Statistical Analysis of Lifestyle Factors Affecting Sleep Parameters

Ohad Zilber & Naor Yichye, Bar-Ilan University, Department of Mathematics Ramat-Gan, Israel

Abstract:

This study investigates the impact of lifestyle factors - alcohol consumption, caffeine intake, and physical activity levels on sleep efficiency and REM sleep percentage. Using a dataset from kaggle, we analyzed various sleep metrics to understand how these factors influence sleep quality. The data was collected as part of a study conducted in Morocco. Previous Research indicates that sleep quality and REM sleep, crucial for cognitive functions and overall health, are influenced by lifestyle factors such as alcohol consumption, and physical activity.

Using statistical test (Mann-Whitney), examination of distribution and other statistical tools we can indicate that increased alcohol consumption negatively affects sleep efficiency, while regular physical activity enhances it. However, the effects of these factors on REM sleep percentage were not statistically significant.

Introduction:

Background: Sleep is a crucial physiological process that affects numerous aspects of human health and well-being. Among the various stages of sleep, REM (Rapid Eye Movement) sleep plays a significant role in cognitive functions such as memory consolidation, emotional regulation, and overall mental health. The quality and efficiency of sleep, which refers to the proportion of time spent asleep relative to the total time spent in bed, are critical metrics in evaluating sleep health. Recent research has focused on various lifestyle factors that might impact sleep, including alcohol consumption, caffeine intake, and physical activity levels.

Literature Review: REM sleep is characterized by rapid eye movements, vivid dreams, and heightened brain activity. It typically occurs in cycles throughout the night and is essential for various cognitive processes. Studies have shown that REM sleep contributes significantly to memory consolidation and emotional processing[1]. Disruptions in REM sleep have been linked to cognitive impairments and mood disorders, highlighting its importance in maintaining overall mental health.

Impact of Alcohol on Sleep: Alcohol consumption has been shown to influence sleep architecture significantly. While alcohol may initially induce sleepiness and help individuals fall asleep faster, it often disrupts the continuity and quality of sleep. Research indicates that alcohol consumption leads to reduced REM sleep at the first half of the night[2]. Chronic alcohol use can exacerbate these effects, leading to sleep disturbances and negative health outcomes[3].

Effect of Caffeine on Sleep: Caffeine is a well-known stimulant found in various beverages such as coffee, tea, and energy drinks. Its consumption, particularly close to bedtime, can delay sleep onset and reduce overall sleep duration. Caffeine's effects on sleep efficiency are also

---

[1] https://itb.biologie.hu- ,Stickgold & Walker, 2005
berlin.de/~kempter/HippoJC/Articles/stickgold05.pdf

[2] /https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6707127 ,Roehrs & Roth, 2001

[3] https://onlinelibrary.wiley.com/doi/abs/10.1111/acer.12006 ,Ebrahim et al., 2013

notable, as it can lead to a reduction in sleep duration and efficiency[4]. The half-life of caffeine varies among individuals, but its impact on sleep remains a significant concern for those with sleep disturbances.

Physical Activity and Sleep Efficiency: Physical activity is often associated with improved sleep quality and efficiency. Regular exercise has been shown to enhance the ability to fall asleep and increase total sleep time[5]. Exercise may also positively influence REM sleep by promoting deeper sleep stages and reducing sleep latency[6]. Conversely, a sedentary lifestyle or lack of physical activity can contribute to poor sleep efficiency and quality.

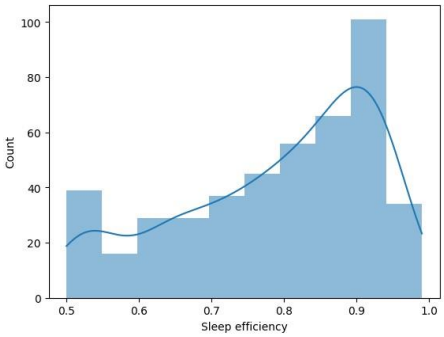
Research Question: This study aims to investigate whether REM sleep and sleep efficiency are affected by increased consumption of alcohol or caffeine or by low levels of physical activity. Understanding the impact of lifestyle factors such as alcohol consumption, caffeine intake, and physical activity on sleep is crucial for developing effective strategies to enhance sleep quality and overall health. By examining these relationships, we seek to contribute to the understanding of how these lifestyle factors influence sleep and to provide insights that may help improve sleep health.

Results:

Hypothesis tests:

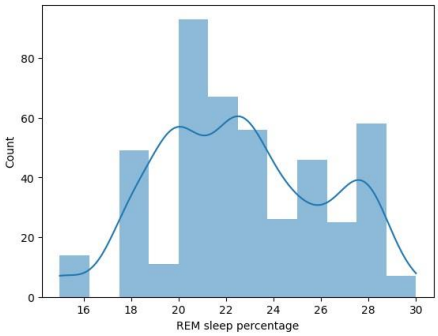
Data Collection and Preparation: The dataset used in this analysis was obtained from a CSV file named Sleep_Efficiency. The data consists of various metrics related to sleep efficiency, including Sleep efficiency, Light sleep percentage, Deep sleep percentage, and REM sleep percentage. The dataset was loaded into a Pandas Data Frame, with the ID column set as the index to uniquely identify each observation.

Data Visualization: We employed histograms to visualize the distributions of the main features. This helped in understanding the spread and shape of the data, identifying any skewness, and checking for potential outliers.



Normality Check: To determine whether the data followed a normal distribution, we defined a function to perform normality tests. These tests helped in identifying the appropriate statistical methods for further analysis.

Sleep Efficiency: The mean sleep efficiency was close to 1, indicating high efficiency on average. The distribution was bounded by 1, with the densest part near 1, suggesting a potential fit to a Weibull distribution.



Deep Sleep Percentage: The histogram indicated a potential inverse relationship with light sleep percentage, as the shapes appeared to be mirror reflections.

Coffee, caffeine, and sleep: A systematic review of epidemiological , studies (Clark & Landolt, 2017)[4]
and randomized controlled trials - PubMed (nih.gov)
Alnawwar MA[Author] - Search Results - PubMed (nih.gov) ,Majd A Alnawwar 2023 [5]

/https://pubmed.ncbi.nlm.nih.gov/20813580 ,Reid et al., 2010 [6]

REM Sleep Percentage: This feature's distribution was the closest to normality compared to the others.

Alcohol Consumption Influence:

Null Hypothesis (*1*): The **REM sleep percentage** is equal between those who consume a lot of alcohol (over 2 glasses daily), those who don't consume so much (not over 2 glasses daily) and those who don't consume alcohol at all.

$$H_0: \mu_{clean} = \mu_{regular} = \mu_{drunk}$$

$$H_1: \mu_{clean} \neq \mu_{regular} \vee \mu_{regular} \neq \mu_{drunk}$$

The confidence level we demand to reject H_0 is $\alpha = 0.05$.

To test this hypothesis, we want to use t-test. However, the normality check for the three groups of samples failed. Therefore, we chose to use Mann-Whitney test. We got the following results:

Statistics=4867.500, p=0.475 (drunk, regular)

Statistics=12160.500, p=0.218 (clean, regular)

Statistics=12225.500, p=0.815 (clean, drunk)

All the p-values are higher than 0.05, so unfortunately, we did not succeed to reject the null hypothesis, saying that the means are equal.

Here we moved to examine another hypothesis.

Null Hypothesis (*2*): The **sleep efficiency** is equal between those who consume a lot of alcohol (over 2 glasses daily), those who don't consume so much (not over 2 glasses daily) and those who don't consume alcohol at all.

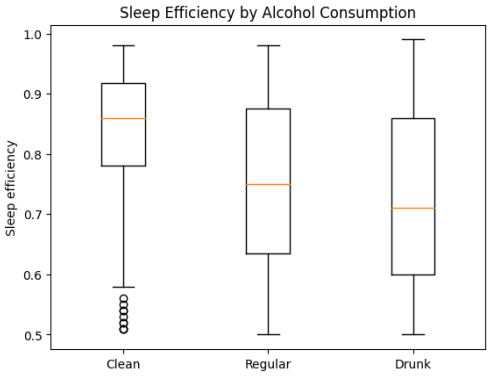
Again, the normality check for the three groups of samples failed. Therefore, we chose to use Mann-Whitney test. We got the following results:

Statistics=4022.000, p=0.136 (drunk, regular)

Statistics=14927.000, p=0.000 (clean, regular)

Statistics=6829.000, p=0.000 (clean, drunk)

We succeeded to reject the null hypothesis, which means the amount of alcohol consumption affects the sleep efficiency.



Caffeine Consumption Influence:

Null Hypothesis (*1*): The **REM sleep percentage** is equal between those who consume a lot of caffeine (over 50 mg daily), and those who don't consume so much (not over 50 mg).

$$H_0: \mu_{low} = \mu_{high}$$

$$H_1: \mu_{low} \neq \mu_{high}$$

To test this hypothesis, we want to use t-test. However, the normality check for one of the groups of samples failed. Therefore, we chose to use Mann-Whitney test. We got the following results:

Statistics=5575.500, p=0.759

The p-value is higher than 0.05, so unfortunately, we did not succeed to reject the null hypothesis, saying that the means are equal.

Null Hypothesis (*2*): The **sleep efficiency** is equal between those who consume a lot of caffeine, those who don't consume so much caffeine.

Again, the normality check for the two groups of samples failed. Therefore, we chose to use Mann-Whitney test. We got the following results:

Statistics=3728.000, p=0.006

We succeeded to reject the null hypothesis, saying that the means are equal, i.e. the amount of caffeine consumption affects the sleep efficiency.

Exercise Frequency Influence:

Null Hypothesis (*1*): The **REM sleep percentage** is equal between those who don't do a lot of exercise (0-1 times a week), and the more active group (2 times a week or more).

$$H_0: \mu_{low} = \mu_{high}$$

$$H_1: \mu_{low} \neq \mu_{high}$$

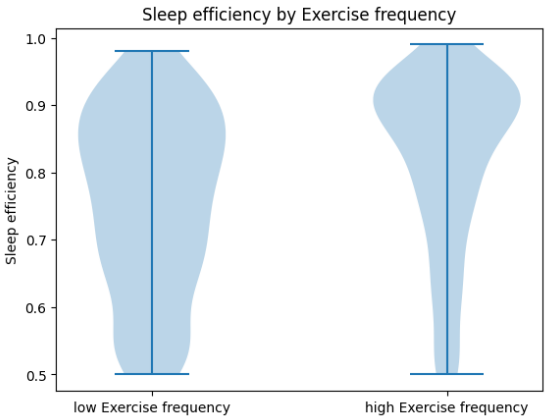
The two groups of samples not normally distributed so we used Mann-Whitney test. Again, we failed to reject the null hypothesis.

Null Hypothesis (*2*): The **sleep efficiency** is equal between those who don't do a lot of exercise, and the more active group.

We used Mann-Whitney test (the "high" samples are not normally distributed).

Statistics=14784.500, p=0.000

We succeeded to reject the null hypothesis, saying that the means are equal, i.e. the exercise frequency affects the sleep efficiency.



Type 2 Error: We checked the probability for Type 2 Error - the probability of failing to reject H_0 when H_0 is incorrect. Since our groups are not normally distributed, we had to do simulations for calculating the power and β (probability for type 2 error). We got the following results: Power=0.9986, β=0.0014.

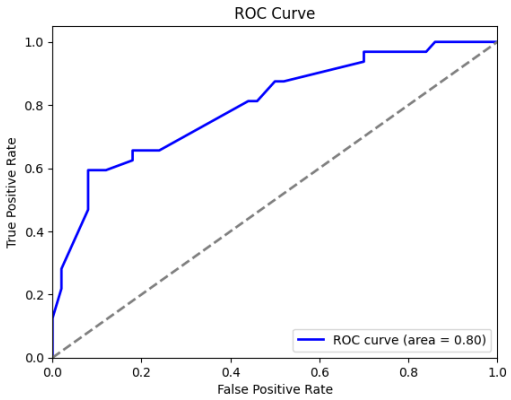
Logistic Regression:

We created logistic regression model, that predicts a person's sleep efficiency based on their exercise frequency, as well as caffeine and alcohol consumption. We got the following results:

Accuracy =0.7804

Confusion Matrix: $\begin{pmatrix} 45 & 5 \\ 13 & 19 \end{pmatrix}$

In 78% of the guesses the model was correct. The classification is not perfect, but this is a satisfying result knowing the wide distributions we saw above. In the ROC curve of our model we got Area Under the Curve (AUC) of 0.80, indicating a strong ability to distinguish between different levels of sleep efficiency based on the selected features.



Confidence Interval:

We did two types of confidence interval.

We computed the confidence interval for the "low" set - people who work out once a week at most. For this group the data is normally distributed. We got the following results:

Mean Sleep Efficiency = 0.76

Confidence interval: (0.74, 0.78)

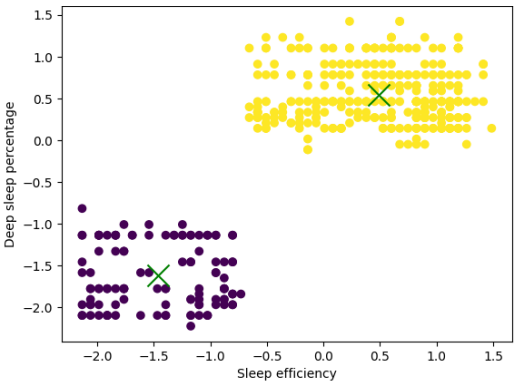
This interval gives us bounds telling us in Confidence of 95% what is the sleep efficiency of people who don't do a lot of exercise.

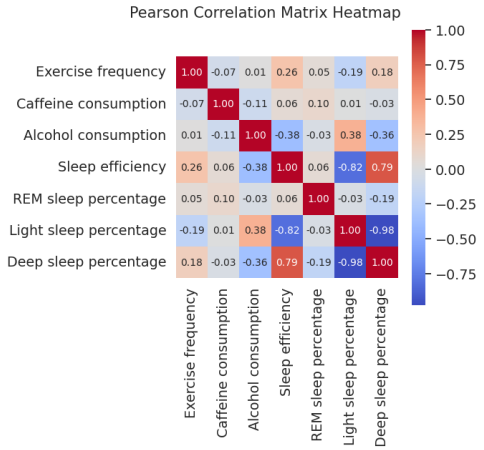
In addition, we computed confidence interval for the proportion - bounds telling us in Confidence of 95% how many people from a sample sleep well (sleep efficiency bigger than 0.86). We got the following results:

Proportion = 0.396

Confidence interval: (0.3480, 0.4430)

Pearson Correlation Matrix: In the correlation matrix (on the next page), we can see the connection between Sleep efficiency to the main features we examined (except caffeine consumption). This stands in contrast to the low correlations shown above between REM sleep percentage and those features. Another thing we can see here is the strong correlation between Sleep efficiency and Deep sleep percentage, which is also shown in the clustering graph below.





Methods:

Normality Check: Shapiro-Wilk and Kolmogorov-Smirnov Tests

To assess the normality of our data, we employed two statistical tests: the Shapiro-Wilk test (we used it when $n < 30$) and the Kolmogorov-Smirnov test ($n \geq 30$).

The Shapiro-Wilk test examines the null hypothesis that a sample $X = \{x_1, x_2, \dots, x_n\}$ is drawn from a normally distributed population. It is particularly powerful for small sample sizes, as it compares the order statistics of the sample to the expected order statistics of a normal distribution. A high p-value suggests that the null hypothesis cannot be rejected, indicating that the data may follow a normal distribution.

The Kolmogorov-Smirnov test is comparing the empirical distribution function (EDF) $F_n(x)$ of the sample data against a specified theoretical cumulative distribution function (CDF) $F(x)$, in this case, the normal distribution. The test statistic is defined by:

$$D_n = \sup_x |F_n(x) - F(x)|$$

A small D_n value indicates that the empirical distribution closely matches the theoretical distribution. A high p-value suggests that the null hypothesis cannot be rejected, indicating that the data may follow a normal distribution. We observed low p-values, leading us to opt for the non-parametric Mann-Whitney test, which does not require the assumption of normality.

Mann-Whitney Test

We used "Mann-Whitney" U test: a non-parametric test used to determine whether there is a significant difference between the distributions of two independent groups.

1. **Ranking**: Combine all data from both groups and rank them from smallest to largest, with ties receiving the average rank.
2. **Calculate U Statistic**: Compute the U statistic for each group using the formula:

$$U = R - \frac{n(n+1)}{2}$$

R is the sum of ranks for a group, and n is the number of observations in that group.

3. **Compare U**: Determine the smaller of the two U values (one for each group). This value is compared to a critical value from the Mann-Whitney U distribution.

Logistic Regression

Logistic regression is a statistical method used for classification problems, where the outcome variable is categorical with two possible outcomes. In our case, Sleep Efficiency above or below 0.86. The primary goal of logistic regression is to model the probability of the occurrence of one of these outcomes based on one or more predictor variables. Mathematically, logistic regression models $P(Y = 1|\ X)$ using the logistic function:

$$P(Y = 1|\ X) = \frac{1}{1 + e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_kX_k)}}$$

Here, β_0 is the intercept, $\beta_1, \beta_2 \ldots \beta_k$ are the coefficients of the predictor variables $X_1, X_2 \ldots X_k$. The term $\beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_kX_k$ is called the logit or the log-odds, which represents the log of the odds ratio of the dependent variable being 1. The logistic function transforms the linear combination of the predictors into a probability value between 0 and 1. This transformation is crucial because it maps any real-valued number into the (0, 1) interval, ensuring that the predicted probabilities are valid.

To estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ logistic regression uses maximum likelihood estimation (MLE). The likelihood function measures how likely the observed data is given a set of coefficients. The coefficients are chosen to maximize this likelihood function, which corresponds to finding the values that best fit the observed data.

Confidence Interval

A confidence interval (CI) provides a range of values within which a population parameter is estimated to lie with a certain level of confidence. For a given sample, a common form of the confidence interval for a mean is expressed as:

$$CI = \bar{X}\ \pm z\frac{\hat{\sigma}}{\sqrt{n}}$$

In our case we use this form (using z distribution and not t distribution), even though the variance is not given, because we have more than 30 samples.

where: $\bar{X}$ is the sample mean, z is the critical value from the standard normal distribution corresponding to the desired confidence level (1.96 for 95% confidence), $\hat{\sigma}$ is the sample standard deviation and n is the sample size. This interval estimates the range within which the true population mean is likely to fall with the specified confidence level.

For confidence interval for proportions, the formula adjusts to:

$$CI = \hat{p}\ \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p}$ is the sample proportion. In this case, there is no requirement of normality.

Pearson Correlation Coefficient

The Pearson correlation coefficient is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of this linear association. The formula for Pearson correlation is:

$$\rho = \frac{Cov(X,Y)}{\hat{\sigma}_x \hat{\sigma}_y}$$

where $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are the estimators for the std of X, Y, respectively.

Conclusions:

Our analysis demonstrated that lifestyle factors such as alcohol consumption, caffeine intake, and physical activity levels have a significant impact on sleep efficiency. Specifically, by looking at the graphs of the distributions we can assess that higher levels of alcohol consumption were associated with reduced sleep efficiency, while increased physical activity was linked to improved sleep efficiency. As shown above, there is strong correlation between sleep efficiency and deep sleep percentage, so we expect that the features we examined will affect the deep sleep percentage as well. These findings align with existing literature that suggests these behaviors influence sleep quality and overall health.

However, our investigation into the effects of these same factors on REM sleep percentage did not yield statistically significant results. The data did not provide strong evidence to suggest that alcohol, caffeine, or physical activity levels significantly alter REM sleep percentage. This suggests that while these lifestyle factors do impact overall sleep quality, their effects on specific stages of sleep, such as REM, may be more nuanced and potentially influenced by other variables not captured in this study.

These conclusions underscore the importance of considering lifestyle choices when examining person's sleep health. Further research is needed to better understand the complex relationships between lifestyle behaviors and different stages of sleep.

Here you can find the dataset, our code and instructions:

https://github.com/NNnaorNN/Statistical-Theory-Project

We would like to extend our heartfelt thanks to our lecturers, Oshrit Shtusel and Or Shkuri, for their guidance, support, and inspiration throughout this semester.