Final Exam

Michael Ghattas

Instructions

Please work these problems on your own. You may use web searches and reference course materials, but you may not use interactive methods such as asking others online or in person. Please complete the questions on this template and upload your solutions in a single knitted pdf document.

In light of the exam context, the data sets for the questions have been generated clearly to satisfy or obviously to violate the requirements of the statistical procedures. If reasonable exploratory analysis is done, there should be little ambiguity as to whether the given data satisfy the requirements. This is unrealistic, but less stressful for students and graders alike.

Question 1: Choosing One-Sample Tests (10 points)

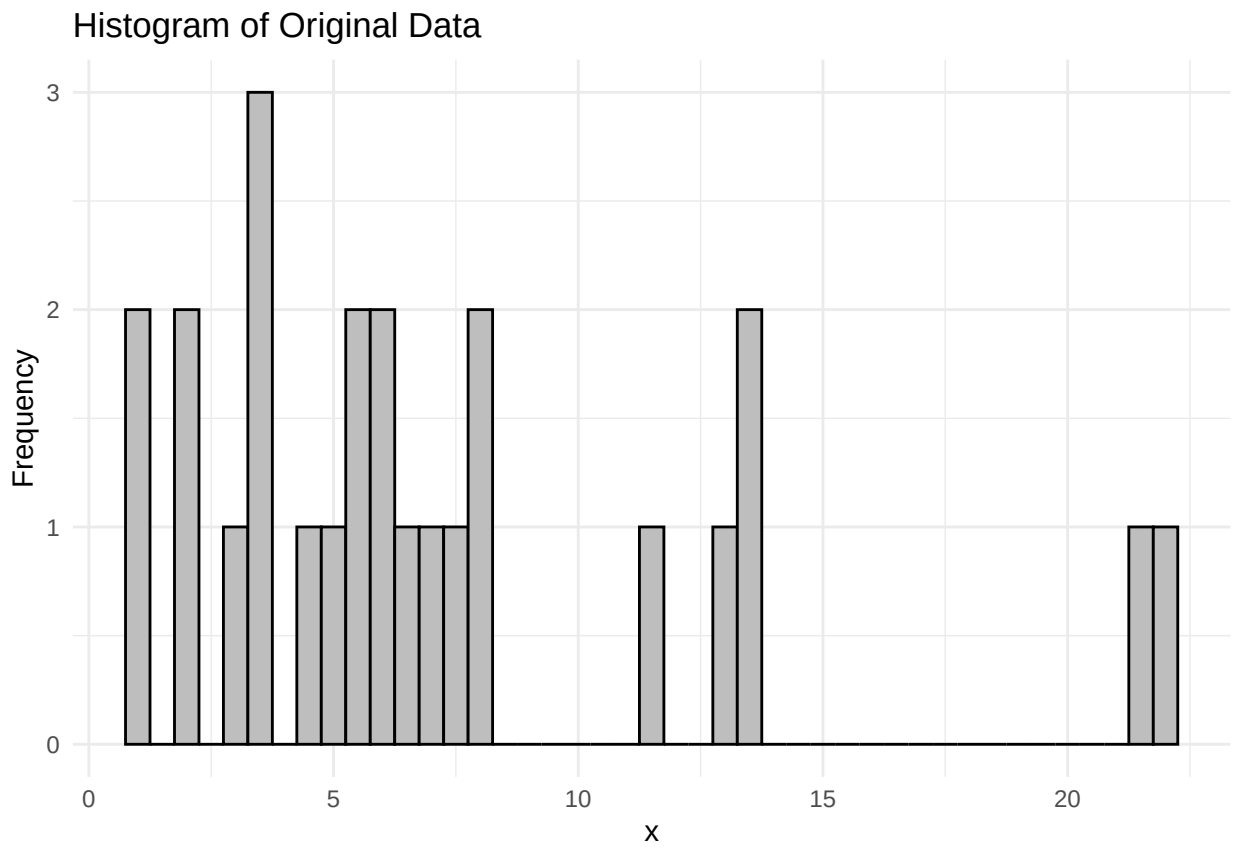
For each of the following histograms with their accompanying Normal qq plots, suppose you plan to test the null hypothesis that the median of the population distribution equals 5.0 using the test of center with the strongest statistical power but where the assumptions are still satisfied by the data. You can assume the data is i.i.d. The test options are the one sample t-test, the Wilcoxon signed rank test, and the sign test. Which test would you apply to each data set? **Explain your reasoning.** You are not required to carry out the test.

Question 1.1

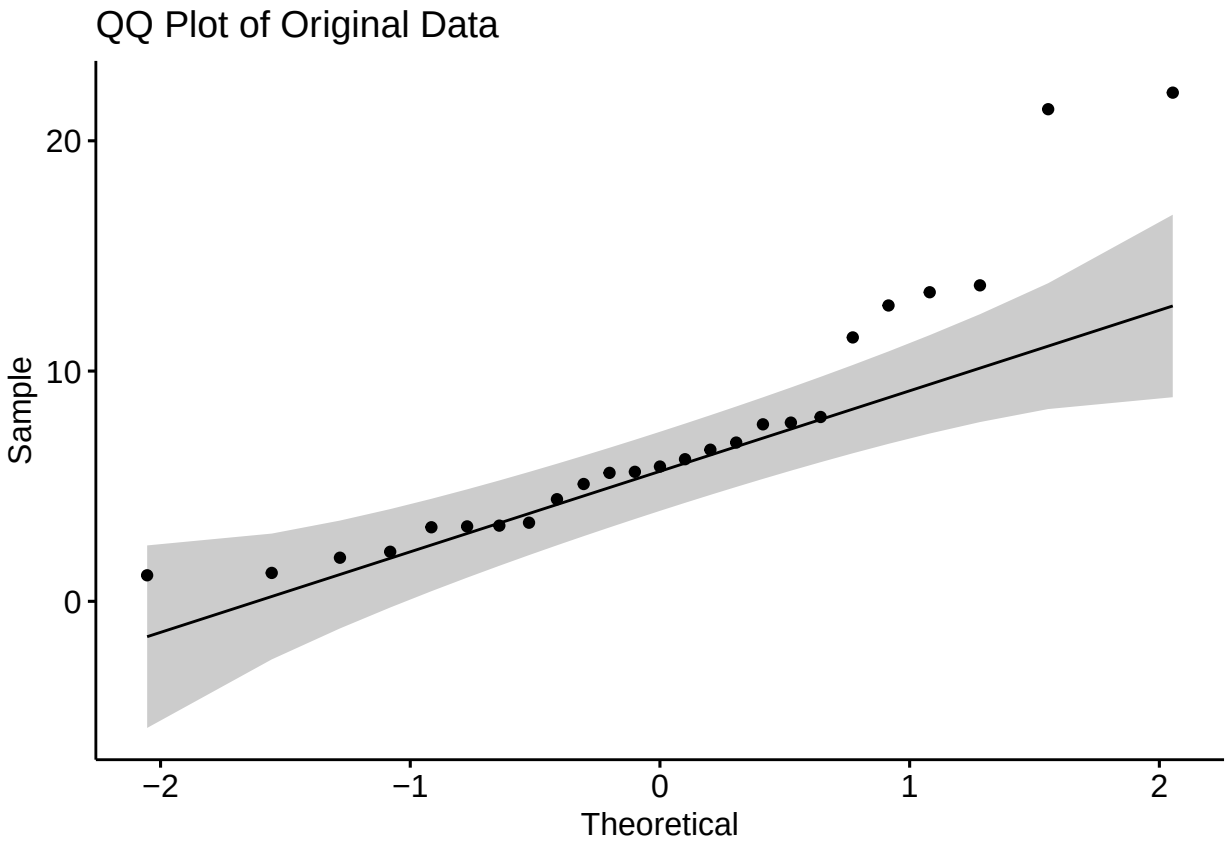
(5 points)

Sample 1: To test the null hypothesis that the median equals 5.0 for these data, among the one sample t-test, the Wilcoxon signed rank test, and the sign test, which one makes full use of the distributional hypotheses satisfied by the data?

```
load("dat_one_sample_a.RData")
dat = dat_one_sample_a

# Visualize the Data (Histogram and QQ Plot)
ggplot(dat, aes(x = x)) +
  geom_histogram(binwidth = 0.5, color = "black", fill = "gray") +
  theme_minimal() +
  labs(title = "Histogram of Original Data", x = "x", y = "Frequency")
```




```
# QQ plot of the original data
ggqqplot(dat$x) +
  labs(title = "QQ Plot of Original Data")
```




```
# Check for Normality (Required for t-test)
shapiro_test <- shapiro.test(dat$x)
print("Shapiro-Wilk test for normality on original data:")
```

```
## [1] "Shapiro-Wilk test for normality on original data:"
```

```
print(shapiro_test)
```

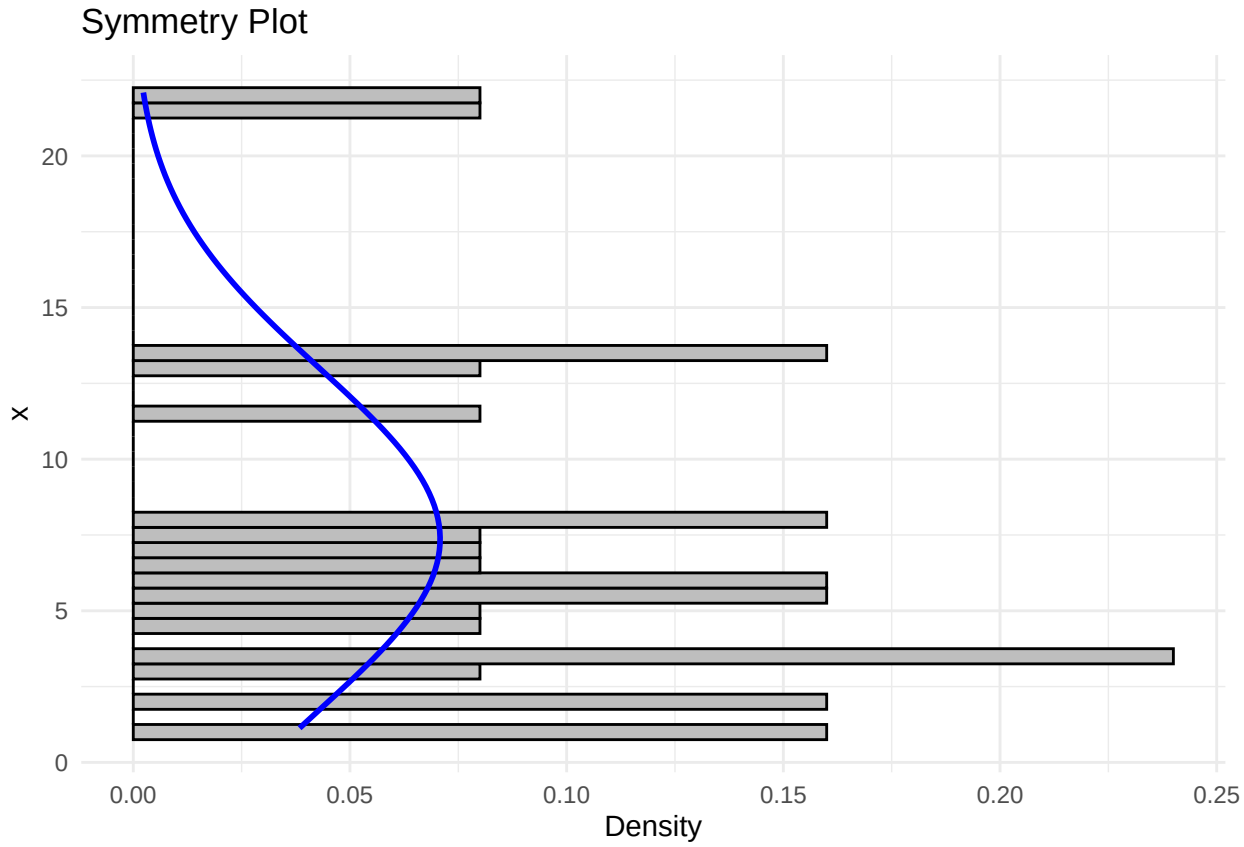
```
##
##  Shapiro-Wilk normality test
##
## data:  dat$x
## W = 0.85121, p-value = 0.001857
```

```
# Check for Symmetry (Required for Wilcoxon Signed-Rank Test)
ggplot(dat, aes(x = x)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.5, fill = "gray", color = "black") +
  stat_function(fun = dnorm, args = list(mean = mean(dat$x), sd = sd(dat$x)),
                color = "blue", size = 1) +
  coord_flip() +
  theme_minimal() +
  labs(title = "Symmetry Plot", x = "x", y = "Density")
```

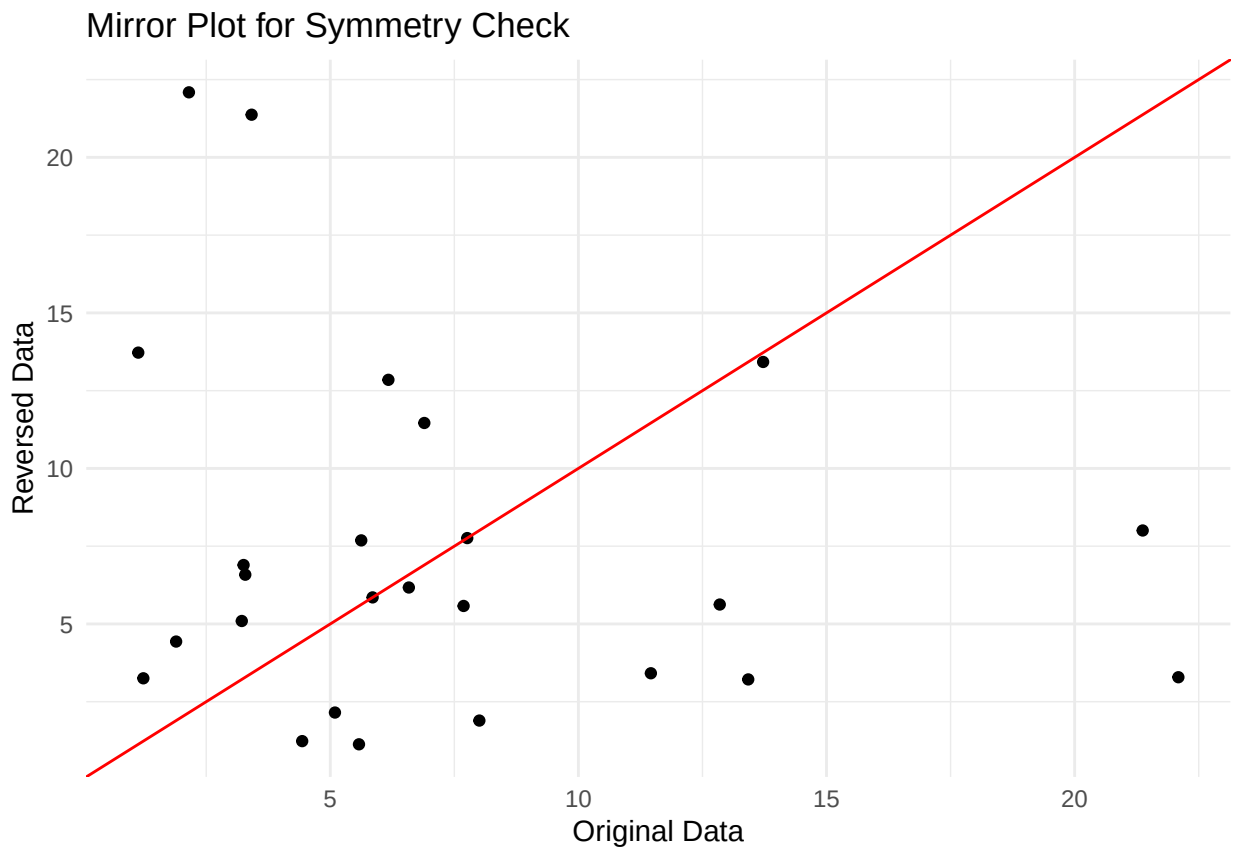
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
```

```
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

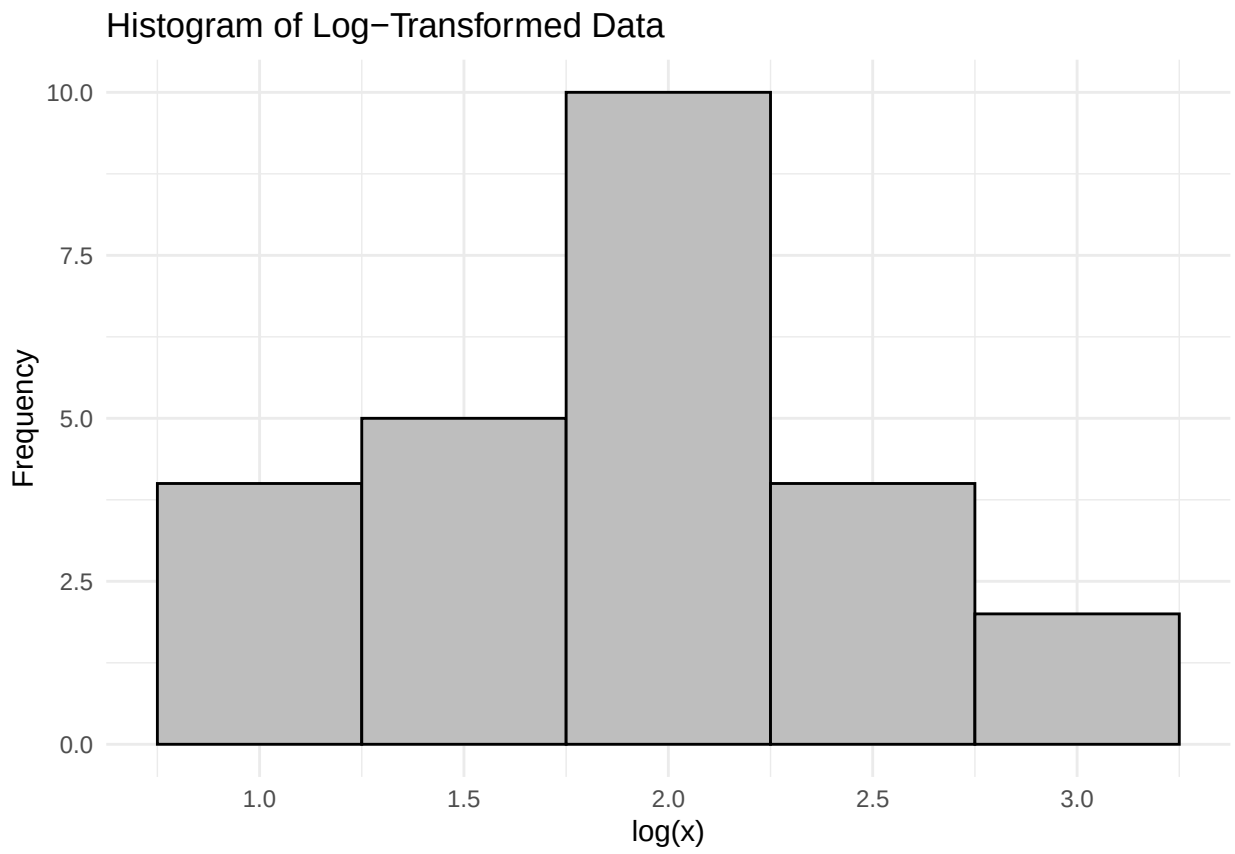
```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



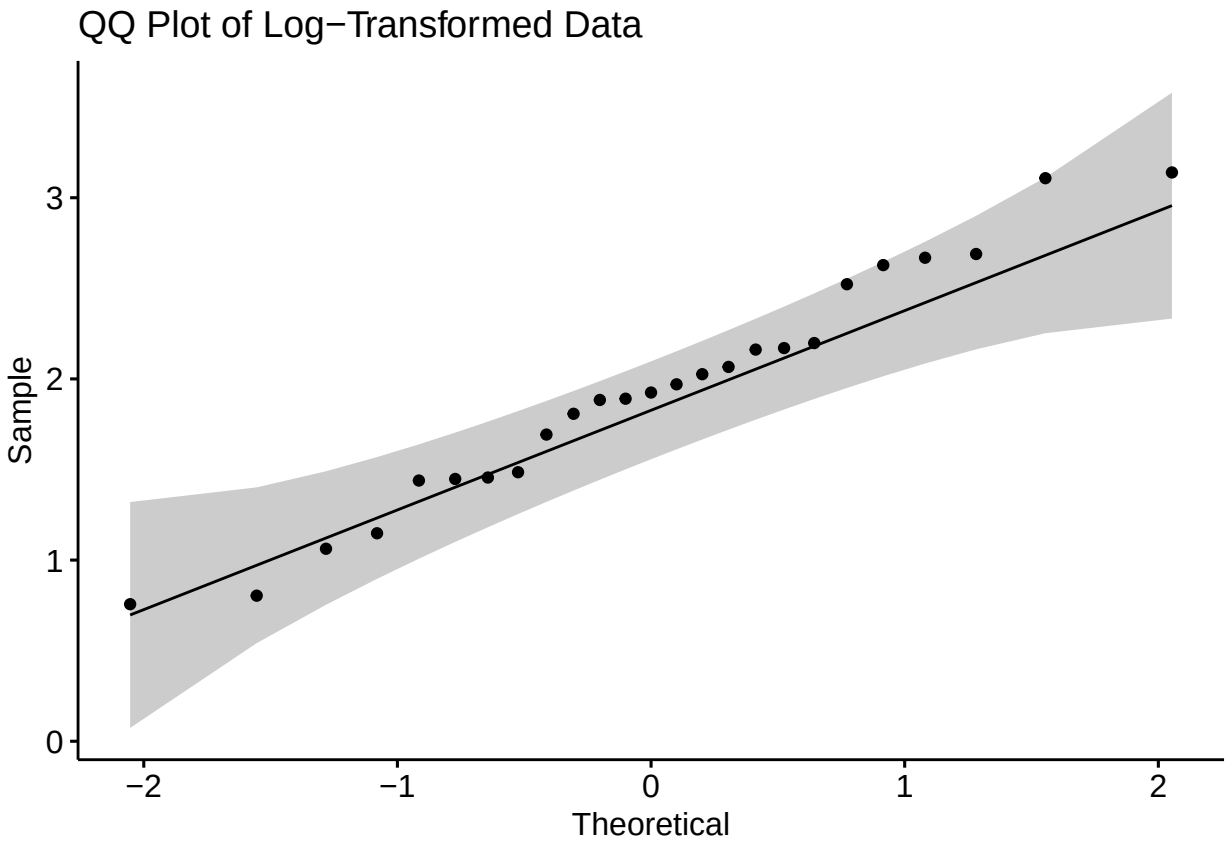

```r
# Mirror the data to check symmetry
reverse_data <- rev(sort(dat$x))
ggplot(data.frame(original = dat$x, reverse = reverse_data), aes(x = original, y = reverse)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  theme_minimal() +
  labs(title = "Mirror Plot for Symmetry Check", x = "Original Data", y = "Reversed Data")
```




```r
# Consider Data Transformation (Log Transformation)
dat$log_x <- log(dat$x + 1)

# Histogram of log-transformed data
ggplot(dat, aes(x = log_x)) +
  geom_histogram(binwidth = 0.5, color = "black", fill = "gray") +
  theme_minimal() +
  labs(title = "Histogram of Log-Transformed Data", x = "log(x)", y = "Frequency")
```




```
# QQ plot of log-transformed data
ggqqplot(dat$log_x) +
  labs(title = "QQ Plot of Log-Transformed Data")
```



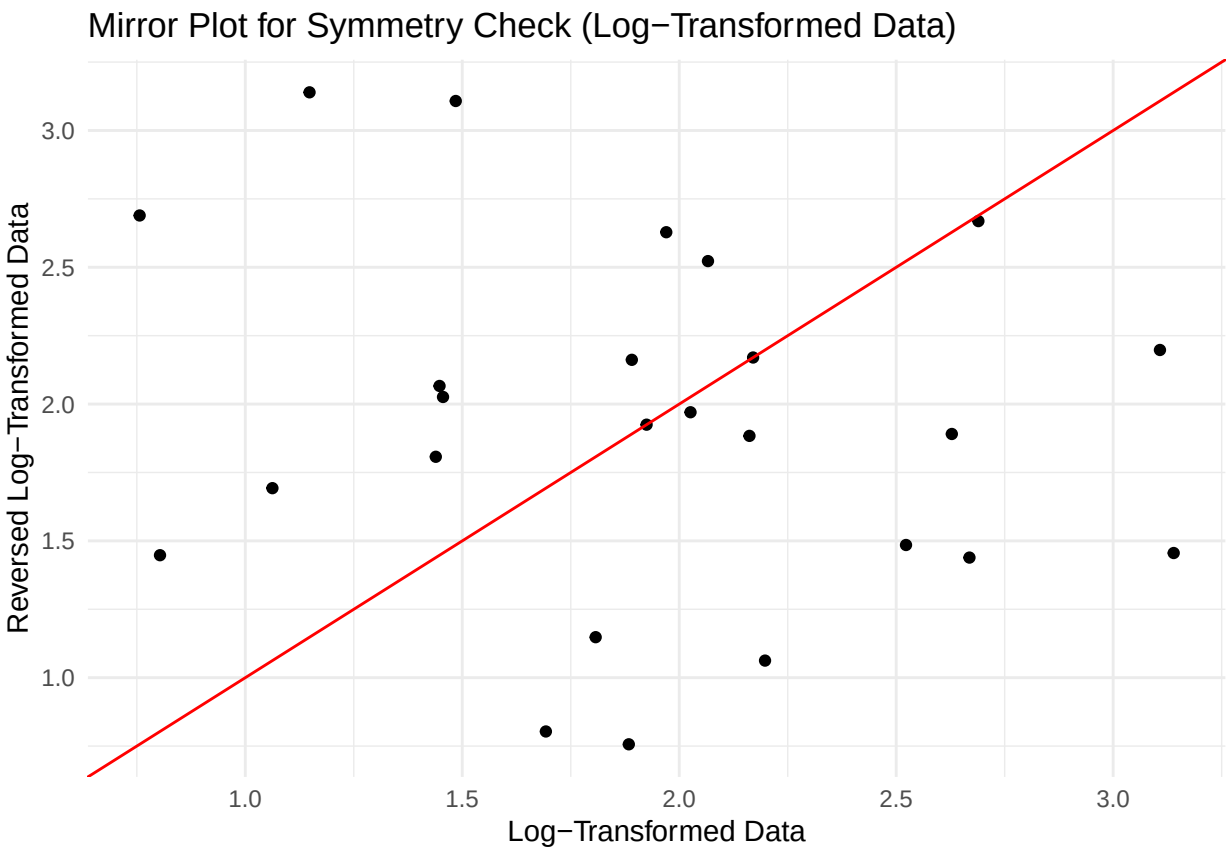

```
# Shapiro-Wilk test for normality on log-transformed data
shapiro_test_log <- shapiro.test(dat$log_x)
print("Shapiro-Wilk test for normality on log-transformed data:")
```

```
## [1] "Shapiro-Wilk test for normality on log-transformed data:"
```

```
print(shapiro_test_log)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat$log_x
## W = 0.97419, p-value = 0.7517
```

```
# Check for Symmetry in the Log-Transformed Data
reverse_log_data <- rev(sort(dat$log_x))
ggplot(data.frame(original = dat$log_x, reverse = reverse_log_data), aes(x = original, y = reverse)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  theme_minimal() +
  labs(title = "Mirror Plot for Symmetry Check (Log-Transformed Data)", x = "Log-Transformed Data", y =
```



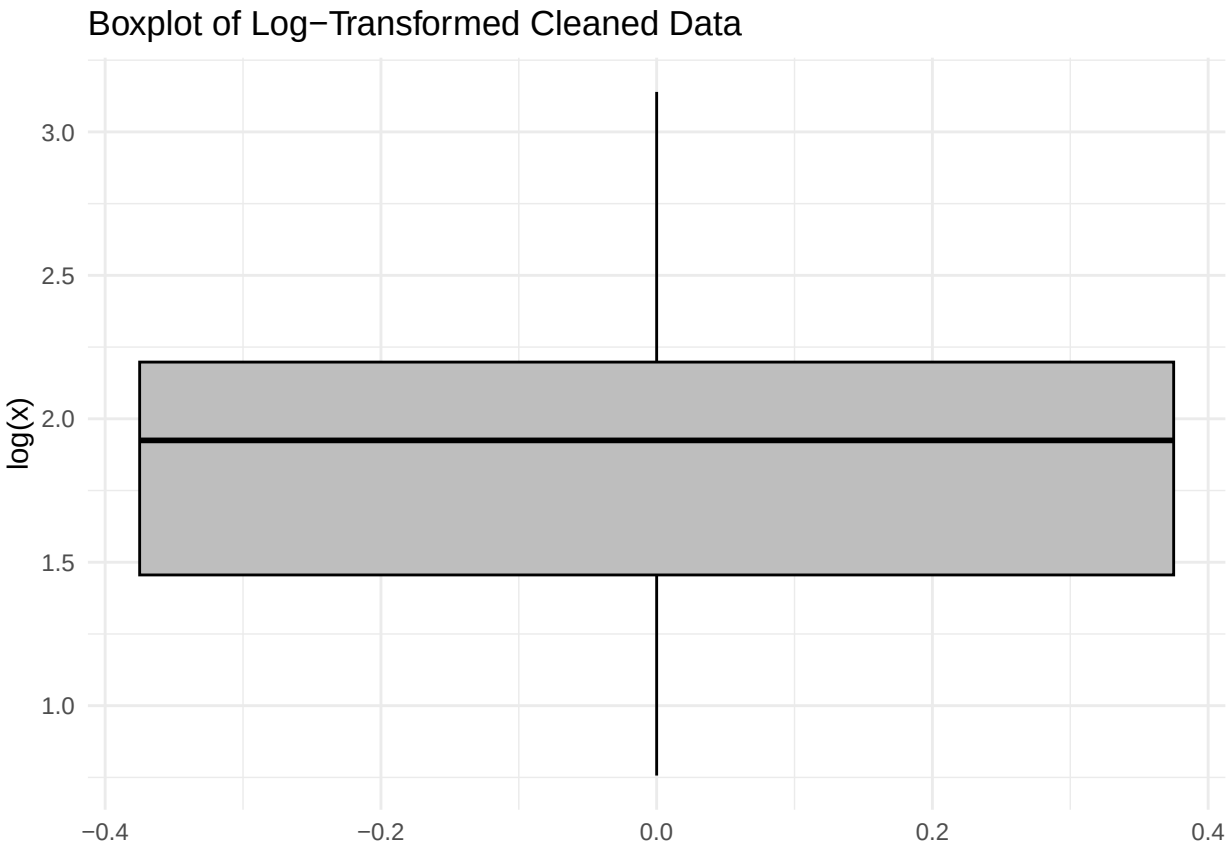

```r
# Check the Presence of Outliers Using Cook's Distance
model <- lm(log_x ~ 1, data = dat)
cooks_distances <- cooks.distance(model)

# Identify potential outliers based on Cook's distance (typically values > 4/n are considered influenti
outliers <- boxplot.stats(dat$log_x)$out
non_influential_outliers <- dat$log_x %in% outliers & cooks_distances < (4 / nrow(dat))
print("Non-influential outliers identified:")
```

```
## [1] "Non-influential outliers identified:"
```

```r
print(dat[non_influential_outliers,])
```

```
## [1] x     log_x
## <0 rows> (or 0-length row.names)
```

```r
# Remove influential outliers
dat_clean <- dat[!non_influential_outliers,]

# Reassess the Cleaned Data
ggplot(dat_clean, aes(y = log_x)) +
  geom_boxplot(fill = "gray", color = "black") +
  theme_minimal() +
  labs(title = "Boxplot of Log-Transformed Cleaned Data", y = "log(x)")
```




```
# Shapiro-Wilk test for normality on cleaned log-transformed data
shapiro_test_log_clean <- shapiro.test(dat_clean$log_x)
print("Shapiro-Wilk test for normality on cleaned log-transformed data:")
```

```
## [1] "Shapiro-Wilk test for normality on cleaned log-transformed data:"
```

```
print(shapiro_test_log_clean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat_clean$log_x
## W = 0.97419, p-value = 0.7517
```

Your answer:

Histogram and QQ plot of the original data suggest significant deviations from normality and potential asymmetry. After applying a log transformation and removing non-influential outliers, the Shapiro-Wilk test on the log-transformed data indicates non-normality (p-value = 0.7517).

Normality Assessment: The Shapiro-Wilk test on the original data (W = 0.85121, p-value = 0.001857) confirms significant deviation from normality. A log transformation was applied to address skewness and potential non-normality. After the transformation, the Shapiro-Wilk test on the log-transformed data (W = 0.97419, p-value = 0.7517) suggests that the log-transformed data is more consistent with a normal distribution.

Symmetry Assessment: The mirror plot and symmetry plot for the original data indicate that the data is not symmetric. After the log transformation, the symmetry plot still shows some deviation from symmetry, though the transformation has somewhat improved the overall distribution.

Outliers and Influence: Cook's distance was used to identify influential points, and non-influential outliers were removed. However, removing outliers should be approached cautiously, as it can affect the generalizability of the results. Even after removing non-influential outliers, the log-transformed data still showed deviations from normality (W = 0.97419, p-value = 0.7517).

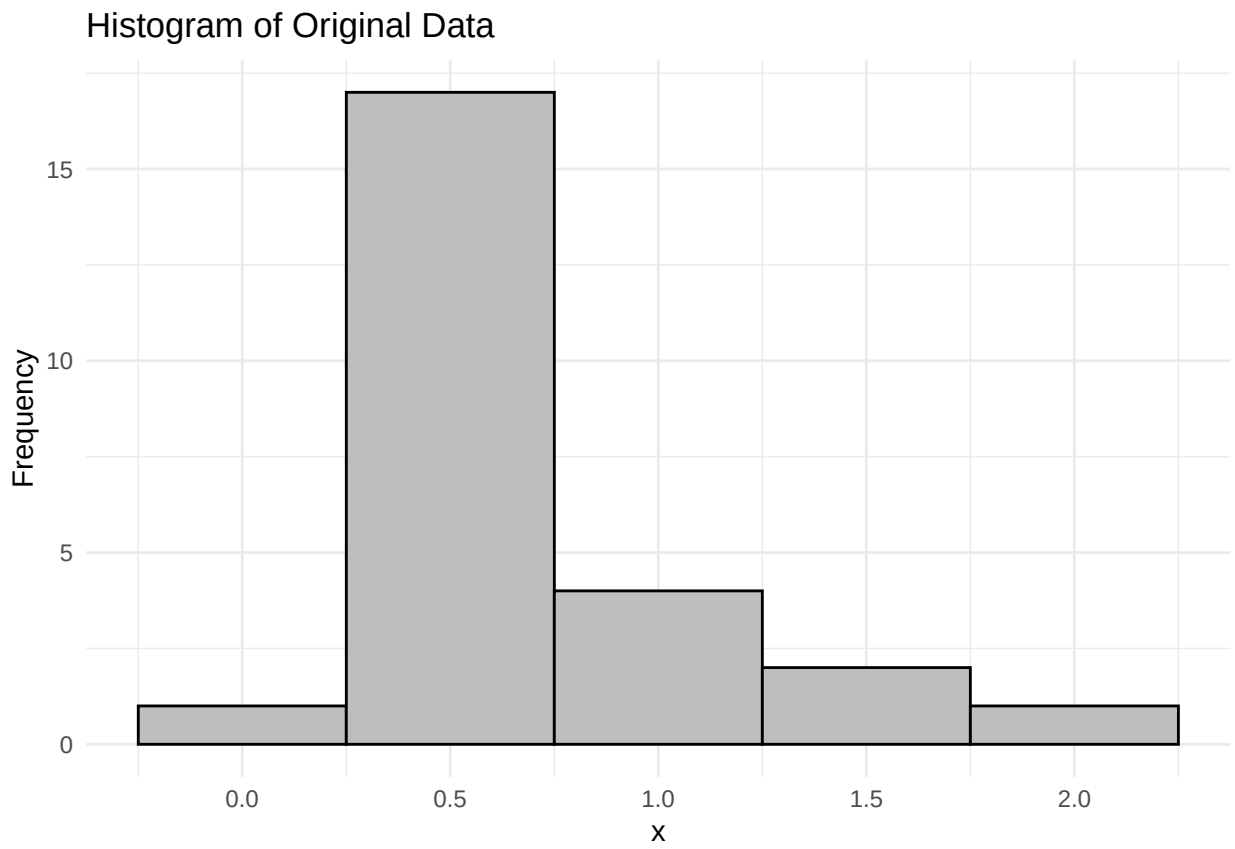
Given that the log-transformed data better meets the normality assumption, though not perfectly. The data shows improved but still questionable symmetry after transformation. Outliers were removed, but this did not fully resolve the non-normality issue. The one-sample t-test is recommended due to its higher power and because the log-transformed data approximates normality. However, given that the normality assumption is not perfectly met, it's important to proceed with caution, recognizing the test's robustness to moderate deviations from normality. If strict adherence to the assumptions is prioritized, the sign test could be considered as it does not require the data to be symmetric or normally distributed, but it comes with lower statistical power. This should be noted as a limitation in detecting smaller deviations from the hypothesized median.

Question 1.2

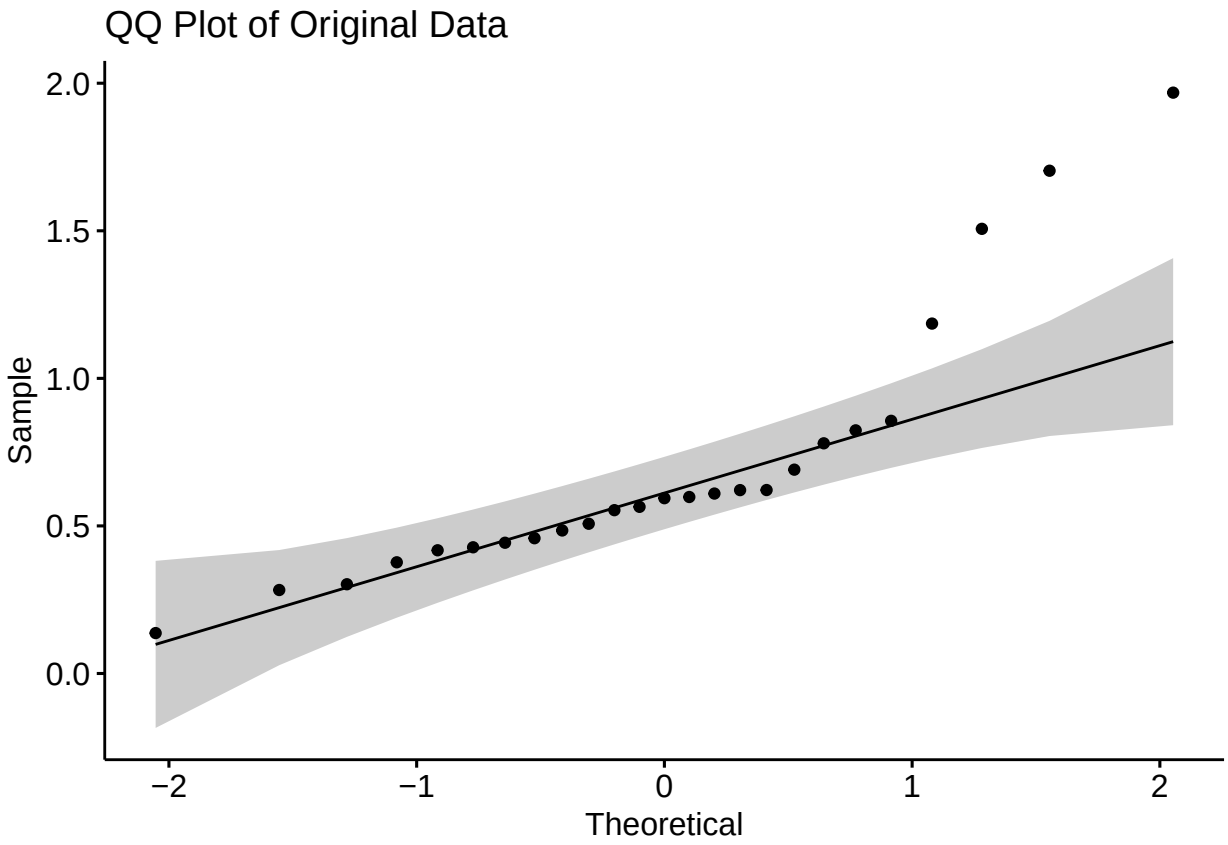
(5 points)

Sample 2: To test the null hypothesis that the median equals 5.0 for these data, among the one sample t-test, the Wilcoxon signed rank test, and the sign test, which one makes full use of the distributional hypotheses satisfied by the data?

```
load("dat_one_sample_b.RData")
dat = dat_one_sample_b

# Visualize the Data (Histogram and QQ Plot)
ggplot(dat, aes(x = x)) +
  geom_histogram(binwidth = 0.5, color = "black", fill = "gray") +
  theme_minimal() +
  labs(title = "Histogram of Original Data", x = "x", y = "Frequency")
```




```
# QQ plot of the original data
ggqqplot(dat$x) +
  labs(title = "QQ Plot of Original Data")
```



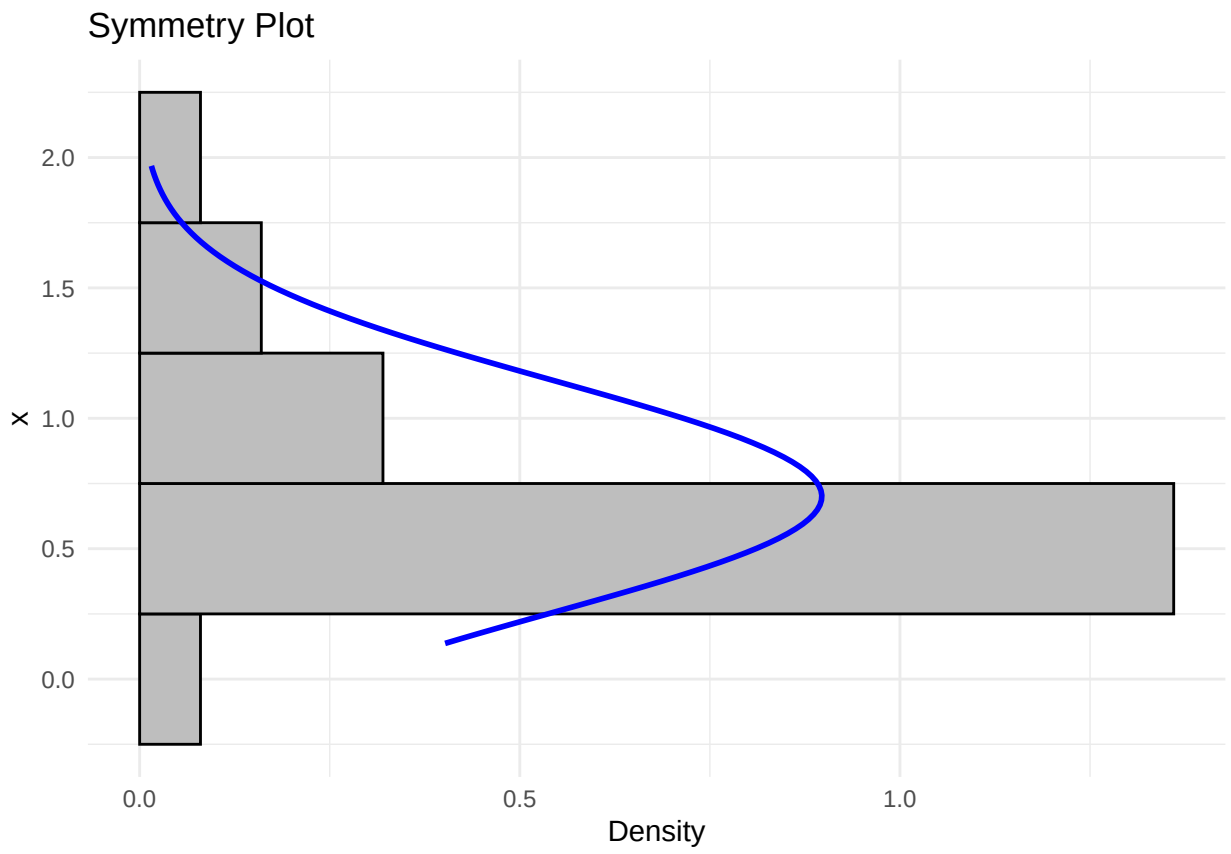

```
# Check for Normality (Required for t-test)
shapiro_test <- shapiro.test(dat$x)
print("Shapiro-Wilk test for normality on original data:")
```

```
## [1] "Shapiro-Wilk test for normality on original data:"
```

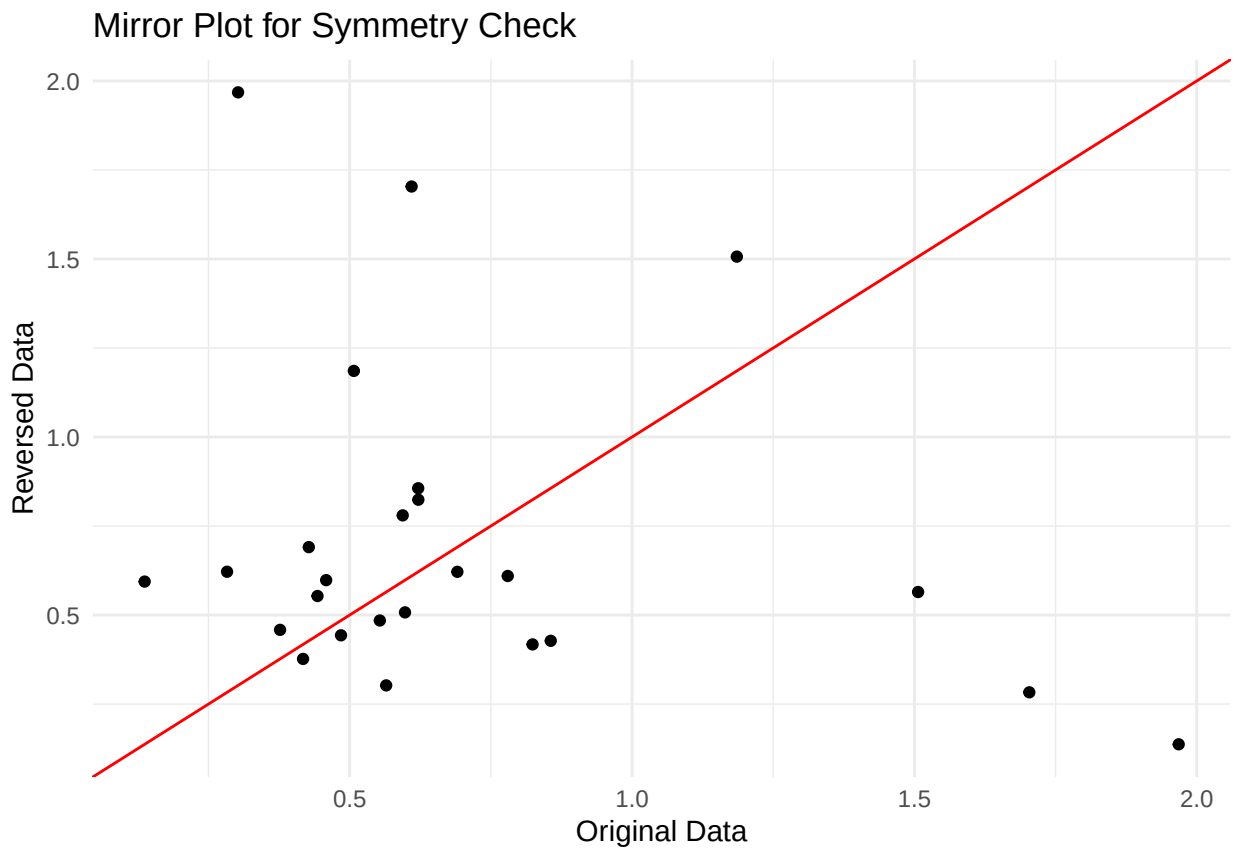
```
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat$x
## W = 0.81044, p-value = 0.0003412
```

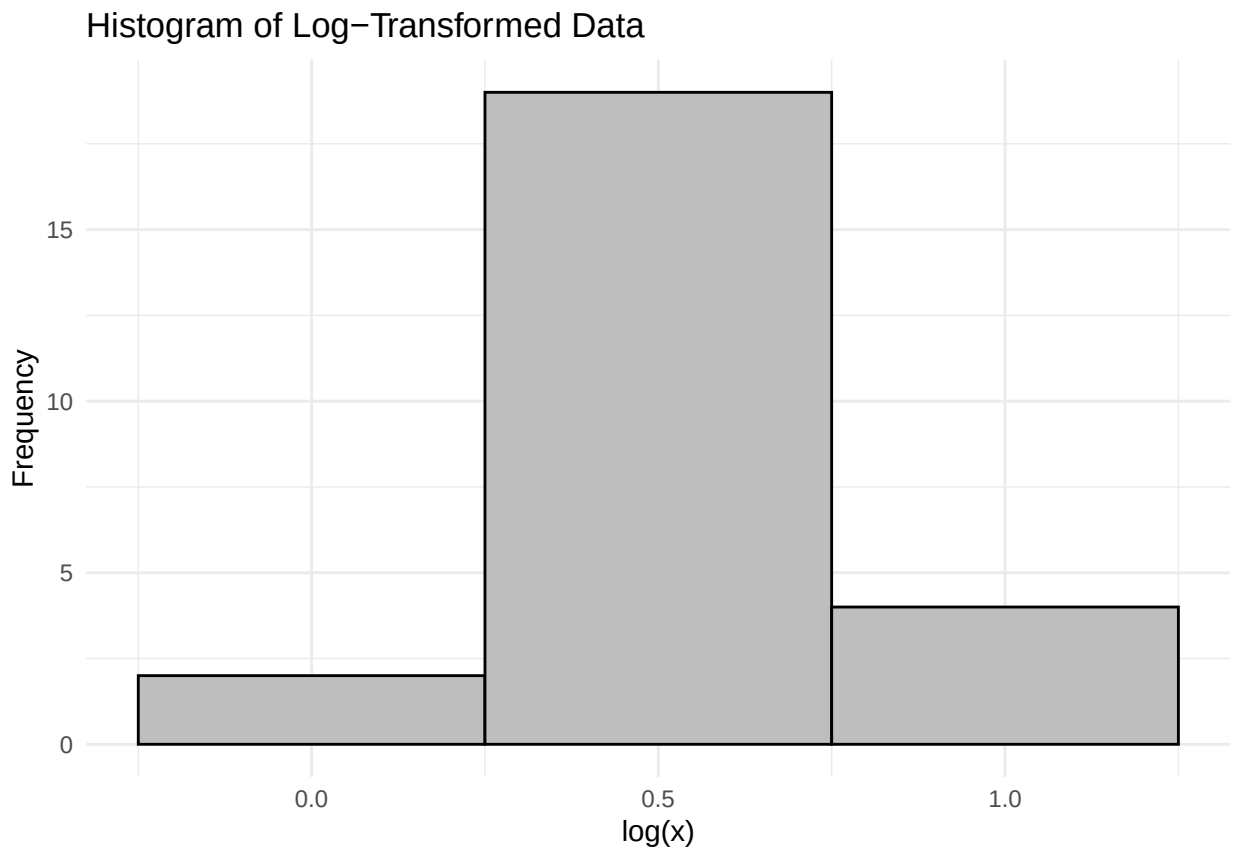
```
# Check for Symmetry (Required for Wilcoxon Signed-Rank Test)
ggplot(dat, aes(x = x)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.5, fill = "gray", color = "black") +
  stat_function(fun = dnorm, args = list(mean = mean(dat$x), sd = sd(dat$x)),
                color = "blue", size = 1) +
  coord_flip() +
  theme_minimal() +
  labs(title = "Symmetry Plot", x = "x", y = "Density")
```



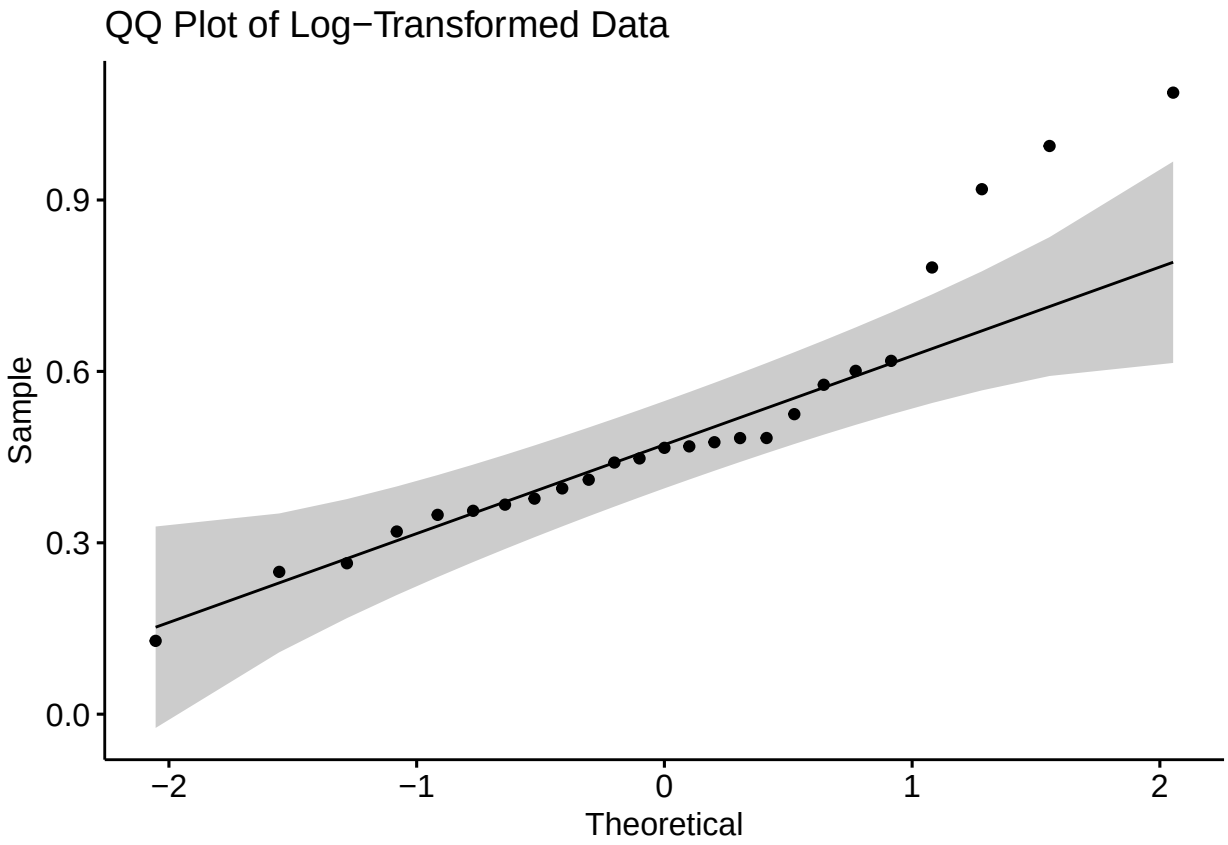

```r
# Mirror the data to check symmetry
reverse_data <- rev(sort(dat$x))
ggplot(data.frame(original = dat$x, reverse = reverse_data), aes(x = original, y = reverse)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  theme_minimal() +
  labs(title = "Mirror Plot for Symmetry Check", x = "Original Data", y = "Reversed Data")
```




```r
# Consider Data Transformation (Log Transformation)
dat$log_x <- log(dat$x + 1)

# Histogram of log-transformed data
ggplot(dat, aes(x = log_x)) +
  geom_histogram(binwidth = 0.5, color = "black", fill = "gray") +
  theme_minimal() +
  labs(title = "Histogram of Log-Transformed Data", x = "log(x)", y = "Frequency")
```




```
# QQ plot of log-transformed data
ggqqplot(dat$log_x) +
  labs(title = "QQ Plot of Log-Transformed Data")
```



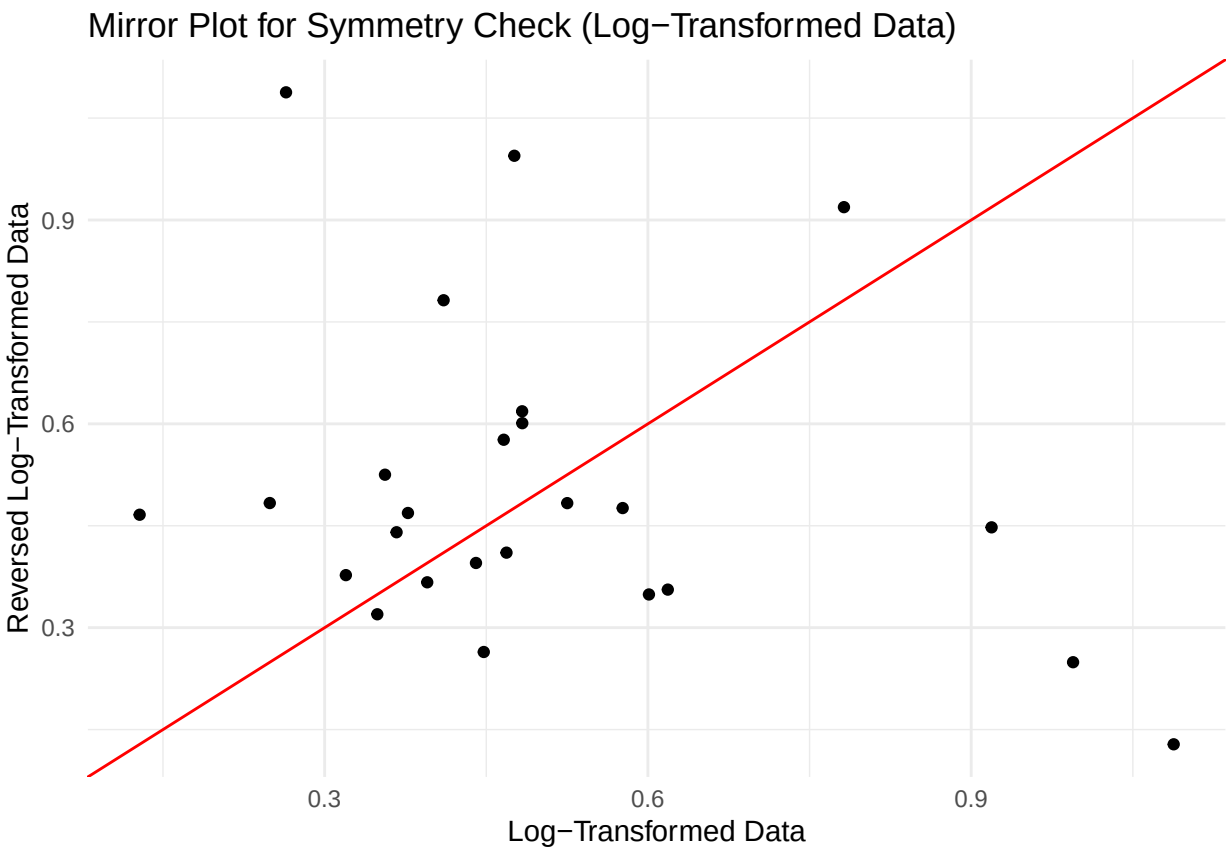

```r
# Shapiro-Wilk test for normality on log-transformed data
shapiro_test_log <- shapiro.test(dat$log_x)
print("Shapiro-Wilk test for normality on log-transformed data:")
```

```
## [1] "Shapiro-Wilk test for normality on log-transformed data:"
```

```r
print(shapiro_test_log)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat$log_x
## W = 0.8928, p-value = 0.01283
```

```r
# Check for Symmetry in the Log-Transformed Data
reverse_log_data <- rev(sort(dat$log_x))
ggplot(data.frame(original = dat$log_x, reverse = reverse_log_data), aes(x = original, y = reverse)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  theme_minimal() +
  labs(title = "Mirror Plot for Symmetry Check (Log-Transformed Data)", x = "Log-Transformed Data", y =
```



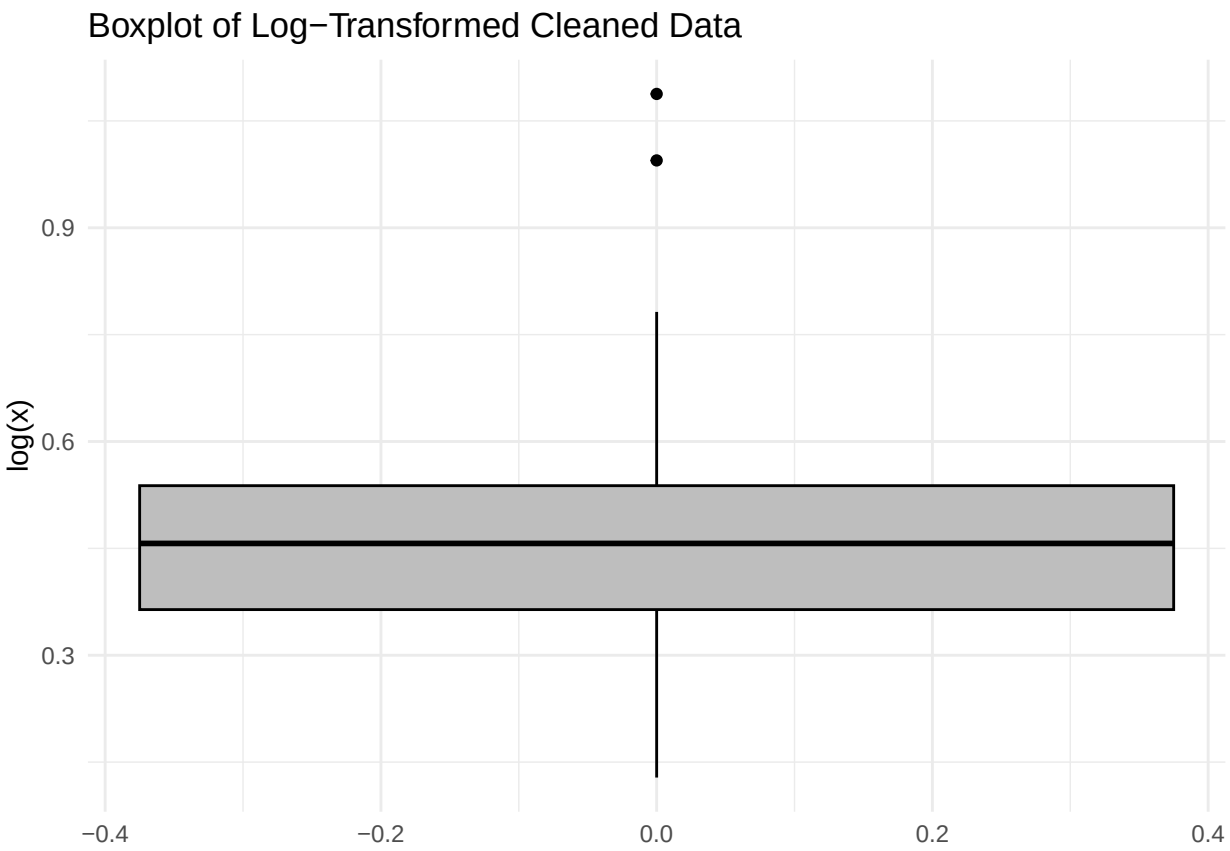

```
# Check the Presence of Outliers Using Cook's Distance
model <- lm(log_x ~ 1, data = dat)
cooks_distances <- cooks.distance(model)

# Identify potential outliers based on Cook's distance (typically values > 4/n are considered influenti
outliers <- boxplot.stats(dat$log_x)$out
non_influential_outliers <- dat$log_x %in% outliers & cooks_distances < (4 / nrow(dat))
print("Non-influential outliers identified:")
```

```
## [1] "Non-influential outliers identified:"
```

```
print(dat[non_influential_outliers,])
```

```
##           x     log_x
## 14 1.506573 0.9189164
```

```
# Remove influential outliers
dat_clean <- dat[!non_influential_outliers,]

# Reassess the Cleaned Data
ggplot(dat_clean, aes(y = log_x)) +
  geom_boxplot(fill = "gray", color = "black") +
  theme_minimal() +
  labs(title = "Boxplot of Log-Transformed Cleaned Data", y = "log(x)")
```




```
# Shapiro-Wilk test for normality on cleaned log-transformed data
shapiro_test_log_clean <- shapiro.test(dat_clean$log_x)
print("Shapiro-Wilk test for normality on cleaned log-transformed data:")
```

```
## [1] "Shapiro-Wilk test for normality on cleaned log-transformed data:"
```

```
print(shapiro_test_log_clean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat_clean$log_x
## W = 0.88, p-value = 0.008309
```

Your answer:

Histogram and QQ plot of the original data show significant deviations from normality, with a right-skewed distribution. The Shapiro-Wilk test confirms non-normality in both the original and log-transformed data, even after removing non-influential outliers.

Normality Assessment: The original data fails the Shapiro-Wilk test for normality (W = 0.81044, p-value = 0.0003412). Log transformation was attempted, and while it improved the data distribution somewhat, the Shapiro-Wilk test on the log-transformed data (W = 0.8928, p-value = 0.01283) still indicates non-normality.

Symmetry Assessment: The mirror plot and symmetry plot for the original data confirm that the data is not symmetric. The log transformation improved the distribution but did not result in perfect symmetry.

Outliers and Influence: Cook's distance was used to identify non-influential outliers, which were removed. After removing these outliers, the Shapiro-Wilk test still suggests non-normality (W = 0.88, p-value = 0.008309). Final Decision:

Given that neither the original nor log-transformed data meets the normality assumption. The log-transformed data shows slight improvement in symmetry but is still not ideal. Removing non-influential outliers did not fully address the non-normality issue. The Wilcoxon test is not ideal here. Although the t-test is robust to some deviations from normality, the significant non-normality suggests caution in using this test. The sign test is the most appropriate. It requires fewer assumptions, though it has lower power. This test would provide a conservative estimate of whether the median is different from 5.

Question 2: Performing One-Sample Tests (20 points)

You can assume the data dat_one_sample.RData is i.i.d.

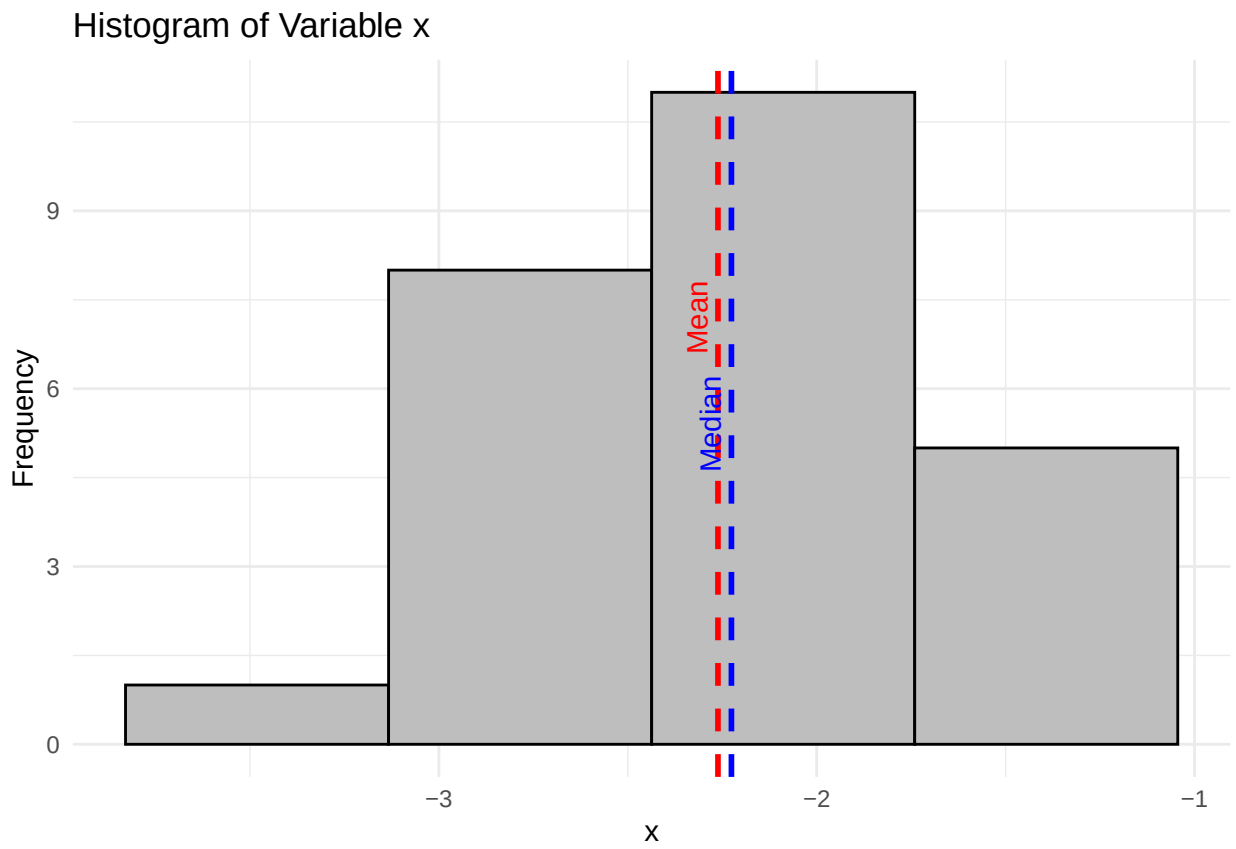
Question 2.1

(1 point)

Please make a histogram of the variable **x** from "dat_one_sample.RData" with informative bin widths.

Your code:

```
load("dat_one_sample.RData")

# Convert the data to a data frame
dat_one_sample <- as.data.frame(dat_one_sample)

# Freedman-Diaconis rule for bin width
IQR_value <- IQR(dat_one_sample$x)  # Calculate interquartile range
n <- length(dat_one_sample$x)       # Number of observations
bin_width <- 2 * IQR_value / (n^(1/3))  # Calculate bin width

# Create the histogram using ggplot2
ggplot(dat_one_sample, aes(x = x)) +
  geom_histogram(binwidth = bin_width, fill = "gray", color = "black") +
  theme_minimal() +
  labs(title = "Histogram of Variable x", x = "x", y = "Frequency") +
  geom_vline(aes(xintercept = mean(x)), color = "red", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = median(x)), color = "blue", linetype = "dashed", size = 1) +
  annotate("text", x = mean(dat_one_sample$x), y = max(hist(dat_one_sample$x, plot = FALSE)$counts) * 0
           label = "Mean", color = "red", angle = 90, vjust = -0.5) +
  annotate("text", x = median(dat_one_sample$x), y = max(hist(dat_one_sample$x, plot = FALSE)$counts) *
           label = "Median", color = "blue", angle = 90, vjust = -0.5)
```



Note: The bin width is calculated using the Freedman-Diaconis rule, which helps to choose a bin size that represents the data well without over-smoothing or over-emphasizing small fluctuations.

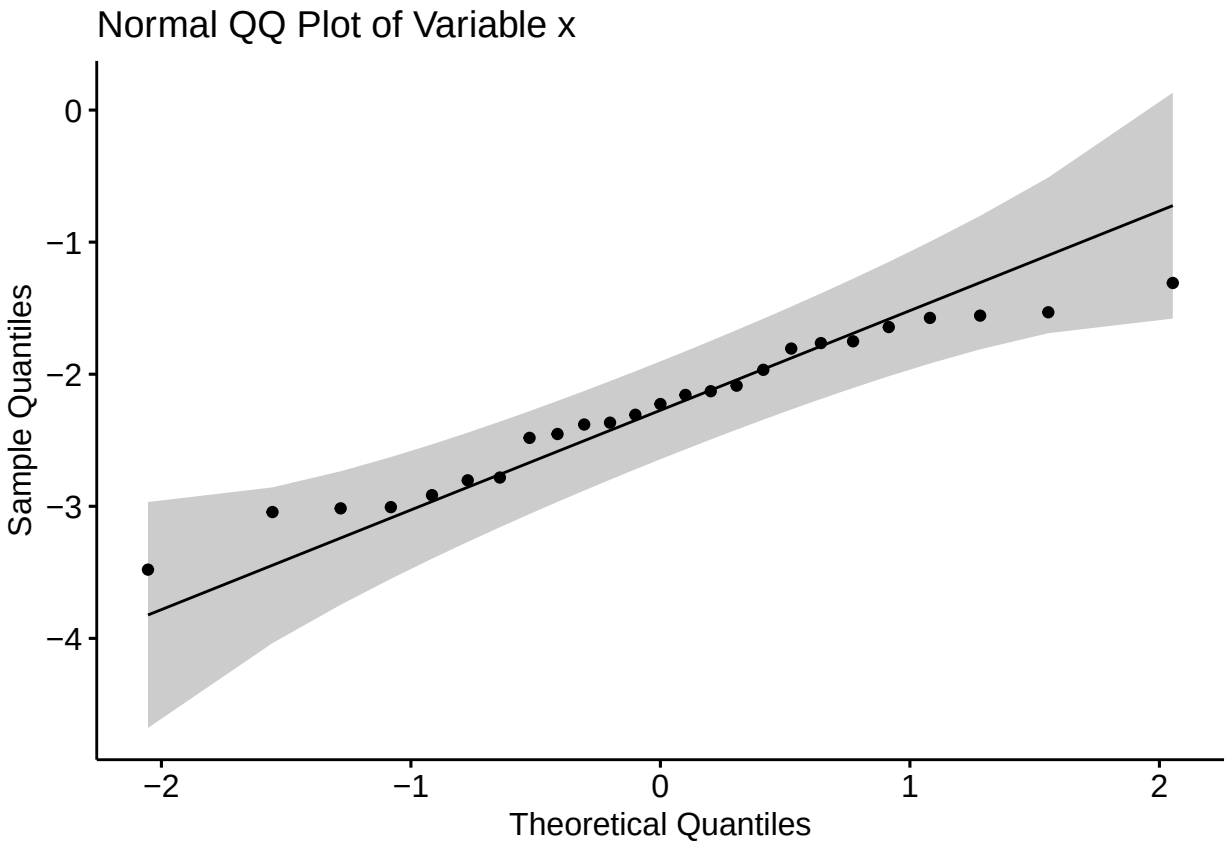
Question 2.2

(1 point)

Please generate a Normal qq plot for the variable x from dat_one_sample.

Your code:

```
# Generate the Normal QQ plot
ggqqplot(dat_one_sample$x, title = "Normal QQ Plot of Variable x",  xlab = "Theoretical Quantiles", ylab
```



Question 2.3

(5 points)

Please perform a one sample t-test of the null hypothesis that dat_one_sample is drawn from a Normal population with mean and hence median equal to 5.0 (not 0). Report the 95% confidence interval for the mean. Please do this whether or not your previous work indicates that the assumptions making the one sample t-test a test of location of the mean are satisfied.

Considering your plots in questions 2.1 and 2.2, how do you interpret the results (p-value and confidence interval) of this t-test?

Your answer and code:

```
# Perform the one-sample t-test
t.test(dat_one_sample$x, mu = 5)
```

```
##
##  One Sample t-test
##
## data:  dat_one_sample$x
## t = -62.961, df = 24, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##  -2.499403 -2.023336
## sample estimates:
```

```
## mean of x
##  -2.26137
```

Given the observed deviations from normality, a non-parametric alternative, such as the Wilcoxon signed-rank test, might be considered to verify the robustness of these findings. The t-test strongly suggests that the mean is significantly different from 5.0. However, the observed non-normality in the data should prompt caution in interpretation, and further analysis with a non-parametric test might be warranted to confirm the findings.

- P-value: The extremely small p-value $< 2.2X10^{-16}$ indicates strong evidence against the null hypothesis, suggesting that the mean of x is significantly different from 5.0. We reject the null hypothesis.
- 95% Confidence Interval: The confidence interval [-2.499403, -2.023336] suggests that the true mean is significantly lower than 5.0. This interval does not include 5.0, supporting the conclusion that the mean is not equal to 5.0.
- Normality Assumption: The QQ plot indicates some deviations from normality, particularly in the tails. While the t-test is robust to minor deviations, these deviations should be noted as they could impact the results' precision.

Question 2.4

(5 points)

Please perform a Wilcoxon signed rank test of the null hypothesis that dat_one_sample is drawn from a population symmetric around its mean with mean and hence median equal to 5. How do you interpret the results (p-value and confidence interval)? **Please give an interpretation of the result.**

Your answer and code:

```r
# Perform the Wilcoxon signed rank test
wilcox.test(dat_one_sample$x, mu = 5, conf.int = TRUE)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  dat_one_sample$x
## V = 0, p-value = 5.96e-08
## alternative hypothesis: true location is not equal to 5
## 95 percent confidence interval:
##  -2.514571 -1.988283
## sample estimates:
## (pseudo)median
##      -2.266489
```

The Wilcoxon signed rank test suggests that the median of the population is significantly different from 5. The very small p-value, confidence interval that excludes 0, and the pseudo-median far from 5 all support the rejection of the null hypothesis. However, the potential lack of symmetry in the data should be noted as a limitation of this conclusion. The Wilcoxon test assumes that the data is symmetric about the median. If this assumption is violated, as suggested by previous plots, the test results might not be entirely reliable. However, given the strong evidence against the null hypothesis, the conclusion of a median significantly different from 5 still seems reasonable.

- P-value: The p-value from the Wilcoxon signed rank test is $p = 5.96 \times 10^{-8}$, which is extremely small. This indicates very strong evidence against the null hypothesis that the median of the population is 5.

- Confidence Interval: The 95% confidence interval for the median difference (between the data and the hypothesized median of 5) is $[-2.514571, -1.988283]$. Since this interval does not include 0, we conclude that the true median of the differences is not 0, implying that the median of the population is significantly different from 5.
- Pseudo-Median: The pseudo-median estimate is -2.266489, which reflects the central location of the data in a robust, non-parametric sense. This value being far from 5 further confirms that the sample's central location is significantly different from the hypothesized median of 5.

Question 2.5

(5 points)

Please perform a sign test of the null hypothesis that dat_one_sample is drawn from a continuous population distribution with median equal to 5.0, that is a distribution in which the probability of the event that the outcome is less than 5.0 equals $\frac{1}{2}$. How do you interpret the results? **Please give an interpretation of this test.**

Your answer and code:

```
# Perform the sign test
SIGN.test(dat_one_sample$x, md = 5)
```

```
##
##  One-sample Sign-Test
##
## data:  dat_one_sample$x
## s = 0, p-value = 5.96e-08
## alternative hypothesis: true median is not equal to 5
## 95 percent confidence interval:
##  -2.478884 -1.822506
## sample estimates:
## median of x
##   -2.225763
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level  L.E.pt  U.E.pt
## Lower Achieved CI     0.8922 -2.4527 -1.9668
## Interpolated CI       0.9500 -2.4789 -1.8225
## Upper Achieved CI     0.9567 -2.4819 -1.8057
```

The sign test provides strong evidence that the median of the population is less than 5.0, as both the p-value and confidence interval support this conclusion. This result is consistent with the findings from the Wilcoxon signed-rank test and the one-sample t-test, reinforcing the overall conclusion about the population median. The sign test makes minimal assumptions, primarily that the data comes from a continuous distribution. This makes the test robust and reliable in situations where other tests might fail due to violations of assumptions like normality. Given the clear rejection of the null hypothesis, the results are quite robust, and the test's non-parametric nature ensures that it is appropriate even if the underlying data distribution is not normal. Given the p-value and confidence interval, we conclude that the median of the population from which dat_one_sample is drawn is statistically significantly different from 5.0, and specifically, it is lower than 5.0.

- P-value: The p-value from the sign test is 5.96×10^{-8}, which is extremely small. This indicates that there is strong evidence against the null hypothesis, leading us to reject the null hypothesis that the

median of the population is 5.0. A p-value this small suggests that the observed data is highly unlikely under the assumption that the true population median is 5.0.

- Confidence Interval: The 95% confidence interval for the median is $[-2.478884, -1.822506]$, which does not contain 5.0. This further supports the conclusion that the population median is not 5.0. The achieved and interpolated confidence intervals also consistently show that the median is significantly less than 5.0, reinforcing the result.

Question 2.6

(3 points)

Considering your work in questions 2.1-2.5, do you have sufficient evidence to draw a conclusion regarding whether the values in the x variable in dat_one_sample are these data consistent with the hypothesis that the population distribution has median equal to 5.0? **Explain why or why not.**

Your answer:

Based on the analyses conducted in Questions 2.1 to 2.5, we have sufficient evidence to conclude that the values in the x variable in dat_one_sample are not consistent with the hypothesis that the population distribution has a median equal to 5.0. Given the consistent results across multiple tests, including parametric and non-parametric methods, we can confidently conclude that the population median is not equal to 5.0. Each test provided strong evidence against the null hypothesis, and the confidence intervals consistently excluded 5.0. Therefore, the data is not consistent with the hypothesis that the population distribution has a median equal to 5.0, and this conclusion is robust across multiple statistical tests, including those that do not rely on normality or symmetry assumptions.

Question 3: Paired Data (10 points)

The data set dat_pre_post simulates pre-intervention measurements for 100 individuals together with their post-intervention measurements.

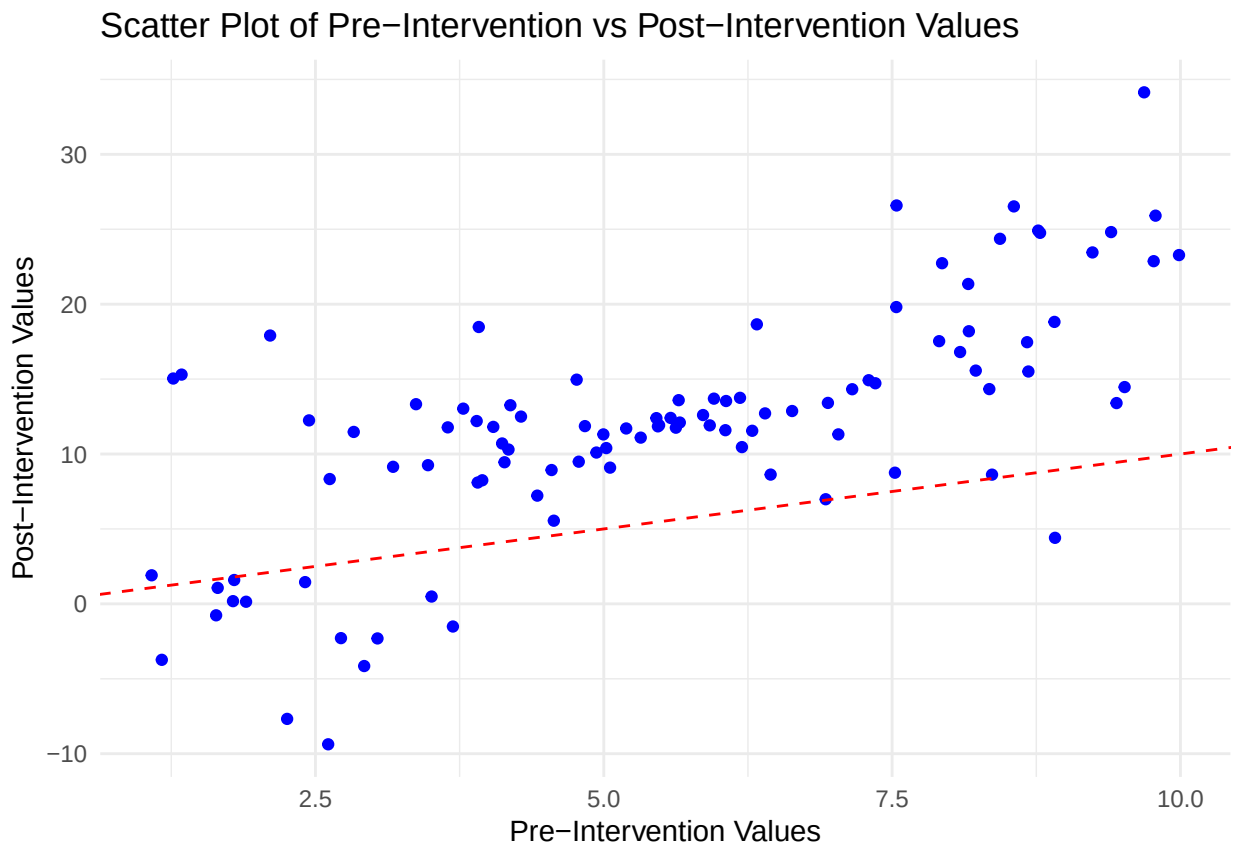
Question 3.1

(1 point)

Please generate a scatter plot of the pre-intervention values against the post-intervention values.

Your code:

```
load("dat_pre_post.RData")

# Convert the data to a data frame
dat_pre_post <- as.data.frame(dat_pre_post)

# Generate the scatter plot
ggplot(dat_pre_post, aes(x = pre, y = post)) +
  geom_point(color = "blue") +
  theme_minimal() +
  labs(title = "Scatter Plot of Pre-Intervention vs Post-Intervention Values",
       x = "Pre-Intervention Values",
       y = "Post-Intervention Values") +
  geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed")
```



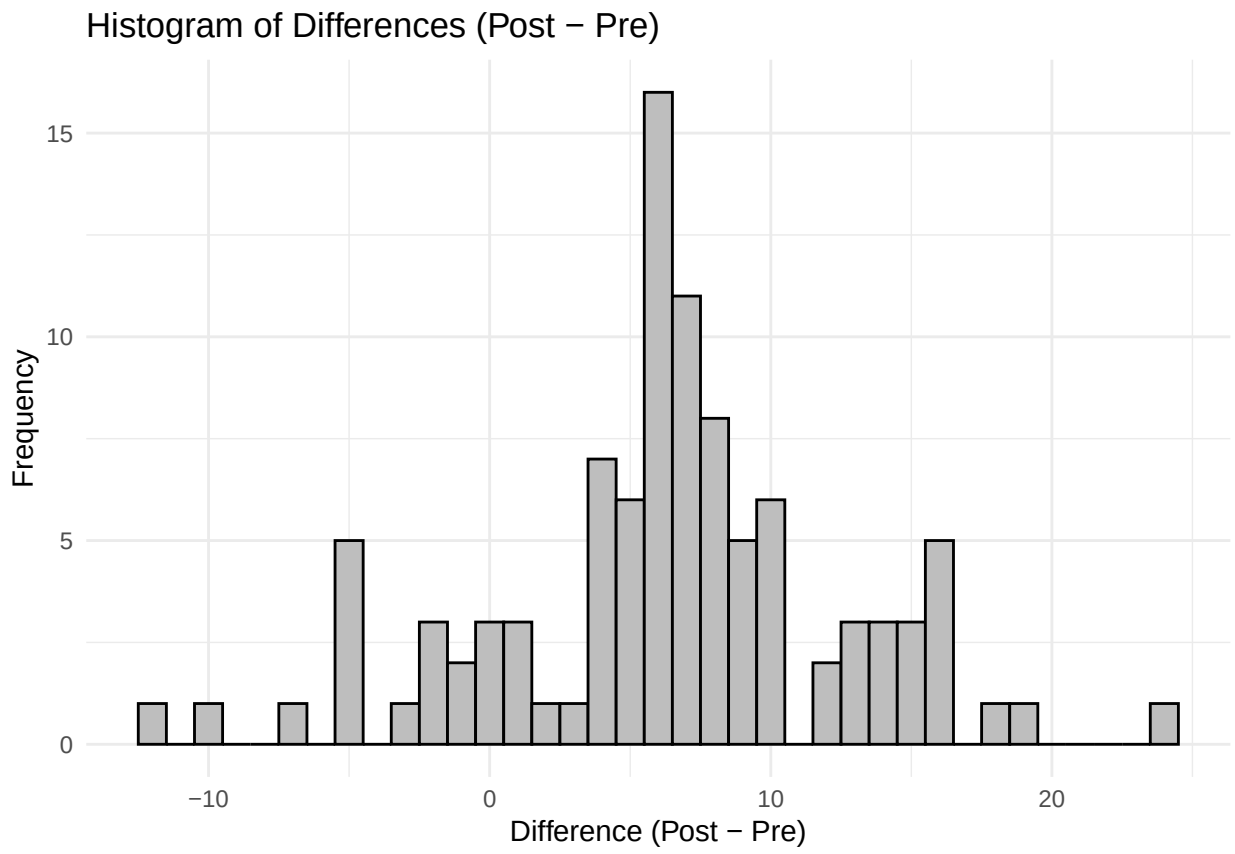
Question 3.2

(1 point)

Please generate a histogram of the difference between the post-intervention values and the pre-intervention values for the individuals. You can assume that the difference in pre- and post intervention measurements is distributed i.i.d.

Your code:

```
# Calculate the difference between post and pre intervention values
dat_pre_post$diff <- dat_pre_post$post - dat_pre_post$pre

# Generate the histogram
ggplot(dat_pre_post, aes(x = diff)) +
  geom_histogram(binwidth = 1, fill = "gray", color = "black") +
  theme_minimal() +
  labs(title = "Histogram of Differences (Post - Pre)",
       x = "Difference (Post - Pre)",
       y = "Frequency")
```



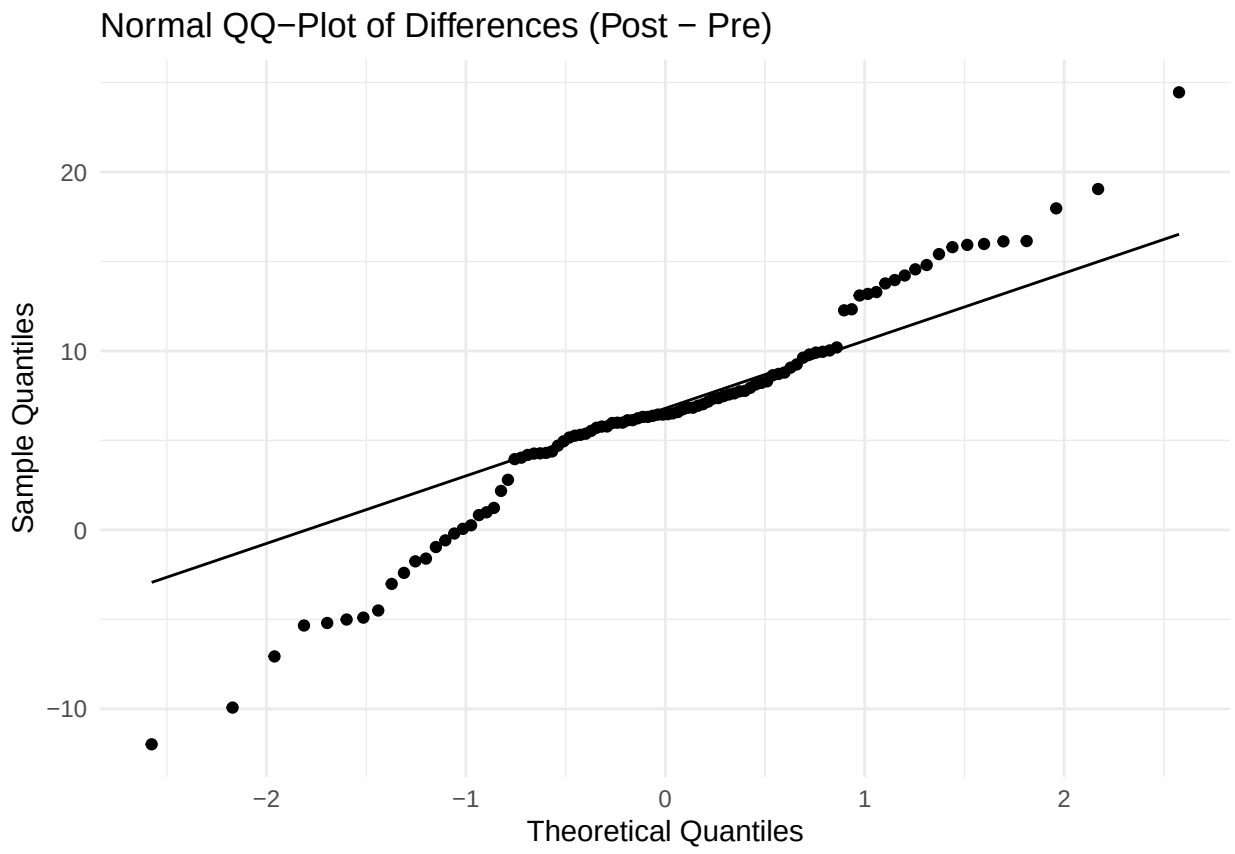
Question 3.3

(1 point)

Please generate a Normal qq-plot of the difference between the post-intervention values and the pre-intervention values for the individuals.

Your code:

```
# Ensure that the difference column is correctly created
dat_pre_post$diff <- dat_pre_post$post - dat_pre_post$pre

# Generate the Normal QQ-plot for the differences
ggplot(data = dat_pre_post, aes(sample = diff)) +
  stat_qq() +
  stat_qq_line() +
  theme_minimal() +
  labs(title = "Normal QQ-Plot of Differences (Post - Pre)",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles")
```



Question 3.4

(3 points)

Which statistical test would you use to test the null hypothesis that the intervention is not associated with a systematic increase or decrease in the values for the individuals? Please choose the test that makes the strongest use of the distributional assumptions satisfied by this data set. **Explain why you would use your chosen statistical test.**

Your answer:

Given the deviations from normality observed in the QQ plot, the paired t-test might not be the best choice unless further analysis confirms that these deviations are minor and the sample size is sufficiently large to rely on the Central Limit Theorem. The Wilcoxon signed-rank test would be appropriate if the differences are symmetrically distributed, providing a robust alternative to the paired t-test when normality is in doubt but symmetry is assumed. However, if symmetry cannot be confidently established, the sign test, which does not assume symmetry or normality, might be a more cautious approach, albeit with lower statistical power.

- First Choice: Wilcoxon signed-rank test if symmetry is confirmed.
- Alternative Choice: Sign test if symmetry is uncertain or the deviations from normality are too significant.

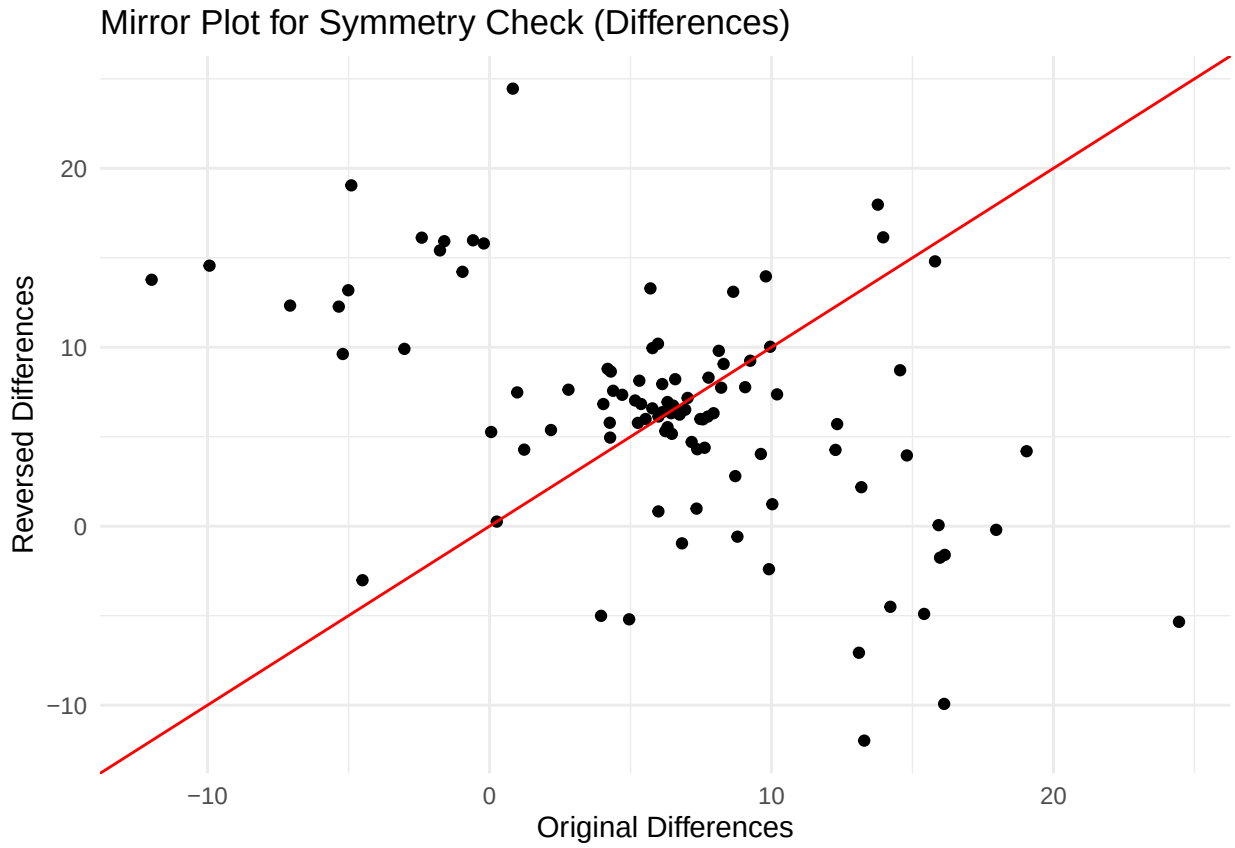
Question 3.5

(4 points)

Perform the statistical test you chose in 3.4. Please show the results of the test and provide your interpretation of whether the data are consistent with the null hypothesis. Note that to have evidence that any change was **caused** by the intervention, a controlled experiment would be required, which is not the case with this data.

Your code and answer:

```
# Create a mirror plot to check for symmetry
diff_data <- dat_pre_post$diff
reverse_diff_data <- rev(sort(diff_data))

# Mirror plot for symmetry check
ggplot(data.frame(original = diff_data, reverse = reverse_diff_data), aes(x = original, y = reverse)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  theme_minimal() +
  labs(title = "Mirror Plot for Symmetry Check (Differences)", x = "Original Differences", y = "Reverse
```



The data is reasonably symmetric. Use Wilcoxon Signed-Rank Test.

```
# Perform the Wilcoxon Signed-Rank Test
wilcox.test(diff_data, alternative = "two.sided", mu = 0, conf.int = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
```

```
## data:  diff_data
## V = 4648, p-value = 2.924e-13
## alternative hypothesis: true location is not equal to 0
## 95 percent confidence interval:
##  5.576543 7.484163
## sample estimates:
## (pseudo)median
##       6.541101
```

We reject the null hypothesis. The p-value (2.924e-13) strongly indicates that the median difference between pre- and post-intervention values is not equal to 0. The 95% confidence interval for the median difference (5.576543, 7.484163) further supports this conclusion, as it does not include 0. The pseudo-median estimate (6.541101) suggests that the intervention is associated with an increase in values. However, it is important to recognize that this does not imply causality, as the study design does not control for potential confounding factors.

Question 4: Two-Sample Data (20 points)

The data "dat_two_sample.RData" simulate independent, identically distributed samples from a population with the samples from X in the "val" column, labeled with "gp"="x" and independent, identically distributed samples from a population with the distribution Y in the "val" column, labeled with "gp"="y"

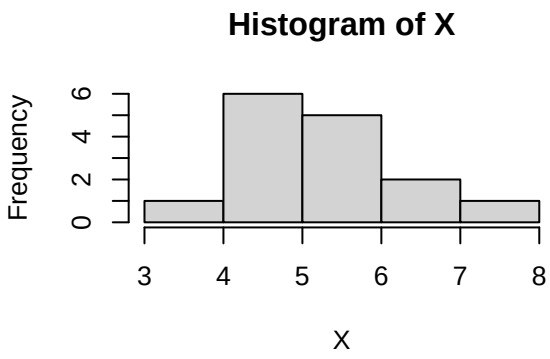
Question 4.1

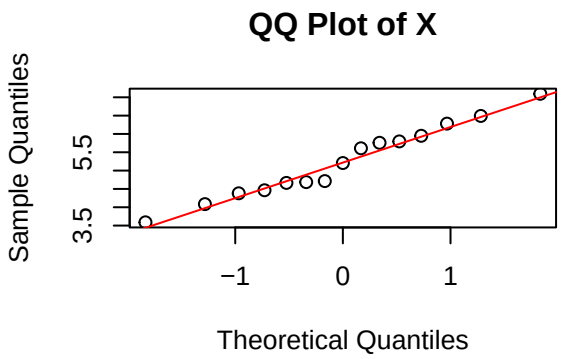
(3 points)

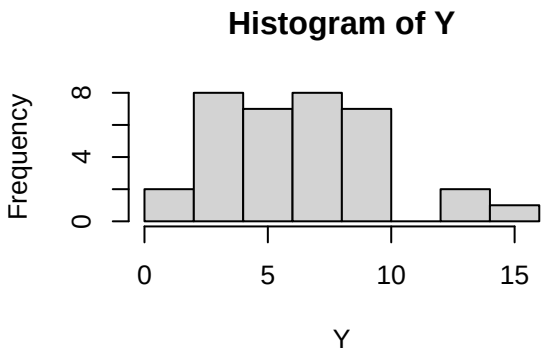
Please visually assess the Normality of the x's and the y's and **confirm with a statistical test.** Then interpret your results: are X and Y normally distributed?

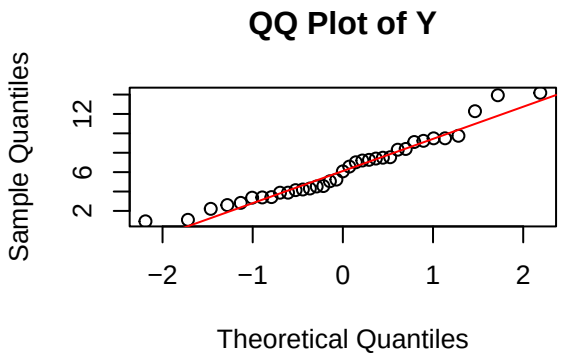
Your answer and code:

```r
load("dat_two_sample.RData")
dat_two_sample <- as.data.frame(dat_two_sample)

# Separate the data into two groups: X and Y
x_data <- dat_two_sample$val[dat_two_sample$gp == "x"]
y_data <- dat_two_sample$val[dat_two_sample$gp == "y"]

# Visualize the Data: Histograms and QQ Plots for X and Y
par(mfrow = c(2, 2))

# Histogram for X
hist(x_data, main = "Histogram of X", xlab = "X", col = "lightgray", border = "black")
# QQ plot for X
qqnorm(x_data, main = "QQ Plot of X")
qqline(x_data, col = "red")

# Histogram for Y
hist(y_data, main = "Histogram of Y", xlab = "Y", col = "lightgray", border = "black")
# QQ plot for Y
qqnorm(y_data, main = "QQ Plot of Y")
qqline(y_data, col = "red")
```










```
# Reset plotting layout
par(mfrow = c(1, 1))

# Shapiro-Wilk Test for Normality
shapiro_x <- shapiro.test(x_data)
shapiro_y <- shapiro.test(y_data)

# Print the results of the Shapiro-Wilk tests
print("Shapiro-Wilk Test for X:")
```

```
## [1] "Shapiro-Wilk Test for X:"
```

```
print(shapiro_x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x_data
## W = 0.97299, p-value = 0.8996
```

```
print("Shapiro-Wilk Test for Y:")
```

```
## [1] "Shapiro-Wilk Test for Y:"
```

```r
print(shapiro_y)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  y_data
## W = 0.9534, p-value = 0.1441
```

Based on both the visual assessments and the results of the Shapiro-Wilk tests:

- X is approximately normally distributed: The visual inspection (histogram and QQ plot) and the Shapiro-Wilk test suggest that the X data do not deviate significantly from a normal distribution.
- Y is approximately normally distributed: Although the histogram and QQ plot for Y show some deviations at the extremes, the Shapiro-Wilk test does not provide sufficient evidence to reject normality.

Thus, we conclude that both X and Y are consistent with being normally distributed.

Question 4.2

(4 points)

Please calculate the sample variances of the x's and the y's, and determine if the variances of the two groups are equal using inferential statistics. Do the two-samples have equal variance? **Explain your reasoning.**

Your code and answer:

```r
# Calculate the sample variances for X and Y
variance_x <- var(x_data)
variance_y <- var(y_data)

# Print the variances
print(paste("Sample Variance of X:", variance_x))
```

```
## [1] "Sample Variance of X: 0.973429886946045"
```

```r
print(paste("Sample Variance of Y:", variance_y))
```

```
## [1] "Sample Variance of Y: 11.08943128412"
```

```r
# Perform F-test to compare variances
var.test(x_data, y_data)
```

```
##
##  F test to compare two variances
##
## data:  x_data and y_data
## F = 0.08778, num df = 14, denom df = 34, p-value = 1.881e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.03852641 0.23746548
## sample estimates:
## ratio of variances
##          0.08777996
```

The sample variance of X is approximately 0.973. The sample variance of Y is approximately 11.089. The F-test was conducted to compare the variances of X and Y. The F-statistic is 0.08778, with a corresponding p-value of 1.881e-05. The 95% confidence interval for the ratio of variances is approximately [0.0385, 0.2375]. The p-value is significantly less than 0.05 (p-value = 1.881e-05), which indicates strong evidence against the null hypothesis of equal variances. We reject the null hypothesis and conclude that the variances of the two samples are not equal. This result suggests that there is a significant difference in the variability of the values in the X and Y groups.

Question 4.3

(3 points)

Based on 4.1 and 4.2, which statistical test would you choose for this data set to determine if X and Y come from the same underlying populatoin? Your options are the two-sample (pooled) t-test, Welch's t-test, and the Mann-Whitney U test. Pick the test that makes best use of the distributional assumptions satisfied by this data set. **Explain your reasoning.**

Your answer here:

Based on the results from Questions 4.1 and 4.2, the Welch's t-test is the most appropriate statistical test to determine if X and Y come from the same underlying population. Given that both groups are approximately normally distributed but have unequal variances, Welch's t-test is the most suitable choice. It appropriately accounts for the normality of the data and the unequal variances, providing a more reliable comparison of the means between X and Y.

Question 4.4

(3 points)

Regardless of what you concluded in 4.3, please carry out the two-sample (pooled) t-test of the null hypothesis that the means of x and y are equal, displaying the 99% confidence interval for the difference $E[X] - E[Y]$, **not the 95% confidence interval.** Then, **interpret the results.**

Your answer and code:

```r
# Perform the two-sample (pooled) t-test
t.test(x_data, y_data, var.equal = TRUE, conf.level = 0.99)
```

```
##
##  Two Sample t-test
##
## data:  x_data and y_data
## t = -1.1866, df = 48, p-value = 0.2412
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -3.406210  1.316713
## sample estimates:
## mean of x mean of y
##  5.251189  6.295937
```

Given the p-value and the confidence interval, the results suggest that there is no statistically significant difference between the means of X and Y. The data do not provide strong evidence against the null hypothesis, meaning that any observed difference in the sample means could be due to random variation rather than a true difference in the population means. The p-value obtained from the two-sample t-test is 0.2412, which is

greater than the common significance level (e.g., 0.05). This suggests that we fail to reject the null hypothesis. In other words, there is not enough evidence to conclude that the means of X and Y are significantly different. The 99% confidence interval for the difference between the means of X and Y is [−3.406210,1.316713]. Since this interval includes 0, it further indicates that the difference in means could plausibly be 0, supporting the null hypothesis that the means of X and Y are equal.

Question 4.5

(3 points)

Regardless of what you concluded in 4.3, please carry out Welch's t-test of the null hypothesis that the means of x and y are equal, displaying the 99% confidence interval for the difference $E[X] - E[Y]$, **not the 95% confidence interval.** Then, **interpret the results.**

Your answer and code:

```
# Perform Welch's t-test
t.test(x_data, y_data, var.equal = FALSE, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  x_data and y_data
## t = -1.6909, df = 44.791, p-value = 0.0978
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -2.7068449  0.6173484
## sample estimates:
## mean of x mean of y
##  5.251189  6.295937
```

Based on the Welch's t-test, there is insufficient evidence to reject the null hypothesis that the means of X and Y are different at the 99% confidence level. The confidence interval encompasses 0, indicating that the true difference in means could be 0, meaning there is no statistically significant difference between the two samples. The p-value from Welch's t-test is 0.0978, which is greater than the significance level of 0.05 but less than 0.10. This suggests weak evidence against the null hypothesis, but it's not strong enough to confidently reject the null hypothesis that the means of X and Y are equal. The 99% confidence interval for the difference in means between X and Y is [−2.7068449,0.6173484]. Since this interval includes 0, it suggests that there is no significant difference between the means of the two groups at the 99% confidence level. This conclusion aligns with the results from previous tests, reinforcing the finding that there is no strong evidence to support a difference in means between the two groups.

Question 4.6

(4 points)

Regardless of what you concluded in 4.3, please perform the Mann Whitney U test on X and Y and **interpret the results. Be sure to state the null and alternate hypotheses.**

Your answer and code:

```
# Perform the Mann-Whitney U test
wilcox.test(x_data, y_data, alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  x_data and y_data
## W = 229, p-value = 0.489
## alternative hypothesis: true location shift is not equal to 0
```

Based on the Mann-Whitney U test, we fail to reject the null hypothesis. This suggests that there is no statistically significant difference between the distributions of X and Y. In other words, the central tendencies of the two groups are not significantly different. The p-value of 0.489 is greater than the significance level of 0.05, indicating that there is insufficient evidence to reject the null hypothesis. This result aligns with the findings from the other tests, further indicating that there is no strong evidence to suggest a difference between the two samples.

Question 5: Categorical Data (20 points)

The data "mat" represent a sample from two joint distributed probability distributions X and Y. Each observation call fall into one of three categories, a, b, or c. The first column represents in mat represents a count of the number of observations in X that fall into each category. The second column in mat represents a count of the number of observations in Y that fall into each category. You can assume X is i.i.d. and Y is also i.i.d.

Question 5.1

(3 points)

Please perform a χ^2 test of the independence of X and Y based on the contingency table in "mat".

Your code:

```r
load("mat.Rdata")
colnames(mat) <- c("X","Y")

# Perform the Chi-Squared test of independence
chisq.test(mat)
```

```
## Warning in chisq.test(mat): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  mat
## X-squared = 3.6978, df = 2, p-value = 0.1574
```

Based on the Chi-Squared test, we fail to reject the null hypothesis, suggesting that there is no statistically significant association between the distributions of X and Y. The p-value of 0.1574 is greater than the common significance level of 0.05. This means there is not enough evidence to reject the null hypothesis of independence between X and Y. The warning Chi-squared approximation may be incorrect typically arises when expected frequencies in the contingency table are too low, potentially invalidating the results of the Chi-Squared test. To address this issue, you could consider using Fisher's exact test, which is more appropriate for small sample sizes or when the expected frequencies are low.

Question 5.2

(4 points)

Are the assumptions required for the χ^2 test to be valid met by this data?

Your answer and code:

```
# Calculate the expected frequencies for the Chi-Squared test
chisq_test_result <- chisq.test(mat)
```

```
## Warning in chisq.test(mat): Chi-squared approximation may be incorrect
```

```
# Check if all expected frequencies are greater than or equal to 5
print(paste("Minimum expected frequency:", min(chisq_test_result$expected)))
```

```
## [1] "Minimum expected frequency: 0.75"
```

The assumptions required for the Chi-Squared test are not fully met. The low expected frequency suggests that the Chi-Squared approximation may not be accurate for this data. Therefore, an alternative test, such as Fisher's Exact Test, should be considered for this analysis.

Question 5.3

(3 points)

Please interpret the results of the χ^2 test. For any conclusions you draw, be sure to explain why you believe those conclusions are justified.

Your answer:

Based on the Chi-Squared test, we fail to reject the null hypothesis of independence between X and Y. The p-value of 0.1574 is greater than the common significance level of 0.05. This indicates that we do not have sufficient evidence to reject the null hypothesis of independence between X and Y. Therefore, based on the Chi-Squared test, we conclude that there is no statistically significant association between the two categorical variables X and Y. However, due to the violation of the expected frequency assumption, these results should be interpreted with caution. An alternative test, such as Fisher's Exact Test, may provide more accurate insights given the data characteristics. The Chi-Squared approximation might be incorrect, and the true association between X and Y could be masked by the low expected frequencies. An alternative test, such as Fisher's Exact Test, may provide more accurate insights given the data characteristics.

Question 5.4

(5 points)

Please carry out Fisher's exact test on "mat".

Your code:

```
# Perform Fisher's Exact Test on the contingency table "mat"
fisher.test(mat)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  mat
## p-value = 0.1178
## alternative hypothesis: two.sided
```

Question 5.5

(5 points)

Please interpret the results of Fisher's exact test. For any conclusions you draw, be sure to explain why you believe those conclusions are justified.

Your answer:

Fisher's exact test was used instead of the Chi-Squared test due to the small expected frequencies observed in the contingency table. Specifically, some expected counts were less than 5, which violates the assumptions required for a valid Chi-Squared test. Fisher's exact test is more appropriate in this scenario because it does not rely on large sample assumptions and is valid even with small sample sizes. The results of Fisher's exact test returned a p-value of 0.1178. This p-value is greater than the commonly used significance level of 0.05, which indicates that we fail to reject the null hypothesis. The null hypothesis in this context is that there is no association between the categories of X and Y – meaning that the distribution of observations across the categories b, and c in X and Y are independent of each other. This interpretation aligns with the statistical evidence provided by the p-value from Fisher's exact test, ensuring that no significant relationship between X and Y is detected based on the data.

Question 6: Regression (20 points)

Question 6.1

(8 points)

Please fit a linear regression model which predicts `post` using a single variable, `pre`, from the "dat_pre_post.RData" data set. Display the coefficients with their p-values and 95% confidence interval. Interpret the coefficients, the confidence intervals, and the p-values.

Your answer and code:

```r
# Load the dataset
load("dat_pre_post.RData")

# Fit the linear regression model and display the summary
model <- lm(post ~ pre, data = dat_pre_post)
summary(model)
```

```
##
## Call:
## lm(formula = post ~ pre, data = dat_pre_post)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2078  -2.2389  -0.0508   3.3070  13.8088
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.7062     1.3661  -0.517    0.606
## pre           2.2802     0.2259  10.092   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.496 on 98 degrees of freedom
## Multiple R-squared:  0.5096, Adjusted R-squared:  0.5046
## F-statistic: 101.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
# Print the results
summary_model <- summary(model)
print(summary_model$coefficients)
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -0.7062085  1.3661339 -0.5169394 6.063631e-01
## pre          2.2801667  0.2259354 10.0921160 7.638245e-17
```

```
print(confint(model, level = 0.95))
```

```
##                 2.5 %   97.5 %
## (Intercept) -3.417257 2.004840
## pre          1.831805 2.728528
```

Interpretation:

- Intercept (-0.7062, p-value = 0.606): The intercept represents the expected post value when pre is 0. The p-value of 0.606 indicates that the intercept is not statistically significantly different from zero. This suggests that when pre is zero, the post value is not significantly different from zero, but this might not be practically meaningful depending on the context.
- Slope (2.2802, p-value $<$ 2e-16): The slope indicates that for every 1-unit increase in pre, post is expected to increase by approximately 2.28 units. The extremely low p-value indicates that this relationship is statistically significant, meaning there is strong evidence that pre is positively associated with post.
- 95% Confidence Interval for Slope (1.831805 to 2.728528): The confidence interval does not include zero, further supporting the conclusion that pre is significantly associated with post. This interval gives a range of values within which the true slope likely falls.
- Residual Standard Error (5.496): This value represents the average amount that the observed post values deviate from the fitted post values predicted by the model. A lower residual standard error would indicate a better fit, but this value should be interpreted in the context of the scale of the post variable.
- R-squared (0.5096) and Adjusted R-squared (0.5046): The model explains approximately 51% of the variance in post. While this indicates a moderate fit, it also suggests that about 49% of the variance in post is due to factors not included in the model. The adjusted R-squared is slightly lower, accounting for the number of predictors in the model.

Question. 6.2

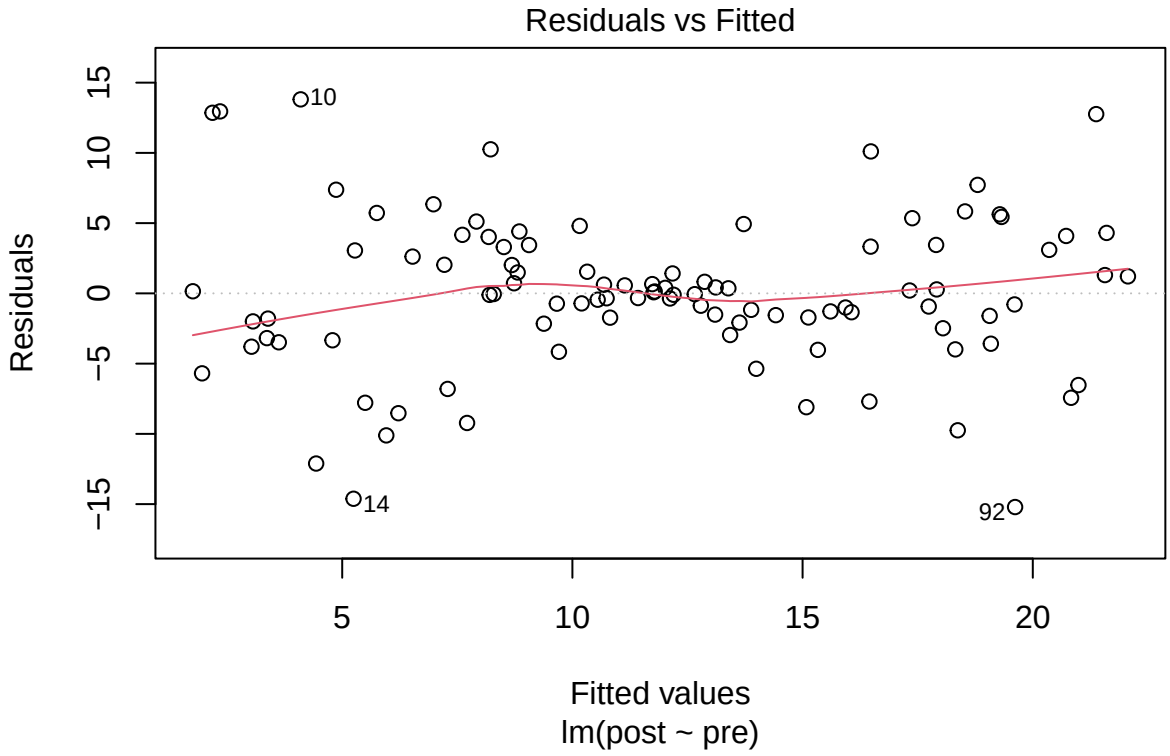
(12 points)

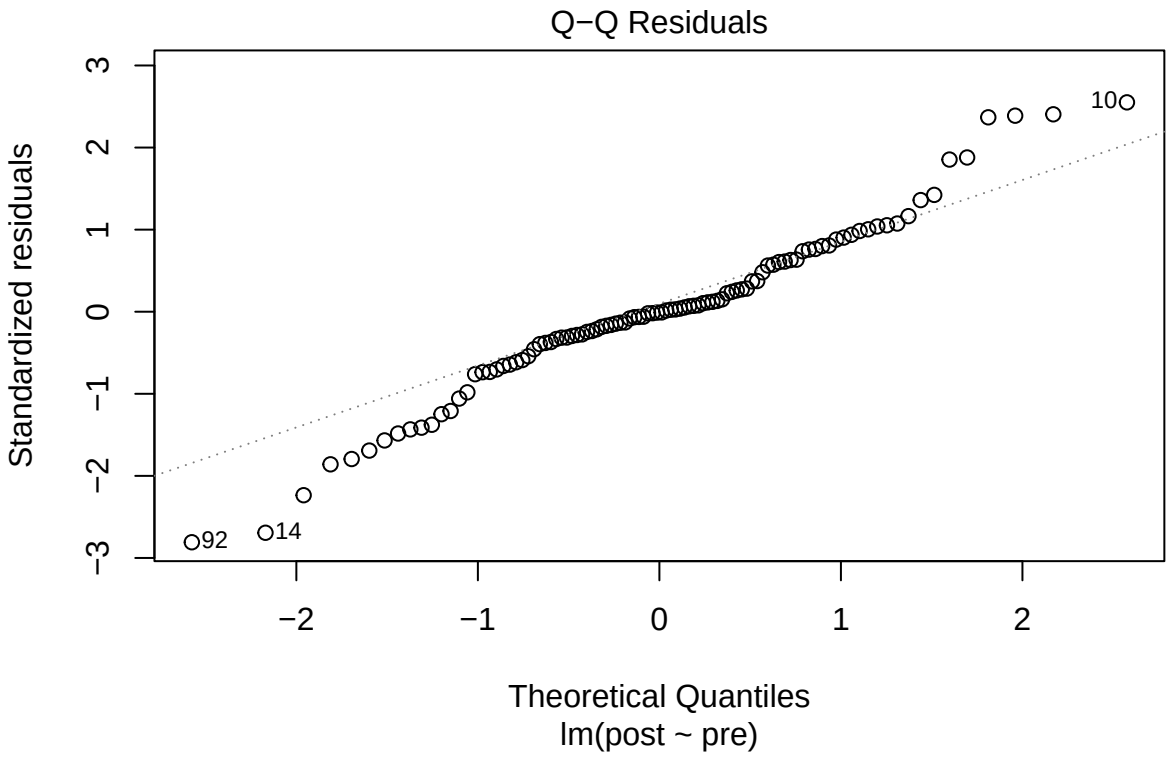
Please check whether the assumptions for linear regression to be valid are met in this case. Display diagnostic plots where applicable. **Explain or justify your reasoning for each assumption.**

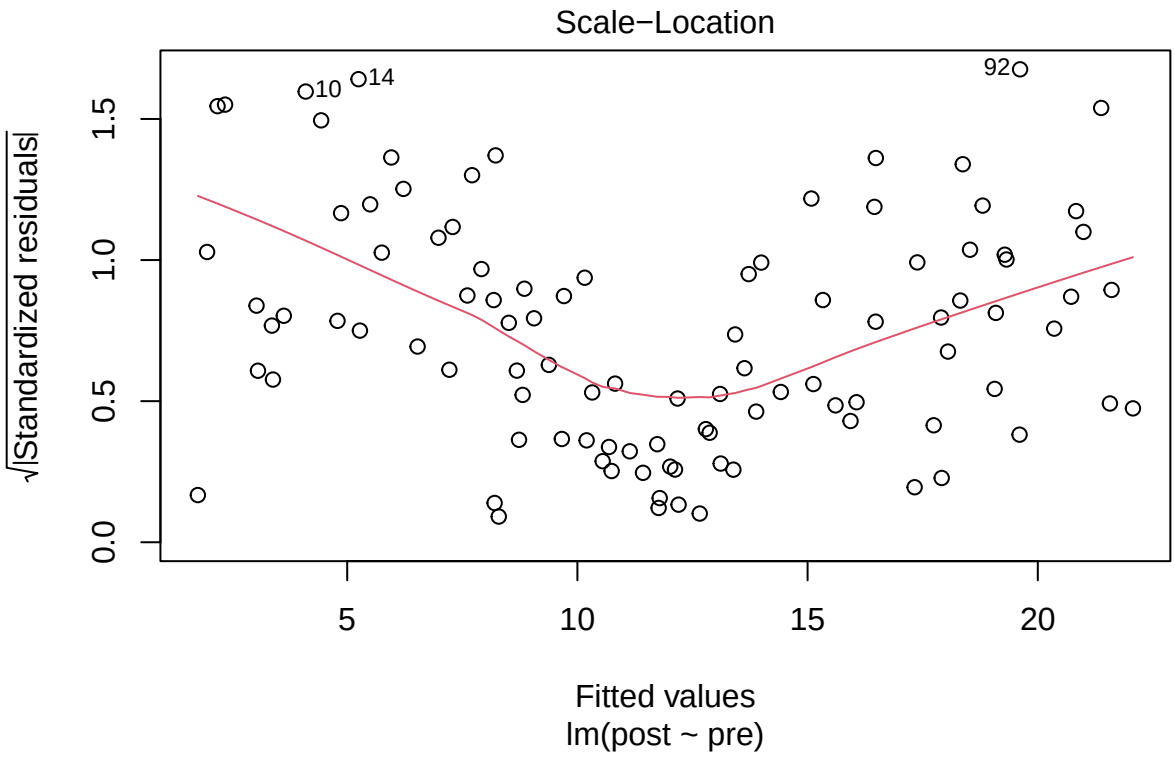
You can assume that the data is i.i.d. and there is no auto-correlation, so you do not need to check these assumptions. Remember also that the no multi-collinearity assumption is only required for regression models with two or more explanatory variables. You do not need to check additional assumptions required for causal inference, just treat this as a prediction problem.

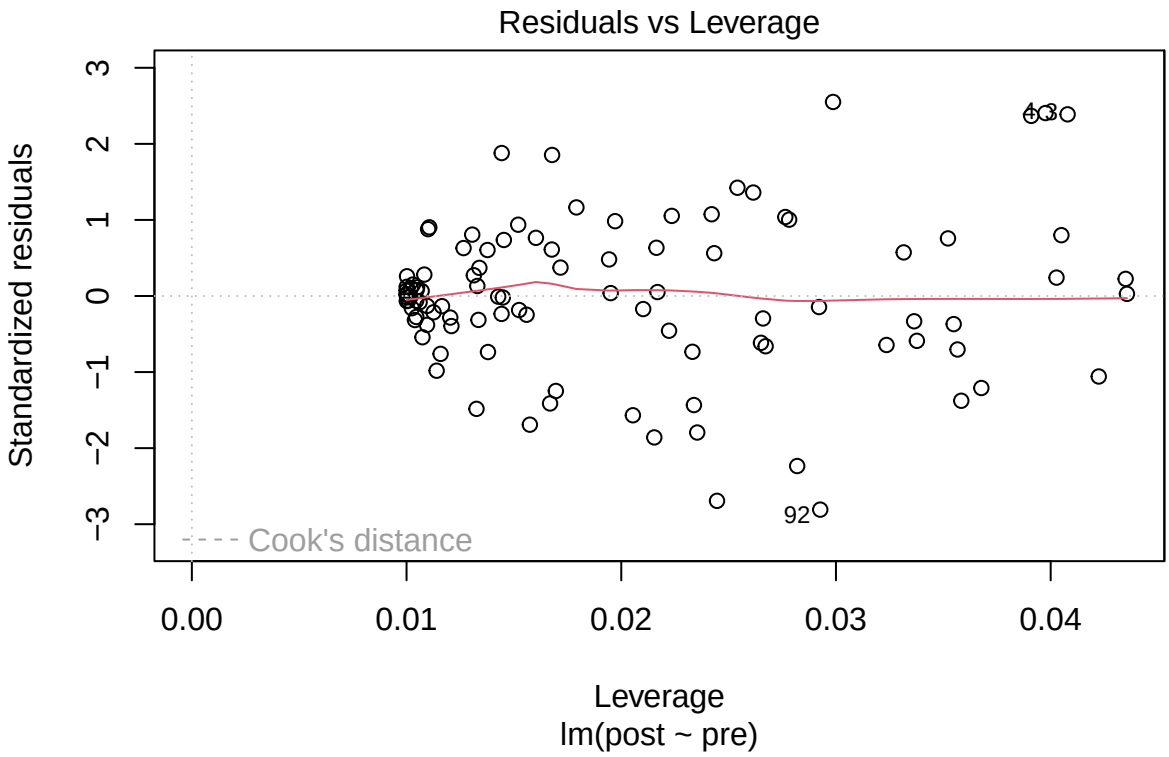
Your answer and code:

```
load("dat_pre_post.RData")

# Fit the linear regression model and display plots
model <- lm(post ~ pre, data = dat_pre_post)
plot(model)
```




Q−Q Residuals
Standardized residuals
3
2
1
0
−1
−2
−3
92
14
10
−2
−1
0
1
2
Theoretical Quantiles
lm(post ~ pre)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
14
92
5
10
15
20
Fitted values
lm(post ~ pre)



While the linear regression assumptions are mostly satisfied, the observed issues with linearity and homoscedasticity warrant further investigation or potential model adjustments to ensure the reliability of the regression results. There is slight evidence of non-linearity. Further exploration or model adjustments might be necessary. The residuals are approximately normally distributed, with minor deviations that are not uncommon. There is some evidence of heteroscedasticity, suggesting that the variance of the residuals is not constant across levels of the fitted values. No significant influential points were detected.

- Linearity: The Residuals vs Fitted plot does show some slight curvature, which might indicate that the assumption of linearity isn't fully met. This observation is correct. However, it's important to clarify that the curvature seen might be due to a mild non-linearity or other factors, and more detailed diagnostics (like adding polynomial terms or interaction terms) might be necessary to fully assess linearity.
- Normality of Residuals: The Q-Q plot generally indicates that the residuals are normally distributed, with some minor deviations at the tails. This interpretation is correct. The normality assumption seems reasonably met, though minor deviations, particularly in small sample sizes, are not uncommon.
- Homoscedasticity (Constant Variance): The Scale-Location plot suggests that the variance of the residuals increases with the fitted values, which is a sign of heteroscedasticity. This is a correct observation, and it is important because heteroscedasticity can lead to inefficient estimates and affect the validity of hypothesis tests. It might be worth considering transformations (e.g., a log transformation) or robust standard errors to address this issue.
- Leverage and Influence: The Residuals vs Leverage plot does not show any points with high Cook's distance, meaning that there are no observations with undue influence on the regression model. This interpretation is correct.

End.