# APPLIED STATISTICS (MA2540)
## FACTORS AFFECTING SLEEP PATTERNS IN ADULTS


Atharv Ramesh Nair — EE20BTECH11006

Akriti — ES21BTECH11005

Balaji Tanushree Singh — ES21BTECH11009

Bhavishya Sirigiri — ES21BTECH11030

Aakarshita Shastri — MA22BTECH11001


August 3, 2024

# 1 Introduction

## Motivation

Understanding sleep patterns and their impact on overall health is crucial in today's fast-paced world. Sleep plays a fundamental role in our physical and mental well-being, influencing various aspects of our daily functioning. However, with the increasing demands of modern life, sleep disorders and inadequate sleep have become prevalent issues, affecting a significant portion of the population.

The motivation behind this project stems from the pressing need to delve deeper into the intricacies of sleep health and its relationship with lifestyle factors. By analyzing a dataset encompassing information from 374 individuals with diverse lifestyles, we aim to unravel valuable insights into sleeping habits and their associated determinants.

## Dataset

The Sleep Health and Lifestyle Dataset comprises 374 rows and 13 columns, covering a wide range of variables related to sleep and daily habits. Shown below are the top 5 rows from the dataset.

| | Person ID | Gender | Age | Occupation | Sleep Duration | Quality of Sleep | Physical Activity Level | Stress Level | BMI Category | Blood Pressure | Heart Rate | Daily Steps | Sleep Disorder |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | Male | 27 | Software Engineer | 6.1 | 6 | 42 | 6 | Overweight | 126/83 | 77 | 4200 | None |
| 1 | 2 | Male | 28 | Doctor | 6.2 | 6 | 60 | 8 | Normal | 125/80 | 75 | 10000 | None |
| 2 | 3 | Male | 28 | Doctor | 6.2 | 6 | 60 | 8 | Normal | 125/80 | 75 | 10000 | None |
| 3 | 4 | Male | 28 | Sales Representative | 5.9 | 4 | 30 | 8 | Obese | 140/90 | 85 | 3000 | Sleep Apnea |
| 4 | 5 | Male | 28 | Sales Representative | 5.9 | 4 | 30 | 8 | Obese | 140/90 | 85 | 3000 | Sleep Apnea |

Figure 1: Sample from Dataset

**Dataset Columns:**

- **Person ID:** An identifier for each individual.
- **Gender:** The gender of the person (Male/Female).
- **Age:** The age of the person in years.
- **Occupation:** The occupation or profession of the person.
- **Sleep Duration (hours):** The number of hours the person sleeps per day.
- **Quality of Sleep (scale: 1-10):** A subjective rating of the quality of sleep, ranging from 1 to 10.
- **Physical Activity Level (minutes/day):** The number of minutes the person engages in physical activity daily.
- **Stress Level (scale: 1-10):** A subjective rating of the stress level experienced by the person, ranging from 1 to 10.
- **BMI Category:** The BMI category of the person (e.g., Underweight, Normal, Overweight).
- **Blood Pressure (systolic/diastolic):** The blood pressure measurement of the person, indicated as systolic pressure over diastolic pressure.
- **Heart Rate (bpm):** The resting heart rate of the person in beats per minute.
- **Daily Steps:** The number of steps the person takes per day.
- **Sleep Disorder:** The presence or absence of a sleep disorder in the person (None, Insomnia, Sleep Apnea).

In this study, we focus on analyzing sleep duration as our primary target variable. Sleep duration, being a numerical variable, serves as a key metric for understanding the average and variability of sleep patterns across various groups. Additionally, we transform categorical variables such as Sleep Disorders, Stress Level, and Physical Activity Levels into binary categories. This allows us to examine proportions and associations within these variables, providing deeper insights into their impact on sleep patterns.

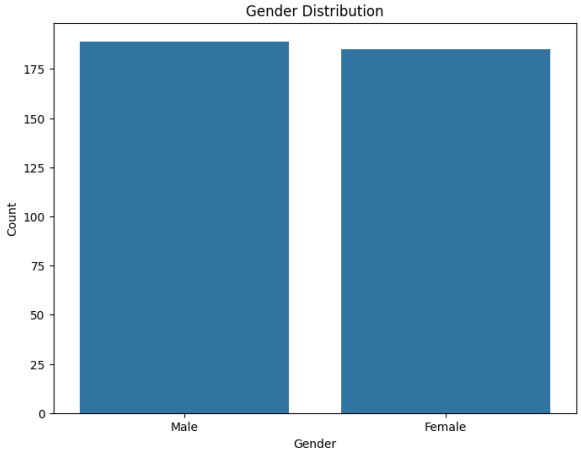
# 2 Data Visualization

Data visualization helps in gaining a better understanding of the nature of the datasets, study the patterns, and extracting meaningful insights.

# Plots:

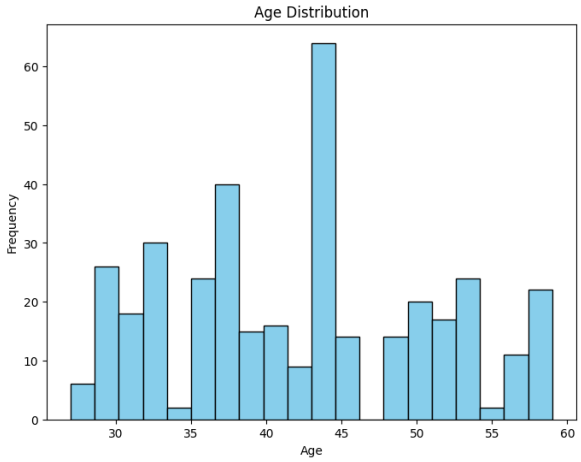
### Bar Plot: Gender Distribution

To examine the gender composition of the sample population and its potential implications for sleep health.



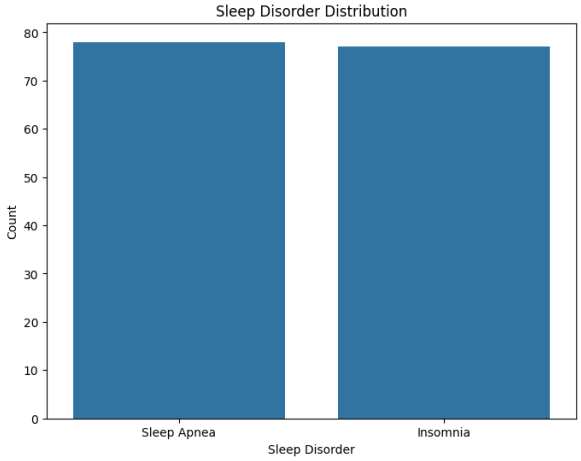

### Histogram: Age Distribution

To visualize the age distribution of the study participants and identify any predominant age groups.



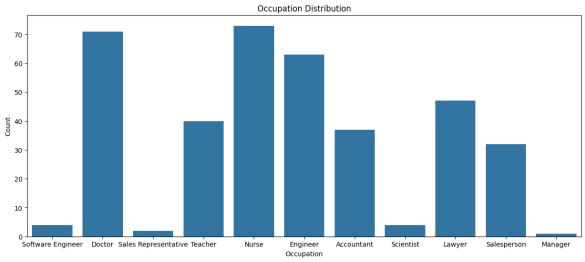

### Bar Plot: Sleep Disorder Distribution

To identify the prevalence of different sleep disorders within the sample population.



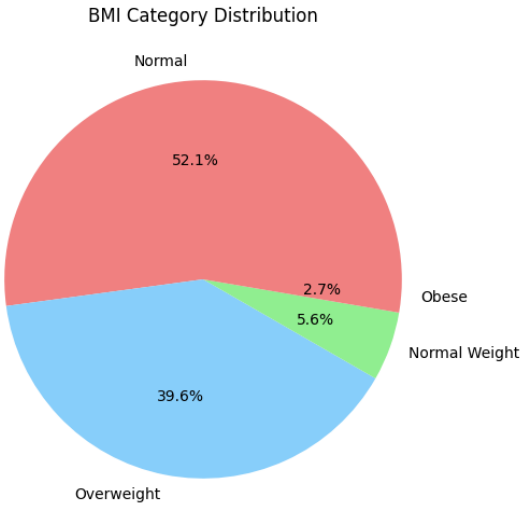
### Bar Plot: Occupation Distribution

To identify the prevalence of people with different occupations within the sample population.



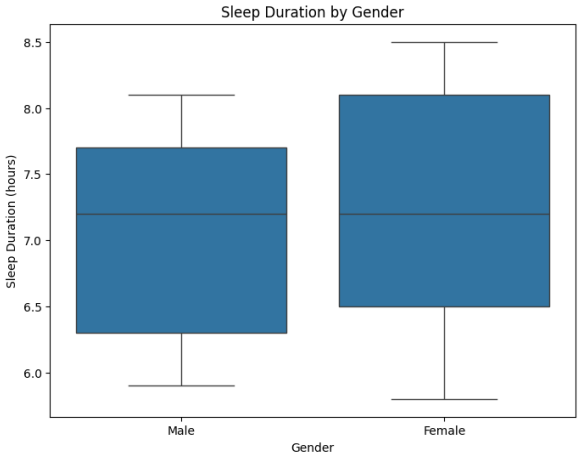

### Pie Chart: BMI Category Distribution

To explore the prevalence of different BMI categories within the study population and assess its relationship with sleep health.



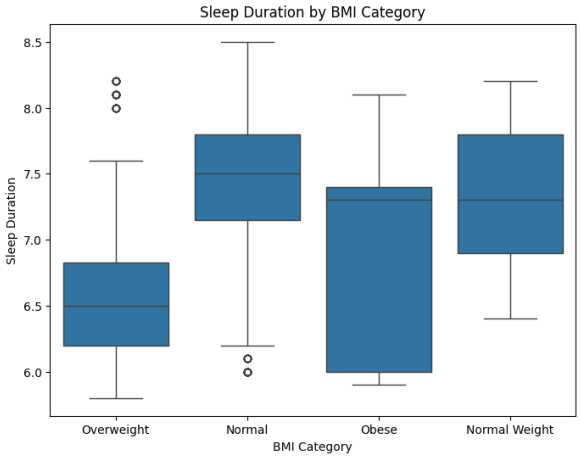

### Box Plot: Sleep Duration by Gender

To investigate potential differences in sleep duration based on gender.



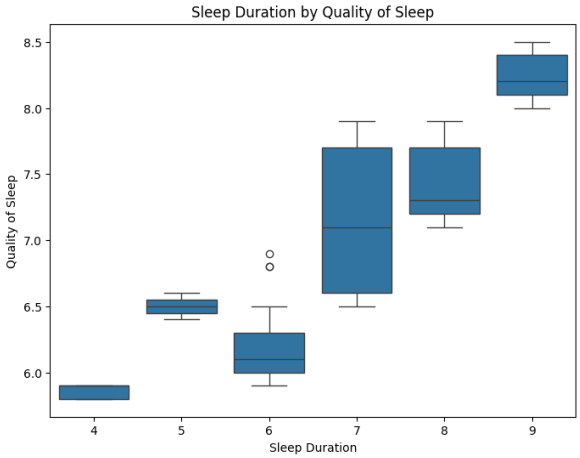

### Box Plot: Sleep Duration by BMI Category

To examine the relationship between BMI and sleep duration.



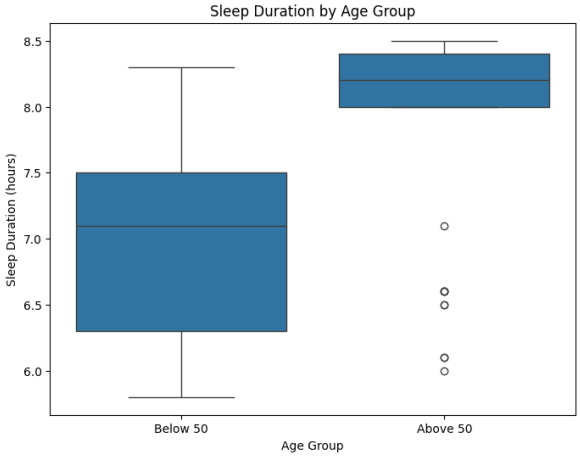

**Box Plot: Sleep Duration by Quality of Sleep**

To explore how sleep duration influences the reported quality of sleep.



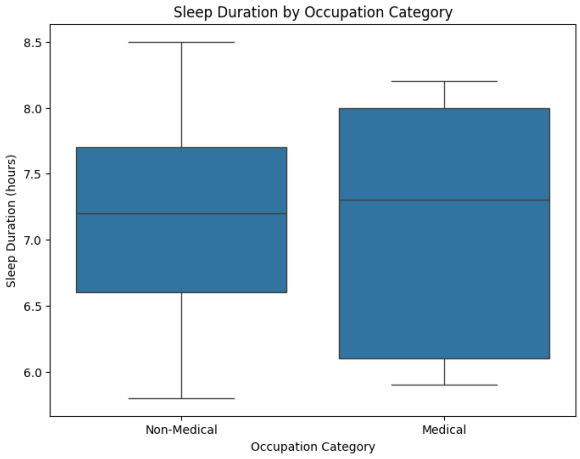

**Box Plot: Sleep Duration by Age**

To explore how sleep duration influences people with age above and below 50.



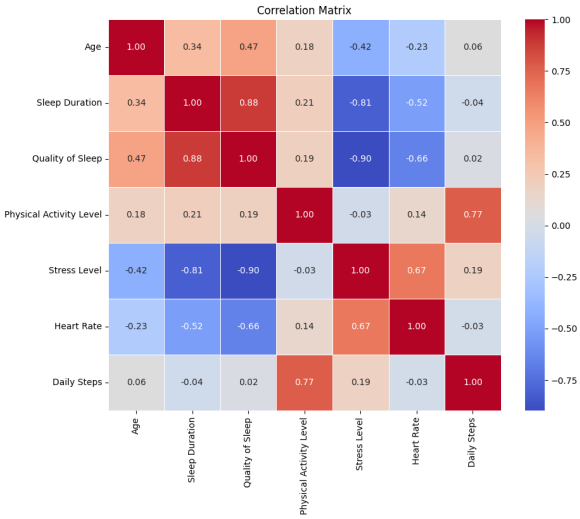

**Box Plot: Sleep Duration by Occupation**

To explore how sleep duration impacts individuals working in medical and non medical fields.



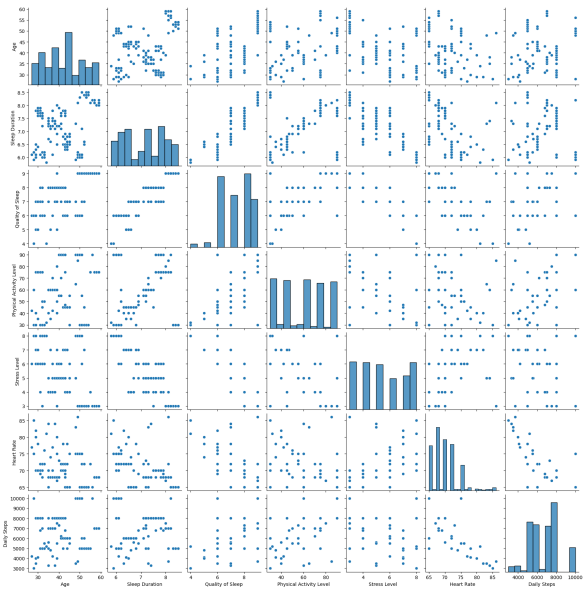

**Correlation Heatmap Matrix**

To identify correlations between sleep-related variables and other demographic or lifestyle numerical factors.




**Pair Plots**

To visualize potential interactions and correlations between different variables.



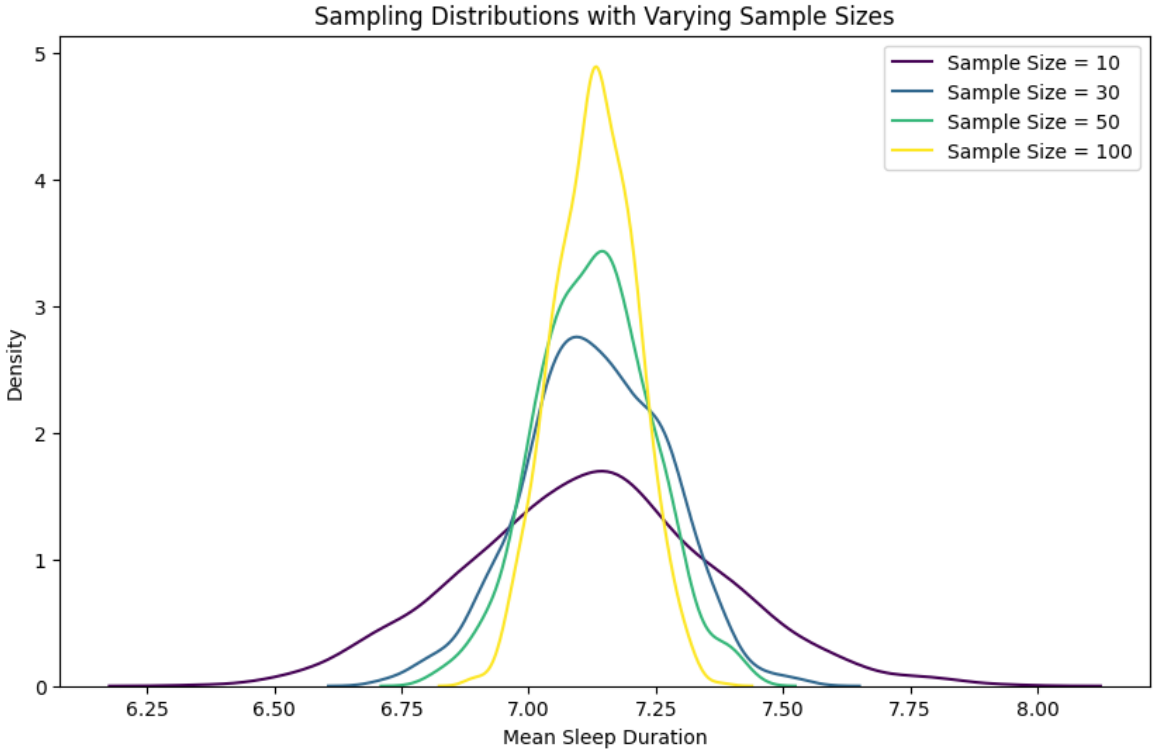

## Conclusions:

- Almost the same number of males and female of varying ages were taken into account for the dataset
- People with different BMIs(obese, overweight and normal weight) and different sleep disorders(Apnea, Insomnia, none) in various occupations are considered
- Females generally sleep for more number of hours than males
- People with normal weight have the best sleep
- Sleep duration and Quality of sleep are highly correlated. They are also highly correlated to the level of stress

# 3 Sampling Distributions

### Description

Sampling distribution is crucial in statistical analysis as it helps in understanding the variability of a sample statistic and making inferences about the population. Here, we have computed sampling distribution using Sleep Duration.
Mean of sample means: 7.132145614687294
Standard deviation of sample means: 0.04113230096455976
Median of sample means: 7.135160427807486




As we increase the sample size, the curve gets narrower (variance decreases). The Sampling Distribution of the sample mean is normal.

# 4 Confidence Interval Estimation

## 4.1 Confidence Interval Estimation of Means

### Problem 1: Average sleep duration of entire sample

#### Description

The objective is to estimate the average sleep duration of individuals by using the data from a sleep health and lifestyle dataset.

#### Formula Used

The confidence interval for a mean is calculated using the formula:

$$\text{CI} = \bar{x} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

#### Results

- Average of the sleep duration: 7.13
- Variance in the sleep duration: 0.63
- degrees of freedom : 373
- $t_{0.025,373}$ : 1.966

- Standard error : 0.04
- Margin Of Error: 0.081
- Confidence interval: (7.05, 7.21)
- Width : 0.1618

### Conclusion

With a confidence level of 95.0%, the mean of sleep duration of the individuals is between 7.05 hrs and 7.21 hrs.

## Problem 2: Difference of average sleep duration between Males and Females

### Description

The objective is to estimate the difference of average sleep duration between Males and Females by using the data from a sleep health and lifestyle dataset.

### Formula Used

The confidence interval for a mean is calculated using the formula:

$$\text{CI} = \bar{x_1} - \bar{x_2} \pm t_{\alpha/2,n+m-2} \times \frac{S_p}{\sqrt{(1/n) + (1/m)}}$$

### Results

- no of females : 185
- no of males : 189
- Average sleep duration difference : 0.193
- Variance in the sleep duration for females : 0.77
- Variance in the sleep duration for males : 0.48
- $\frac{S^2_{\text{females}}}{S^2_{\text{males}}}$ : $1.61 < 4$
- $S_p^2 = \frac{(f-1)\times S^2_{\text{females}} + (m-1)\times S^2_{\text{males}}}{f+m-2}$
- $S_p^2 = 0.625$
- $t_{0.025,372}$ : 1.966
- Standard error : 0.082
- Margin Of Error: 0.16
- Confidence interval: (0.032, 0.35)
- Width : 0.32

### Conclusion

With a confidence level of 95.0%, the average of difference of sleep duration of females and males is between 0.032 hrs and 0.35 hrs.

## Problem 3: Difference of average sleep duration between individuals with high stress level and low stress level

### Description

The objective is to estimate the difference of average sleep duration between individuals with high stress level ($>= 5$) and low stress level($< 5$) by using the data from a sleep health and lifestyle dataset.

### Formula Used

The confidence interval for a mean is calculated using the formula:

$$\text{CI} = \bar{x_1} - \bar{x_2} \pm t_{\alpha/2,n+m-2} \times \frac{S_p}{\sqrt{(1/n) + (1/m)}}$$

### Results

- No of individuals with stress level ($>= 5$) : 233
- No of individuals with stress level ($< 5$) : 141
- Average sleep duration difference : 0.803
- Variance in the sleep duration for high stress level people : 0.494
- Variance in the sleep duration for low stress level people : 0.463
- $\frac{S^2_{\text{high stress}}}{S^2_{\text{low stress}}}$ : $1.068 < 4$
- $S_p^2 = \frac{(h-1)\times S^2_{\text{high}} + (l-1)\times S^2_{\text{low}}}{h+l-2}$
- $S_p^2 = 0.48$
- $t_{0.025,372}$ : 1.966
- Standard error : 0.074
- Margin Of Error: 0.146
- Confidence interval: (0.658, 0.949)
- Width : 0.29

### Conclusion

With a confidence level of 95.0%, the average of difference of sleep duration of individuals with high stress level and low stress level is between 0.658 hrs and 0.949 hrs.

## Problem 4: Difference of average sleep duration between individuals in medical field and individuals in non-medical occupations with the Physical Activity Levels between 50 and 70

### Description

The objective is to estimate the difference of average sleep duration between medicals(Doctor and nurse) and non-medicals with Physical Activity Levels lying between 50 and 70 by using the data from a sleep health and lifestyle dataset.

### Formula Used

The confidence interval for a mean is calculated using the formula:

$$\text{CI} = \bar{x_1} - \bar{x_2} \pm t_{\alpha/2,r} \times \sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}$$

### Results

- No of medicals : 6
- No of non-medicals : 79
- $r = \frac{(\frac{S^2_{\text{medicals}}}{m} + \frac{S^2_{\text{non-medicals}}}{nm})^2}{\frac{(\frac{S^2_{\text{medicals}}}{m})^2}{m-1} + \frac{(\frac{S^2_{\text{non-medicals}}}{nm})^2}{nm-1}}$
- Degrees of freedom : 5
- Average sleep duration difference : 0.261
- Variance in the sleep duration for medicals : 0.338
- Variance in the sleep duration for non-medicals : 0.0458
- $\frac{S^2_{\text{medicals}}}{S^2_{\text{non-medicals}}}$ :$7.38 > 4$
- $S = \sqrt{\frac{S^2_{\text{medicals}}}{m} + \frac{S^2_{\text{non-medicals}}}{nm}}$
- $S = 0.239$
- $t_{0.025,5}$ : 2.555
- Standard error : 0.238
- Margin Of Error: 0.61
- Confidence interval: (-0.348, 0.872)
- Width : 1.22

### Conclusion

With a confidence level of 95.0%, the average of difference of sleep duration of individuals medical occupation and non-medical occupation is between 0.08hrs and 0.289hrs.

## 4.2 Confidence Interval Estimation of Variance

### Problem 1: Estimation of variance of sleep duration of entire sample

### Description

The objective is to estimate the variance of sleep duration of individuals by using the data from a sleep health and lifestyle dataset.

### Formula Used

The confidence interval for a variance is calculated using the formula:

$$\text{CI} : (\frac{(n-1)}{b}S^2, \frac{(n-1)}{a}S^2)$$

- $a = \chi^2_{1-\alpha/2,n-1}$
- $b = \chi^2_{\alpha/2,n-1}$

### Results

- Variance in the sleep duration: 0.63
- degrees of freedom : 373
- $\chi^2_{0.975,373} = 321.38$
- $\chi^2_{0.025,373} = 428.40$
- Confidence interval: (0.55, 0.73)
- Width : 0.18

### Conclusion

With a confidence level of 95.0%, the variance of sleep duration of the individuals vary between 0.55 hrs and 0.73 hrs.

### Problem 2: Estimation of variance of sleep duration between individuals in medical field and individuals in other occupations

### Description

The objective is to estimate the variance range of sleep duration between medicals(Doctor and nurse) and non-medicals of population by using the data from a sleep health and lifestyle dataset.

**Formula Used**

The confidence interval for a variance is calculated using the formula:

$$\text{CI} = (\frac{1}{F_{\alpha/2}(n-1, m-1)} \times \frac{S_x^2}{S_y^2}, F_{\alpha/2}(m-1, n-1) \times \frac{S_x^2}{S_y^2}))$$

**Results**

- No of medicals : 144
- No of non-medicals : 230
- Variance in the sleep duration for medicals : 0.86
- Variance in the sleep duration for non-medicals : 0.48
- $F_{0.025}(229, 143)$ : 1.288
- $F_{0.025}(143, 229)$ : 1.277
- Confidence interval: (0.433, 0.712)
- Width : 0.279

**Conclusion**

With a confidence level of 95.0%, the variance of sleep duration of individuals medical occupation and non-medical occupation vary between 0.433hrs and 0.712hrs.

## 4.3 Confidence Interval Estimation of Proportions

**Problem 1: Estimation of Sleep Apnea Prevalence**

**Description**

The objective is to estimate the proportion of individuals experiencing sleep apnea using data from a sleep health and lifestyle dataset.

**Formula Used**

The confidence interval for a single proportion was calculated using the formula:

$$\text{CI} = \hat{p} \pm z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

**Results**

- Count of individuals experiencing sleep apnea: 78
- Total number of individuals: 374
- Proportion of individuals experiencing sleep apnea ($\hat{p}$): 0.209

- Standard error: 0.021
- Margin of error: 0.041
- Confidence interval lower bound: 16.72%
- Confidence interval upper bound: 24.99%

### Conclusion

With a confidence level of 95.0%, the proportion of individuals experiencing sleep apnea is estimated to be between 16.72% and 24.99%.

## Problem 2: Estimation of High Stress Levels Prevalence

### Description

The goal is to estimate the proportion of individuals with high stress levels using data filtered based on stress criteria.

### Formula Used

The confidence interval for a single proportion was computed using the formula mentioned earlier.

### Results

- Number of individuals meeting the criteria: 166
- Total number of individuals in the dataset: 374
- Proportion of individuals meeting the criteria ($\hat{p}$): 0.444
- Standard error: 0.025
- Margin of error: 0.050
- Confidence interval lower bound: 39.33%
- Confidence interval upper bound: 49.43%

### Conclusion

With a confidence level of 95.0%, the proportion of individuals with high stress levels is estimated to be between 39.33% and 49.43%.

## Problem 3: Difference in Insomnia Proportions between Age Groups

### Description

The aim is to estimate the difference in proportions of individuals with insomnia between two age groups: 30-45 and 45-60.

### Formula Used

The confidence interval for a single proportion was calculated using the formula:

$$\text{Difference in Proportions} = \hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

### Results

- Proportion of individuals with insomnia aged 30-45: 0.261
- Proportion of individuals with insomnia aged 45-60: 0.091
- Difference in proportions: 0.170
- Sample size of individuals aged 30-45: 245
- Sample size of individuals aged 45-60: 110
- Difference in proportions: 0.1703
- Standard error: 0.0392
- Margin of error: 0.0778
- Lower bound of the confidence interval: 9.2562
- Upper bound of the confidence interval: 24.8069

### Conclusion

With a confidence level of 95.0%, the difference in proportions of individuals with insomnia between individuals aged 30-45 and 45-60 is estimated to be between 9.26% and 24.80%.

## Problem 4: Difference in Insomnia Proportions between Medical and Non-Medical Occupations

### Description

The objective is to estimate the difference in proportions of individuals with insomnia between medical (doctors or nurses) and non-medical occupations using data from a sleep health and lifestyle dataset.

### Formula Used

The confidence interval for the difference in proportions was calculated using the formula mentioned earlier.

### Results

- Proportion of individuals with insomnia in medical occupations (doctors or nurses): 0.042
- Proportion of individuals with insomnia in non-medical occupations: 0.309
- Sample size of medical occupations: 144
- Sample size of non-medical occupations: 230
- Difference in proportions: 0.267
- Standard error: 0.035
- Margin of error: 0.068
- Confidence interval lower bound: 19.84%
- Confidence interval upper bound: 33.56%

### Conclusion

With a confidence level of 95.0%, the difference in proportions of individuals with insomnia between non-medical occupations and medical (doctors or nurses) is estimated to be between 19.84% and 33.56%.

## Problem 5: Difference in Physical Activity Levels between Medical and Non-Medical Professions

### Description

The goal is to estimate the difference in proportions of individuals with low physical activity levels between medical professions (doctors or nurses) and non-medical professions using data from a sleep health and lifestyle dataset.

### Formula Used

The confidence interval for the difference in proportions was calculated using the formula mentioned earlier.

### Results

- Proportion of individuals with low physical activity levels in medical professions (doctors or nurses): 0.257
- Proportion of individuals with low physical activity levels in non-medical professions: 0.539
- Sample size of medical professions: 144
- Sample size of non-medical professions: 230
- Difference in proportions: 0.282

- Standard error: 0.049
- Margin of error: 0.097
- Confidence interval lower bound: 18.5224%
- Confidence interval upper bound: 37.9148%

**Conclusion**

With a confidence level of 95.0%, the difference in proportions of individuals with low physical activity levels in non-medical professions compared to those in medical professions is estimated to be between 18.52% and 37.91%.

# 5 Hypothesis Testing

## 5.1 Average Sleep Duration of the entire population

We want to investigate the mean sleep duration of the population using the sample data The confidence interval obtained was [7.05,7.21] hrs

- Let $\bar{x}$ be the sample Mean of Sleep Duration
- Let $s$ be sample Standard Deviation of Sleep Duration
- Let $\mu$ be the population mean and $n$ be the sample Size

Our Hypothesis (left-tailed) is as follows:

$$H_0 : \mu \geq 7 \quad H_a : \mu < 7$$

From the data, we have

$$\begin{aligned} \bar{x} &= 7.13 \\ s^2 &= 0.63 \\ n &= 374 \end{aligned}$$

Since, population variance is unknown, we use the t-test. The test statistics is as follows:

$$t^* = \frac{\bar{x} - \mu}{s/\sqrt{n}} = 3.21$$

According to the rejection region method, $H_0$ is rejected when the calculated test statistic $t^*$ falls below the critical value $-t_{\alpha,n-1}$ (where $\alpha = 0.05$ and $n = 374$). With $t_{\alpha,n-1} = 1.649$, and given that $t^* = 3.21 > -1.649$, we conclude that there is no statistical evidence to reject the null hypothesis $H_0$. Consequently, we cannot say that the population's mean sleep duration is less than 7 hours.

## 5.2 Difference in Average Sleep Duration for Men and Women

We aim to explore potential disparities in average sleep duration between men and women. Understanding such differences could shed light on gender-specific sleep patterns, which have implications for overall health and well-being. The confidence interval obtained previously was [0.032, 0.35] hrs.

- $\bar{x_1}$, $s_1^2$, $\mu_1$, and $n_1$: Sample mean, sample variance, population mean, and sample size for men, respectively.
- $\bar{x_2}$, $s_2^2$, $\mu_2$, and $n_2$: Sample mean, sample variance, population mean, and sample size for women, respectively.

Our Hypothesis (two-tailed) is as follows:

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 \neq 0$$

From the data, we have

$$\bar{x_1} = 7.03 \quad \text{and} \quad \bar{x_2} = 7.23$$

$${s_1}^2 = 0.48 \quad \text{and} \quad {s_2}^2 = 0.77$$

$$n_1 = 189 \quad \text{and} \quad n_2 = 185$$

Since, population variance is unknown and since $\frac{s_1}{s_2} \in [0.5, 2]$, we use the pooled t-test. The pooled variance $S_p^2$ is given by:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = 0.625$$

The test statistic is then given by:

$$t^* = \frac{\bar{x_1} - \bar{x_2} - 0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = -2.36$$

According to the rejection region approach, we reject the null hypothesis $H_0$ when $|t^*| > t_{\alpha/2,df}$, where $\alpha = 0.05$ and $df = n_1 + n_2 - 2 = 372$. With $t_{0.025,372} = 1.966$, and given that the calculated test statistic yields 2.36, exceeding the critical value, we reject the null hypothesis $H_0$. We have enough statistical evidence to infer that the average sleep durations of men and women are indeed different.

## 5.3 Difference in average sleep duration between medical and non-medical professionals engaging in moderate physical activity levels

We aim to examine whether there exists a difference in the average sleep duration between two distinct groups: medical professionals (comprising doctors and nurses) and non-medical professionals. We focus particularly on individuals engaging in moderate physical activity levels, ranging between 50 and 70.

- Let $\bar{x_1}$ and $\bar{x_2}$ represent the sample means of sleep duration for medical and non-medical professionals, respectively, both engaged in moderate physical activities.
- Similarly, let $\bar{s_1}$ and $\bar{s_2}$ denote their corresponding sample standard deviations, and $\mu_1$ and $\mu_2$ be the respective population means
- Finally, let $n_1$ and $n_2$ denote the sample sizes for the medical and non-medical professional groups, respectively.

Our Hypothesis (two-tailed)is as follows:

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 \neq 0$$

From the data, we have

$$\begin{aligned} \bar{x_1} = 6.93 \quad \text{and} \quad \bar{x_2} = 7.19 \\ s_1 = 0.582 \quad \text{and} \quad s_2 = 0.2141 \\ n_1 = 6 \quad \text{and} \quad n_2 = 79 \end{aligned}$$

Since, population variance is unknown and since $\frac{s_1}{s_2} = 2.72$, we use the Welch's t-test.

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \approx 5$$

The test statistic is then given by:

$$t^* = \frac{\bar{x_1} - \bar{x_2} - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -1.095$$

According to the rejection region approach, $H_0$ is rejected if $t^* > |t_{\alpha/2,df}|$, with $\alpha = 0.05$ and $df = 5$. As $t_{0.025,5} = 2.57$, and given that $1.095 < 2.57$, we fail to reject the null hypothesis $H_0$. Hence, we lack sufficient statistical evidence to say that there exists a significant difference in the mean sleep duration between medical and non-medical professionals engaging in moderate physical activity.

## 5.4 The variance of Sleep Duration for the population

We want to examine the consistency in sleep durations among individuals.We aim to determine whether the variance in sleep duration exceeds a threshold of 1 hour. The confidence interval obtained was $[0.55, 0.73] hrs^2$

- Let $s^2$ a represent the sample variances of sleep duration for people with high-stress levels
- Let $\sigma_0^2 = 1$ be the population variance and let n denote the sample size

Our Hypothesis is as follows:

$$H_0 : \sigma^2 \leq 1 \quad \sigma^2 > 1$$

From the data, we have

$$s^2 = 0.63$$
$$n = 374$$

Using the chi-square test, the test statistic is then given by:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = 236.13$$

In the rejection region approach, $H_0$ is rejected if $\chi^2 > \chi_U^2$, where $\chi_U^2$ represents the upper-tail value for significance level $\alpha$ (degrees of freedom $df = n - 1 = 69$). Here, $\chi_U^2 = 419.03$. Given that $236.13 < 419.03$, we lack sufficient statistical evidence to reject the null hypothesis $H_0$. Thus, we cannot conclude that the variance in sleep duration exceeds 1 hour.

## 5.5 The ratio of variances in Sleep Duration for people from medical and non-medical professions

We aim to explore differences in the variance of sleep durations between individuals in medical and non-medical professions. We aimed to gain insight into the consistency of sleep durations across different occupational categories. The confidence interval obtained was $[0.433, 0.712] hrs^2$

- Let $s_1^2$ and $s_2^2$ represent the sample variances of sleep duration for medical and non-medical professionals respectively.
- Let $\sigma_1^2$ and $\sigma_2^2$ denote the corresponding population variances respectively and $n_1$ and $n_2$ denote the sample sizes for the medical and non-medical professional groups, respectively

. Our Hypothesis (two-tailed) is as follows:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad H_a : \sigma_1^2 \neq \sigma_2^2$$

From the data, we have

$$s_1^2 = 0.86 \quad \text{and} \quad s_2 = 0.48$$

$$n_1 = 144 \quad \text{and} \quad n_2 = 230$$

We use the F-Test with degree of freedoms $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$.
The test statistic is given by:

$$F^* = \frac{s_1^2}{s_2^2} = 1.79$$

In the rejection region approach, we reject $H_0$ if $F^* \leq F_{1-\alpha/2,df_1,df_2}$ or $F^* \geq F_{\alpha/2,df_1,df_2}$ where $\alpha = 0.05, df_1 = 144, df_2 = 230$. Here $F_{1-\alpha/2,df_1,df_2} = 0.74$ and $F_{\alpha/2,df_1,df_2} = 1.34$. Since $1.79 > 1.34$ and $1.32 > 0.74$, we reject the null hypothesis $H_0$. We have enough statistical evidence to say that the ratio of variances in sleep duration across these professions is different.

## 5.6 The proportion of individuals experiencing sleep apnea in the population

We aim to explore the proportion of individuals experiencing sleep apnea in the population. We want to see if there is a significant deviation in this proportion from 0.2 (according to National Sleep Foundation). The confidence interval was obtained as [16.72%, 24.99%.].

- Let the proportion of individuals experiencing sleep apnea be p
- Let the sample size be n

Our Hypothesis (Two-tailed) is as follows:

$$H_0 : p = 0.2 \quad H_a : p \neq 0.2$$

From the data, the Number of individuals with sleep apnea is 78

$$\hat{p} = \frac{78}{374} \approx 0.208 \quad \text{and} \quad n = 374$$

The test statistic is given by:

$$Z^* = \frac{\hat{p} - p}{\sqrt{\frac{p \times (1-p)}{n}}} \approx 0.4136$$

In the rejection region approach, we reject $H_0$ if $|Z^*| \geq Z_{\alpha/2,n-1}$ where $\alpha = 0.05, n = 374$. Here $Z_{\alpha/2,n-1} = 1.96$. Since $0.4136 < 1.96$, we fail to reject the null hypothesis $H_0$. Hence, we do not have enough evidence to conclude that the proportion of individuals experiencing sleep apnea in the population is different from 0.2.

## 5.7 The proportion of people experiencing high stress levels

We aim to explore the proportion of individuals experiencing high-stress levels (¿5). We want to see if there is a significant deviation in this proportion from 0.5

- Let the proportion of individuals experiencing high-stress levels be p
- Let the sample size be n

Our Hypothesis (Two-tailed) is as follows:

$$H_0 : p = 0.5 \quad H_a : p \neq 0.5$$

From the data, the number of such individuals is 166

$$\hat{p} = \frac{166}{374} \approx 0.444 \quad \text{and} \quad n = 374$$

The test statistic is given by:

$$Z^* = \frac{\hat{p} - p}{\sqrt{\frac{p \times (1-p)}{n}}} \approx -2.17$$

As mentioned in the previous section, Since $2.17 > 1.96$, we reject the null hypothesis $H_0$. Hence there is sufficient evidence to conclude that the proportion of individuals experiencing high stress levels is different from 0.5.

## 5.8 The difference in proportions of individuals with insomnia between male and female adults

We aim to compare the prevalence of insomnia in male and female adults. By examining these proportions, we aim to shed light on potential disparities in the occurrence of insomnia across gender demographics

- Let the sample size for males be $n_{male}$
- Let the sample size for females be $n_{female}$
- Let the proportions of males and females with insomnia be $p_{male}$ and $p_{female}$ respectively

Our Hypothesis (Two-tailed) is as follows:

$$H_0 : p_{\text{male}} - p_{\text{female}} = 0 \quad H_a : p_{\text{male}} - p_{\text{female}} \neq 0$$

From the data, the number of individuals with insomnia are 41 and 36 for males and females respectively

$$n_{\text{male}} = 189 \quad \text{and} \quad n_{\text{female}} = 185$$

$$\hat{p}_{\text{male}} = \frac{41}{189} \approx 0.2169 \quad \text{and} \quad \hat{p}_{\text{female}} = \frac{36}{185} \approx 0.1945$$

The standard error ($SE$) of the difference in proportions:

$$SE = \sqrt{\frac{\hat{p}_{\text{male}}(1-\hat{p}_{\text{male}})}{n_{\text{male}}} + \frac{\hat{p}_{\text{female}}(1-\hat{p}_{\text{female}})}{n_{\text{female}}}}$$

$$\approx 0.0417$$

The test statistic is given by :

$$z^* = \frac{\hat{p}_{\text{male}} - \hat{p}_{\text{female}}}{SE}$$

$$\approx \frac{0.2169 - 0.1945}{0.0417} \approx 0.53$$

Since $|Z^*| = |0.53| < 1.96$, we fail to reject the null hypothesis. Hence we do not have enough statistical evidence to conclude that there is a difference in the proportions of individuals with insomnia between male and female adults.

## 5.9 The difference in proportions of individuals with insomnia between two occupation groups(medical and non-medical)

We aimed to examine the difference in the proportions of individuals experiencing insomnia between two distinct occupational groups: medical and non-medical. We seek to shed light on potential differences in the prevalence of insomnia among individuals from different professional backgrounds

- Let the sample size for individuals in medical and non-medical occupation groups be $n_{medical}$ and $n_{non-medical}$ respectively.
- Let the proportion of individuals in the medical and non-medical occupation groups be $p_{medical}$ and $p_{non-medical}$ respectively

Our Hypothesis (Two-tailed) is as follows:

$$H_0 : p_{\text{medical}} - p_{\text{non-medical}} = 0 \quad H_a : p_{\text{medical}} - p_{\text{non-medical}} \neq 0$$

From the data, the number of such individuals with both groups are 6 and respectively

$$n_{\text{medical}} = 144 \quad \text{and} \quad n_{\text{non-medical}} = 230$$

$$\hat{p}_{\text{medical}} = \frac{6}{144} \approx 0.041 \quad \text{and} \quad \hat{p}_{\text{non-medical}} = \frac{71}{230} \approx 0.308$$

The standard error ($SE$) of the difference in proportions:

$$SE = \sqrt{\frac{\hat{p}_{\text{medical}}(1-\hat{p}_{\text{medical}})}{144} + \frac{\hat{p}_{\text{non-medical}}(1-\hat{p}_{\text{non-medical}})}{230}}$$

$$SE \approx 0.0347$$

The test statistic is given by:

$$z^* = \frac{(\hat{p}_{\text{medical}} - \hat{p}_{\text{non-medical}})}{SE}$$

$$z^* = \frac{-0.26}{0.0347} \approx -7.69$$

Since $|Z^*| = 7.69 > 1.96$, we reject the null hypothesis. Hence we have enough evidence to conclude that there is a difference in the proportions of individuals with insomnia between medical and non-medical occupation groups.

# 6 Conclusion

We have conducted a thorough study of different factors affecting sleep duration, disorders, etc. We have done a parallel study of similar problems for both confidence interval estimation and hypothesis testing. In the confidence interval estimation case, we have conducted some additional studies. Based on this, we were able to understand sleep patterns better and how it varies in different demographics.

# 7 Contribution

Akriti(ES21BTECH11005):

- Data Visualization
- Sampling Distribution
- Preparing ppt

Bhavishya Sirigiri(ES21BTECH11030)

- Finding Confidence intervals for sample means and variances
- Report for the respective work

Balaji Tanushree Singh(ES21BTECH11009)

- Finding Confidence intervals for proportions
- Report for the respective work

Atharv Ramesh Nair(EE20BTECH11006)

- Hypothesis Testing for means and variances
- Report for the respective work

Aakarshita Shastri(MA22BTECH11001)

- Hypothesis Testing for proportions
- Report for the respective work