SDS 220 - Exploratory Data Analysis (Part 1)

IMS Chapters 4 and 5

1. **[IMS 5.7] Days off at a mining plant** . Workers at a particular mining site receive an average of 35 days paid vacation, which is lower than the national average. The manager of this plant is under pressure from a local union to increase the amount of paid time off. However, he does not want to give more days off to the workers because that would be costly. Instead, he decides he should fire 10 employees in such a way as to raise the average number of days off that are reported by his employees. In order to achieve this goal, should he fire employees who have the most number of days off, least number of days off, or those who have about the average number of days off?

   In order to increase the average number of days off, the manager should fire any 10 employees whose average number of days off is between the minimum and the mean number of days off for the entire workforce at this plant. However, firing the 10 employees with the minimum number of days off will have the biggest impact on the average.

2. Suppose you work at a primate conservancy and have recently taken a survey of your n = 15 female orangutans. Their weights, measured in pounds, are given below:

   $$87, 91, 86, 85, 82, 97, 86, 77, 75, 85, 80, 89, 91, 91, 95.$$

   (a) Create a histogram *by hand* that displays the distribution of female orangutan weights in your conservancy. Use bins of [75, 80), [80, 85), [85, 90), [90, 95), and [95, 100), where the notation [a, b) indicates that the lower limit, a, should be included in the bin and the upper limit, b, should not.

   (b) Calculate *by hand* the sample mean and sample variance for your orangutan weights. What are the units for each of these measures?

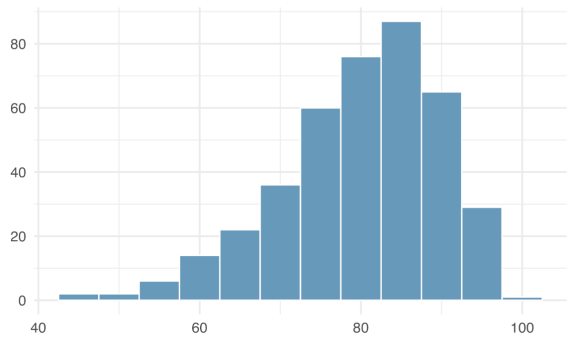
   (c) Create a boxplot *by hand* of the female orangutan weights in your conservancy

   (d) Suppose you had recorded the orangutan weights in kilograms instead (note that one pound is equal to 0.45 kilograms). What do you anticipate would happen to the measures of center—mean and median—that you calculated on the previously after a multiplicative change? What do you anticipate would happen to the measures of spread that you calculated—variance and IQR?

   (e) Now suppose instead that, when you weighed the orangutans, your scale was miscalibrated and mistakenly added 2 pounds to every orangutan's weight. What do you anticipate would happen to the measures of center after an additive change? What do you anticipate would happen to the measures of spread?

   See other answer sheet.

3. **[IMS 5.12] Median vs. mean**. Estimate the median for the 400 observations shown in the histogram and note whether you expect the mean to be higher or lower than the median.



Median is around 80. We would expect the mean to be slightly lower since the distribution is left skewed.

4. **[IMS 5.19] Midrange**. The midrange of a distribution is defined as the average of the maximum and the minimum of that distribution. Is this statistic robust to outliers and extreme skew? Explain your reasoning.

   No, the outliers are likely the maximum and the minimum of the distribution so a statistic based on these values cannot be robust to outliers.