Assignment 1

Student Name: Kevin Yu

Student Number: 1462539

Subject Code: MAST20005

Subject Name: Statistics

August 21, 2024

Exercise 1

Exercise 2

Part 1

Please note that I have abbreviated $\sum_{i=1}^{m}$ as $\sum$ at some places to avoid overcrowding

$$
\begin{aligned}
L(\sqrt{p}) &= \prod_{i=1}^{m} \binom{n}{x_i} (\sqrt{p})^{x_i} (1-\sqrt{p})^{n-x_i} \\
&= \prod_{i=1}^{m} \left[\binom{n}{x_i} \right] (\sqrt{p})^{\sum x_i} (1-\sqrt{p})^{\sum (n-x_i)} \\
\implies \ln L(\sqrt{p}) &= \ln \prod_{i=1}^{m} \left[\binom{n}{x_i} \right] + \sum_{i=1}^{m} (x) \ln \sqrt{p} + \sum_{i=1}^{m} (n-x_i) \ln (1-\sqrt{p}) \\
\implies \frac{d}{d\sqrt{p}} [\ln L(\sqrt{p}))] &= \frac{\sum x_i \cdot \frac{1}{2} p^{1/2}}{\sqrt{p}} + \frac{(mn - \sum x_i) \cdot -\frac{1}{2} p^{1/2}}{1-\sqrt{p}} \\
&= \frac{(1-\sqrt{p}) p^{-1/2} \sum x_i - mn + \sum x_i}{2\sqrt{p}(1-\sqrt{p})}
\end{aligned}
$$

Setting $\frac{d}{d\sqrt{p}} [\ln L(\sqrt{p}))] = 0$ yields,

$$
\begin{aligned}
& (1-\sqrt{p}) p^{-1/2} \sum_{i=1}^{m} x_i - mn + \sum_{i=1}^{m} x_i = 0 \\
\implies & (p^{-1/2} - 1) \sum_{i=1}^{m} x_i - mn + \sum_{i=1}^{m} x_i = 0 \\
\implies & p^{-1/2} \sum_{i=1}^{m} x_i - mn = 0 \\
\implies & p^{-1/2} = \frac{mn}{\sum x_i} = \frac{n}{\bar{X}_m} \\
\implies & p^{1/2} = \frac{\bar{X}_m}{n} \\
& \therefore p = \frac{\bar{X}_m^2}{n^2} \qquad \blacksquare
\end{aligned}
$$

Part 2

$$
\bar{x}_3 = \frac{1}{3}
$$

Exercise 3

Part 2

In addition to the variance, we can calculate

- the range (max - min) to measure the spread of the data,
- the interquartile range (IQR), the difference between the third and first quartiles, which is less sensitive to outliers,
- the median as a measure of central tendency that is less sensitive to outliers.

Now, let $y_{(1)}, y_{(2)}, \ldots, y_{(9)}$ and $z_{(1)}, z_{(2)}, \ldots, z_{(9)}$ be the ordered samples from Sample 1, Sample 2 respectively.

For Sample 1, $y_{(1)} = 1.333$ and $y_{(9)} = 1.684$ so the range is $1.684 - 1.333 = 0.351$.

For Sample 2, $z_{(1)} = 1.333$ and $z_{(9)} = 1.523$ so the range is $1.523 - 1.333 = 0.190$.

Now to calculate the interquantile range, we recall from the lecture that having $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ as the ordered observations; let the p-th quantile of the observations be denoted by $\hat{c}_p$ where $0 < p < 1$. Then, letting $k = 1+(n-1)p$ and t and w be the whole and fractional part of k respectively, (i.e. $t = \lfloor k \rfloor$ and $w = k - t$),

$$\hat{c}_p = x_{(t)} + w(x_{(t+1)} - x_{(t)}).$$

Therefore, for Sample 1, the first quartile "position" is at $1 + (9-1)(0.25) = 3$ so the first quartile is $\hat{q}_1 = \hat{c}_{0.25} = y_{(3)} = 1.447$. The third quartile "position" is at $1 + (9-1)(0.75) = 7$ so the third quartile is $y_{(7)} = 1.577$. Hence, the IQR is $1.577 - 1.447 = 0.130$.

For Sample 2, the first and third quartile "position" is the same as Sample 1. Therefore, the first quartile is $z_{(3)} = 1.333$ and the third quartile is $z_{(7)} = 1.333$. Hence, the IQR is $1.333 - 1.333 = 0$. We also note that since IQR of Sample 2 is 0, so $z_{(9)} = 1.523$ is an extreme outlier.

Finally, the median of Sample 1 is $y_{(5)} = 1.529$ and the median of Sample 2 is $z_{(5)} = 1.333$.

Ultimately, the range and IQR of Sample 1 are greater than those of Sample 2. This suggests that Sample 1 has a greater variability, while Sample 2 has the same value for the first, second (medium) and third quartiles, indicating that the data is more concentrated around a single value. ■

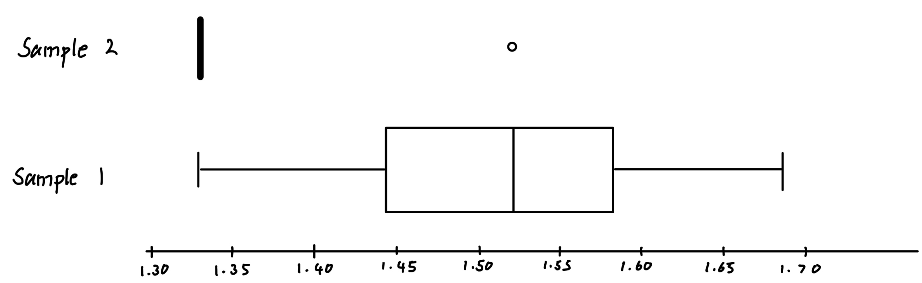

Figure 1: Boxplot of Sample 1 and Sample 2

Exercise 4

Exercise 5