Math 275 Lab 1 Intro to Data

Zachary Goodsell

9/24/2019

Question 1

There are 20,000 cases and 9 variables for each of these cases. Data types - genhlth:categorical, exerany:binary, hlthplan:binary,smoke100:binary, height:continuous,weight:continuous,wtdesire:continuous, age:continuous,gender:binary

```
names(cdc)
```

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

Question 2

The interquartile range of height is 6. The interquartile range of age is 26. The gender frequency is 47.8% for males and 52.1% for females. The exerany frequency is 25.4% for 0 and 74.6% for 1. 0 is for no monthly exercise and 1 is for monthly exercise. There are 9569 males in the sample. For participants with excellent health the percent is 23.3%.

```
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
```

```
iqr_height=summary(cdc$height)[5]-summary(cdc$height)[2]
iqr_height
```

```
## 3rd Qu.
##       6
```

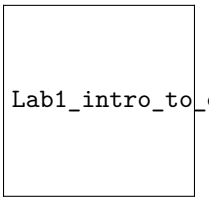
```
#Interquartile range of height is 6
iqr_age=summary(cdc$age)[5]-summary(cdc$age)[2]
iqr_age
```

```
## 3rd Qu.
##      26
```

```r
#Interquartile range of age is 26
gender=table(cdc$gender)
genderFreq = gender/20000
genderFreq
```

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

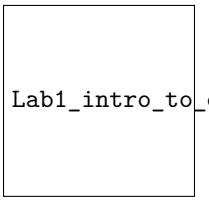
```r
barplot(genderFreq)
```




```r
exerany=table(cdc$exerany)
exeranyFreq = exerany/20000
exeranyFreq
```

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

```r
barplot(exeranyFreq)
```




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

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

```r
numMales = table(cdc$gender)[1]
numMales
```

```
##    m
## 9569
```

```r
table(cdc$genhlth)/20000
```

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

```
table(cdc$genhlth)[1]/20000
```

```
## excellent
##   0.23285
```

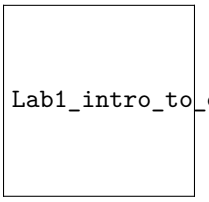
Question 3

The mosaic shows the difference between long term smoking in men and women. We see that more men smoke 100 or more cigarettes in there lifetime when compared to woman.

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

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

```
mosaicplot(table(cdc$gender,cdc$smoke100),main="Mosaic Plot of Gender vs. smoke100")
title(ylab="Person Has Smoked More Than 100 Cigarettes in Lifetime",
      xlab="Persons Gender")
```

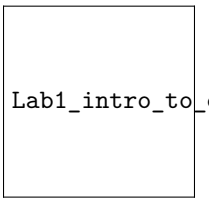


Question 4

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

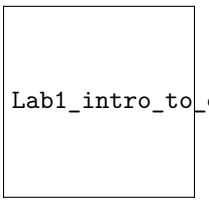
```
##       genhlth exerany hlthplan smoke100 height weight wtdesire age gender
## 13  excellent       1        0        1     66    185      220  21      m
## 37  very good       1        0        1     70    160      140  18      f
## 96  excellent       1        1        1     74    175      200  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
```

Question 5



The box plot above shows the relationship between the persons BMI and overall health.

```
boxplot(bmi ~ cdc$exerany,
        main="Relationship Between BMI and Monthly Exercise",xlab="Whether or Not Person Has Exercised
```



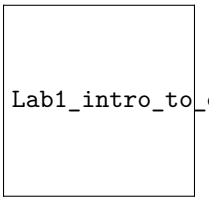
I chose monthly exercise because those who don't exercise tend to have a high BMI.The box plot shows that BMI is slightly lower in those who exercise.

Own Questions

Question 1

The relationship shows how close the peoples weights are to their desired weights. a perfect 1/1 slop would represent everyone being at there desired weight. However, it is seen that the points aren't all at that slope.

```
plot(cdc$weight, cdc$wtdesire,
     xlab="Weight",ylab="Desired Weight",main="Plot of Persons Actual Weights vs. Desired Weights")
```



Question 2

```
wdiff = abs(cdc$wtdesire-cdc$weight)
head(wdiff)
```

```
## [1]  0 10  0  8 20  0
```

Question 3

wdiff is quatitative data. If it is 0 that means that the persons actual weight and desired weight are the same. If wdiff isn't 0 that is fow far the person is from there desired weight.

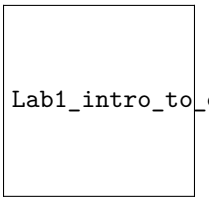
Question 4

The histogram is right skewed. The mean of 10 shows that 50% of the people are with in 10 pounds of there desired weight. Further more, the histogram shows that roughly 75% of the people are within 25 pounds of there desired weight.

```
summary(wdiff)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00   10.00   17.11   25.00  500.00
```

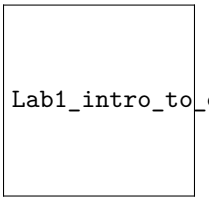
```
hist(wdiff)
```




```
wdiff_under100=subset(wdiff, wdiff<=100)
summary(wdiff_under100)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00   10.00   15.99   23.00  100.00
```

```
hist(wdiff_under100)
```



Question 5

The plot shows that men are closer to there desired weight. The women have a bigger spread.

```
cdc_under100=subset(cdc,
                    abs(cdc$wtdesire-cdc$weight)<=100)
cdc_under100_male=subset(cdc_under100, gender=='m')
cdc_under100_female=subset(cdc_under100, gender=='f')
summary(abs(cdc_under100_male$wtdesire-cdc_under100_male$weight))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00   10.00   13.96   20.00  100.00
```

```r
sd(abs(cdc_under100_male$wtdesire-cdc_under100_male$weight))
```

```
## [1] 17.24185
```

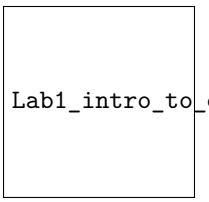
```r
summary(abs(cdc_under100_female$wtdesire-cdc_under100_female$weight))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    2.00   10.00   17.87   25.00  100.00
```

```r
sd(abs(cdc_under100_female$wtdesire-cdc_under100_female$weight))
```

```
## [1] 19.73895
```

```r
boxplot(abs(cdc_under100$wtdesire-cdc_under100$weight) ~ cdc_under100$gender,
        main="Difference Between Actual and Desired Weight, by Gender",
        xlab="Gender",ylab="Difference Between Actual and Desired Weight")
```



Question 6

70.76% of peoples weight fall within one SD of the mean weight

```r
weight_mean = mean(cdc$weight)
weight_sd = sd(cdc$weight)
weight_mean
```

```
## [1] 169.683
```

```r
weight_sd
```

```
## [1] 40.08097
```

```r
sum(cdc$weight>=(weight_mean-weight_sd) & cdc$weight<=(weight_mean+weight_sd))/nrow(cdc)
```

```
## [1] 0.7076
```