STA 100 Homework 1 Solution

Yidong Zhou

1. Please find below answers and justifications.

   (a) Observational. This study should be observational because the researcher is observing and analyzing data that already exist without any manipulation or intervention. It is not feasible or ethical to assign individuals to smoking or non-smoking groups.

   (b) Experimental. This study should be experimental because the researcher is actively manipulating the treatment assignment by randomly assigning participants to different groups and then comparing the outcomes between the groups.

   (c) Observational. This study should be observational because the researcher is observing and measuring variables without any intervention or manipulation of the participants' exercise habits.

   (d) Experimental. This study should be experimental because the researcher is actively manipulating the teaching method by assigning students to different classes and assessing the impact of the new method on academic performance.

   (e) Observational. This study should be observational because the researcher is observing and comparing individuals in different environmental conditions, without any active manipulation or intervention on their exposure to pollution.

2. These variables are

   (a) discrete

   (b) discrete/continuous

   (c) nominal

   (d) ordinal

3. (a) The average number of times a student exercised is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{100}(0 \times 20 + 1 \times 40 + 2 \times 24 + 3 \times 14 + 10 \times 2) = 1.5.$$

   (b) The median of the number of times a student exercised is the average of the 50th and 51th sorted observations, both of which are equal to 1. The median is thus $(1+1)/2 = 1$.

   (c) The variance of the number of times a student exercised is

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum (X_i - \bar{X})^2 \\ &= \frac{1}{100-1}[20 \times (0-1.5)^2 + 40 \times (1-1.5)^2 + 24 \times (2-1.5)^2 + 14 \times (3-1.5)^2 + 2 \times (10-1.5)^2] \\ &= 2.39. \end{aligned}$$

   (d) The standard deviation of the number of times a student exercised is $s = \sqrt{s^2} = \sqrt{2.39} = 1.55$.

   (e) The first quartile of the number of times a student exercised is the average of the 25th and 26th sorted observations, both of which are equal to 1. The first quartile is thus $Q_1 = (1+1)/2 = 1$.

   (f) The third quartile of the number of times a student exercised is the average of the 75th and 76th sorted observations, both of which are equal to 2. The third quartile is thus $Q_3 = (2+2)/2 = 2$.

(g) Interquartile range is $\text{IQR} = Q_3 - Q_1 = 2 - 1 = 1$, so the lower fence is $\text{LF} = Q_1 - 1.5\text{IQR} = 1 - 1.5 \times 1 = -0.5$.

(h) The upper fence is $\text{UF} = Q_3 + 1.5\text{IQR} = 2 + 1.5 \times 1 = 3.5$.

(i) The outliers in the data set are the two 10's since they are greater than the upper fence.

4. (a) FALSE. The standard deviation measures the spread of the dataset, whereas the mean measures the center. We could have a dataset that has a very small spread but a large center. For example, a dataset consists of ten values of 1000 has a mean of 1000 and a standard deviation of 0.

   (b) FALSE. The range of a dataset is defined as the difference between the maximum and minimum values of the dataset, and both the maximum and minimum could potentially be outliers (extreme points outside the main bulk of the data). So having extreme points outside the main bulk of the data versus not having those extreme points will have a strong influence on the range of the dataset.

   (c) TRUE. By definition, the qth percentile of a dataset is the value for which $q\%$ of the data is below it and $(100 - q)\%$ is above. Here $q = 90$.

   (d) TRUE. Since the mean of a dataset is the average of all the values in the dataset, of course including its outliers, the values of those extreme points have a strong influence on the mean.

5. Let S_1 = Species 1, S_2 = Species 2, and I = Infected.

   (a) $P(S_1) = \frac{54}{109} = 0.4954$.

   (b) $P(I) = \frac{58}{109} = 0.5321$.

   (c) $P(S_1 \cap I) = \frac{38}{109} = 0.3486$.

   (d) $P(S_2 \cap I^C) = \frac{35}{109} = 0.3211$.

6. (a) $P(I|S_1) = \frac{P(S_1 \cap I)}{P(S_1)} = \frac{38/109}{54/109} = 0.7037$.

   (b) $P(I|S_2) = \frac{P(S_2 \cap I)}{P(S_2)} = \frac{20/109}{55/109} = 0.3636$.

   (c) $P(S_1|I) = \frac{P(S_1 \cap I)}{P(I)} = \frac{38/109}{58/109} = 0.6552$.

   (d) $P(S_2|I) = \frac{P(S_2 \cap I)}{P(I)} = \frac{20/109}{58/109} = 0.3448$.

   (e) Since $P(I|S_1) = 0.7037 \neq P(I) = 0.5321$, they are not independent.

7. Let + and − denote the events testing positive and negative, respectively. Define D as the event that someone has the disease and D^C its complement (someone does not have the disease).

   (a) $P(+ \cap D) = P(+|D)P(D) = 0.95 \times 0.04 = 0.038$.

   (b)

$$\begin{aligned} P(+) &= P(+ \cap D) + P(+ \cap D^C) \\ &= P(+|D)P(D) + P(+|D^C)P(D^C) \\ &= 0.95 \times 0.04 + (1 - 0.99) \times (1 - 0.04) \\ &= 0.0476. \end{aligned}$$

   (c)

$$P(D|+) = \frac{P(+ \cap D)}{P(+)} = \frac{0.038}{0.0476} = 0.7983.$$

(d)

$$\begin{aligned}
P(D^C|-) &= \frac{P(- \cap D^C)}{P(-)} \\
&= \frac{P(-|D^C)P(D^C)}{P(-|D^C)P(D^C) + P(-|D)P(D)} \\
&= \frac{0.99 \times (1 - 0.04)}{0.99 \times (1 - 0.04) + (1 - 0.95) \times 0.04} \\
&= 0.9979.
\end{aligned}$$

8. (a) The number of the hairs Yidong can expect to lose on a given morning is the expectation of X.

$$\begin{aligned}
E(X) &= \sum_{i=1}^{5} i \times P(X = i) \\
&= 1 \times 0.05 + 2 \times 0.1 + 3 \times 0.5 + 4 \times 0.3 + 5 \times 0.05 \\
&= 3.2.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{i=1}^{5} (i - E(X))^2 \times P(X = i) \\
&= (1 - 3.2)^2 \times 0.05 + (2 - 3.2)^2 \times 0.1 + (3 - 3.2)^2 \times 0.5 + (4 - 3.2)^2 \times 0.3 + (5 - 3.2)^2 \times 0.05 \\
&= 0.76.
\end{aligned}$$

The standard deviation of X is thus $\sqrt{\mathrm{Var}(X)} = \sqrt{0.76} = 0.87$.

(b) Let A denote the event that Yidong loses at 2 hairs on a given morning. Let B denote the event that Yidong loses at least 2 hairs per day over 10 days. If follows that

$$P(A) = 1 - P(A^C) = 1 - P(X = 1) = 1 - 0.05 = 0.95.$$

Note that each of the days are independent. One has

$$P(B) = P(A)^{10} = 0.95^{10} = 0.60.$$

(c) Let Y denote the number of the days that Yidong loses 5 hairs over 10 days. Then Y is binomial distributed with parameters $n = 10$ and $p = P(X = 5) = 0.05$. The probability that Yidong loses 5 hairs at least one day over 10 days is

$$P(Y \geq 1) = 1 - P(Y = 0) = 1 - 0.95^{10} = 0.40.$$

9. Here $X \sim B(12, 0.37)$.

(a)

$$P(X = 0) = \binom{12}{0} 0.37^0 (1 - 0.37)^{12} = 0.04.$$

(b)

$$\begin{aligned}
P(X \geq 11) &= P(X = 11) + P(X = 12) \\
&= \binom{12}{11} 0.37^{11} (1 - 0.37)^1 + \binom{12}{12} 0.37^{12} (1 - 0.37)^0 \\
&= 0.00014.
\end{aligned}$$

(c)

$$\begin{aligned}P(4 < X \leq 6) &= P(X = 5) + P(X = 6)\\ &= \binom{12}{5} 0.37^5 (1 - 0.37)^7 + \binom{12}{6} 0.37^6 (1 - 0.37)^6\\ &= 0.365.\end{aligned}$$

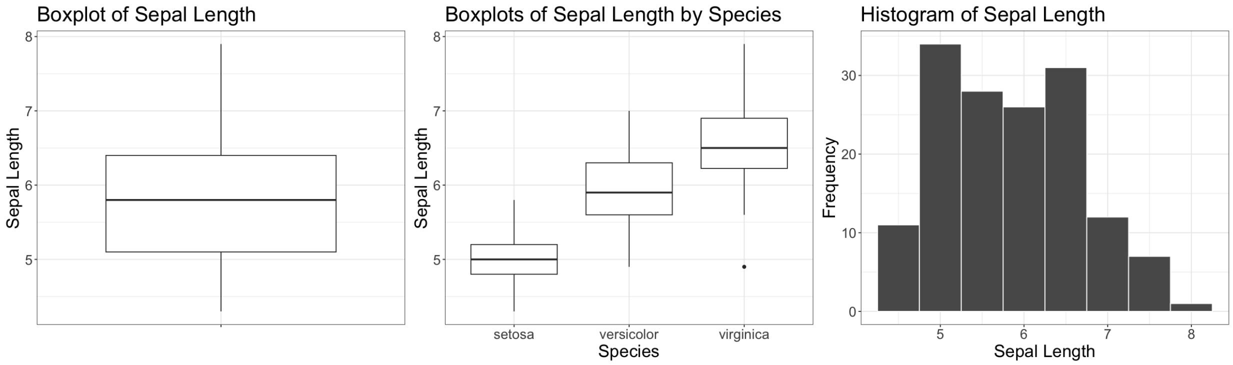
(d)

$$E(X) = np = 12 \times 0.37 = 4.44.$$

$$\mathrm{Var}(X) = np(1 - p) = 12 \times 0.37 \times (1 - 0.37) = 2.7972.$$

In [3]:

```
library(ggplot2)

# Load the Iris dataset
data(iris)

# (a)
head(iris)

# (b)
# Calculate mean, standard deviation, and variance for sepal length
mean_sepal_length <- mean(iris$Sepal.Length)
sd_sepal_length <- sd(iris$Sepal.Length)
var_sepal_length <- var(iris$Sepal.Length)

# (c)
# Find the five-number summary and calculate IQR for sepal length
summary_sepal_length <- summary(iris$Sepal.Length)
iqr_sepal_length <- 6.4 - 5.1

# (d)
# Plot a boxplot of sepal length
g1 <- ggplot(iris, aes(x = "", y = Sepal.Length)) +
    geom_boxplot() +
    labs(title = "Boxplot of Sepal Length",
         x = "", y = "Sepal Length") +
    theme_bw() +
    theme(text = element_text(size = 20))

# (e)
# Plot side-by-side boxplots of sepal length split by species
g2 <- ggplot(iris, aes(x = Species, y = Sepal.Length)) +
    geom_boxplot() +
    labs(title = "Boxplots of Sepal Length by Species",
         x = "Species", y = "Sepal Length") +
    theme_bw() +
    theme(text = element_text(size = 20))

# (f)
# Plot a histogram of sepal length over all species
g3 <- ggplot(iris, aes(x = Sepal.Length)) +
    geom_histogram(binwidth = 0.5, color = "white") +
    labs(title = "Histogram of Sepal Length",
         x = "Sepal Length", y = "Frequency") +
    theme_bw() +
    theme(text = element_text(size = 20))
# Additional comments and observations can be made
# on the generated plots and statistics.

library(patchwork)
options(repr.plot.width=20, repr.plot.height=6)
g1 + g2 + g3
```



(a) The sepal length, petal width, and species of the 6th observation in the dataset are 5.4, 0.4, setosa, respectively.

(b) The mean, standard deviation, and variance for sepal length are 5.84, 0.83, 0.69, respectively.

(c) The five number summary is

```
Min.    1st Qu. Median  Mean    3rd Qu. Max.
4.300   5.100   5.800   5.843   6.400   7.900
```

and the IQR is thus $6.4 - 5.1 = 1.3$.

(d) There are no outliers in the data.

(e) The sepal length for virginica is in general larger than that for versicolor, which is in general larger than that for setosa. The spread of the sepal length for setosa is much smaller compared to the other two species, versicolor and virginica. And there is one outlier of sepal length for virginica, which is an extremely low value relative to the rest of the data points in the same group.

(f) According to the histogram, the sepal length data is unimodal and slightly right-skewed.