Project 2

Akal Ustat Singh, Phillip Gudov, Frank Kutsar
STATS 50-05, Spring 2025

1. Dissolved Oxygen from American River samples

a. Normality of Dissolved Oxygen Data

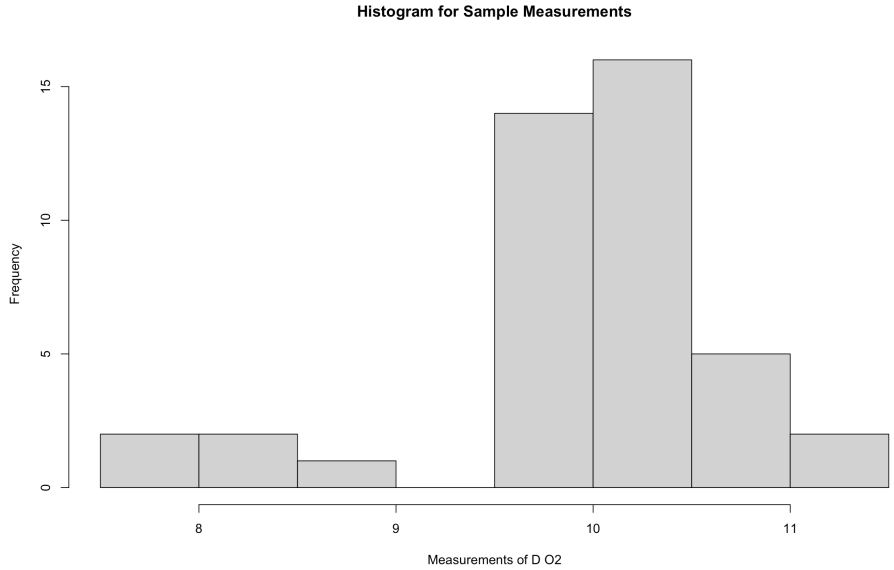


Figure 1: Histogram of dissolved oxygen data in the American River.

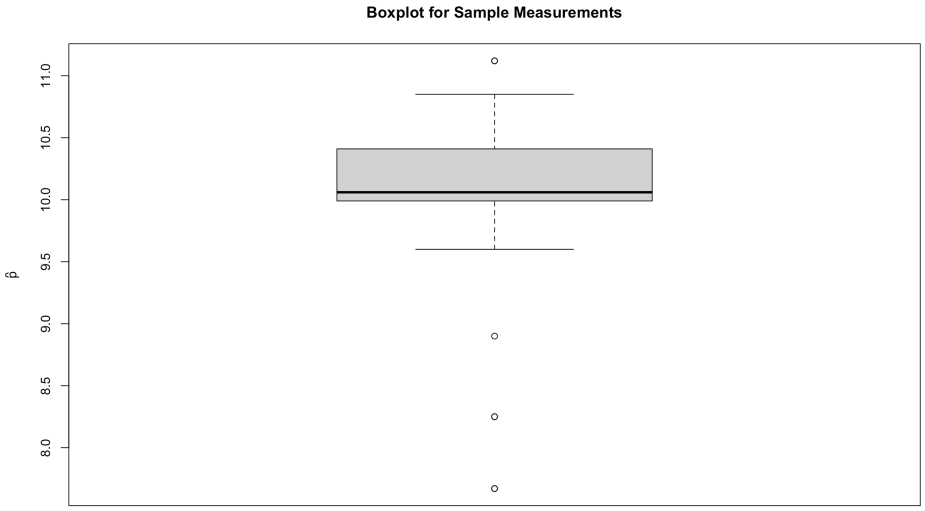


Figure 2: Boxplot of dissolved oxygen data in the American River.

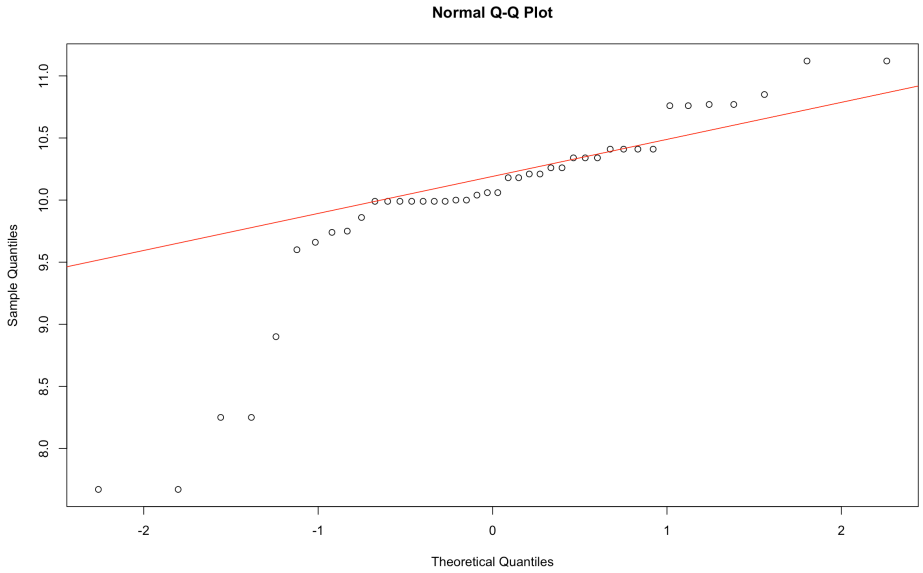


Figure 3: Normality plot of dissolved oxygen data in the American River.

Histogram: The distribution is roughly mound-shaped but slightly right-skewed. Most data cluster between 9.8 – 10.5 mg/L, with a thin tail produced by a few lower (≈ 7.7 – 8.3 mg/L) and higher (≈ 11 mg/L) values.

Box-plot: The central 50 % of the data (IQR) lies tightly between ≈ 10.0 and 10.4 mg/L. Four points plot outside the whiskers and are flagged as outliers (two low ≈ 7.7 mg/L and two high ≈ 11 mg/L).

Normality plot: As we can see, the data is quite different from the normality line. Thus, we have probable reason to suspect that this data is not from a normal population.

ks.Tests

```
 Asymptotic one-sample Kolmogorov-Smirnov test

data:  d_oxygen_num
D = 0.26227, p-value = 0.006191
alternative hypothesis: two-sided
```

Since $p < 0.05$, we have a statistically significant reason to eliminate the null hypothesis. We accept the alternative hypothesis (that the data is not normally distributed.)

b. Sample Statistics and Tests

$\overline{x} = 9.989$

η (sample median): 10.06

$s = 0.789$

According to the histogram, we have four outliers. We cannot therefore use the t-test. It is not good to use the z-test: $n < 50$, so it might not be reasonable to use s to estimate σ; the data is quite likely not even be normal.

2. Dissolved Oxygen from tap water samples

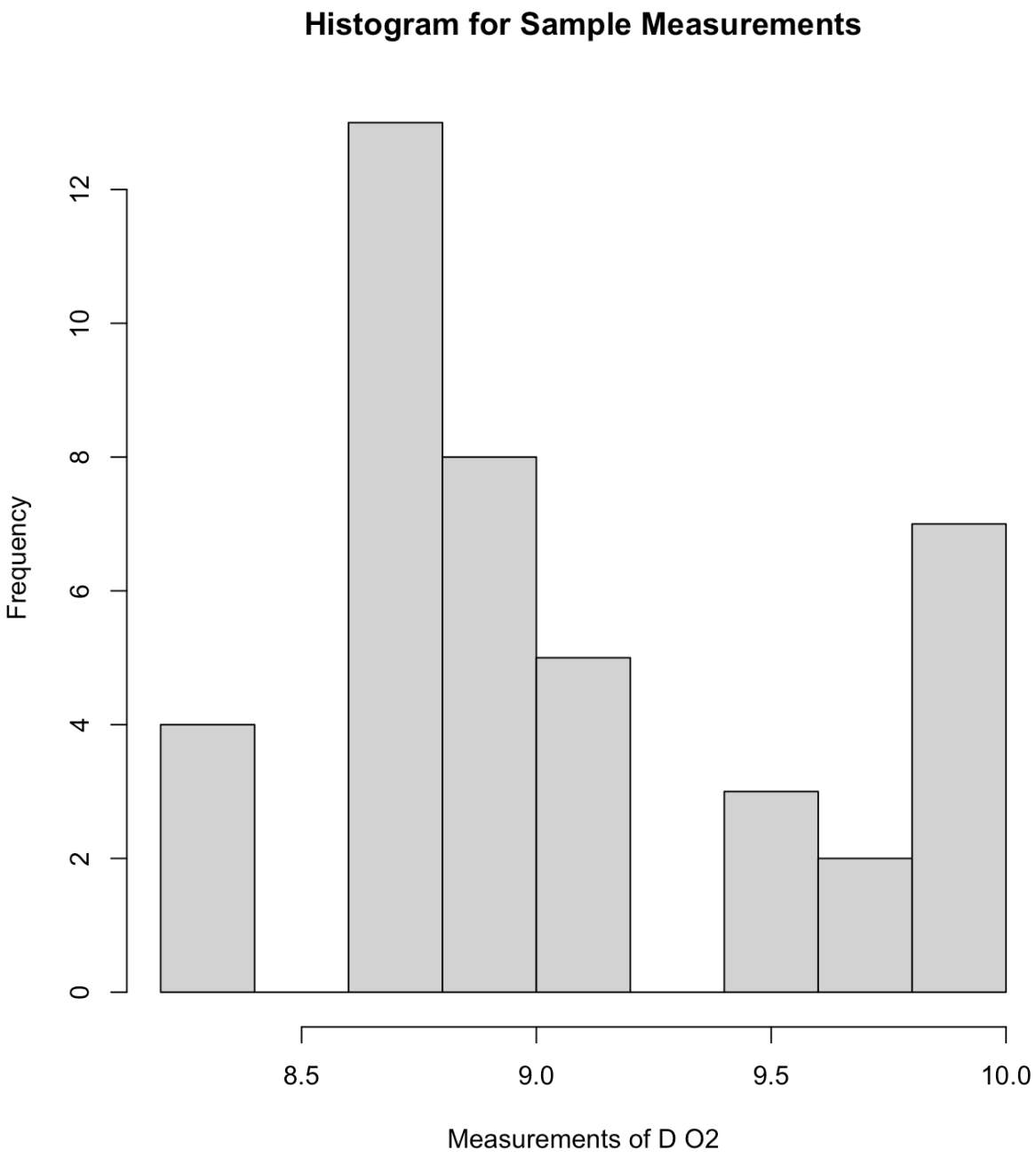


Figure 4: Histogram of dissolved oxygen data in tap water.

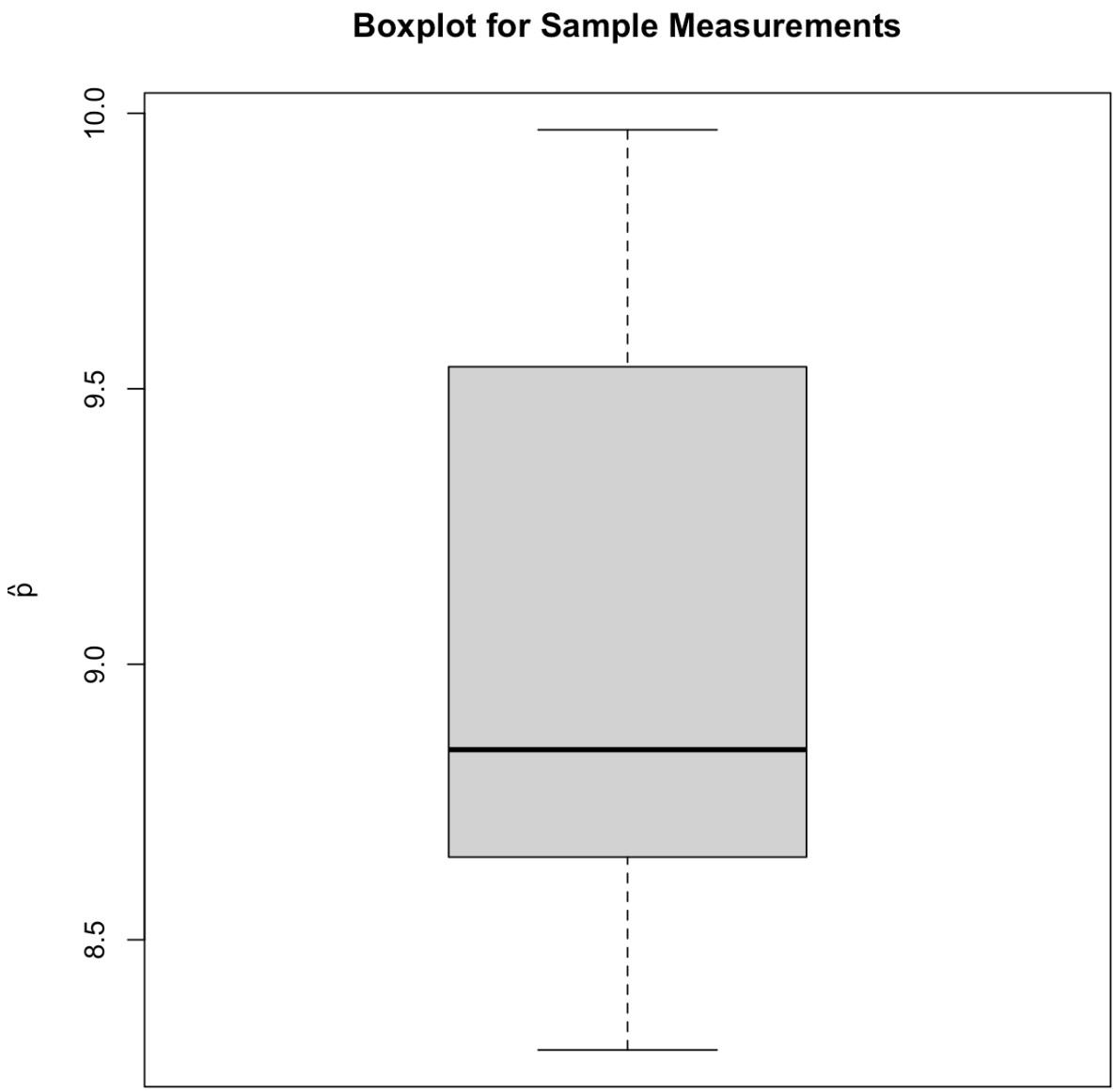


Figure 5: Boxplot of dissolved oxygen data in tap water.

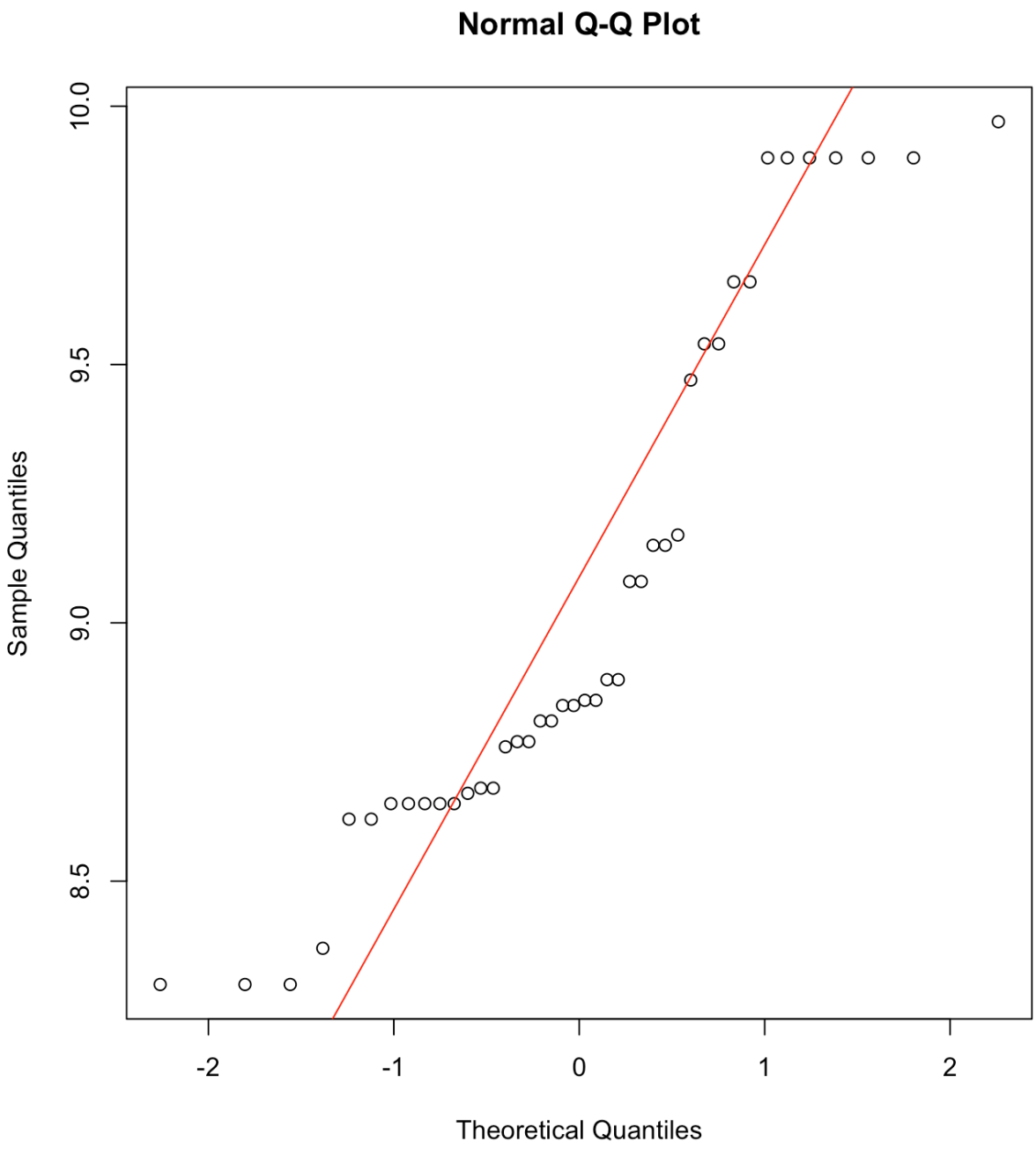


Figure 6: Normality plot of dissolved oxygen data in tap water.

The normality plot seems skewed like the previous question, so the data might not be from a normal population.

ks.Tests

```
data:  d_oxygen_num
D = 0.21059, p-value = 0.04821
alternative hypothesis: two-sided
```

Since $p < 0.05$, the results are statistically significant, therefore the null hypothesis must be rejected. Thus, the data might not be normally distributed.

b. Sample Statistics and Tests

$\overline{x} = 9.041$

η (sample median): 8.845

$s = 0.516$

We have no outliers, as shown in the histogram. Since $n > 30$, the sample is random, and we know s, we can use the t-test.

t-test

At 95% confidence, $\alpha = 0.05$. We have $df = n - 1 = 41$.

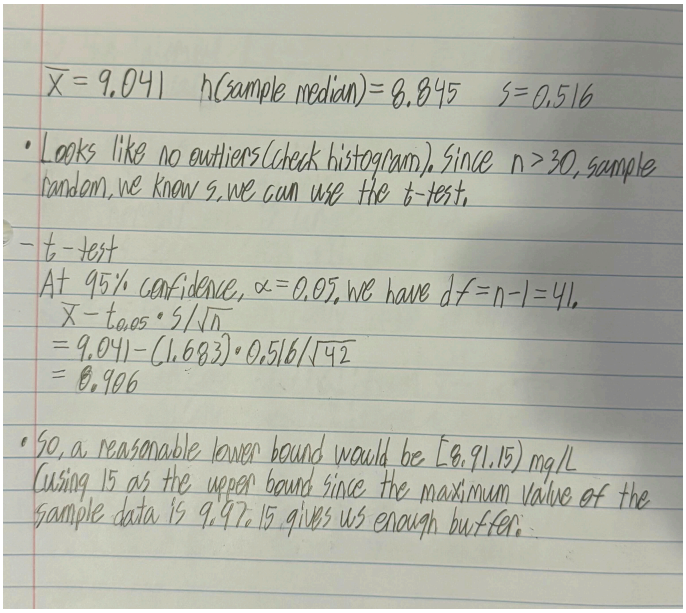
$\overline{x} - t_{0.05} \cdot \frac{s}{\sqrt{n}}$

$= 9.041 - (1.683) \cdot \frac{0.516}{\sqrt{42}}$

$= 8.906$.

So, a reasonable lower bound would be [8.91, 15) mg/L (using 15 as the upper bound since the maximum value of the sample data is 9.97. 15 gives us enough buffer).

Checking work on paper



3. Extra Credit

a. Scatterplots

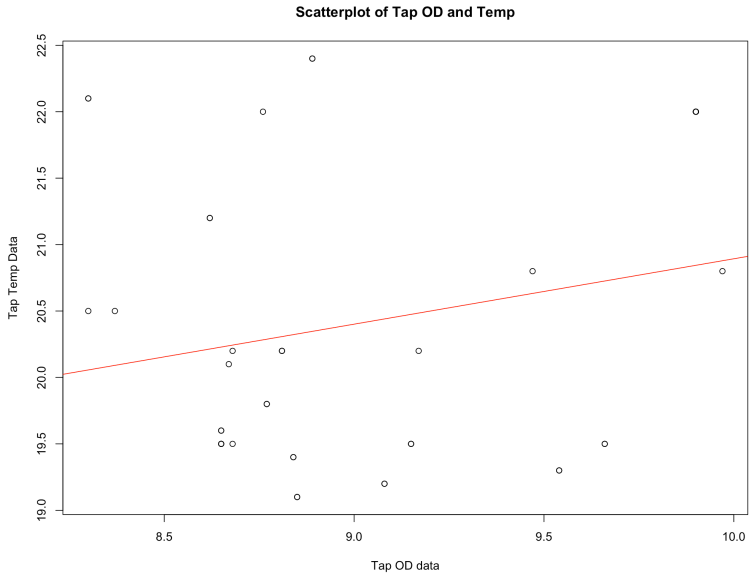


Figure 8: Scatter plot of OD data to temp data for Tap.

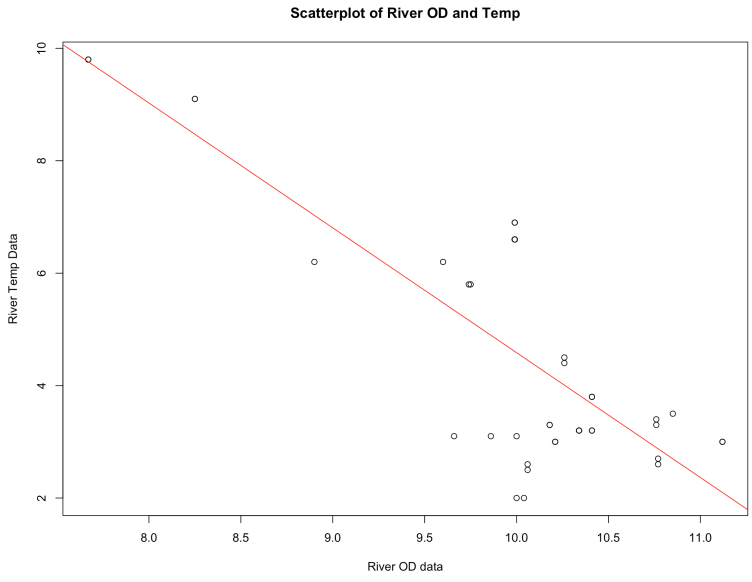


Figure 9: Scatter plot of OD data to temp data for River.

As we can see visually, the Tap water has a rather weak correlation (which may relate to environmental factors, namely that the rangen of temperatures is rather narrow and the fact that tap water is treated), while river water has a strong negative correlation between temperature and OD data. Here are the numerical values:

```
> tap_cor
[1] 0.224143
```

```
> river_cor
[1] -0.7959663
```

For the tap water measurements, we get the follow linear regression data:

```
lm(formula = tap_temp_num ~ tap_od_num)

Coefficients:
(Intercept)   tap_od_num
    15.9707       0.4923
```

which gives us the following linear regression equation: $y = 0.4923x + 15.9707$.

The river water data is computed similarly to give us the equation $y = -2.221x + 26.794$.

```
Call:
lm(formula = river_temp_num ~ river_od_num)

Coefficients:
 (Intercept)  river_od_num
      26.794        -2.221
```

According to Navidi Chapter 7.4 ("Checking Assumptions and Transforming Data), with outliers, we should first attempt to determine why we have these outliers. If an outlier is caused by recording or equipment errors, they can be deleted from the data set. Outliers that do not affect the least squares line or estimated standard deviations of slope or intercept, they can stay. If deletion of an outlier significantly changes the regression line, then the coefficients should be reported as an interval.

b. Confidence Interval 2b in Context

Yes, we can reasonably conclude that the average dissolved oxygen concentration in tap water is above 7mg/L. In fact, we are 95% certain (our range is so much above 7).

c. Pressure measurements, temperature, OD measurements.

For scatterplots and regression/correlation testing between OD and temperature, please see the answer to 3a.

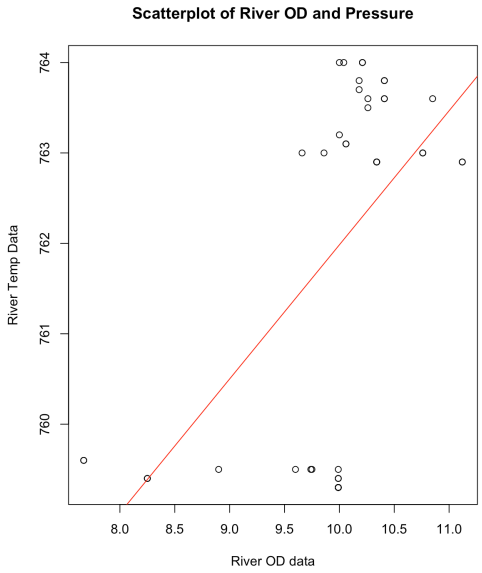


Figure 10: Scatter plot of OD data to Pressure data for River.

As we can see, there is some positive correlation between pressure and OD data. Using regression, we get the following equation: $y = 1.483x + 747.153$.

```
Call:
lm(formula = river_press_num ~ river_od_num)

Coefficients:
 (Intercept)  river_od_num
     747.153         1.483
```

```
> press_cor
[1] 0.5927292
```

So, we also have a relatively strong positive correlation with $r = 0.593$.

Appendix A: R Code for Question 1

Akal Ustat Singh, Phillip Gustov, Frank Kutsar
STATS 50-05, Spring 2025

```r
library(readxl)
file_path <- "./project2/AmericanRiver-Sp2024.xlsx"

# Read the first range of cells
range1 <- read_excel(file_path, range = "D12:D27", col_names = FALSE)

# Read the second range of cells
range2 <- read_excel(file_path, range = "D31:D44", col_names = FALSE)

# Read the third range of cells
range3 <- read_excel(file_path, range = "D48:D59", col_names = FALSE)

dissolved_oxygen <- rbind(range1, range2, range3)
d_oxygen_num <- c(as.numeric(unlist(dissolved_oxygen)))

hist(
    d_oxygen_num,
     main = "Histogram for Sample Measurements",
     xlab = "Measurements of D O2",
)

boxplot(
  x = d_oxygen_num,
  main = "Boxplot for Sample Measurements",
  ylab = expression(hat(p), "value")
)

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


sample_mean = mean(d_oxygen_num)
median = median(d_oxygen_num)
std_dev = sd(d_oxygen_num)
ks.test(d_oxygen_num, 'pnorm', sample_mean, std_dev)
```

Appendix B: R Code for Question 2

Akal Ustat Singh, Phillip Gustov, Frank Kutsar
STATS 50-05, Spring 2025

```r
library(readxl)
file_path <- "./project2/AmericanRiver-Sp2024.xlsx"

# Read the first range of cells
range1 <- read_excel(file_path, range = "I12:I27", col_names = FALSE)

# Read the second range of cells
```

```r
range2 <- read_excel(file_path, range = "I31:I44", col_names = FALSE)

# Read the third range of cells
range3 <- read_excel(file_path, range = "I48:I59", col_names = FALSE)

dissolved_oxygen <- rbind(range1, range2, range3)
d_oxygen_num <- c(as.numeric(unlist(dissolved_oxygen)))

hist(
  d_oxygen_num,
  main = "Histogram for Sample Measurements",
  xlab = "Measurements of D O2",
)

boxplot(
  x = d_oxygen_num,
  main = "Boxplot for Sample Measurements",
  ylab = expression(hat(p), "value")
)

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


sample_mean = mean(d_oxygen_num)
median = median(d_oxygen_num)
std_dev = sd(d_oxygen_num)
ks.test(d_oxygen_num, 'pnorm', sample_mean, std_dev)
t_value <-qt(p=0.05, df = 41, lower.tail=FALSE)

lower_bound = 9.041 - t_value * (std_dev / sqrt(42))
max_value = max(d_oxygen_num)
```

Appendix C: R Code for Question 3

Akal Ustat Singh, Phillip Gustov, Frank Kutsar
STATS 50-05, Spring 2025

```r
library(readxl)
file_path <- "./project2/AmericanRiver-Sp2024.xlsx"

# Read the first range of cells
tap_od_range1 <- read_excel(file_path, range = "I12:I27", col_names = FALSE)

# Read the second range of cells
tap_od_range2 <- read_excel(file_path, range = "I31:I44", col_names = FALSE)

# Read the third range of cells
tap_od_range3 <- read_excel(file_path, range = "I48:I59", col_names = FALSE)

tap_od <- rbind(tap_od_range1, tap_od_range2, tap_od_range3)
tap_od_num <- c(as.numeric(unlist(tap_od)))
```

```r
# Read the first range of cells
tap_temp_range1 <- read_excel(file_path, range = "G12:G27", col_names = FALSE)

# Read the second range of cells
tap_temp_range2 <- read_excel(file_path, range = "G31:G44", col_names = FALSE)

# Read the third range of cells
tap_temp_range3 <- read_excel(file_path, range = "G48:G59", col_names = FALSE)

tap_temp <- rbind(tap_temp_range1, tap_temp_range2, tap_temp_range3)
tap_temp_num <- c(as.numeric(unlist(tap_temp)))

# Read the first range of cells
river_od_range1 <- read_excel(file_path, range = "D12:D27", col_names = FALSE)

# Read the second range of cells
river_od_range2 <- read_excel(file_path, range = "D31:D44", col_names = FALSE)

# Read the third range of cells
river_od_range3 <- read_excel(file_path, range = "D48:D59", col_names = FALSE)

river_od <- rbind(river_od_range1, river_od_range2, river_od_range3)
river_od_num <- c(as.numeric(unlist(river_od)))


# Read the first range of cells
river_temp_range1 <- read_excel(file_path, range = "B12:B27", col_names = FALSE)

# Read the second range of cells
river_temp_range2 <- read_excel(file_path, range = "B31:B44", col_names = FALSE)

# Read the third range of cells
river_temp_range3 <- read_excel(file_path, range = "B48:B59", col_names = FALSE)

river_temp <- rbind(river_temp_range1, river_temp_range2, river_temp_range3)
river_temp_num <- c(as.numeric(unlist(river_temp)))

plot(x=tap_od_num, y=tap_temp_num, main="Scatterplot of Tap OD and Temp", xlab="Tap OD
data", ylab="Tap Temp Data")
tap_fit <- lm(tap_temp_num ~ tap_od_num)
abline(tap_fit,col='red')

plot(x=river_od_num, y=river_temp_num, main="Scatterplot of River OD and Temp",
xlab="River OD data", ylab="River Temp Data")
river_fit <- lm(river_temp_num ~ river_od_num)
abline(river_fit,col='red')

tap_cor <- cor(tap_od_num, tap_temp_num)
river_cor <-cor(river_od_num, river_temp_num)
```

```r
# code for 3c
# # Are the pressure measurements related to the temperature or OD measurements in the
Excel Öle (river
# or tap water)? Use scatterplots, correlations etc. to support your argument.

# Read the first range of cells
river_press_range1 <- read_excel(file_path, range = "C12:C27", col_names = FALSE)

# Read the second range of cells
river_press_range2 <- read_excel(file_path, range = "C31:C44", col_names = FALSE)

# Read the third range of cells
river_press_range3 <- read_excel(file_path, range = "C48:C59", col_names = FALSE)

river_press <- rbind(river_press_range1, river_press_range2, river_press_range3)
river_press_num <- c(as.numeric(unlist(river_press)))

od_to_press_fit <- lm(river_press_num ~ river_od_num)
plot(x=river_od_num, y=river_press_num, main="Scatterplot of River OD and Pressure",
xlab="River OD data", ylab="River Temp Data")
abline(lm(river_press_num ~ river_od_num),col='red')

od_to_press_fit
press_cor <- cor(river_od_num, river_press_num, use = "complete.obs")
```