Homework 2

Kyu Park

2021 1 20

1.

a) Population parameter
b) Observation
c) Variable
d) Sample statistics
e) The sample selection may have been biased. The population for this survey is the entire adults in the United States regardless of their availability of phones. However, in this survey, only the responses of adults who were able to respond to respond over phone were recorded. People may not have phone due to many reasons; one could be because they could not pay electricity bills, and if that is the case, people who did not watch CNN prime time may have been underestimated.

2.

Selecting students at random at an elementary school may bias the average. This is because there are many families with no children at all that were not counted in this research. Because their number and statistics cannot be included due to the lack of elementary school child who could respond, this may lead to the overestimation of the average number of children in a family.

3.

Let $f(w)$ be the mean squared difference of w, then

$$f(w) = \sum_{i=1}^{n} (x_i - w)^2 = \sum_{i=1}^{n} ({x_i}^2 - 2x_i w + w^2) = \sum_{i=1}^{n} ({x_i}^2) - 2w\sum_{i=1}^{n} x_i + \text{nw}^2$$

The goal is to minimize $f(w)$, therefore $f'(w) = 0$ and $f''(w) > 0$,

$$f'(w) = -2\sum_{i=1}^{n} x_i + 2nw = 0$$

$f'(w) = 0$ only if

$$-2\sum_{i=1}^{n} x_i = -2nw$$

and solving for w gives

$$w = \frac{\sum_{i=1}^{n} x_i}{n}$$

which equals to the sample average.

Since second derivative $f''(w)$ is positive,

$$f''(w) = 2n$$

at $w = \frac{\sum_{i=1}^{n} x_i}{n}$ the Mean Square Deviation is minimized.

Therefore, when w equals to the sample average, the Mean Square Deviation is minimized.

4.

a) $s_y = \sqrt{\frac{\sum_{i=1}^{N}((ax_i+b)-(a\bar{x} + b))^2}{N - 1}} = \sqrt{\frac{\sum_{i=1}^{N}(ax_i-a\bar{x})^2}{N - 1}} = |a|\sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N - 1}} = |a|\, s_x$

b) Say there are n amounts of elements in $X = (x_1, \ ..., x_n)$ and median value is x_m. The linear transformation will not change the order of elements in the set X, therefore the median after the linear transformation is $ax_m + b$, which is $a \times Median(X) + b$,

$$Median(Y) = a \times Median(X) + b$$

c) If $a > 0$, $Q1(Y) = aQ1(X) + b$ and $Q3(Y) = aQ3(x) + b$. This does not necessarily hold true if $a < 0$, since the order of the elements in the set will flip, as the max of X times a will become min of Y after the linear transformation. Likewise, Q1 and Q3 will flip since all the orders are flipped. For $a < 0$, $Q1(Y) = aQ3(x) + b$ and $aQ3(Y) = aQ1(x) + b$.

5.

a) $0.2 + 0.7 - 0.1 = 0.8$

b) $0.2 * 0.7 \neq 0.1$ Dependent Events

c) $0.1/0.7 = 1/7$

d) i) $0.2 * 0.7 = 0.14$
   ii) $0.2 + 0.7 - 0.14 = 0.76$
   iii) $0.14/0.2 = 0.7$

6.

a) (2), This is a very typical bell curve with unimodal distribution, skewed to nowhere. The histogram is roughly symmetric.
b) (3), This histogram shows uniform distribution with skewedness to nowhere.
c) (1), The histogram is skewed to the right, and is unimodal.

7.

Let x be the portfolio value current year and $\mu(x)$ be the expected return of the portfolio value next year. Then,

$$\textit{expected return} \ = \ \mu(x) = \ \ 0.1(0.12) \ +0.4(0.05) \ - \ 0.5(0.1) \ = \ -0.018 = \ -1.8\%$$

The expected return is -1.8% of the portfolio value current year.

The variance of the return could be described as,

$$\sum_{j=1}^{k} (x_j - \mu)^2 \, P\,(X = x_j) = \ 0.1(0.12 \ - \ (-0.018))^2 + 0.4(0.05 \ - \ (-0.018))^2 + 0.5(-0.1 \ - \ (-0.018)^2 \ = \ 0.007116$$

The variance σ^2 is 0.007116.

8.

a) The histogram of the median value of owner occupied houses is bell curved, and its distribution is skewed to the right.

b) The geological distribution of the median value of owner occupied houses is apparent in the map but not in the histogram. One could easily locate which geological area has approximately which median value of owner occupied houses through the map.
The numerical distribution of the median value of owner occupied houses is apparent in the histogram but not in the map. The percentage of each categories of mean value of owner occupied houses could be identified through the histogram but not through the map. Also, the mean, median and mode value of the median value of owner occupied houses could be identified easily through the histogram, which isn't comparably apparent in the map

c) Splitting the data by median could be done by

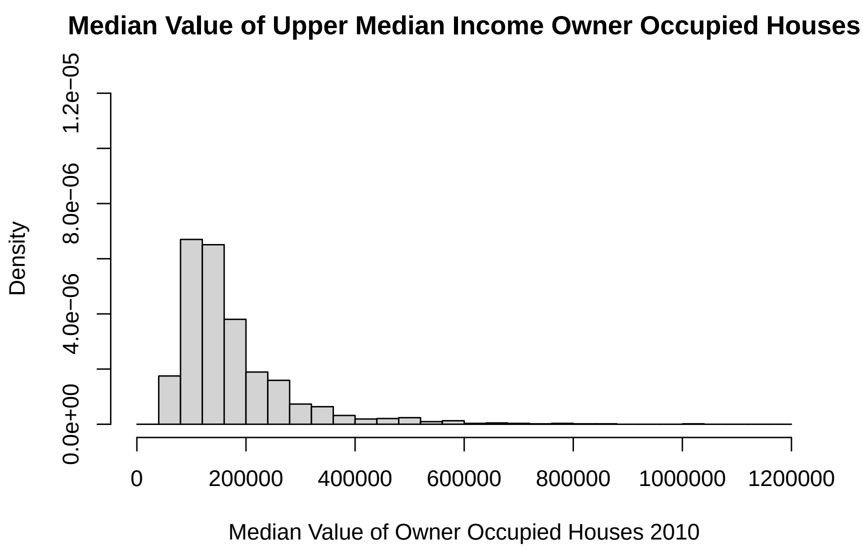
```
library(openintro)
```

```
## Loading required package: airports

## Loading required package: cherryblossom

## Loading required package: usdata
```

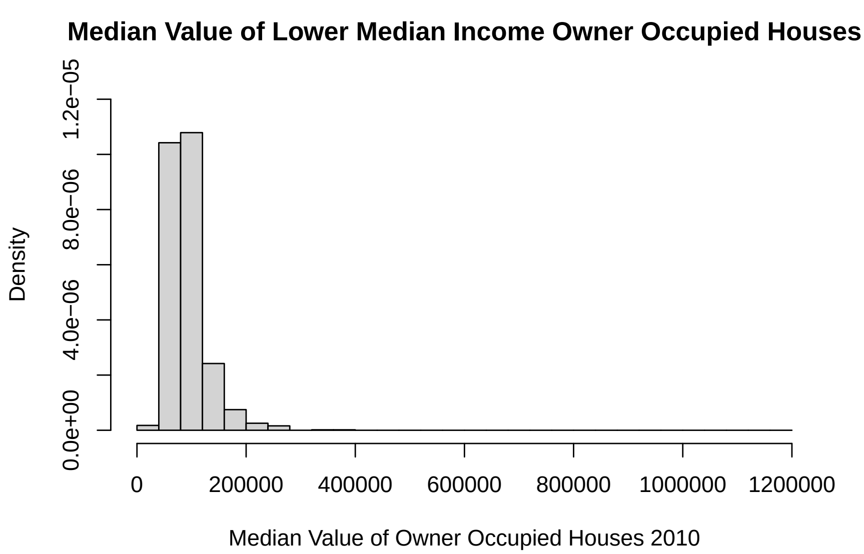
```
data(county_complete)
upperMedian <- county_complete[county_complete$median_household_income_2010 >
                                 median(county_complete$median_household_income_2010), ]
lowerMedian <- county_complete[county_complete$median_household_income_2010 <
                                 median(county_complete$median_household_income_2010), ]
```

Histogram for these two subset are shown below

```
hist(upperMedian$median_val_owner_occupied_2010,
     main = "Median Value of Upper Median Income Owner Occupied Houses",
     xlab = "Median Value of Owner Occupied Houses 2010",
     ylab = "Density", freq = FALSE, ylim = range(0, 0.000012),
     xlim = range(0, 1200000), breaks = seq(0, 1200000, 40000))
```




```
hist(lowerMedian$median_val_owner_occupied_2010,
     main = "Median Value of Lower Median Income Owner Occupied Houses",
     xlab = "Median Value of Owner Occupied Houses 2010",
     ylab = "Density", freq = FALSE, ylim = range(0, 0.000012),
     xlim = range(0, 1200000), breaks = seq(0, 1200000, 40000))
```



d) These two histograms differ in the distribution. The histogram of the upper median income owner is more evenly distributed, while the histogram of the lower median income owner is less distributed and more concentrated in the center. Also, it is notable that there are no median value of houses exceeding 40000 for lower median income owner, while there are a few for upper median income owner.