Hypothesis Testing

In this project, we will work on a dataset that has the laptop screen time data of a person. The goal of this project is to see whether the person uses his laptop more on weekdays or on weekends. We will run a Two-sample Test for Means to conduct a hypothesis test to see whether we have any statistically significant difference between laptop usage on weekdays vs. weekends. The dataset contains two months of laptop usage data. There are 3 columns, 'Date', 'Day' and 'Usage'. Screentime is recorded in hours (e.g. 6.40 hours).

In [ ]:

```
# importing relevant libraries
import pandas as pd
import numpy as np
from scipy import stats
```

In [ ]:

```
path = 'screentime.csv'
```

In [ ]:

```
# Loading and checking the dataset
df = pd.read_csv(path)
df.head(5)
```

Out[ ]:

	Date	Day	Usage
0	5/1/2023	Monday	6.40
1	5/2/2023	Tuesday	6.61
2	5/3/2023	Wednesday	5.16
3	5/4/2023	Thursday	5.26
4	5/5/2023	Friday	8.87

Exploratory Data Analysis and Data Cleaning

In [ ]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 61 entries, 0 to 60
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   Date    61 non-null     object 
 1   Day     61 non-null     object 
 2   Usage   61 non-null     float64
dtypes: float64(1), object(2)
memory usage: 1.6+ KB
```

Date column is recorded as 'object' and not 'datetime'. But we do not need to change that for this project.

In [ ]:
```
# basic statistics of screentime
df['Usage'].describe()
```

Out[ ]:
```
count    61.000000
mean      4.471279
std       1.547388
min       1.780000
25%       3.210000
50%       4.560000
75%       5.550000
max       8.870000
Name: Usage, dtype: float64
```

Let's convert usage time from hours to minutes.

In [ ]:
```
df['usage_minutes'] = np.round(df['Usage'] * 60)
```

In [ ]:
```
df['usage_minutes'].describe()
```

```
Out[ ]: count     61.000000
        mean     268.459016
        std       92.838493
        min      107.000000
        25%      193.000000
        50%      274.000000
        75%      333.000000
        max      532.000000
        Name: usage_minutes, dtype: float64
```

```
In [ ]: # let's divide the days in 'weekday' or 'weekend' in a new column
        df['day_type'] = np.where((df['Day'] == 'Saturday') | (df['Day'] == 'Sunday'), 'Weekend', 'Weekday')
```

```
In [ ]: # checking row count for the new column
        df['day_type'].value_counts()
```

```
Out[ ]: Weekday    45
        Weekend    16
        Name: day_type, dtype: int64
```

```
In [ ]: # let's check the whole dataframe again
        df.head()
```

Out[ ]:

	Date	Day	Usage	usage_minutes	day_type
0	5/1/2023	Monday	6.40	384.0	Weekday
1	5/2/2023	Tuesday	6.61	397.0	Weekday
2	5/3/2023	Wednesday	5.16	310.0	Weekday
3	5/4/2023	Thursday	5.26	316.0	Weekday
4	5/5/2023	Friday	8.87	532.0	Weekday

Now, we will filter the dataset. We will create two new dataframes, one will have all weekday records and the other will have all weekend records.

In [ ]:

```
# weekday data
weekdays = df[df['day_type'] == 'Weekday']
```

In [ ]:

```
# weekend data
weekends = df[df['day_type'] == 'Weekend']
```

Observed Mean

Now, let's check the observed mean for both the new datasets.

In [ ]:

```
print('Weekday observed mean:  ',weekdays['usage_minutes'].mean())
print('Weekend observed mean:  ',weekends['usage_minutes'].mean())
```

```
Weekday observed mean:   272.5111111111111
Weekend observed mean:   257.0625
```

In [ ]:

```
print('Difference in observed mean: ', weekdays['usage_minutes'].mean() - weekends['usage_minutes'].mean())
```

```
Difference in observed mean:   15.448611111111112
```

We see that there is a difference in the observed mean of the screentime for different types of days. But, this observed difference might simply be due to chance - rather than an actual difference in the corresponding population means. A hypothesis test can help us determine whether or not our results are statistically significant.

Hypothesis Test

In a two-sample t-test, the null hypothesis states that there is no difference between the means of our two groups. The alternative hypothesis states the contrary claim: there is a difference between the means of our two groups.
In this case our hypotheses are:

- H_0: There is no difference in the mean usage time between weekdays and weekends
- H_A: There is a difference in the mean usage time between weekdays and weekends

Our significance level for this test is the standard 5% or 0.05.

In [ ]:

```python
# hypothesis test
stats.ttest_ind(a = weekdays['usage_minutes'], b = weekends['usage_minutes'], equal_var = False)
```

Out[ ]:

```
Ttest_indResult(statistic=0.5852580741708376, pvalue=0.563075537220598)
```

Based on our sample data, the difference between the mean laptop usage weekends and weekdays is 15.44 minutes.
Our p-value is 0.56 or around 56%. It means that there is 56% probability of observing a difference in mean usage time as extreme as or more extreme than the observed difference of 15.44 minutes, if the null hypothesis is true. In other words, it’s likely that the difference in the two means is due to chance.

Result

Since our p-values is greater than the significance level (0.56 > 0.05), we fail to reject the null hypothesis. We conclude that there is not a statistically significant difference between the mean laptop usage time during weekdays and weekends.