Project 1

Akal Ustat Singh

STATS 50-05, Spring 2025

For the purposes of this project, I have used the R programming language. The exact code is attached as an appendix to this document.

1. Binomial $\hat{p}$ Estimation

$p \simeq 0.677 \mp 0.067$

We are given the number of success from 200 numbers from a $X \sim \text{Bin}(40, p)$ population. For each number, we are also given the sample proportions $\hat{p} = \frac{X}{n}$. Given this information. we need to:

- (a) describe the shape of the $\hat{p}$ values using both summary measures and visual evidence
- (b) estimate the standard error for $\hat{p}$ values.

Let us first analyze the shape and summary measures of $\hat{p}$.

a. Summary Measures

We can evaluate the mean and median numerically.

Since we can estimate X using $\overline{X} = X_1 + X_2 + ... + X_{200}$ and all values $\hat{p_i} = \frac{X_i}{n} = \frac{X}{40}$, $\mu_{\hat{p}} = \frac{1}{200}\sum_{k=1}^{200} \hat{p_k} \simeq 0.676625 \simeq 0.677$. Since $E[\hat{p}] = p$, we can say that $p \simeq 0.677$.

Additionally, the standard deviation can be computed $\sigma_{\hat{p}} \simeq 0.06776574$. We can also calculate the standard error: $\sqrt{\frac{1}{200} \cdot \hat{p}(1-\hat{p})} \simeq 0.03307594 \simeq 0.033$.

$M_{\hat{p}} \simeq 0.675$ (since $M_{\hat{p}} > \mu_{\hat{p}}$, the data is left-skewed), range $= 0.325$, and IQR $= 0.1$.

b. Visual Evidence

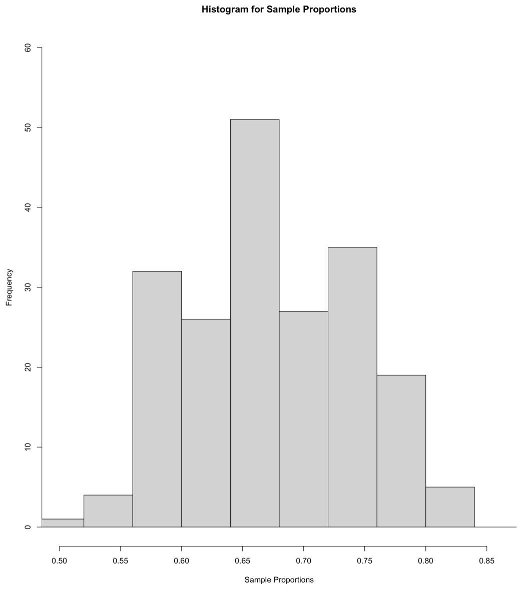

Figure 1: The Histogram of the $\hat{p}$ values of 200 random numbers from the population $X \sim \text{Bin}(40, p)$.

As we can see in the histogram (and using the mean and median calculated before), the data is left-skewed. This is further shown using the box plot below:

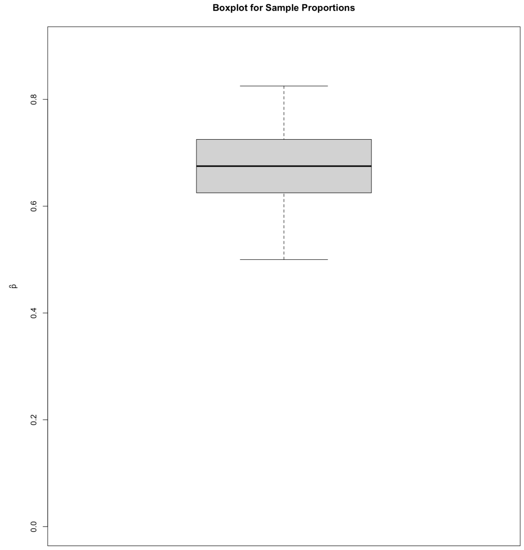

Figure 2: The boxplot of the $\hat{p}$ values of 200 random numbers from the population $X \sim \text{Bin}(40, p)$.

2. Oxygen Concetration Levels

a. Visual Plots

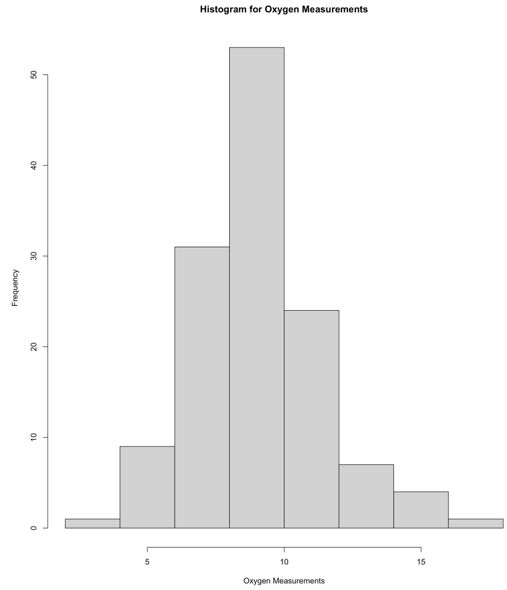

Figure 3: Histogram of the Oxygen level data.

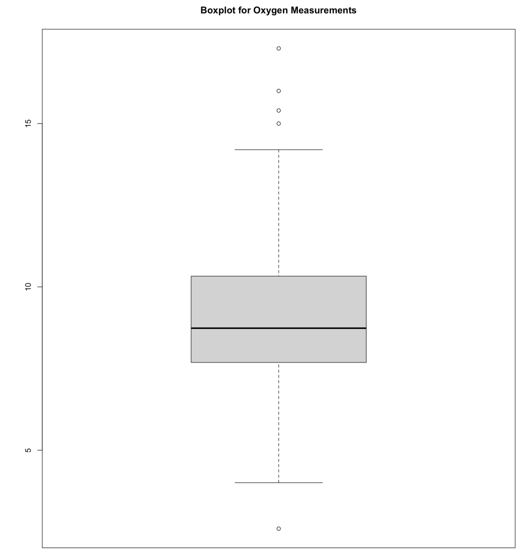

Figure 4: Boxplot of the Oxygen level data.

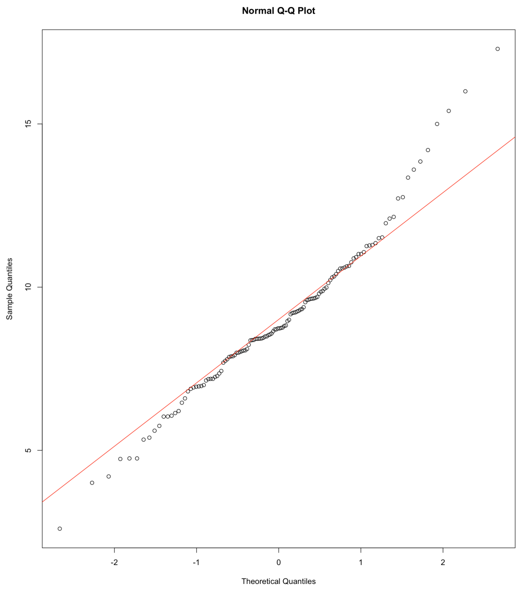

Figure 5: Normality Plot of the Oxygen level data.

As visible from the Normality plot, the data is approximately normal. However, the deviations at the tails, especially visible at the upper tail, indicate slight skewness. This is supported by the number of outliers visible in the boxplot and and the very slight skew in the histogram.

b. Mean and Std. Dev.

Letting the population $\overline{X} = X_1 + X_2 + ... + X_{130}$, we can caluclate the mean, standard deviation, and standard error.

Doing the calculations in the R programming language, we have the following result:

$\mu_{\overline{X}} \simeq 9.003598 \simeq 9.00$ mg/L.

$\sigma_{\overline{X}} = \simeq 2.414338 \simeq 2.41$ mg/L.

Standard error $\simeq 0.2117514 \simeq 0.21$ mg/L.

Given this information, we can conclude that the average oxygen concentration in the river is above 5 mg/L. Firstly, the estimated mean is about 9mg/L, and furthermore, since the standard deviation is approximately 2.41 mg, 5 mg/L is above 1 standard deviation from the mean (close to 1.5).

Appendix A: R Code for Q1

Akal Ustat Singh
STATS 50-05, Spring 2025

```
library(readxl)
data <- read_excel(
  "./project1/Data_Project1.xlsx",
  sheet = "binom",
  col_names = FALSE
)
successes <- data[, 1]
proportions <- data[, 2]
proportions_as_numeric <- c(as.numeric(unlist(proportions)))

break_intervals <- seq(
  floor(
    min(
      proportions
    )
  ), ceiling(max(proportions)),
  by = 0.04
)

hist(proportions_as_numeric,
  breaks = break_intervals,
  main = "Histogram for Sample Proportions",
  xlab = "Sample Proportions",
  xlim = c(0.5, 0.86),
  ylim = c(0, 60)
)

boxplot(
  x = proportions_as_numeric, ylim = c(0.0, 0.9),
  main = "Boxplot for Sample Proportions",
  ylab = expression(hat(p), " value")
)
mean_p <- mean(proportions_as_numeric)
std_dev_p <- sd(proportions_as_numeric)
sigma_p <- sqrt(1 / 200 * (mean_p * (1 - mean_p)))
median_p <- median(proportions_as_numeric)
range_p <- max(proportions_as_numeric) - min(proportions_as_numeric)
iqr_p <- IQR(proportions_as_numeric)
```

Appendix B: R Code for Q2

Akal Ustat Singh
STATS 50-05, Spring 2025

```
library(readxl)
data <- read_excel(
  "./project1/Data_Project1.xlsx",
  sheet = "Oxygen",
```

```r
  col_names = FALSE
)
measurements <- data[,1]
measurements_as_numeric <- c(as.numeric(unlist(measurements)))

hist(measurements_as_numeric,
     main = "Histogram for Oxygen Measurements",
     xlab = "Oxygen Measurements",
)
boxplot(
  x = measurements_as_numeric,
  main = "Boxplot for Oxygen Measurements"
)

qqnorm(measurements_as_numeric)
qqline(measurements_as_numeric, col = "red")

mean_msmt <- mean(measurements_as_numeric)
sigma_msmt <- sd(measurements_as_numeric)
std_error <- sigma_msmt / sqrt(length(measurements_as_numeric))
```