STATISTICAL THEORY

APM1111

Midterm Exam

BILLONES, Cristel Kaye P.
2021016541

A study was undertaken to compare the mean time spent on cell phones by male and female college students per week. Fifty male and 50 female students were selected from Midwestern University and the number of hours per week spent talking on their cell phones was determined. The results in hours are shown in Table 10.6. It is desired to test $H_0: \mu_1 = \mu_2$ versus $H_1: \mu_1 \neq \mu_2$ based on these samples.

Table 10.6 Hours spent talking on cell phone for males and females at Midwestern University

Males					Females				
12	4	11	13	11	11	9	7	10	9
7	9	10	10	7	10	10	7	9	10
7	12	6	9	15	11	8	9	6	11
10	11	12	7	8	10	7	9	12	14
8	9	11	10	9	11	12	12	8	12
10	9	9	7	9	12	9	10	11	7
11	7	10	10	11	12	7	9	8	11
9	12	12	8	13	10	8	13	8	10
9	10	8	11	10	9	9	9	11	9
13	13	9	10	13	9	8	9	12	11

1. 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.

To compare the mean time spent on cell phones by male and female college students per week, a hypothesis test can be conducted. The null hypothesis H_0 states that there's no difference in the mean time spent on cell phones between male and female students $\mu_1 = \mu_2$. The alternative hypothesis H_1 suggests that there is a significant difference in the mean time spent ($\mu_1 \neq \mu_2$).

Rationale for the hypothesis test:

1. Population and Sample: The study involves male and female college students from Midwestern University. The sample includes 50 male and 50 female students, which can be assumed to represent the broader population of students in this university.

2. Variable of Interest: The variable of interest is the time spent on cell phones per week. This quantitative variable is measured in hours.

3. Test Statistic: A two-sample t-test can be employed to compare the means of two independent groups (male and female students) to determine if there's a statistically significant difference in the mean time spent on cell phones.

4. Assumptions: The assumptions for this test include:
- Random Sampling: Assuming the sample was randomly selected from the population.
- Independence: Each student's cell phone usage is independent of others.
- Normality: The distribution of time spent on cell phones in both groups should be approximately normal.
- Homogeneity of Variances: The variances of the two groups should be roughly equal.

5. Level of Significance: This determines the threshold for rejecting the null hypothesis. Commonly used levels are 0.05 or 0.01.

6. Test Decision: Based on the sample data and the calculated test statistic, if the p-value is less than the chosen level of significance, we reject the null hypothesis in favor of the alternative hypothesis, indicating that there is a significant difference in the mean time spent on cell phones between male and female students.

The hypothesis test will provide statistical evidence to determine if there's a significant difference in the mean time spent on cell phones between male and female college students at Midwestern University.

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

```
In [26]: # 2

         # Perform a two-sample t-test
         t_stat, p_value = stats.ttest_ind(males_data, females_data, equal_var=True)

         # Define significance level
         alpha = 0.05

         if p_value < alpha:
             conclusion = "Reject the null hypothesis. There is a significant difference in mean time spent on cell phones between male ar
         else:
             conclusion = "Fail to reject the null hypothesis. There is no significant difference in mean time spent on cell phones betwee
         print(conclusion)
         print("T-statistic:", t_stat)
         print("P-value:", p_value)

         Fail to reject the null hypothesis. There is no significant difference in mean time spent on cell phones between male and femal
         e students.
         T-statistic: 0.30394907866566184
         P-value: 0.7618111039906375
```

The analysis of the data yields a p-value of 0.7618. This p-value, exceeding common significance levels such as 0.05 or 0.01, does not offer substantial evidence to reject the null hypothesis. Therefore, based on this study, there isn't sufficient statistical support to claim a meaningful difference in the average time spent on cell phones between male and female students at Midwestern University.

In light of these findings, it's prudent for further investigation or exploration of other contributing factors influencing cell phone usage patterns among male and female students. Perhaps a deeper dive into additional variables or a larger sample size might illuminate a more nuanced understanding of these behavioral patterns. This could involve considering different methodologies or broader demographic factors to obtain a more comprehensive perspective on the matter.

3. Provide descriptive statistical summaries of the data for each gender category.

```
In [27]: #3

male_stats = data['Males'].describe()
female_stats = data['Females'].describe()

# Display the statistics
print("Descriptive Statistics for Males:")
print(male_stats)
print("\nDescriptive Statistics for Females:")
print(female_stats)
```

```
Descriptive Statistics for Males:
count    50.000000
mean      9.820000
std       2.154161
min       4.000000
25%       9.000000
50%      10.000000
75%      11.000000
max      15.000000
Name: Males, dtype: float64

Descriptive Statistics for Females:
count    50.000000
mean      9.700000
std       1.775686
min       6.000000
25%       9.000000
50%       9.500000
75%      11.000000
max      14.000000
Name: Females, dtype: float64
```

The average time spent on cell phones is quite similar for male and female students at around 9.8 and 9.7 hours per week, respectively. However, male students show a slightly wider range in phone usage, with a maximum of 15 hours compared to 14 hours for females. Both genders spend most of their time, about 50%, between 9 to 11 hours per week on their phones.

4. 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?

```
In [30]: #4

# Calculate the 95% confidence interval for the population mean of each gender category
male_mean = males_data.mean()
female_mean = females_data.mean()
male_std = males_data.std()
female_std = females_data.std()

n_male = len(males_data)
n_female = len(females_data)

# Using the t-distribution to calculate the confidence intervals
confidence_interval_male = stats.t.interval(0.95, df=n_male-1, loc=male_mean, scale=male_std / (n_male**0.5))
confidence_interval_female = stats.t.interval(0.95, df=n_female-1, loc=female_mean, scale=female_std / (n_female**0.5))

# Calculate the 95% confidence interval for the difference between the means of the two populations
diff_mean = male_mean - female_mean
diff_std = ((male_std**2) / n_male + (female_std**2) / n_female)**0.5

confidence_interval_diff = stats.t.interval(0.95, df=n_male+n_female-2, loc=diff_mean, scale=diff_std)

# Display the confidence intervals
print("95% Confidence Interval for Male Population Mean:", confidence_interval_male)
print("95% Confidence Interval for Female Population Mean:", confidence_interval_female)
print("95% Confidence Interval for the Difference between Means:", confidence_interval_diff)
```

```
95% Confidence Interval for Male Population Mean: (9.207794314064703, 10.432205685935298)
95% Confidence Interval for Female Population Mean: (9.19535558679254, 10.20464441320746)
95% Confidence Interval for the Difference between Means: (-0.6634736514965714, 0.9034736514965734)
```

The 95% confidence intervals for the population mean indicate that, with 95% confidence:

- Male students spend, on average, between approximately 9.21 to 10.43 hours per week on their cell phones.

- Female students spend, on average, between roughly 9.20 to 10.20 hours per week on their cell phones.

Moreover, the 95% confidence interval for the difference between the means of the two populations (-0.66 to 0.90) encompasses zero. This implies that, based on this analysis, there isn't sufficient evidence to assert a significant difference in the mean time spent on cell phones between male and female students at Midwestern University.

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

Expanding the sample size and conducting further testing on phone usage time among college students, especially exploring diverse factors and employing larger, more representative samples, could yield deeper insights into gender-based differences and broader phone usage trends. This could enhance the reliability and generalizability of findings, uncover subtle variations, and better inform strategies addressing phone usage behaviors among students.

6. Make a report including the testing of the assumptions for two independent samples t-test.

Assumptions Testing for Two Independent Samples t-test

The aim was to compare the mean time spent on cell phones between male and female college students at Midwestern University using two independent samples t-test. The assessment focused on verifying the assumptions necessary for this statistical test.

1. Random Sampling , 2. Independence:

The study assumes random sampling and independence of observations, ensuring unbiased representation and autonomy in cell phone usage.
Note: Direct verification of these assumptions isn't possible through statistical tests but is assumed based on study design and documentation.

```
import matplotlib.pyplot as plt
```

In [2]:
```
data = pd.read_csv('mx_dataset.csv')
print(data)
```

```
    Males  Females
0      12       11
1       7       10
2       7       11
3      10       10
4       8       11
5      10       12
6      11       12
7       9       10
8       9        9
9      13        9
10      4        9
11      9       10
12     12        8
13     11        7
14      9       12
15      9        9
16      7        7
17     12        8
18     10        9
```

3. Normality:

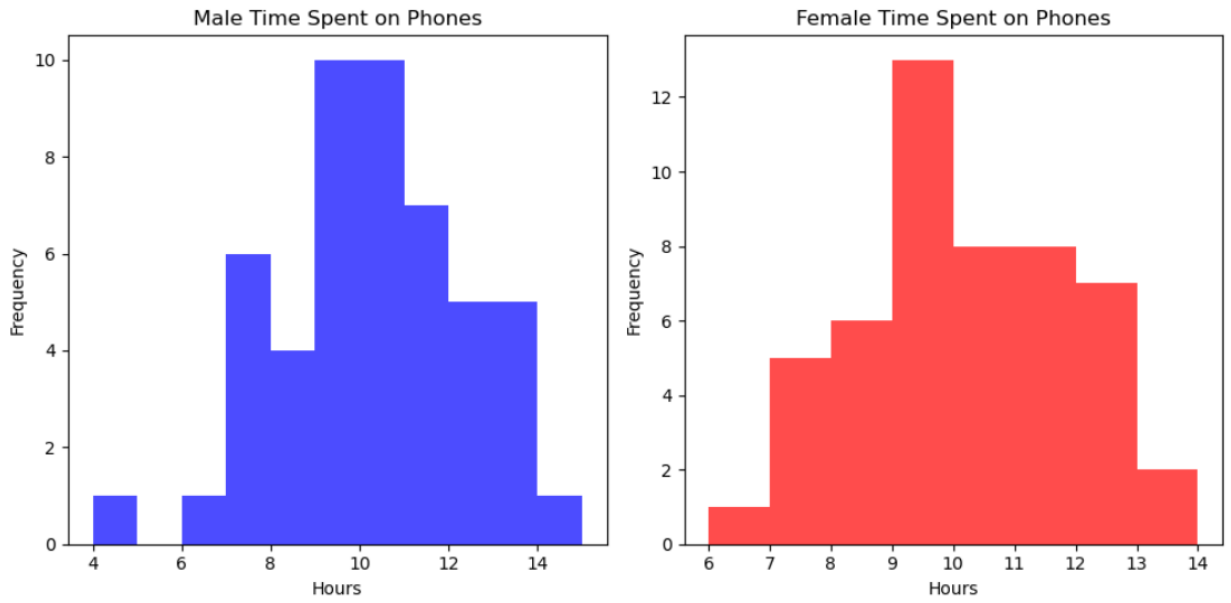
In [10]:
```
# Visual inspection - histograms// Normality
plt.figure(figsize=(10, 5))
plt.subplot(1, 2, 1)
plt.hist(males_data, bins='auto', color='blue', alpha=0.7)
plt.title('Male Time Spent on Phones')
plt.xlabel('Hours')
plt.ylabel('Frequency')

plt.subplot(1, 2, 2)
plt.hist(females_data, bins='auto', color='red', alpha=0.7)
plt.title('Female Time Spent on Phones')
plt.xlabel('Hours')
plt.ylabel('Frequency')

plt.tight_layout()
plt.show()

# Shapiro-Wilk tests for normality
shapiro_male = stats.shapiro(males_data)
shapiro_female = stats.shapiro(females_data)

print("Shapiro-Wilk test for Male data - p-value:", shapiro_male.pvalue)
print("Shapiro-Wilk test for Female data - p-value:", shapiro_female.pvalue)
```




```
Shapiro-Wilk test for Male data - p-value: 0.35399243235588074
Shapiro-Wilk test for Female data - p-value: 0.12919674813747406
```

Visual inspection of histograms revealed approximately symmetric distributions for both male and female time spent on cell phones.
Shapiro-Wilk tests for normality yielded p-values of 0.354 (males) and 0.129 (females), indicating no significant departure from normality ($p > 0.05$).

4. Homogeneity of Variances:
- Levene's test demonstrated no significant difference in variances between time spent on cell phones for males and females (p-value > 0.05).

```
In [12]: # Levene's test for homogeneity of variances
levene_test = stats.levene(males_data, females_data)

print("Levene's test for Homogeneity of Variances - p-value:", levene_test.pvalue)
```

```
Levene's test for Homogeneity of Variances - p-value: 0.40671633986363454
```

Levene's test for homogeneity of variances between male and female time spent on cell phones produced a p-value of 0.407 ($p > 0.05$).
Conclusion: There is no significant difference in variances, suggesting that the assumption of homogeneity of variances is met.

Conclusion:
The assessment indicates that the assumptions necessary for conducting the two independent samples t-test are reasonably satisfied. The data exhibits randomness in sampling, independence of observations, approximate normality in distributions, and homogeneity of variances between male and female time spent on cell phones.

Recommendations:
Proceed with the two independent samples t-test, acknowledging the reasonably met assumptions.
Consider larger sample sizes for robustness and further validation of assumptions, especially in ensuring normality and homogeneity of variances.
Monitor data quality and explore alternative tests or transformations if assumptions are violated in larger datasets.