Preyasi Gaur
705704939

Statistics 10
Homework 2

1. a. In general, students are happy with their body weight as approximately 66%, i.e 600 out of 900 of the surveyed students have responded “About right.”

   b. If the researchers want to compare the body image differences in males and females, they can plot a side-by-side histogram.

   c. We can make the comparison between males feeling happy about their body weight and females feeling happy about their body weight, we can calculate the relative frequencies of the same:
   Relative frequency of women who feel happy about their weight = $\frac{Women\ who\ responded\ "about\ right"}{Total\ Number\ of\ women\ surveyed} = \frac{310}{470} \approx 0.6576$
   Relative frequency of men who feel happy about their weight = $\frac{Men\ who\ responded\ "about\ right"}{Total\ Number\ of\ men\ surveyed}$
   $= \frac{290}{430} \approx 0.6744$
   Thus, we can conclude that male students are more likely to feel they are “about right.”

   d. Calculating the relative frequencies:
   Respondents who feel overweight = $\frac{Number\ of\ people\ who\ responded\ "overweight"}{Total\ Number\ of\ people\ surveyed} = \frac{130+68}{900} = \frac{198}{900}$
   $= 0.22$
   Respondents who feel underweight = $\frac{Number\ of\ people\ who\ responded\ "underweight"}{Total\ Number\ of\ people\ surveyed} =$
   $\frac{30+72}{900} = \frac{102}{900} \approx 0.11$
   Thus, clearly people are more likely to feel overweight than overweight.
   Relative frequency of women who feel overweight about their weight = $\frac{Women\ who\ responded\ "overweight"}{Total\ Number\ of\ women\ surveyed} = \frac{130}{470} \approx 0.28$
   Relative frequency of men who feel happy about their weight = $\frac{Men\ who\ responded\ "overweight"}{Total\ Number\ of\ men\ surveyed}$
   $= \frac{68}{430} \approx 0.16$
   Clearly, thus women are more likely to feel that they are overweight than men.

2. a. In this case,
   Explanatory Variable : Start Median Salary
   Response Variable : Mid-Career Median Salary

   b. We use median instead of mean as median is:
   (i) Mean is very sensitive to outliers, while median is much less affected by outliers
   (ii) Median also helps us determine that midpoint of the data and thus the division of the data

   c. Using the equation of the least squares regression line, we can predict that the median mid-career salary
   Median Mid-Career Salary = -7.699 + 1.989(Starting Median Salary)
   = -7.699 + 1.989(60,000)
   = 119332.301
   Thus, the median mid-career salary for a median starting salary of $60,000 is $119332.301

   d. Median Mid-Career Salary = -7.699 + 1.989(Starting Median Salary)
   = -7.699 + 1.989(100,000)
   = 198892.301
   Thus, the median mid-career salary for a median starting salary of $60,000 is $198892.301

4. a. Yes, if the doctor places the most severe patients on antidepressants then he will not be able to compare the effectiveness of “talk therapy” and antidepressants, as he is affecting the randomization of the dataset. Thus, the results he gets will not be accurate.

b. It is not acceptable for the doctor to know which patient is receiving which treatment as it would lead to bias when he draws his conclusions. Clearly the doctor strongly believes that antidepressants work better than “talk therapy” and thus if he knows what patient is getting what treatment, he would be biased to draw the conclusion that he is inclined on. The best case would be a double-blinded study where we prevent bias when doctors evaluate the patients outcomes. This helps improve the reliability of the trials.

c. Some improvements I would suggest would be to conduct a double blinded study where neither the doctor nor the patients know what treatment they’ll get. The patients should be chosen to get treatment randomly. This will help reduce bias and help to draw more accurate results.

3. a. $b = r\frac{s_y}{s_x}$

$= 0.95\left(\frac{46.34}{2.32}\right)$

$= 18.975$

$a = \bar{y} - b\bar{x}$

$= 141.67 - (18.975)(11.03)$

$= -67.624$

Thus, the slope is 18.975 and the intercept is -67.624.

b. Equation of regression line $\Rightarrow \hat{y} = -67.624 + 18.975x$.

This tells us that for every 1% increase in % alcohol (x), leads to an increase of 18.975 calories.

When the % alcohol equals to 0, the calorie ct. is -67.624. However, in reality this is not possible

not possible.

c. The coefficient of determine $= R^2 = 0.9025$

R^2 measures the variance in response variable y is explained by predictor x $0 \leq R^2 \leq 1$. In our case, R^2 is very large implying smaller amount of variation about the regression line. 90.25% of the variation in y is explained by x.

d. The slope of the line will decrease and the value of r will also decrease.