SA 1 -APM1111

PAUL JOAQUIN DELOS SANTOS

2024-10-09

This study compares the mean time spent on cell phones by male and female college students per week. We will use a two-sample t-test to determine if there is a significant difference between the mean times for males and females.

(10 points) Formulate and present the rationale for a hypothesis test that the researcher could use to compare the mean time spent on cell phones by male and female college students per week.

We aim to test the following hypotheses:

- Null Hypothesis (H0): There is no difference in the mean time spent on cell phones by males and females, i.e., $H_0 : \mu_1 - \mu_2 = 0$.
- Alternative Hypothesis (Ha): There is a difference in the mean time spent on cell phones by males and females, i.e., $H_a : \mu_1 - \mu_2 \neq 0$.

This two-sample t-test will help determine whether the mean time spent by males and females on cell phones is significantly different, or if the observed differences are due to random chance.

Data

We collected data from 50 male and 50 female students at Midwestern University. The number of hours per week spent talking on their cell phones is provided below:

```
# Input data
males <- c(12, 4, 11, 13, 11, 7, 9, 10, 10, 7, 7, 12, 6, 9, 15, 10, 11, 12, 7, 8,
           8, 9, 11, 10, 9, 10, 9, 9, 7, 9, 11, 7, 10, 10, 11, 9, 12, 12, 8, 13,
           9, 10, 8, 11, 10, 13, 13, 9, 10, 13)

females <- c(11, 9, 7, 10, 9, 10, 10, 7, 9, 10, 11, 8, 9, 6, 11, 10, 7, 9, 12, 14,
             11, 12, 12, 8, 12, 12, 9, 10, 11, 7, 12, 7, 9, 8, 11, 10, 8, 13, 8, 10,
             9, 9, 9, 11, 9, 9, 8, 9, 12, 11)
```

(30 points) Analyze the data to provide the hypothesis testing conclusion. What is the p-value for your test? What is your recommendation for the researcher?

We will perform a two-sample t-test to analyze the data and calculate the p-value.

```
# Perform two-sample t-test
t_test_result <- t.test(males, females, alternative = "two.sided", var.equal = TRUE)

# Extract relevant values from t-test result
test_statistic <- t_test_result$statistic
p_value <- t_test_result$p.value
confidence_interval <- t_test_result$conf.int
mean_difference <- t_test_result$estimate[1] - t_test_result$estimate[2]
degrees_of_freedom <- t_test_result$parameter

# Create a summary table for the t-test results
t_test_summary <- data.frame(
```

```
  Statistic = c("t-statistic", "p-value", "Mean Difference", "95% Confidence Interval Lower",
                "95% Confidence Interval Upper", "Degrees of Freedom"),
  Value = c(test_statistic, p_value, mean_difference, confidence_interval[1],
            confidence_interval[2], degrees_of_freedom)
)

# Display the summary table
kable(t_test_summary, caption = "Summary of Two-Sample T-Test Results")
```

Summary of Two-Sample T-Test Results

Statistic	Value
t-statistic	0.3039491
p-value	0.7618111
Mean Difference	0.1200000
95% Confidence Interval Lower	-0.6634737
95% Confidence Interval Upper	0.9034737
Degrees of Freedom	98.0000000

If the p-value is less than 0.05, we will reject the null hypothesis. Otherwise, we fail to reject it.

```
# Conclusion based on p-value
if (p_value < 0.05) {
  conclusion <- "There is a significant difference in the mean times spent on cell phones by ma
les and females."
} else {
  conclusion <- "There is no significant difference in the mean times spent on cell phones by m
ales and females."
}

cat("P-value: ", p_value, "\n")
```

```
## P-value:  0.7618111
```

```
cat("Conclusion: ", conclusion)
```

```
## Conclusion:  There is no significant difference in the mean times spent on cell phones by ma
les and females.
```

Recommendations

Based on the results of the two-sample t-test, we can make the following recommendation:

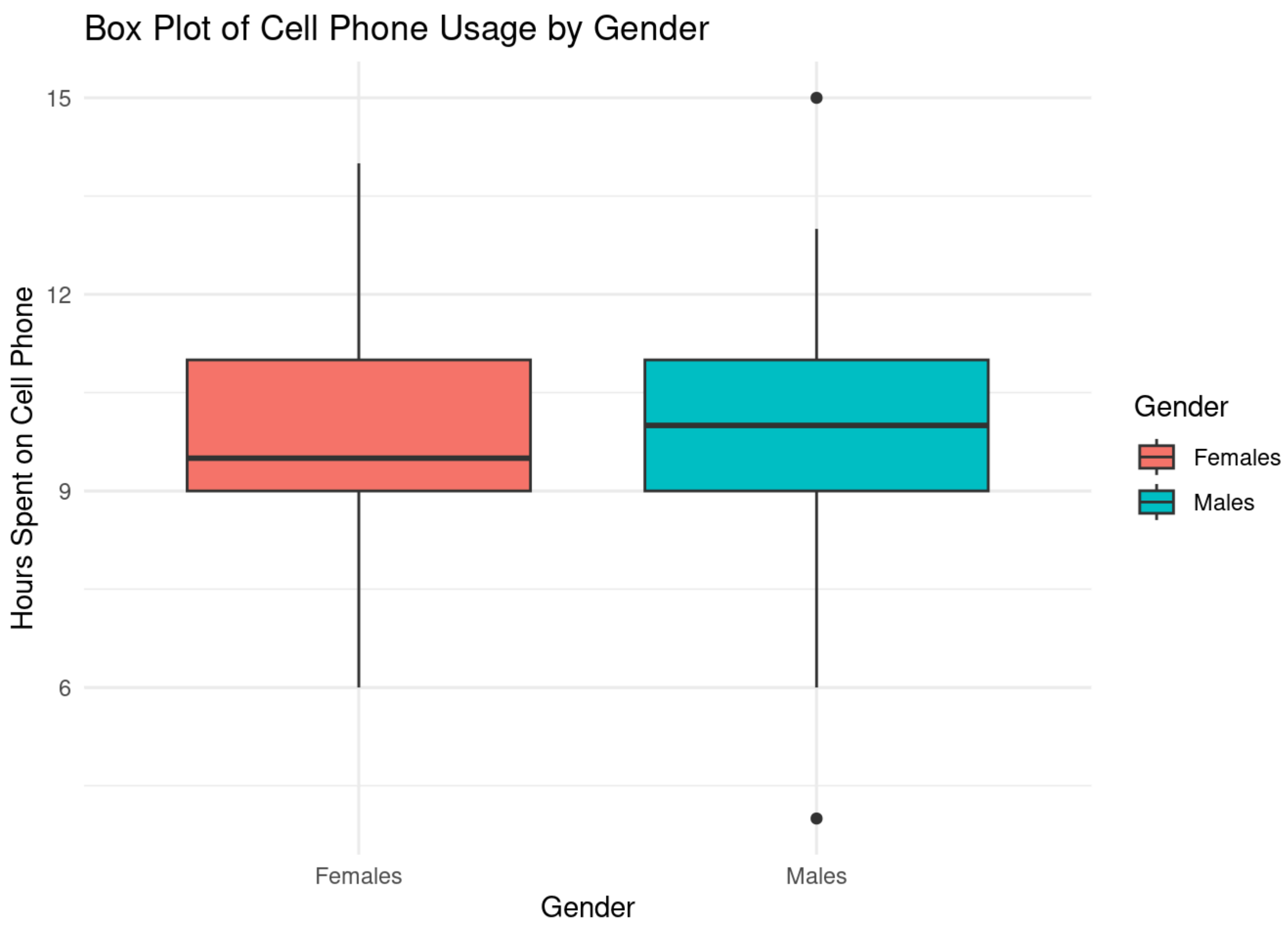
If the p-value is below the significance level of 0.05, the researcher should reject the null hypothesis and conclude that there is a statistically significant difference in the mean time spent on cell phones by male and female students. This would suggest that gender has an impact on cell phone usage.

If the p-value is above 0.05, the researcher should fail to reject the null hypothesis, concluding that there is no statistically significant difference in the mean time spent on cell phones between males and females. This implies that gender may not be a major factor influencing the time spent on cell phones.

(10 points) Provide descriptive statistical summaries of the data for each gender category.

We begin by calculating the descriptive statistics (mean, standard deviation, and sample size) for each group:

```
# Create a data frame for plotting
data <- data.frame(
  Gender = rep(c("Males", "Females"), each = 50),
  Hours = c(males, females)
)

# Create box plot
ggplot(data, aes(x = Gender, y = Hours, fill = Gender)) +
  geom_boxplot() +
  labs(title = "Box Plot of Cell Phone Usage by Gender",
       x = "Gender",
       y = "Hours Spent on Cell Phone") +
  theme_minimal()
```




```
# Calculate descriptive statistics for males
mean_males <- mean(males)
sd_males <- sd(males)
n_males <- length(males)

# Calculate descriptive statistics for females
mean_females <- mean(females)
sd_females <- sd(females)
n_females <- length(females)

# Create a summary table for descriptive statistics
descriptive_stats <- data.frame(
  Gender = c("Males", "Females"),
```

```
  Mean = c(mean_males, mean_females),
  SD = c(sd_males, sd_females),
  N = c(n_males, n_females)
)
kable(descriptive_stats, caption = "Descriptive Statistics by Gender")
```

Descriptive Statistics by Gender

Gender	Mean	SD	N
Males	9.82	2.154161	50
Females	9.70	1.775686	50

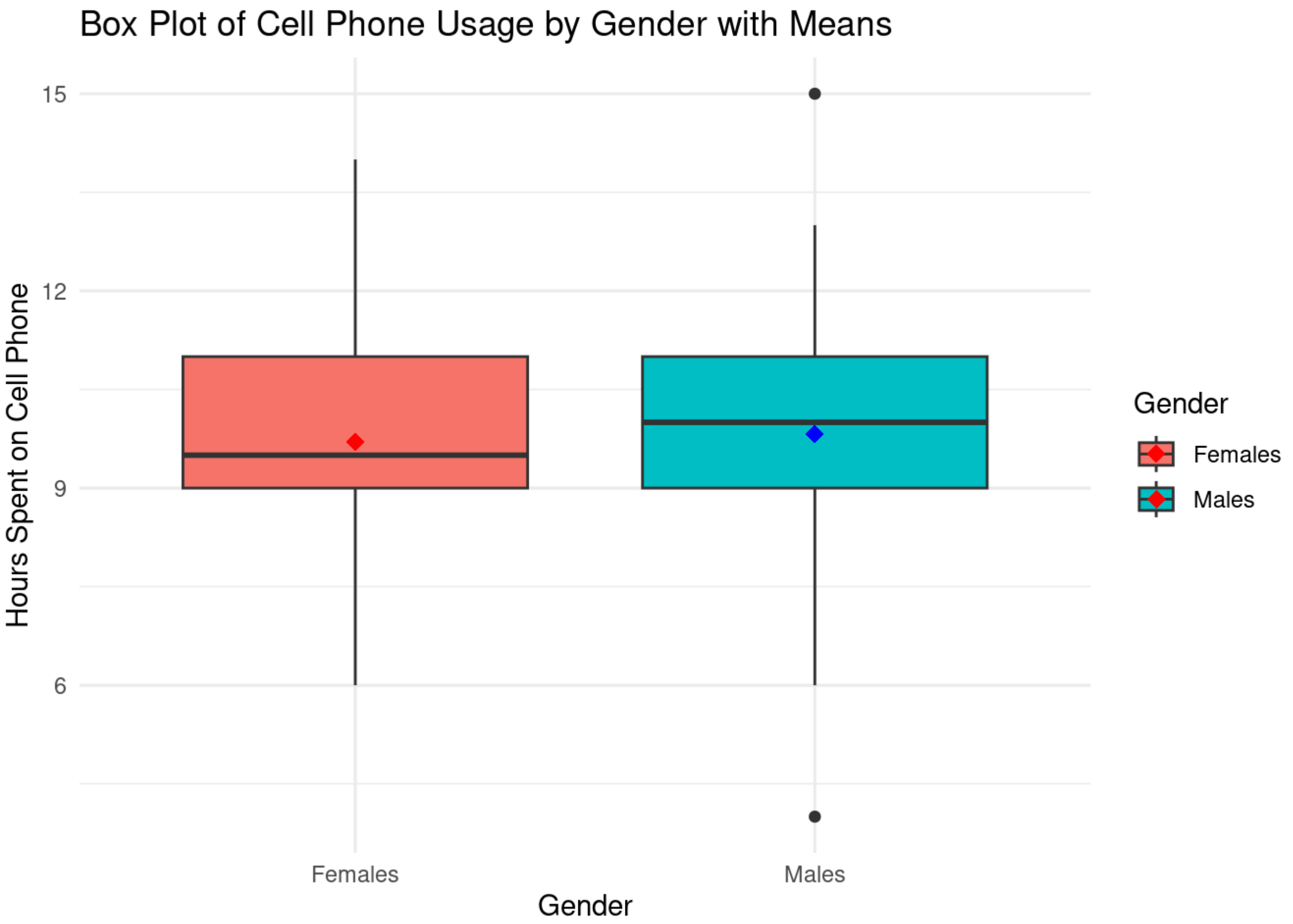
(10 points) What is the 95% confidence interval for the population mean of each gender category, and what is the 95% confidence interval for the difference between the means of the two populations?

```
# Calculate 95% confidence interval for males
ci_males <- t.test(males)$conf.int

# Calculate 95% confidence interval for females
ci_females <- t.test(females)$conf.int

# Create a summary table for confidence intervals
confidence_intervals <- data.frame(
  Gender = c("Males", "Females"),
  `95% Confidence Interval` = c(
    paste(round(ci_males[1], 2), "-", round(ci_males[2], 2)),
    paste(round(ci_females[1], 2), "-", round(ci_females[2], 2))
  )
)

kable(confidence_intervals, caption = "95% Confidence Intervals by Gender")
```

95% Confidence Intervals by Gender

Gender	X95..Confidence.Interval
Males	9.21 - 10.43
Females	9.2 - 10.2

```
# Create box plot with aligned means
ggplot(data, aes(x = Gender, y = Hours, fill = Gender)) +
  geom_boxplot() +
  geom_point(aes(x = "Males", y = mean_males), color = "blue", size = 3, shape = 18) +
  geom_point(aes(x = "Females", y = mean_females), color = "red", size = 3, shape = 18) +
  labs(title = "Box Plot of Cell Phone Usage by Gender with Means",
       x = "Gender",
       y = "Hours Spent on Cell Phone") +
  theme_minimal()
```



The 95% confidence interval for the difference between the means is calculated as part of the two-sample t-test:

```
# Confidence interval for the difference between the means
# Perform two-sample t-test
t_test_result <- t.test(males, females, alternative = "two.sided", var.equal = TRUE)

# Extract the confidence interval from the t-test result
confidence_interval <- t_test_result$conf.int

# Create a summary table for confidence intervals
conf_interval_table <- data.frame(
  `Confidence Interval` = paste(round(confidence_interval[1], 2), "-", round(confidence_interva
l[2], 2))
)

kable(conf_interval_table, caption = "Confidence Interval from t-test Results")
```

Confidence Interval from t-test Results

Confidence.Interval
-0.66 - 0.9

(20 points) Do you see a need for larger sample sizes and more testing with the time spent on cell phones? Discuss.

While 50 male and 50 female participants provide a reasonable sample size, increasing the sample size would lead to more precise estimates of the population means and narrower confidence intervals. Larger samples would reduce variability and provide more reliable insights into differences between genders. Additionally, a larger sample would make the results more generalizable to other populations of college students.

(20 points) Make a report including the testing of the assumptions for two independent samples t-test.

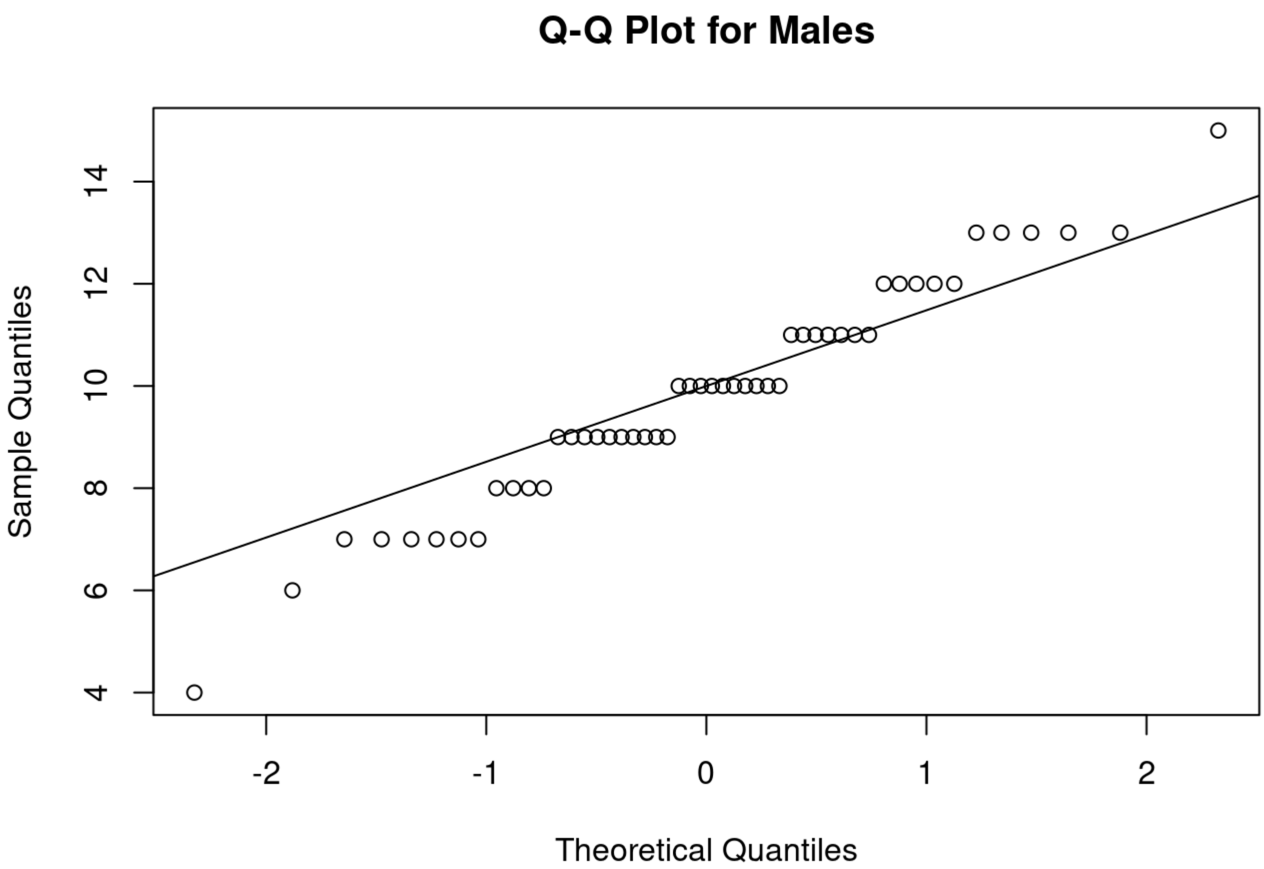
For a two-sample t-test, the following assumptions must be satisfied:

1. Independence: The observations for males and females are independent.
2. Normality: The data in each group should be approximately normally distributed.
3. Equal Variances: The variances of the two groups should be roughly equal.

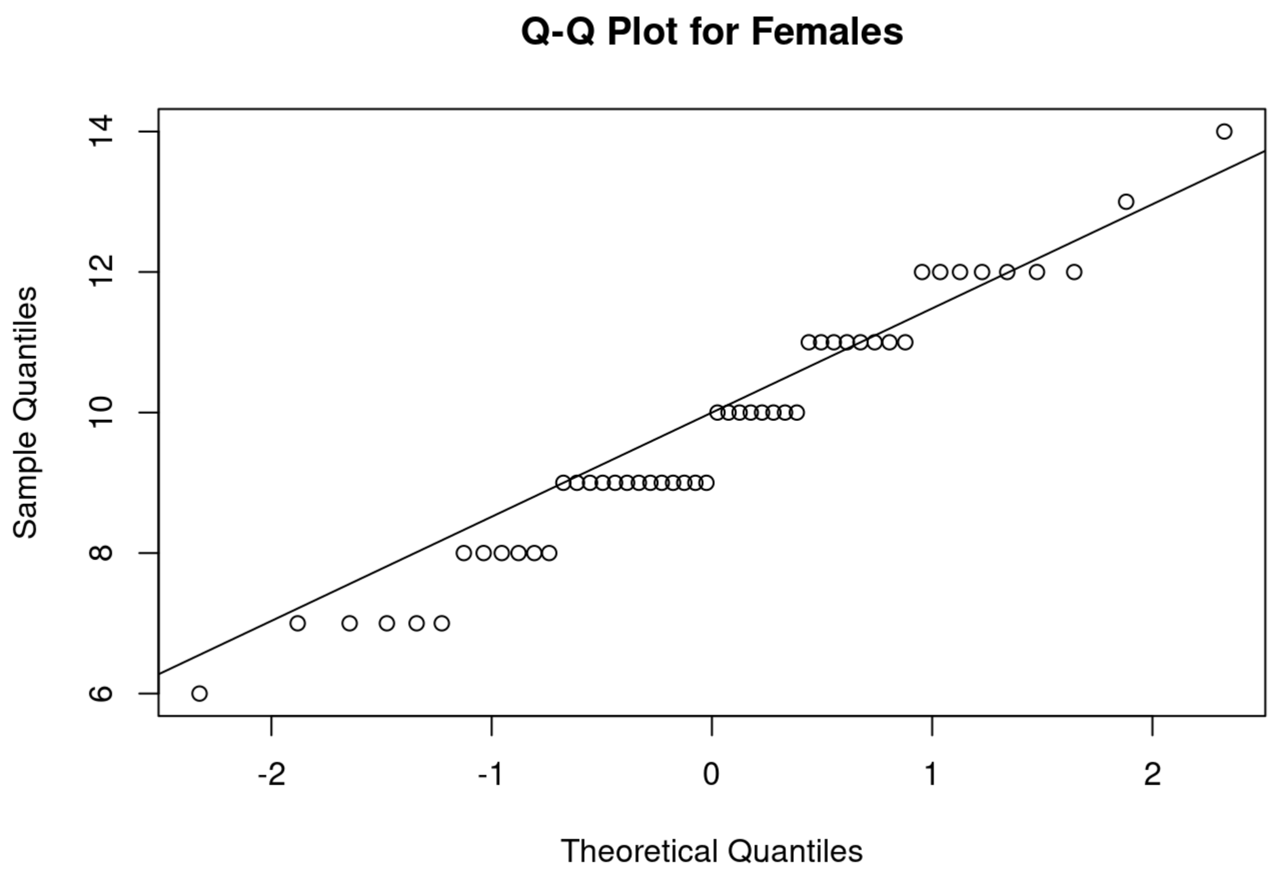
Testing for Normality

We check normality using Q-Q plots and the Shapiro-Wilk test.

```
# Q-Q plot for males
qqnorm(males, main = "Q-Q Plot for Males")
qqline(males)
```



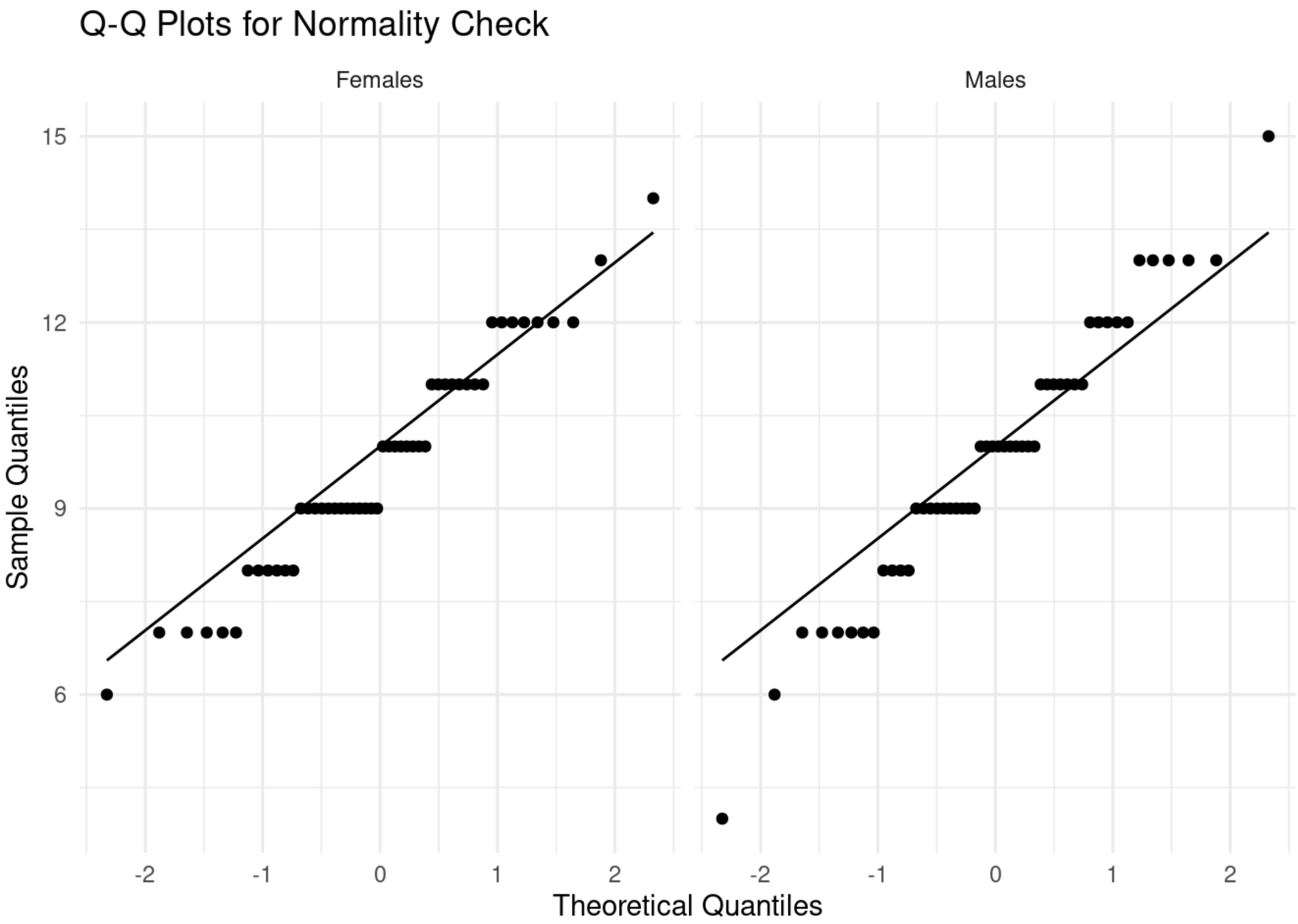

```
# Q-Q plot for females
qqnorm(females, main = "Q-Q Plot for Females")
qqline(females)
```




```
# Shapiro-Wilk test for normality
shapiro_males <- shapiro.test(males)
shapiro_females <- shapiro.test(females)

# Display the results
data.frame(
  Gender = c("Males", "Females"),
  Shapiro_W_P_Value = c(shapiro_males$p.value, shapiro_females$p.value)
)
```

```
##    Gender Shapiro_W_P_Value
## 1   Males         0.3539890
## 2 Females         0.1291972
```

```
# Q-Q plots as an additional normality check
ggplot(data, aes(sample = Hours)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~Gender) +
  labs(title = "Q-Q Plots for Normality Check",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  theme_minimal()
```



If the p-value from the Shapiro-Wilk test is greater than 0.05, we can assume that the data is approximately normally distributed.

Testing for Equal Variances

We test for equal variances using an F-test.

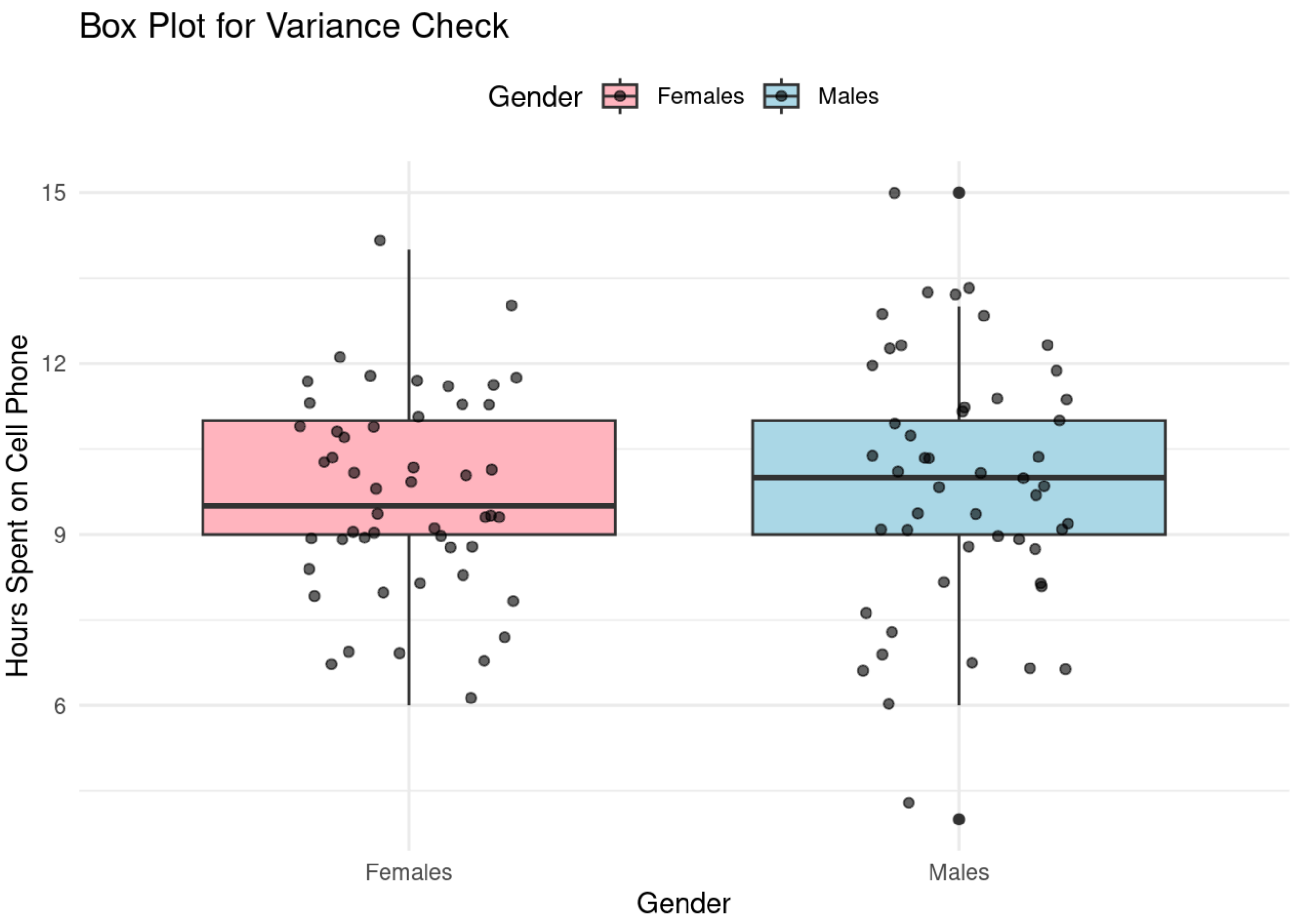
```
# Variance test
var_test <- var.test(males, females)

# Display the result
var_test
```

```
##
##  F test to compare two variances
##
## data:  males and females
## F = 1.4717, num df = 49, denom df = 49, p-value = 0.1798
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8351633 2.5934397
## sample estimates:
## ratio of variances
##           1.471715
```

If the p-value is greater than 0.05, we assume equal variances.

```
# Create box plot to check for equal variances with colors and points
ggplot(data, aes(x = Gender, y = Hours, fill = Gender)) +
  geom_boxplot() +
  geom_jitter(position = position_jitter(0.2), color = "black", size = 1.5, alpha = 0.6) + # Ad
d points
```

```
  labs(title = "Box Plot for Variance Check",
       x = "Gender",
       y = "Hours Spent on Cell Phone") +
  theme_minimal() +
  scale_fill_manual(values = c("Males" = "lightblue", "Females" = "lightpink")) + # Custom colors
  theme(legend.position = "top")  # Position the legend at the top
```



Summary

In this analysis, we tested the hypothesis that the mean time spent on cell phones per week by male and female students is the same. Descriptive statistics were provided for each gender, and a two-sample t-test was conducted. The results, including the p-value, confidence intervals, and assumptions of normality and equal variances, were evaluated. The final recommendation was provided based on the results of the hypothesis test.