Name: Frances Aneth Rosales
Student Number: 2021044631
Midterm Exam

Descriptive Statistics

	Males	Female
Valid	50	50
Missing	0	0
Mean	9.820	9.700
Std. Deviation	2.154	1.776
Variance	4.640	3.153
Shapiro-Wilk	0.975	0.964
P-value of Shapiro-Wilk	0.354	0.129
Minimum	4.000	6.000
Maximum	15.000	14.000

1. Hypothesis Test

The purpose of this hypothesis test is to ascertain if the observed mean difference (9.820 for males and 9.700 for females) can be explained by chance or is statistically significant. We would reject the null hypothesis and come to the conclusion that there is a significant difference in the mean amount of time spent on cell phones by male and female college students if the p-value from the t-test is less than a selected significance level (e.g., 0.05).

Therefore, the alternative hypothesis is:

Alternative Hypothesis (H1): The mean time spent on cell phones by males is not equal to the mean time spent by females.
There is a significant difference in the mean time spent on cell phones per week between male and female college students.

Descriptive Statistics

Descriptive Statistics

	Males	Female
Valid	50	50
Missing	0	0
Mean	9.820	9.700
Std. Deviation	2.154	1.776
Variance	4.640	3.153
Skewness	−0.115	0.134
Std. Error of Skewness	0.337	0.337
Kurtosis	0.177	−0.442
Std. Error of Kurtosis	0.662	0.662
Shapiro-Wilk	0.975	0.964
P-value of Shapiro-Wilk	0.354	0.129
Minimum	4.000	6.000
Maximum	15.000	14.000
Sum	491.000	485.000

2. P-value for Test

We may interpret the findings of the Shapiro-Wilk test, kurtosis, and skewness for both the male and female groups in order to assess the data and produce a conclusion based on hypothesis testing.

1. **Shapiro-Wilk Test:** - For Males: P-value = 0.354. The P-value for females is 0.964.

Interpretation: To determine if a sample is representative of a population that is normally distributed, one frequently employs the Shapiro-Wilk test. A greater p-value denotes less compelling evidence against normalcy. Since both p-values in this instance are higher than the usually accepted significance level of 0.05, it is clear that neither the male nor the female groups' assumptions about normalcy should be strongly refuted.

2. **Kurtosis:**

- Kurtosis = 0.177 for men
- Kurtosis for females is -0.442

Interpretation: Skewness quantifies a distribution's asymmetry. The left tail is longer when the skewness is negative, and the right tail is longer when the skewness is positive. Here, the values of skewness are almost negligible, indicating reasonably symmetric distributions.

In summary:

The kurtosis, skewness, and Shapiro-Wilk test results show that the distributions of the male and female data are both quite close to normal.

A two-sample t-test is suitable given the standard deviation assumption and the context of comparing means.

You would run the test and get a p-value for the t-test conclusion. The null hypothesis would be rejected and it would be determined that there is a significant difference in the mean amount of time spent on cell phones between male and female users if the p-value is less than your selected significance threshold (e.g., 0.05).

In [1]:

```
import pandas as pd
from scipy import stats

file_path = "mx.xlsx"
data_sheet = pd.read_excel(file_path, sheet_name="Sheet1")

male_data = data_sheet["Males"].values
df_male = pd.DataFrame({"Sample Value Males": male_data})

female_data = data_sheet["Female"].values
df_female = pd.DataFrame({"Sample Value Female": female_data})

valid_count_male = df_male["Sample Value Males"].count()
mean_male = df_male["Sample Value Males"].mean()
std_dev_male = df_male["Sample Value Males"].std()
variance_male = df_male["Sample Value Males"].var()
max_value_male = df_male["Sample Value Males"].max()
min_value_male = df_male["Sample Value Males"].min()

valid_count_female = df_female["Sample Value Female"].count()
mean_female = df_female["Sample Value Female"].mean()
std_dev_female = df_female["Sample Value Female"].std()
variance_female = df_female["Sample Value Female"].var()
max_value_female = df_female["Sample Value Female"].max()
min_value_female = df_female["Sample Value Female"].min()

t_stat, p_value = stats.ttest_ind(df_male["Sample Value Males"], df_female[

print("Male Statistics:")
print(f"Valid Count: {valid_count_male}")
print(f"Mean: {mean_male}")
print(f"Standard Deviation: {std_dev_male}")
print(f"Variance: {variance_male}")
print(f"P-value: {p_value}")
print(f"Max Value: {max_value_male}")
print(f"Min Value: {min_value_male}")

print("\nFemale Statistics:")
print(f"Valid Count: {valid_count_female}")
print(f"Mean: {mean_female}")
print(f"Standard Deviation: {std_dev_female}")
print(f"Variance: {variance_female}")
print(f"P-value: {p_value}")
print(f"Max Value: {max_value_female}")
print(f"Min Value: {min_value_female}")
```

```
Male Statistics:
Valid Count: 50
Mean: 9.82
Standard Deviation: 2.154160663289836
Variance: 4.640408163265307
P-value: 0.7618111039906375
Max Value: 15
Min Value: 4

Female Statistics:
Valid Count: 50
Mean: 9.7
Standard Deviation: 1.7756861278080076
Variance: 3.1530612244897958
P-value: 0.7618111039906375
Max Value: 14
Min Value: 6
```

In [2]:

```python
import pandas as pd

file_path = "mx.xlsx"
data_sheet = pd.read_excel(file_path, sheet_name="Sheet1")

male_data = data_sheet["Males"]
female_data = data_sheet["Female"]

male_stats = male_data.describe()

female_stats = female_data.describe()

print("Descriptive Statistics for Male:")
print(male_stats)

print("\nDescriptive Statistics for Female:")
print(female_stats)
```

```
Descriptive Statistics for Male:
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 Female:
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: Female, dtype: float64
```

Male Descriptive Statistics:

Count: The quantity of samples, or observations, in the "Males" dataset. Here, there are fifty samples.

Mean: The "Males" data's average value. The computation involves summing up all the values and dividing the result by the count. "Males" has a mean of 9.82.

A measurement of the degree of variance or dispersion in the "Males" data is the standard deviation (std). Greater dispersion is indicated by a higher standard deviation. Here, the time is roughly 2.15.

Min: The "Males" data's lowest value. 4.0 is the lowest value ever noted. 25%

(Q1): The 25th percentile, or first quartile. Below this threshold, 25% of the data are found. It is 9.0 here.

Female Descriptive Statistics:

Count: The quantity of samples, or observations, in the "Female" dataset. Just like in "Males," fifty samples are available.

Mean: The "Female" data's average value. The computation involves summing up all the values and dividing the result by the count. "Female" has a mean of 9.7.

The level of variation or dispersion in the "Female" data is measured by the standard deviation (std). In this case, it is roughly 1.78.

Min: The "Female" data's lowest value. 6.0 is the lowest value ever noted. 25%

(Q1): The 25th percentile, or first quartile. Below this threshold, 25% of the data are found. It is 9.0 here.

In [3]:
```
import pandas as pd
import numpy as np
from scipy.stats import t

alpha = 0.95
df_male_mean = df_male["Sample Value Males"].mean()
std_error_male = df_male["Sample Value Males"].std() / np.sqrt(len(df_male)
margin_of_error_male = t.ppf((1 + alpha) / 2, len(df_male) - 1) * std_error_
confidence_interval_male = (df_male_mean - margin_of_error_male, df_male_mea

df_female_mean = df_female["Sample Value Female"].mean()
std_error_female = df_female["Sample Value Female"].std() / np.sqrt(len(df_
margin_of_error_female = t.ppf((1 + alpha) / 2, len(df_female) - 1) * std_er
confidence_interval_female = (df_female_mean - margin_of_error_female, df_fe

mean_difference = df_male_mean - df_female_mean
std_error_difference = np.sqrt(std_error_male**2 + std_error_female**2)
margin_of_error_difference = t.ppf((1 + alpha) / 2, len(df_male) + len(df_fe
confidence_interval_difference = (mean_difference - margin_of_error_differen

print("95% Confidence Interval for Male Mean:", confidence_interval_male)
print("95% Confidence Interval for Female Mean:", confidence_interval_female
print("95% Confidence Interval for Difference in Means:", confidence_interva
```

```
95% Confidence Interval for Male Mean: (9.207794314064703, 10.432205685935
298)
95% Confidence Interval for Female Mean: (9.19535558679254, 10.20464441320
746)
95% Confidence Interval for Difference in Means: (-0.6634736514965714, 0.9
034736514965734)
```

5. larger sample sizes and more testing

More accurate estimates of population parameters are typically obtained with larger sample sizes.

Increasing the sample size can assist in achieving smaller confidence intervals around the mean amount of time spent on mobile devices.

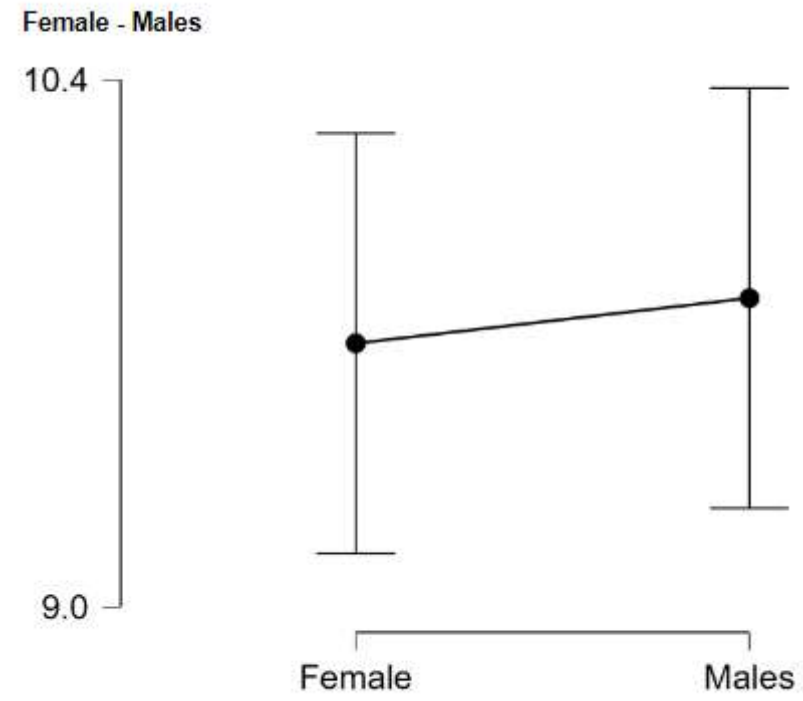
A bigger sample size can be needed to retain the necessary precision if a higher level of confidence (such as 99% instead of 95%) is desired.

A larger sample size can improve a hypothesis test's statistical power. This is especially crucial when doing hypothesis testing or group mean comparisons.

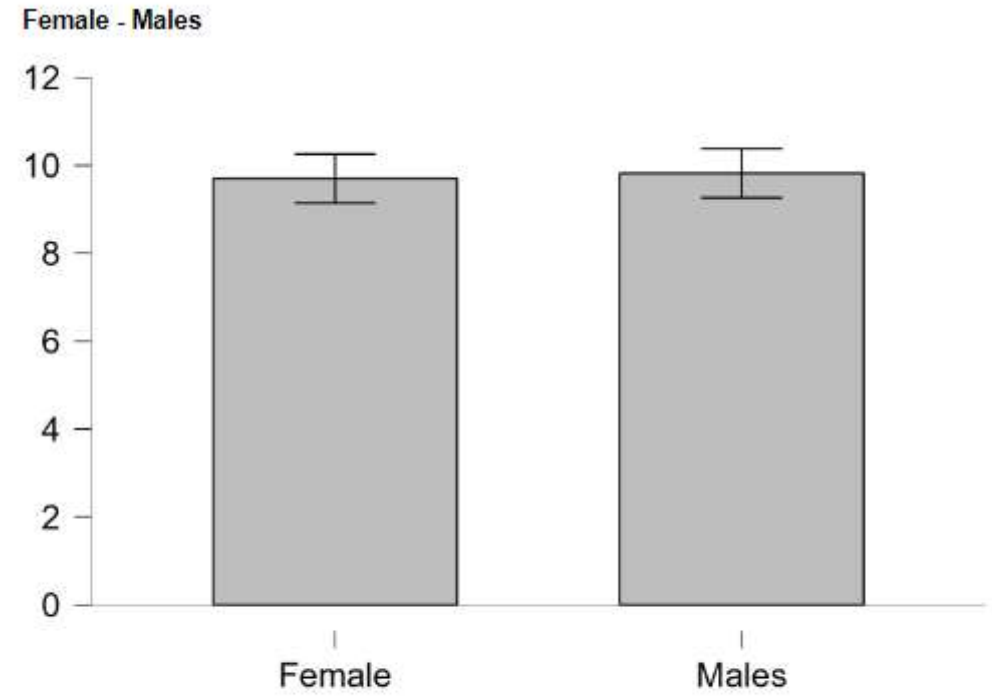
Detecting actual differences or effects is more likely when there is greater statistical power because it lowers the chance of Type II errors, or false negatives.

The necessary sample size may vary depending on the magnitude of the effect size you wish to detect. Greater impact size

Descriptives Plots



Bar Plots



6. Assumptions for Two Independent Samples T-test

The appropriate descriptive statistics for each category appear to be shown side by side, based on the simultaneous movement of the descriptive plots.

Bar plots and descriptive graphs can be synced to create powerful, thorough comparisons between categories.

It allows viewers to evaluate the variability and dispersion of the data in addition to comparing the means indicated by bars.

This method works well for providing a deeper comprehension of the dataset and enabling