da460\_assignment1\_grahn

jason grahn

9/27/2018

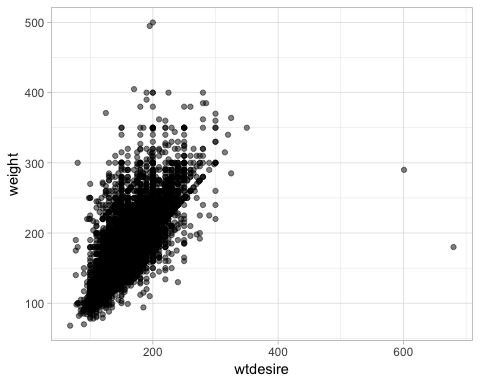
# Part1

Based on Handout 1 R (use the same source data), apply R to answer the following questions. Make sure you include the command line/code, then paste relevant output/results, and also comment on the output/results as needed (to answer the questions).

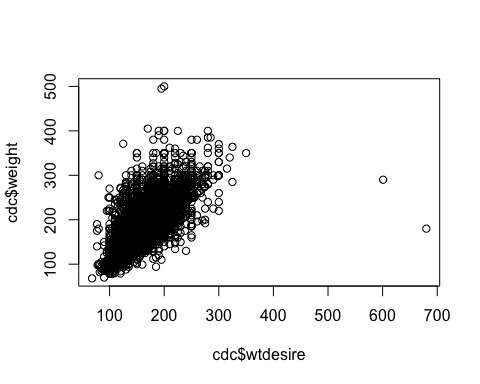
source('http://www.openintro.org/stat/data/cdc.R')

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

cdc %>%   
 ggplot(aes(wtdesire, weight)) +  
 geom\_point(alpha = 0.5) +  
 theme\_light()



#or in base-R  
plot(cdc$wtdesire, cdc$weight, type = "p")

 There appears to be a generally strong positive relationship between weight and wtdesire. As weight increases, the wtdesire of the individual increases as well.

## 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.

cdc <- cdc %>%   
 mutate(wdiff = wtdesire - weight)  
  
#or the less useful way in base-R  
head(wdiff <- (cdc$wtdesire - cdc$weight),10)

## [1] 0 -10 0 -8 -20 0 -9 -10 -20 -10

It’s a bit silly to make these a new object all on their own. Using mutate() we can create this variable inside the pre-existing dataframe.

## 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?

Wdiff is continuous and numerical. If an observation of wdiff is zero, it means the sample is at their desired weight. If wdiff is negative, it means the observation wants to lose weight, a positive wdiff means the observation wants to gain weight.

## 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?

#first some stats around wdiff  
cdc %>%   
 select(wdiff) %>%   
 describe()

## vars n mean sd median trimmed mad min max range skew  
## X1 1 20000 -14.59 24.05 -10 -11.41 14.83 -300 500 800 -1.45  
## kurtosis se  
## X1 21.61 0.17

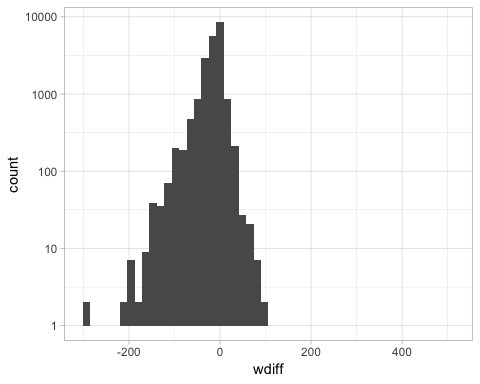
#those min and max are pretty huge numbers. confirm?  
cdc %>%   
 filter((wdiff == max(wdiff) | wdiff == min(wdiff)))

## genhlth exerany hlthplan smoke100 height weight wtdesire age gender  
## 1 poor 1 1 0 74 500 200 45 m  
## 2 fair 1 1 1 69 495 195 32 f  
## 3 good 0 1 0 69 180 680 24 m  
## wdiff  
## 1 -300  
## 2 -300  
## 3 500

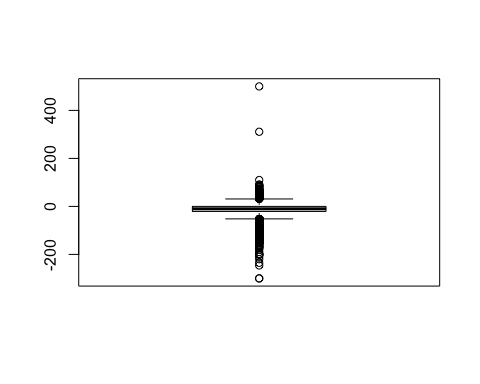
#looking at distribution through a histogram  
cdc %>%   
 ggplot() +   
 geom\_histogram(aes(wdiff), bins = 50) +  
#needing to use logscale for Y becase the high peaks  
 scale\_y\_log10() +  
 theme\_light()

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 24 rows containing missing values (geom\_bar).



#or, that boxplot in base-R  
boxplot(wdiff)



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

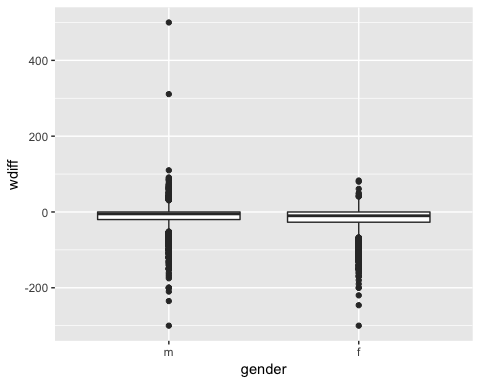
#for males  
cdc %>%   
 filter(gender == "m") %>%   
 select(wdiff) %>%   
 describe(IQR = TRUE)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 9569 -10.71 23.49 -5 -8.41 7.41 -300 500 800 -0.6 37.44  
## se IQR  
## X1 0.24 20

#for females  
cdc %>%   
 filter(gender == "f") %>%   
 select(wdiff) %>%   
 describe(IQR = TRUE)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 10431 -18.15 24 -10 -14.2 14.83 -300 83 383 -2.26 9.22  
## se IQR  
## X1 0.23 27

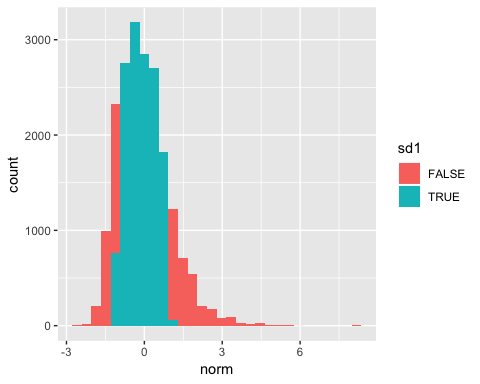
#and the side-by-side boxplot  
cdc %>%   
 ggplot() +   
 geom\_boxplot(aes(x = gender, y = wdiff))

 With minor exception, females generally want to lose weight. Males have an larger propensity for gaining weight.

## 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.

normweight <- cdc %>%   
 select(weight) %>%   
 mutate(mean = mean(weight),  
 sd = sd(weight),  
 norm = round((weight - mean)/sd,2),  
 sd1 = if\_else(norm >-1 & norm <1, TRUE, FALSE)  
 )  
   
ggplot(normweight) +   
 geom\_histogram(aes(x = norm, fill = sd1))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



table(normweight$sd1)/nrow(normweight)\*100

##   
## FALSE TRUE   
## 29.24 70.76

The mean weight is 169.68295 and the standard deviation is 40.08097. 70.76% of the weights fall within 1 standard deviation of the mean.