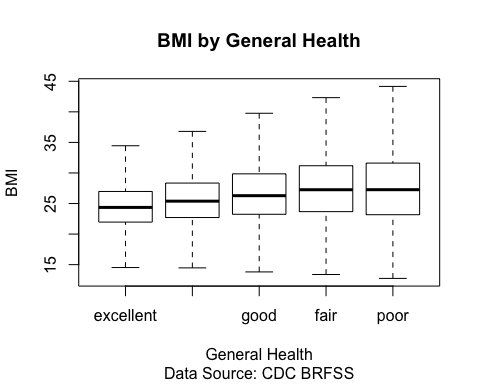
boxplot(cdc$height ~ cdc$gender, main = "Height by Gender",   
 xlab = "Gender", ylab = "Height",  
 outline = FALSE,  
 sub="Data Source: CDC BRFSS")



# calculate BMI  
# IQR range is larger in fair. Poor shows higher median BMI  
cdc$bmi <- bmi <- (cdc$weight / cdc$height^2) \* 703  
# compare the bmi and general health of men and women  
boxplot(bmi ~ cdc$genhlth, main="BMI by General Health",   
 ylab = "BMI", xlab="General Health",   
 sub="Data Source: CDC BRFSS")



boxplot(bmi ~ cdc$genhlth, main="BMI by General Health",   
 ylab = "BMI", xlab="General Health",   
 outline = FALSE,  
 sub="Data Source: CDC BRFSS")



# summary stats  
summary(cdc$bmi[cdc$genhlth == "excellent"])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.40 21.96 24.36 24.84 26.97 54.81

summary(cdc$bmi[cdc$genhlth == "very good"])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 14.47 22.71 25.39 26.02 28.35 60.22

summary(cdc$bmi[cdc$genhlth == "good"])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 13.81 23.24 26.28 26.97 29.86 62.64

summary(cdc$bmi[cdc$genhlth == "fair"])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 13.39 23.67 27.26 28.12 31.18 73.09

summary(cdc$bmi[cdc$genhlth == "poor"])

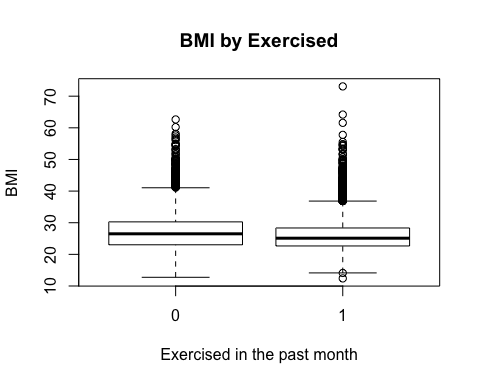
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.75 23.17 27.26 28.39 31.61 64.19

### Exercise 5

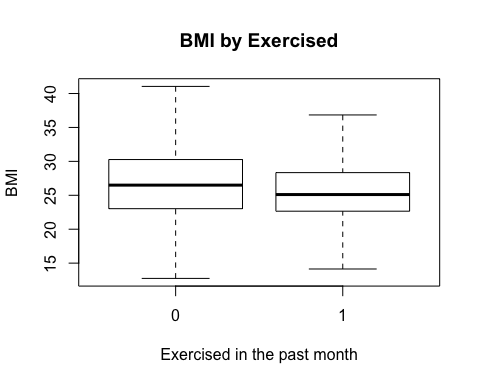
What does this box plot show? Pick another categorical variable from the data set and see how it relates to BMI. List the variable you chose, why you might think it would have a relationship to BMI, and indicate what the figure seems to suggest.

Hypothesis: Respondent that exercised in the past month will have lower BMI on average compared to those that do not. This figure seems to suggest that excercise is a significant factor. Conclusion: People with lower BMI have excersiced recently.

# IQR range is larger in those who exercised in the past month. Those that di not exercise in the past month shows higher median BMI  
boxplot(bmi ~ cdc$exerany, main="BMI by Exercised",   
 ylab = "BMI", xlab="Exercised in the past month")



boxplot(bmi ~ cdc$exerany, main="BMI by Exercised",   
 ylab = "BMI", xlab="Exercised in the past month",  
 outline = FALSE)



summary(cdc$bmi[cdc$exerany == 1])

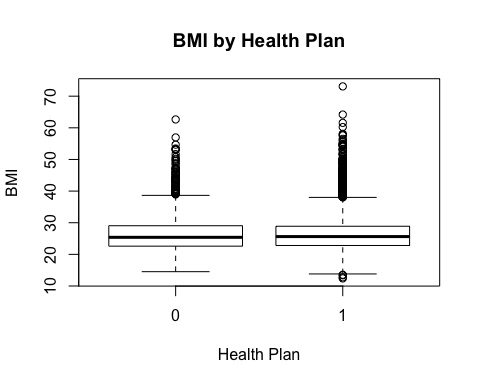
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.40 22.66 25.11 25.98 28.34 73.09

summary(cdc$bmi[cdc$exerany == 0])

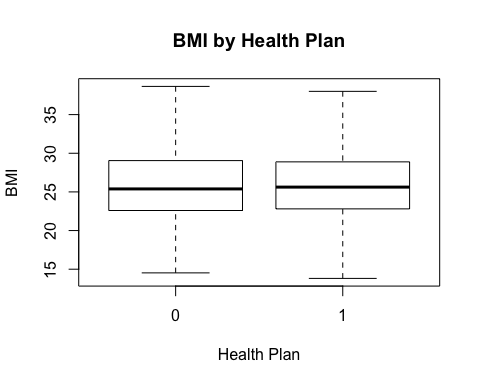
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.75 23.03 26.50 27.27 30.27 62.64

Hypothesis: Respondent that have some form of health coverage will have similar BMI on average compared to those that do not. This figure seems to suggest that having some form of health coverage is not significant factor.

# IQR range and median BMI is larger in those that have some form of health coverage.  
boxplot(bmi ~ cdc$hlthplan, main="BMI by Health Plan",   
 ylab = "BMI", xlab="Health Plan")



boxplot(bmi ~ cdc$hlthplan, main="BMI by Health Plan",   
 ylab = "BMI", xlab="Health Plan",  
 outline = FALSE)



summary(cdc$bmi[cdc$hlthplan == 1])

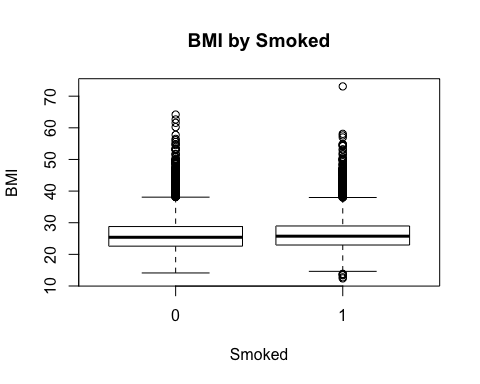
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.40 22.80 25.62 26.30 28.89 73.09

summary(cdc$bmi[cdc$hlthplan == 0])

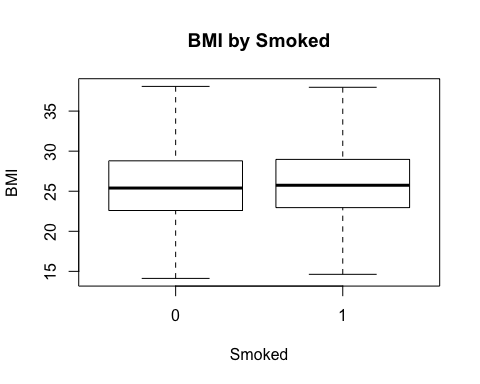
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 14.52 22.59 25.39 26.34 29.05 62.64

Conclusion: People with lower BMI have not smoked at least 100 cigarettes in their entire life.

# IQR range and median BMI is larger in those that smoked at least 100 cigarettes in their entire life.  
boxplot(bmi ~ cdc$smoke100, main="BMI by Smoked",   
 ylab = "BMI", xlab="Smoked")



boxplot(bmi ~ cdc$smoke100, main="BMI by Smoked",   
 ylab = "BMI", xlab="Smoked",  
 outline = FALSE)



summary(cdc$bmi[cdc$smoke100 == 1])

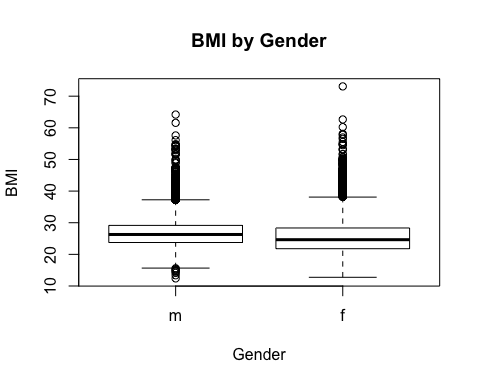
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.40 22.95 25.74 26.39 28.97 73.09

summary(cdc$bmi[cdc$smoke100 == 0])

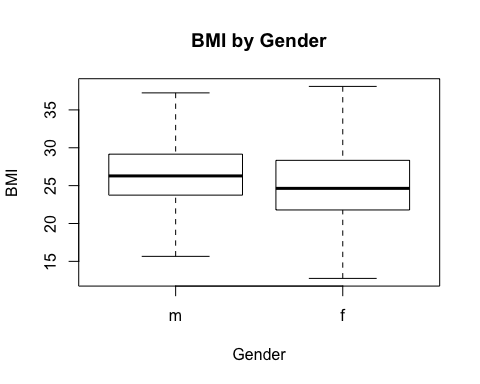
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 14.12 22.59 25.40 26.24 28.79 64.19

Hypothesis: Respondent that are male will have lower BMI on average compared to those that do not. This figure seems to suggest that gender is a significant factor.

# IQR range is larger in female. Male shows higher median BMI  
boxplot(bmi ~ cdc$gender, main="BMI by Gender",   
 ylab = "BMI", xlab="Gender")



boxplot(bmi ~ cdc$gender, main="BMI by Gender",   
 ylab = "BMI", xlab="Gender",  
 outline = FALSE)



summary(cdc$bmi[cdc$gender == "m"])

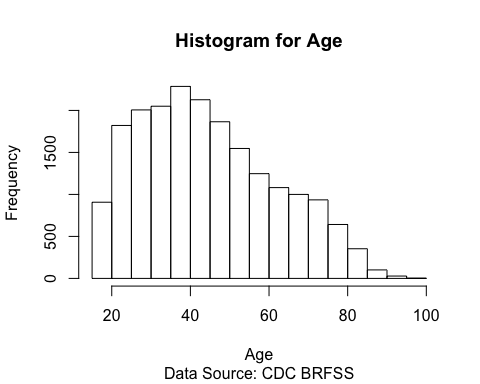
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.40 23.75 26.28 26.92 29.16 64.19

summary(cdc$bmi[cdc$gender == "f"])

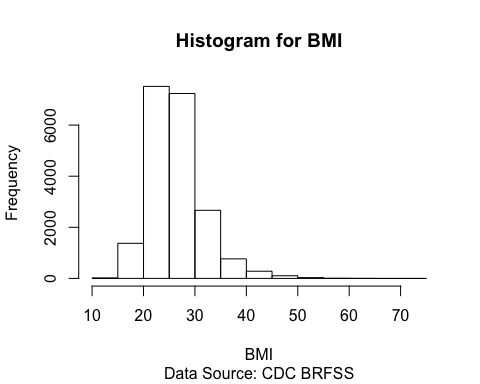
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.75 21.79 24.64 25.74 28.34 73.09

The second histogram appears to be a better graph to analyze for distribution. Histograms are generally a very good way to see the shape of a single distribution, but the shape can change depending on how the data is split between the various bins. The first plot makes a default histogram of bmi and the second plot contains 50 breaks.

# histogram for the age of respondents  
hist(cdc$age, main = "Histogram for Age", xlab = "Age",  
 sub="Data Source: CDC BRFSS")



# histogram for the bmi ofrespondents  
hist(bmi, main = "Histogram for BMI", xlab = "BMI",   
 sub="Data Source: CDC BRFSS")



# histogram for the bmi ofrespondents with 50 breaks  
hist(bmi, breaks = 50, main = "Histogram for BMI", xlab = "BMI",   
 sub="Data Source: CDC BRFSS")

