

**Analysis of the impact of social media usage on the number of hours of sleep**

**Statistical Analysis**

2021/2022

1st Semester

André Antunes Oliveira

**m20211253**

**Professor**

Ana Cristina Costa

**Index**

[1. Introduction 3](#_Toc89431304)

[2. Methodology 3](#_Toc89431305)

[3. Results 5](#_Toc89431306)

[3.1 Exploratory data analysis 5](#_Toc89431307)

[3.2 Distribution fitting tests 7](#_Toc89431308)

[3.3 Tests for equality of variances 9](#_Toc89431309)

[3.4 Analysis of variance (ANOVA) 10](#_Toc89431310)

[3.5 Multiple comparison tests 10](#_Toc89431311)

[4. Conclusion 11](#_Toc89431312)

[References 11](#_Toc89431313)

[Appendices 12](#_Toc89431314)

[Appendix A – Outputs from Python code 12](#_Toc89431315)

[Appendix B –Python code 15](#_Toc89431316)

**Index of Figures**

[Figure 1 – Right-sided test. 4](#_Toc89431320)

[Figure 2 - Histogram of each sample by social media usage intensity. 6](#_Toc89431321)

[Figure 3 - Violin plots of each sample by social media usage intensity. 6](#_Toc89431322)

[Figure 4 - Normal Q-Q plots for each social media usage intensity. 9](#_Toc89431323)

**Index of Tables**

[Table 1 - Descriptive statistics of the dataset. 5](#_Toc89431326)

[Table 2 - Confidence intervals for the parameters of the populations. 7](#_Toc89431327)

[Table 3 - Shapiro-Wilk test results provided by Python. [5] 8](#_Toc89431328)

[Table 4 - Levene's test results provided by Python. [6] 9](#_Toc89431329)

[Table 5 - ANOVA test results provided by Python. [7] 10](#_Toc89431330)

[Table 6 - Tukey's HSD test results provided by Python. [8] [7] 10](#_Toc89431331)

# **1. Introduction**

The main objective of this project is to study if the intensity of social media usage influences the number of hours of sleep. For this purpose, 4 distinct groups of 20 people were requested to characterize their social media usage (low, moderate, high, very high) and report their average number of hours of sleep.

Therefore, the methodological approach is based on testing the equality of the 4 populations means and, in case they differ, evaluate their difference.

# **2. Methodology**

As previously referred, the main objective of this work is to test the equality of the 4 populations means to understand if the intensity of social media usage influences the average number of hours of sleep. To achieve this goal, it is required to evaluate several aspects of the different populations and corresponding samples. The populations under study are the following:

* – Average number of hours of sleep, for people reporting low social media usage.
* – Average number of hours of sleep, for people reporting moderate social media usage.
* – Average number of hours of sleep, for people reporting high social media usage.
* – Average number of hours of sleep, for people reporting very high social media usage.

A sample of 20 observations was collected for each social media usage level. It is assumed that the 4 samples are independent of each other. In addition, it is considered a significance level of 5% for all the tests conducted, thus . The corresponding statistical tables were the ground truth to get the critical values of the tests performed. The analysis was developed using Python – one of the top programming languages for the purpose (code available in Appendix A).

Initially, the analysis was based on a preliminary assessment of the descriptive statistics of the dataset, allowing to evaluate several aspects such as the sample mean, sample variance, confidence intervals, among others.

In order to use ANOVA to test the equality of the 4 populations means, it is fundamental to guarantee 3 requirements [1]:

* The samples and the observations used are independent.
* The samples are originated from normal populations.
* The variance of the 4 populations is the same (homoscedasticity).

In case any of the assumptions above is not verified, it will require the usage of another test.

Regarding the first ANOVA requirement, as previously referred, it is assumed that the samples are independent of each other and that the observations are also independent.

In order to test if the samples come from normal populations and check if the second ANOVA assumption is satisfied, a distribution fitting test needs to be performed. For this purpose, the Shapiro-Wilk test was the selected method, as the populations parameters are unknown. Therefore, the test is based on the following hypotheses [2]:

* : the sample comes from a normal population with µ and σ unknown.
* : the sample does not come from a normal population.

The Shapiro-Wilk test will be performed for each sample (out of the 4 samples available) and the final decision should consider that must be rejected if , where is the observed value of the test statistic and is the critical value of the test. [2]

Regarding the third ANOVA requirement, the Levene's test (centered at the mean of each group) was performed in order to check the homoscedasticity of the different populations. The test is centered at the mean as the normality assumption has been proven (as shown in the next section), otherwise it would be beneficial to perform the Levene’s test using the median, which is less sensitive to variations. This test is based on the following hypotheses [3]:

The decision based on the Levene's test must consider that should be rejected if , as it is a right-sided test (shown in Figure 1). Therefore, if falls in the rejection region, must be rejected. [3]

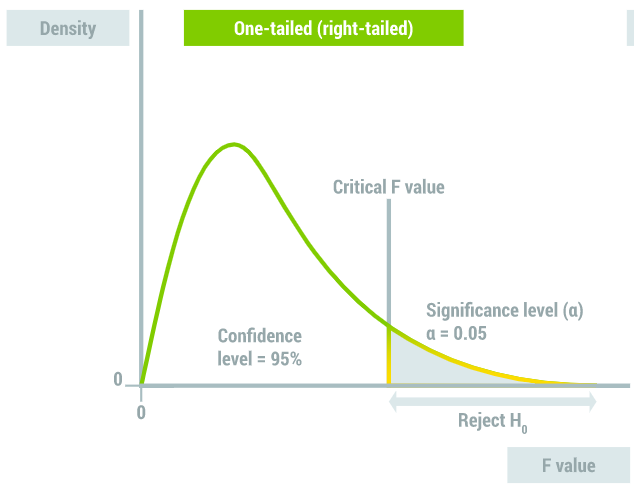


Figure 1 – Right-sided test.

Finally, after verifying every assumption of the One-way ANOVA with fixed effects, it is possible to perform the test. Considering the dataset provided and the study goal, the factor of the one-way ANOVA is the social media usage intensity, encompassing 4 levels (low, moderate, high, very high), and the experimental unit is the average number of hours of sleep. The test is based on the following hypotheses [1]:

The decision based on the test considers that should be rejected if , as it is a right-sided test (shown in Figure 1), likewise the Levene’s test. As a result, if falls in the rejection region, must be rejected. [1]

After the One-way ANOVA test, the results will show that not all the populations have the same mean, but it is inconclusive about which means are unequal. Therefore, it requires performing a multiple comparison test. For this purpose, the Tukey's HSD test is the most appropriate, as the sample size is the same for all social media usage intensities (), allowing a deep understanding of which pair/pairs of populations have different means. This test is based on the following hypotheses [4]:

The decision based on the Tukey's HSD test considers that should be rejected if . [4]

# **3. Results**

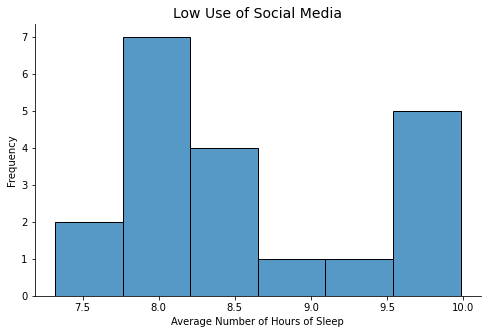
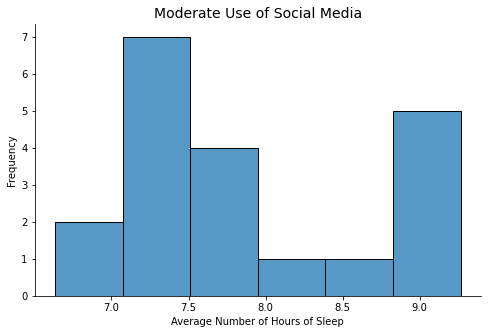
## **3.1 Exploratory data analysis**

As previously referred, the dataset used for this study encompasses 4 samples of 4 different populations. When analyzing the descriptive statistics of the dataset, presented in Table 1, it is possible to verify that the samples have the same size (), as well as the mean and median values differ among samples. In addition, it is clear that the standard deviations of the samples with low and high usage of social media are equal, as so it happens in the case of samples with moderate and very high social media exposure. Considering the extreme values, the sample from low usage of social media presents the highest records of average hours of sleep, while the sample from very high social media intensity has the lowest average sleep time observed.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Low Use | Moderate Use | High Use | Very High Use |
| N | 20 | 20 | 20 | 20 |
| mean | 8.586 | 7.885 | 6.986 | 6.085 |
| Median | 8.344 | 7.646 | 6.744 | 5.846 |
| standard deviation | 0.826 | 0.815 | 0.826 | 0.815 |
| VARIANCE | 0.682 | 0.665 | 0.682 | 0.665 |
| minimum | 7.319 | 6.633 | 5.719 | 4.833 |
| maximum | 9.984 | 9.265 | 8.384 | 7.465 |

Table 1 - Descriptive statistics of the dataset.

The histograms of each sample are represented in Figure 2, allowing a deeper understanding of the dataset. It is possible to verify that the data doesn’t seem far from a normal distribution, even though there are some spikes for higher values of average hours of sleep.

** **

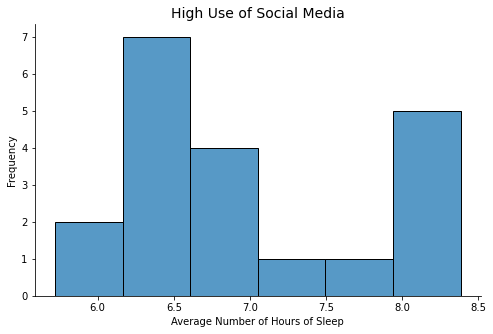
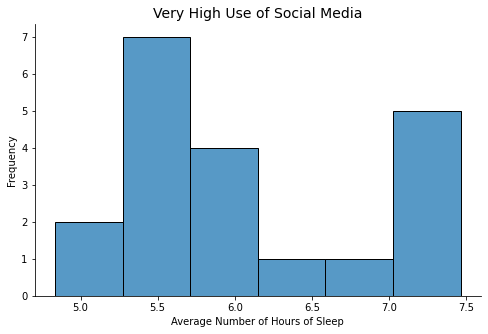
** **

Figure 2 - Histogram of each sample by social media usage intensity.

The previous fact is visible on the violin plots shown in Figure 3, where the density curves and box plots are represented. In addition, it is possible to confirm that the medians of the samples are different and decreasing with the increase of social media usage. The box plots also confirm that the data is not entirely symmetrical, given the bigger gap between the median and the third quartiles. Finally, it is evident that there are no outliers in the data, otherwise there would be points falling outside the plot.

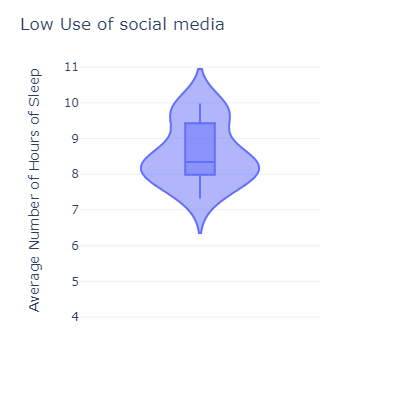
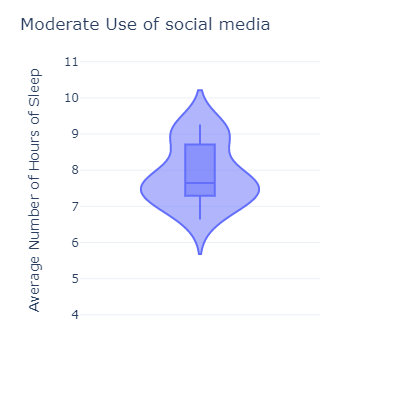
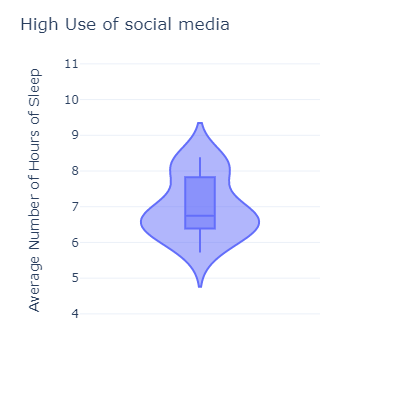
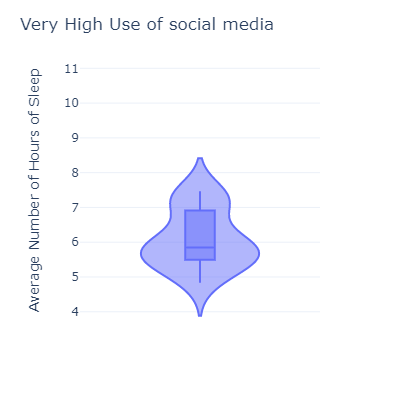
   

Figure 3 - Violin plots of each sample by social media usage intensity.

The last step of the exploratory data analysis was calculating the confidence intervals for the mean, standard deviation and variance of the populations. It was assumed the test statistic for normal populations of size and unknown, as the results of the Shapiro-Wilk test were already available.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| PARAMETER | | statistic | | 95% CONFIDENCE INTERVAL | | |  |
| LOW USE | **MEAN** | 8.586 | 8.199 | | 8.972 |
| **STANDARD DEVIATION** | 0.826 | 0.418 | | 1.191 |
| **VARIANCE** | 0.682 | 0.295 | | 1.068 |
| MODERATE USE | **MEAN** | 7.885 | 7.503 | | 8.266 |
| **STANDARD DEVIATION** | 0.815 | 0.413 | | 1.176 |
| **VARIANCE** | 0.665 | 0.283 | | 1.046 |
| High Use | **MEAN** | 6.986 | 6.599 | | 7.372 |
| **STANDARD DEVIATION** | 0.826 | 0.418 | | 1.191 |
| **VARIANCE** | 0.682 | 0.295 | | 1.068 |
| Very High Use | **MEAN** | 6.085 | 5.703 | | 6.466 |
| **STANDARD DEVIATION** | 0.815 | 0.413 | | 1.176 |
| **VARIANCE** | 0.665 | 0.283 | | 1.046 |

Table 2 - Confidence intervals for the parameters of the populations.

When analyzing the confidence intervals of the populations’ parameters presented in Table 2, it is possible to conclude:

* **Low use of social media** – it can be said with 95% confidence that the population mean is between 8.199 and 8.972. In addition, its standard deviation is, with 95% confidence, between 0.418 and 1.191. As a result, there is 95% confidence that the variance is somewhere between 0.295 and 1.068.
* **Moderate use of social media** – it can be said with 95% confidence that the population mean is between 7.503 and 8.266. In addition, its standard deviation is, with 95% confidence, between 0.413 and 1.176. As a result, there is 95% confidence that the variance is somewhere between 0.283 and 1.046.
* **High use of social media** – it can be said with 95% confidence that the population mean is between 6.599 and 7.372. In addition, its standard deviation is, with 95% confidence, between 0.418 and 1.191. As a result, there is 95% confidence that the variance is somewhere between 0.295 and 1.068.
* **Very high use of social media** – it can be said with 95% confidence that the population mean is between 5.703 and 6.466. In addition, its standard deviation is, with 95% confidence, between 0.413 and 1.176. As a result, there is 95% confidence that the variance is somewhere between 0.283 and 1.046.

## **3.2 Distribution fitting tests**

As previously mentioned in section 2, the Shapiro-Wilk test was performed to check if the samples come from populations with normal distribution. The test output provided by Python is presented below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | STATISTIC | P-VALUE | CRITICAL VALUE |
| Low Use | 0.911 | 0.066 | 0.905 |
| Moderate Use | 0.911 | 0.066 | 0.905 |
| High Use | 0.911 | 0.066 | 0.905 |
| Very High Use | 0.911 | 0.066 | 0.905 |

Table 3 - Shapiro-Wilk test results provided by Python. [5]

When analyzing the Shapiro-Wilk test results presented in Table 3, it is possible to take the following conclusions:

* **Low use of social media**:
  + as
  + - as
  + Therefore, should not be rejected for . As a result, there is evidence that the sample data of low usage of social media comes from a normal distribution.
* **Moderate use of social media**:
  + as
  + - as
  + Therefore, should not be rejected for . As a result, there is evidence that the sample data of moderate usage of social media comes from a normal distribution.
* **High use of social media**:
  + as
  + - as
  + Therefore, should not be rejected for . As a result, there is evidence that the sample data of high usage of social media comes from a normal distribution.
* **Very high use of social media**:
  + as
  + - as
  + Therefore, should not be rejected for . As a result, there is evidence that the sample data of very high usage of social media comes from a normal distribution.

The results stated above can be confirmed by observing the Q-Q (quantile-quantile) plots presented in Figure 4. When comparing the distribution of data against the normal distribution, namely the existing quantiles versus the normal theoretical quantiles, it is visible that the fit is not perfect[[1]](#footnote-2) for all values but close to it.

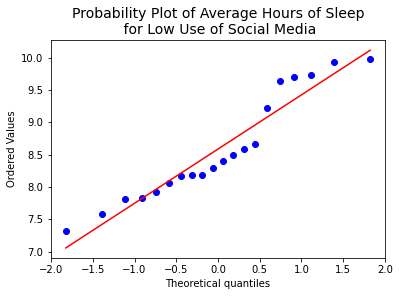
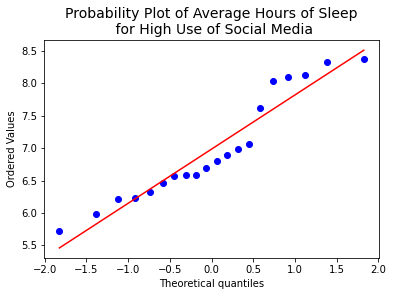
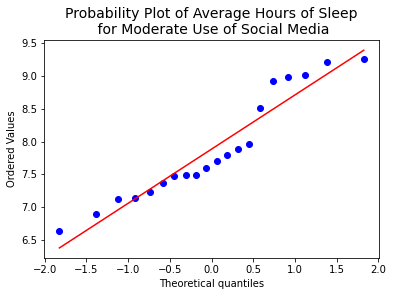
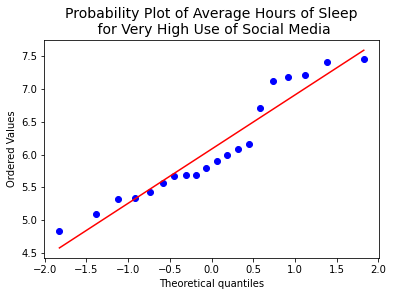
  

Figure 4 - Normal Q-Q plots for each social media usage intensity.

## **3.3 Tests for equality of variances**

As previously mentioned in section 2, after verifying the normality assumption of ANOVA, the next step is checking the equality of variances with the Levene’s test:

|  |  |  |  |
| --- | --- | --- | --- |
|  | STATISTIC | P-VALUE | CRITICAL VALUE |
| Levene's test | 0.0024 | 0.9998 | 2.7249 |

Table 4 - Levene's test results provided by Python. [6]

When analyzing the Levene’s test results presented in Table 4, it is possible to conclude:

* as
* - as
* Hence, falls in the acceptance region and should not be rejected for . As a result, there is evidence that all populations have the same variance (homoscedastic).

## **3.4 Analysis of variance (ANOVA)**

After verifying all assumptions of the One-way ANOVA with fixed effects, there right conditions are met to proceed with the test.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | DF | SUM OF SQUARES | mEAN SQUARES |  | P-VALUE |  |
| mODEL | 3 | 70.83427 | 23.61142 | 35.067 | 2.474×10-14 | 2.7249 |
| eRROR | 76 | 51.1731 | 0.67333 |  |  |  |
| tOTAL | 79 | 122.0074 |  |  |  |  |

Table 5 - ANOVA test results provided by Python. [7]

When analyzing the ANOVA F-test results presented in Table 5, it is possible to draw the following conclusions:

* as
* - as
* Hence, falls in the rejection region and should be rejected for . As a result, there is evidence that at least one of the populations means differs from the others.

## **3.5 Multiple comparison tests**

Considering the results of the ANOVA test, it is required to understand which populations means are different by performing the Tukey’s HSD test.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | MEAN DIFF | STANDARD ERROR | STATISTIC | P-value[[2]](#footnote-5) | CONFIDENCE INTERVAL | | REJECT |
| **LOWER** | **UPPER** |
| Low Use-High Use | 1.60000 | 0.259486 | -6.166034 | 0 | 1.083189 | 2.116811 | True |
| Moderate Use-High Use | 0.89893 | 0.259486 | -3.464270 | 0.004768 | 0.382118 | 1.415741 | True |
| Very High Use-High Use | -0.90107 | 0.259486 | 3.472518 | 0.004647 | -1.417882 | -0.384259 | True |
| Moderate Use-Low Use | -0.70107 | 0.259486 | 2.701764 | 0.041463 | -1.217882 | -0.184259 | True |
| Very High Use-Low Use | -2.50107 | 0.259486 | 9.638552 | 0 | -3.017882 | -1.984259 | True |
| Very High Use-Moderate Use | -1.80000 | 0.259486 | 6.936788 | 0 | -2.316811 | -1.283189 | True |

Table 6 - Tukey's HSD test results provided by Python. [8] [7]

When analyzing the Tukey’s HSD test results presented in Table 6, it is possible to take the following conclusions:

* with
* - for every pair of population means.
* There is evidence that should be rejected, considering , for every pair of population means.
* Thus, there is a statistically significant difference between the means of every population of social usage intensity.

# **4. Conclusion**

Finally, after completing all statistical tests, it can be concluded that there is clear evidence that the intensity of social media usage influences the number of hours of sleep. In addition, it can be referred that the average sleep time tends to decrease with the increase of social media usage intensity – meaning that people who spend more time on Facebook/Instagram/Twitter will eventually sleep less. This is corroborated by the fact that higher social media usage levels originated lower mean values of average hours of sleep, and vice-versa.

From a physical and biological perspective, it can be reasoned that people are prioritizing their desire to be connected to each other and the world, over their basic human needs. As each day is composed by 24 hours, it seems that people are not managing their time effectively to achieve a balance between rest and social engagement. Nowadays, it looks like people are more prone to focus on their emotional needs rather than their physical ones.

# **References**

|  |  |
| --- | --- |
| [1] | "Freie Universität Berlin," One-way ANOVA Hypothesis Test. [Online]. Available: https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/ANOVA/One-way-ANOVA-Hypothesis-Test/index.html. [Accessed 28 November 2021]. |
| [2] | Z. Hanusz, J. Tarasinska and W. Zielinski, "Shapiro–Wilk Test with Known Mean," *REVSTAT – Statistical Journal,* vol. 14, no. 1, p. 89–100, February 2016. |
| [3] | "Levene Test for Equality of Variances," NIST/SEMATECH e-Handbook of Statistical Methods, [Online]. Available: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm. [Accessed 29 November 2021]. |
| [4] | "Tukey’s HSD," University of Dayton, [Online]. Available: https://academic.udayton.edu/gregelvers/psy217/labs2014/TukeyMC.pdf. [Accessed 30 November 2021]. |
| [5] | "How to Perform a Shapiro-Wilk Test in Python," 30 November 2021. [Online]. Available: https://www.statology.org/shapiro-wilk-test-python/. |
| [6] | "How to Perform Levene’s Test in Python," 30 November 2021. [Online]. Available: https://www.statology.org/levenes-test-python/. |
| [7] | E. Marsja, "Four Ways to Conduct One-Way ANOVA with Python," [Online]. Available: https://www.marsja.se/four-ways-to-conduct-one-way-anovas-using-python/. [Accessed 2 Dezembro 2021]. |
| [8] | "How to Perform Tukey’s Test in Python," 30 November 2021. [Online]. Available: https://www.statology.org/tukey-test-python/. |
| [9] | "psturng," Kite, [Online]. Available: https://www.kite.com/python/docs/statsmodels.stats.libqsturng.psturng. [Accessed 2 December 2021]. |

# **Appendices**

## **Appendix A – Outputs from Python code**

**Dataset Info**

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 20 entries, 0 to 19

Data columns (total 4 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Low Use 20 non-null float64

1 Moderate Use 20 non-null float64

2 High Use 20 non-null float64

3 Very High Use 20 non-null float64

dtypes: float64(4)

memory usage: 768.0 bytes

**Dataset Description**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Low Use | Moderate Use | High Use | Very High Use |
| count | 20 | 20 | 20 | 20 |
| mean | 8,585617 | 7,884546 | 6,985617 | 6,084546 |
| std | 0,825712 | 0,81539 | 0,825712 | 0,81539 |
| min | 7,318528 | 6,633297 | 5,718528 | 4,833297 |
| 25% | 8,021216 | 7,3272 | 6,421216 | 5,5272 |
| 50% | 8,344326 | 7,646272 | 6,744326 | 5,846272 |
| 75% | 9,319812 | 8,609565 | 7,719812 | 6,809565 |
| max | 9,983507 | 9,264963 | 8,383507 | 7,464963 |

**Median Values**

Median of Low Use: 8.344

Median of Moderate Use: 7.646

Median of High Use: 6.744

Median of Very High Use: 5.846

**Variance Values**

Variance of Low Use: 0.682

Variance of Moderate Use: 0.665

Variance of High Use: 0.682

Variance of Very High Use: 0.665

**95% Confidence Intervals**

--- Population Mean ---

Low Use

95 percent confidence interval: (8.199 , 8.972)

Moderate Use

95 percent confidence interval: (7.503 , 8.266)

High Use

95 percent confidence interval: (6.599 , 7.372)

Very High Use

95 percent confidence interval: (5.703 , 6.466)

--- Population Standard Deviation ---

Low Use

95 percent confidence interval: (0.418 , 1.191)

Moderate Use

95 percent confidence interval: (0.413 , 1.176)

High Use

95 percent confidence interval: (0.418 , 1.191)

Very High Use

95 percent confidence interval: (0.413 , 1.176)

--- Population Variance ---

Low Use

95 percent confidence interval: (0.295 , 1.068)

Moderate Use

95 percent confidence interval: (0.283 , 1.046)

High Use

95 percent confidence interval: (0.295 , 1.068)

Very High Use

95 percent confidence interval: (0.283 , 1.046)

**Shapiro-Wilk Test**

Low Use

stat=0.911, p=0.066

Shapiro-Wilk Test

The sample comes from a normal population with µ and σ unknown.

Moderate Use

stat=0.911, p=0.066

Shapiro-Wilk Test

The sample comes from a normal population with µ and σ unknown.

High Use

stat=0.911, p=0.066

Shapiro-Wilk Test

The sample comes from a normal population with µ and σ unknown.

Very High Use

stat=0.911, p=0.066

Shapiro-Wilk Test

The sample comes from a normal population with µ and σ unknown.

**Levene's test**

stat=0.0024, p=0.9998

Levene's Test centered at the mean

The variances are equal across all samples/groups.

**ANOVA**

sum\_sq df F PR(>F) EtaSq mean\_sq

group 70.834270 3.0 35.066627 2.474053e-14 0.580574 23.611423

Residual 51.173104 76.0 NaN NaN NaN 0.67333

Total 122.007374 79.0 NaN NaN NaN NaN

**Tukey's HSD test**

***Statsmodels***

summary: Multiple Comparison of Means - Tukey HSD, FWER=0.05

=================================================================

group1 group2 meandiff p-adj lower upper reject

-----------------------------------------------------------------

High Use Low Use 1.6 0.001 0.9184 2.2816 True

High Use Moderate Use 0.8989 0.0048 0.2173 1.5806 True

High Use Very High Use -0.9011 0.0046 -1.5827 -0.2194 True

Low Use Moderate Use -0.7011 0.0415 -1.3827 -0.0194 True

Low Use Very High Use -2.5011 0.001 -3.1827 -1.8194 True

Moderate Use Very High Use -1.8 0.001 -2.4816 -1.1184 True

-----------------------------------------------------------------

mean diffs: [ 1.6 0.89892979 -0.90107021 -0.70107021 -2.50107021 -1.8 ]

std pairs: [0.18348438 0.18348438 0.18348438 0.18348438 0.18348438 0.18348438]

groups unique: ['High Use' 'Low Use' 'Moderate Use' 'Very High Use']

df total: 76

Unadjusted p values: [0.001 0.00476841 0.00464727 0.04146347 0.001 0.001 ]

***Pingouin***

A B mean(A) mean(B) diff se \

0 High Use Low Use 6.985617 8.585617 -1.60000 0.259486

1 High Use Moderate Use 6.985617 7.884546 -0.89893 0.259486

2 High Use Very High Use 6.985617 6.084546 0.90107 0.259486

3 Low Use Moderate Use 8.585617 7.884546 0.70107 0.259486

4 Low Use Very High Use 8.585617 6.084546 2.50107 0.259486

5 Moderate Use Very High Use 7.884546 6.084546 1.80000 0.259486

T p-tukey hedges

0 -6.166034 0.001000 -1.911132

1 -3.464270 0.004768 -1.073733

2 3.472518 0.004647 1.076290

3 2.701764 0.041463 0.837399

4 9.638552 0.001000 2.987422

5 6.936788 0.001000 2.150023

**Critical Value of Studentized Range**

Critical Value of Studentized Range: 0.9

## **Appendix B –Python code**

#!/usr/bin/env python

# coding: utf-8

# # Analysis of the impact of social media usage on the number of hours of sleep

# A study pretends to analyze if the intensity of social media usage influences the number of hours of sleep. With this purpose, four distinct groups were selected, each characterizing a level of intensity of social media usage: Low usage, moderate usage, high usage, and very high usage. Each one of these groups is composed of a sample of 20 people who were firstly asked how they would characterize their social media usage (between the four options available) and later asked their average number of hours of sleep.

# \_\_\_\_\_\_\_\_\_\_\_\_

# In[72]:

# Setup

import pandas as pd

import numpy as np

import statistics

import seaborn as sns

import matplotlib.pyplot as plt

from matplotlib import ticker

import seaborn as sns

import plotly.express as px

import pylab

import scipy.stats as st

from scipy.stats import shapiro

from scipy.stats import levene

import statsmodels.api as sm

from statsmodels.formula.api import ols

from statsmodels.stats.libqsturng import psturng

from statsmodels.stats.multicomp import pairwise\_tukeyhsd

import pingouin as pg

# In[3]:

# Plot settings

subPlots\_Title\_fontSize = 12

subPlots\_xAxis\_fontSize = 10

subPlots\_yAxis\_fontSize = 10

subPlots\_label\_fontSize = 10

plots\_Title\_fontSize = 14

plots\_Title\_textColour = 'black'

plots\_Legend\_fontSize = 12

plots\_Legend\_textColour = 'black'

# In[4]:

# Import data

df = pd.read\_excel('Series51.xlsx')

# In[5]:

df.head()

# In[6]:

# Get dataset info

df.info()

# ## Exploratory Data Analysis

# In[7]:

# Get dataset statistics

df.describe()

# In[8]:

# Calculate median

for i in df.columns:

    print(f"Median of {i}: %.3f " % (statistics.median(df[i])))

# In[9]:

# Calculate variance

for i in df.columns:

    print(f"Variance of {i}: %.3f " % (statistics.variance(df[i])))

# In[10]:

# Plot histograms

def plot\_histogram(df,col):

    # Draw

    fig, ax = plt.subplots(figsize=(8,5))

    g = sns.histplot(df[col], kde=False)

    # Decoration

    fmt = "{x:,.0f}"

    tick = ticker.StrMethodFormatter(fmt)

    ax.yaxis.set\_major\_formatter(tick)

    sns.despine()

    plt.title(col +' of Social Media', fontsize=plots\_Title\_fontSize)

    plt.xlabel('Average Number of Hours of Sleep')

    plt.ylabel("Frequency")

    plt.rc('axes', labelsize=subPlots\_label\_fontSize)

# In[11]:

for i in df.columns:

    plot\_histogram(df,i)

# In[12]:

# Violin plot

def violin\_plot(ds,col,width,height):

    fig = px.violin(ds, y=col, box=True, points= False)

    fig.update\_layout(height=height, width=width, title\_text=i + ' of social media', template = "plotly\_white")

    fig.update\_yaxes(title\_text='Average Number of Hours of Sleep')

    fig.show()

# In[13]:

for i in df.columns:

    violin\_plot(df,i,400,400)

# In[14]:

## Create 95% confidence intervals using the Normal Distribution

## Performed after the Shapiro-Wilk Normality Test

alpha=0.95

# Population mean

print('--- Population Mean ---')

for i in df.columns:

    c1,c2 = st.t.interval(alpha=alpha, df=len(df)-1, loc=np.mean(df[i]), scale=st.sem(df[i]))

    print(i)

    print(f"95 percent confidence interval: (%.3f , %.3f)\n" % (c1,c2))

# Population standard deviation

print('--- Population Standard Deviation ---')

for i in df.columns:

    c1,c2 = st.t.interval(alpha=alpha, df=len(df)-1, loc=np.std(df[i]), scale=st.sem(df[i]))

    print(i)

    print(f"95 percent confidence interval: (%.3f , %.3f)\n" % (c1,c2))

# Population variance

print('--- Population Variance ---')

for i in df.columns:

    c1,c2 = st.t.interval(alpha=alpha, df=len(df)-1, loc=statistics.variance(df[i]), scale=st.sem(df[i]))

    print(i)

    print(f"95 percent confidence interval: (%.3f , %.3f)\n" % (c1,c2))

# ## Testing

# #### Normality

# In[15]:

# Shapiro-Wilk Normality Test

def normality\_test(data):

    '''H0: the sample comes from a normal population with µ and σ unknown.

    H1: the sample does not come from a normal population.'''

    stat, p = shapiro(data)

    print('stat=%.3f, p=%.3f' % (stat, p))

    print('Shapiro-Wilk Test')

    if p > 0.05:

        print('The sample comes from a normal population with µ and σ unknown.')

    else:

        print('The sample does not come from a normal population.')

    print('\n')

# In[16]:

for i in df.columns:

    print(i)

    normality\_test(df[i])

# In[17]:

# Q-Q plot

for i in df.columns:

    st.probplot(df[i], dist="norm", plot=pylab)

    plt.title('Probability Plot of Average Hours of Sleep\n for '+i+' of Social Media', fontsize=plots\_Title\_fontSize)

    pylab.show()

# #### Homoscedasticity

# In[18]:

#Levene's test centered at the mean

def variance\_test(df):

    '''H0: the variances are equal across all samples/groups.

    H1: the variances are not equal across all samples/groups.

    '''

    stat, p =  levene(df.iloc[:, 0], df.iloc[:, 1], df.iloc[:, 2], df.iloc[:, 3] , center='mean')

    print('stat=%.4f, p=%.4f' % (stat, p))

    print("Levene's Test centered at the mean")

    if p > 0.05:

        print('The variances are equal across all samples/groups.')

    else:

        print("The variances are not equal across all samples/groups.")

# In[19]:

variance\_test(df)

# #### ANOVA

# In[24]:

## Analysis of Variance Test

# Store values of each sample

vals = []

for i in range(0,len(df.columns)):

    col\_vals = df.iloc[:, i].tolist()

    vals = vals + col\_vals

data = pd.DataFrame({'weight': vals,

                   'group': np.repeat(df.columns.to\_list(), repeats=len(df))})

mod = ols('weight ~ group',

                data=data).fit()

aov\_table = sm.stats.anova\_lm(mod, typ=2)

# Effect sizes

esq\_sm = aov\_table['sum\_sq'][0]/(aov\_table['sum\_sq'][0]+aov\_table['sum\_sq'][1])

aov\_table['EtaSq'] = [esq\_sm, 'NaN']

# Totals

aov\_table.loc['Total']= aov\_table.sum(numeric\_only=True, axis=0)

aov\_table.at['Total', 'F'] = None

aov\_table.at['Total', 'PR(>F)'] = None

# Mean Square

mean\_sqr\_0 = aov\_table['sum\_sq'][0]/aov\_table['df'][0]

mean\_sqr\_1 = aov\_table['sum\_sq'][1]/aov\_table['df'][1]

aov\_table['mean\_sq'] = [mean\_sqr\_0, mean\_sqr\_1,'NaN']

print(aov\_table)

# #### Multiple comparison test

# In[71]:

# Store values of each sample

vals = []

for i in range(0,len(df.columns)):

    col\_vals = df.iloc[:, i].tolist()

    vals = vals + col\_vals

#create DataFrame to hold data

df\_tukey = pd.DataFrame({'score': vals,

                   'group': np.repeat(df.columns.to\_list(), repeats=len(df))})

# perform Tukey's test

res2 = pairwise\_tukeyhsd(endog=df\_tukey['score'],

                          groups=df\_tukey['group'],

                          alpha=0.05)

print("summary:", res2.summary())

print("mean diffs:", res2.meandiffs)

print("std pairs:",res2.std\_pairs)

print("groups unique: ", res2.groupsunique)

print("df total:", res2.df\_total)

p\_values = psturng(np.abs(res2.meandiffs / res2.std\_pairs), len(res2.groupsunique), res2.df\_total)

print()

print("Unadjusted p values:", p\_values)

# In[74]:

# perform Tukey's test using pingouin to get statistics

pt = pg.pairwise\_tukey(dv='weight', between='group', data=data)

pt

# In[84]:

# Calculate Critical Value of Studentized Range

print('Critical Value of Studentized Range: ', psturng(0.5, len(res2.groupsunique), res2.df\_total))

1. For a perfect normal distribution, the observations should all occur on the 45-degree straight line of the Q-Q plot. [↑](#footnote-ref-2)
2. Statsmodels function for the p-value has a lower bound of 0.001. As a result, the p-values obtained as 0.001 were considered to be approximately 0. [9] [↑](#footnote-ref-5)