STA 2210 Midterm Review Exercises

The following data represents the results of a survey of general health for a random sample of 20000 American adults (we used this data during Lab 1).

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

Directions: A researcher wants to investigate 3 research questions, labeled Q1-Q3 below, pertaining to the data and they need your help. Help the researcher by completing all the exercises below. The exercises are listed below the research question to which they are relevant. When typing up your solutions, be sure to include any R codes and figures/plots used.

1. **Identify the cases.**

20,000 observations for 9 variables

**Q1**. **Is there a relationship between gender and perception of one’s general health?**

1. **Identify the variables and their types.**

Genhlth (categorical), Gender (categorical).

1. **Create a contingency table for the variables.**

table(cdc$genhlth, cdc$gender)

m f

excellent 2298 2359

very good 3382 3590

good 2722 2953

fair 884 1135

poor 283 394

1. **What percent of the sample reports being in excellent health?**

Table(cdc$genhlth) = 4657 excellent; 4657/20,000 = 0.233 = 23.3%

1. **What percent of females report being in excellent health?**

females <- subset(cdc, cdc$gender == 'f') = 10,431 females

table(females$genhlth)

= 2359 excellent; 2,359 / 10,431 = 0.226 = 22.6%

1. **Is believing one is in “excellent” health independent of one’s gender? Justify your answer.**

Female excellent health = 22.6%

males <- subset(cdc, cdc$gender == 'm') = 9569 males

table(males$genhlth) = 2298 excellent

Male excellent health = 2298/9569 = 24.0%

Total excellent health = 4657 / 20,000 = 23.3%

Based on the data, believing in having excellent health is dependent on gender because the probabilities are different between males and females, and different for the overall relationship. Knowledge of one’s gender can give a small idea on what one would lean towards for their general health.

**Q2**. **Is there a difference of the average weight of those who exercise and those who do not exercise?**

1. **Identify the variables and their types.**

Weight (numerical continuous), exerany (categorical)

1. **Let denote the random variable that equals “1” if a randomly selected case exercises, and “0” if they do not exercise. Find the standard deviation of .**

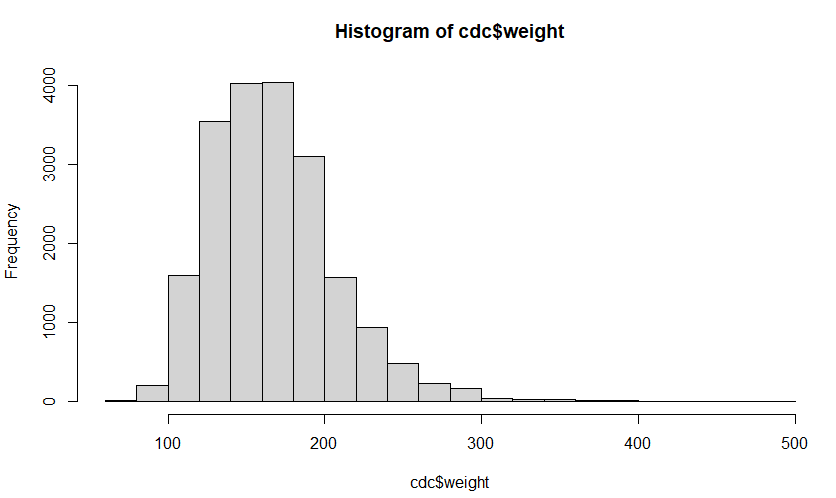
exercise <- subset(cdc, cdc$exerany ==1)

sd(exercise$weight)

= 38.8333

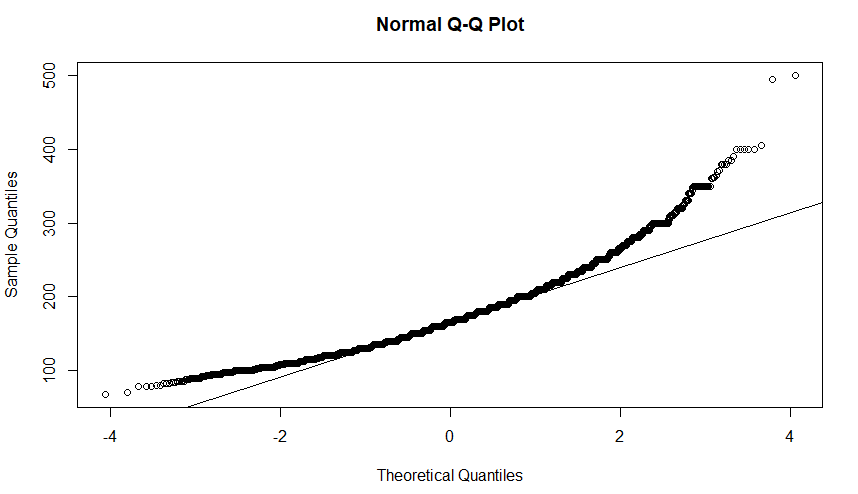
1. **Using at least 3 different methods, determine if weight follows a normal distribution.**

hist(cdc$weight)

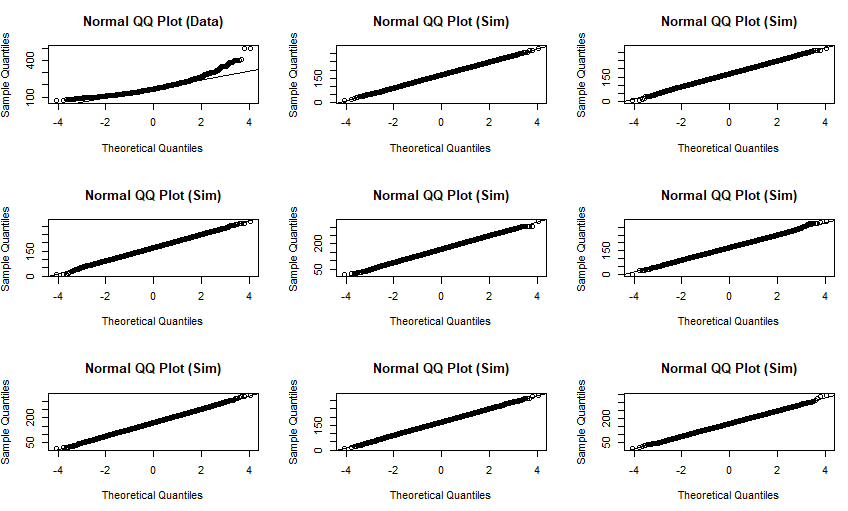


qqnorm(cdc$weight)

qqline(cdc$weight)



qqnormsim(cdc$weight)



The histogram shows a unimodal graph which appears to be right skewed, as more data is on the left side of the graph. The qq plot also curves up at the right end of the graph, which indicates a right skew. The real plot also appears to be different from the simulated qq graphs, which indicates weight does not follow much of a normal distribution.

1. **Calculate the probability that a randomly selected adult weighs less than 150 lbs**
2. empirically, using the data.

sum(cdc$weight < 150) / length(cdc$weight)

= 0.323

1. theoretically, using a normal distribution.

pnorm(q = 150, mean(cdc$weight), sd(cdc$weight))

= 0.312

1. **Calculate the probability that a randomly selected adult weighs more than 200 lbs**
2. empirically, using the data.

sum(cdc$weight > 200) / length(cdc$weight)

= 0.174

1. theoretically, using a normal distribution.

1 - pnorm(q = 200, mean(cdc$weight), sd(cdc$weight))

= 0.225

1. **Using numerical summaries and a side-by-side boxplot, determine if there is a difference in the average weight of the two groups.**

noexercise <- subset(cdc, cdc$exerany == 0)

exercise <- subset(cdc, cdc$exerany ==1)

mean(exercise$weight)

= 169.0387

mean(noexercise$weight)

= 171.5722

summary(exercise$weight)

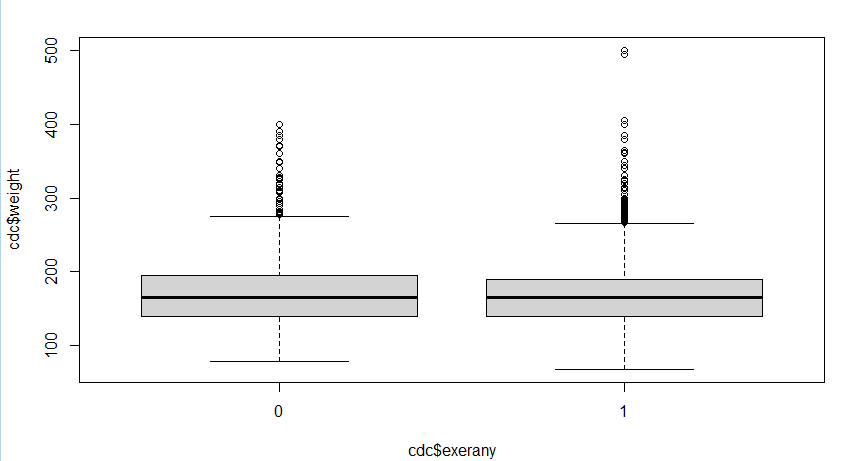
Min. 1st Qu. Median Mean 3rd Qu. Max.

68 140 165 169 190 500

summary(noexercise$weight)

Min. 1st Qu. Median Mean 3rd Qu. Max.

78.0 140.0 165.0 171.6 195.0 400.0

boxplot(cdc$weight ~ cdc$exerany)

There is a very slight difference in average weight between those who say they exercise and those who do not. Those who exercise have a mean that is 2.6 lbs. lower than non-exercisers, and a smaller 75th percentile. However, the differences are small, as both have the same 25th percentile and median. But with the exercise group having 3 times the number of observations, having a lower mean with a much larger sample size can show that there is a difference in average weights between the 2 variables.

**Q3**. **Do heavier people tend to have a larger desired weight?**

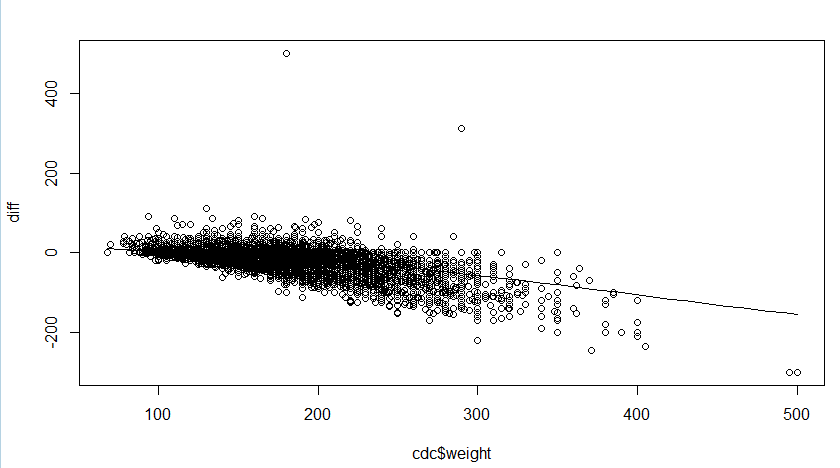
1. **Identify the variables and their types? Which is the explanatory variable?**

Weight (numerical continuous), wtdesire (numerical continuous). Explanatory = weight

1. **Describe the relationship between the variables. Justify your answer with a plot and be sure to discuss any unusual observations.**

diff <- cdc$wtdesire - cdc$weight

scatter.smooth(cdc$weight, diff)



I created a value called ‘diff’ which took the difference between the desired and actual weight of each observation, and plotted the actual weights vs. diff. As shown in the scatterplot, there is a downward trend for diff as the weights increase. As the graph increases the weight, more people’s diff value decreases, which means they want to lose weight. This shows that the heavier a person is, the more they tend to want to lose weight and have a lower desired weight. There are some exceptions, such as a few heavier people over 300 pounds wanting to gain weight, and two major outliers wanting to gain 300+ pounds (maybe errors). Overall, there is a negative correlation between actual weight and desired weight.