FPT QUY NHON UNIVERSITY
ARTIFICIAL INTELLIGENCE

FPT Education
FPT UNIVERSITY

MAS291

COMPUTER PROJECT

Topic:

ANALYZING HOW SLEEP DURATION INFLUENCES STRESS LEVELS IN MALES AND FEMALES.

Class:	AI18C	
Students:	Nguyen Tan Thang	QE180019
	Huynh Quoc Trung	QE180038
	Huynh Huu Hieu	QE180067
	Tran Tuan Kiet	QE180152
	Nguyen Huu Hoai	QE180083

Quy Nhon, 7-2024

Contents

1. Determine the topic

1.1 Problem statement

- Sleep is a crucial factor affecting human health and mental well-being. In modern society, many people struggle to maintain a proper sleep schedule due to work pressure, studies, and other factors. This can lead to mental health issues, particularly elevated stress levels. Understanding the relationship between the number of hours slept and stress levels can help us develop solutions to improve mental health and quality of life.

1.2 Research topic

- This study aims to analyze data to determine the relationship between sleep duration and stress levels, with the goal of clarifying these relationships to contribute to improving mental and physical health within the community.

2. Data source

- Data link: https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset/data.

- The dataset has 374 records and includes 13 attributes related to sleep health and lifestyle, but for this project, only attributes such as ID, Gender, Sleep Duration, and Stress Level are needed.

3. Descriptive statistics

- There are a total of 374 records, with 189 males and 185 females.

- The statistics regarding Sleep Duration and Stress Level:

	Sleep Duration	Stress Level
Count	374	374
Mean	7.132086	5.385027
Standard deviation	0.795657	1.774526
Min	5.8	3
25%	6.4	4
50%	7.2	5
75%	7.8	7
Max	8.5	8

4. Solving the problems

4.1 Test a hypothesis and construct confidence intervals for the mean and the proportion of a population.

4.1.1 Test a hypothesis and construct confidence intervals for the mean of a population.

Can we conclude that the average sleep duration is less than 7 hours? The first 10 values extracted from the data are as follows:

Sleep Duration									
6.1	6.2	6.2	5.9	5.9	5.9	6.3	7.8	7.8	7.8

Answer:

- Hypothesis testing:

(H_0): $\mu = 7$ (Average sleep time is not less than 7 hours)
(H_1): $\mu < 7$ (Average sleep time is less than 7 hours).

From the sample data we obtained:

Sample size: n = 10

Sample mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \approx 6.59$$

Sample variance:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \approx 0.717$$

$\Rightarrow$ Sample standard deviation: $s \approx 0.846$

We have test statistics:

$$t_0 = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = \frac{6.59 - 7}{0.846/\sqrt{10}} \approx -1.532$$

We have: $\alpha = 1 - 0.95$, degree freedom $= 10 - 1 \Rightarrow -t_{\alpha,n-1} = -t_{0.05,9} \approx -1.833$

Since the value $t_0 = -1,532 > -t_{\alpha,n-1} = -1,833$, we fail to reject the hypothesis H_0 with a significance level of 0.05. Therefore, there is not enough evidence to conclude sleep duration is less than 7 hours.

- The 95% confident interval for the average sleep duration is:

$$\text{CI} = \bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$
$$\text{CI} = 6.59 \pm 1.833 \cdot \frac{0.846}{\sqrt{10}}$$
$$\text{CI} = (6.09, 7.08)$$

The 95% confidence interval for the population mean shows that the population mean sleep duration ranges from 6.09 to 7.08 hours.

4.1.2 Test a hypothesis and construct confidence intervals for the proportion of a population.

Is there the proportion of people with sleep duration > 7 hours is higher than 50%? Given that out of a random sample of 100 person, 60 person have sleep duration > 7 hours.

Answer:

- Hypothesis testing:

H_0: p = 0.5 (proportion of people with sleep duration not greater than 7 hours)
H_1: p > 0.5 (proportion of people with sleep duration greater than 7 hours)

From the sample data we obtained:

Hypothesized proportion: $p_0 = 0.5$

Total number of people: n = 100

Proportion of people with sleep duration greater than 7: $\hat{p} = \frac{60}{100} = 0.6$

Test statistic:

$$z_0 = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx 1.99$$

Compare the z-value to the critical z-value $z_\alpha \approx 1.645$ for a significance level $\alpha = 0.05$. Since z-value is greater than 1.645 so we reject the null hypothesis H_0

Therefore, there is sufficient evidence to conclude that the proportion of people with sleep duration > 7 hours is higher than 50%

- The 95% confident interval for p > 7:

With $\alpha = 0.05$, we have $z_{\alpha/2} \approx 1.96$ Confidence Interval:

$$CI = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
$$CI \approx (0.54, 0.696)$$

Therefore, 95% confident that the true proportion of people with sleep duration > 7 hours lies between 50.4% and 69.6%.

4.2 Test a hypothesis and construct a confidence interval for the difference in mean and proportions of two populations.

4.2.1 Test a hypothesis and construct a confidence interval for the difference in mean of two populations.

Is there a significant difference in the mean sleep duration with stress level greater than 6 between males and females? The random 15 values extracted from the data are as follows:

Male Sleep Duration

6.5	6	6.1	6.3	6	6.5	6.2	6.8	6.1	6.5	6.3	6	6.4	6.1	6.5

Female Sleep Duration

6.2	6.1	6.5	6.1	6.8	6.1	6.1	6.6	6	6.2	5.9	6.1	6	6.2	6.1

Answer:

- Hypothesis testing:

$H_0 : \mu_1 = \mu_2$ (There is no significant difference in mean sleep duration between males and females with stress level > 6.)

$H_1 : \mu_1 \neq \mu_2$ (There is a significant difference in mean sleep duration between males and females with stress level > 6.)

From the sample data we obtained:

Number of males: $n_1 = 15$

Number of females: $n_2 = 15$

Sample mean and variance of male stress levels:

$$\bar{x_1} = \frac{\sum_{i=1}^{n_1} x_i}{n_1} \approx 6.29$$

$$s_1^2 = \tfrac{1}{n_1-1} \sum_{i=1}^{n} (x_i - \bar{x_1})^2 \approx 0.06$$

Sample mean and variance of female stress levels:

$$\bar{x_2} = \frac{\sum_{i=1}^{n_2} x_i}{n_2} = 6.2$$

$$s_2^2 = \tfrac{1}{n_2-1} \sum_{i=1}^{n} (x_i - \bar{x_2})^2 \approx 0.06$$

Pooled variance:

$$s_p = \sqrt{\tfrac{s_1^2 \cdot (n_1-1) + s_2^2 \cdot (n_2-1)}{n_1+n_2-2}} \approx 0.06$$

Test statistic:

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \approx 4.11$$

We have: α = 1-0.95, degree freedom = 15+15-2

$\Rightarrow$ Critical values for a two-tailed test with $t_{\alpha/2,28} \approx \pm 2.048$ (from the t-distribution table)

Since $t_0 > 2.048$. Therefore, we reject the null hypothesis H_0. There is a significant difference in the mean sleep duration between males and females with stress level > 6.

- Construct a 95% confidence interval:

The 95% confidence interval for the difference in mean between the two populations is calculated as:

$$CI = \bar{x}_1 - \bar{x}_2 - t_{\alpha/2,df} \cdot \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,df} \cdot \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$CI \approx (-0.66, 0.84)$$

The 95% confidence interval for the difference in mean sleep duration with stress levels greater than 6 between males and females is from -0.66 to 0.84. This indicates that there is a statistically significant difference in mean sleep duration with stress levels greater than 6 between males and females.

4.2.2 Test a hypothesis and construct a confidence interval for the difference in proportions of two populations.

Is there a significant difference in the proportion of people with Stress Level > 6 between males and females? Given that out of a random sample of 60 males, 24 have Stress Level > 6; out of a random sample of 50 females, 15 have Stress Level > 6.

Answer:

- Hypothesis testing:

$H_0 : p_1 = p_2$ (The proportion of people with stress levels > 6 between male and female is equal)

$H_1 : p_1 \neq p_2$ (The proportion of people with stress levels > 6 between male and female is not equal)

From the data we obtained:

Total number of males: $n_1 = 60$

Total number of females: $n_2 = 50$

Number of males with Stress Level > 6: $x_1 = 24$

Number of females with Stress Level > 6: $x_2 = 15$

Calculate the sample proportion for each group:

$$\hat{p_1} = \frac{x_1}{n_1} = 0.4$$

$$\hat{p_2} = \frac{x_2}{n_2} = 0.3$$

Calculate $\hat{p}$ (the pooled proportion estimate from both samples):

$$\hat{p} = \frac{x_1+x_2}{n_1+n_2} \approx 0.355$$

Calculate z-value :

$$z = \frac{\hat{p_1}-\hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}} \approx 1.09$$

We have: $\alpha = 1 - 0.95 \pm z_{\alpha/2} \approx \pm 1.96$ for a significance level $\alpha = 0.05$.

Since -1.96 < z-value < 1.96, we fail to reject the null hypothesis H_0

There is no significant difference in the proportion of individuals with a stress level > 6 between males and females at the 0.05 significance level.

- Construct a 95% confidence interval:

  The 95% confidence interval for the difference in proportions between the two populations is calculated as:

$$CI = (\hat{p_1} - \hat{p_2}) \pm z_{\alpha/2} \times \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$
$$CI \approx (-0.077, 0.277)$$

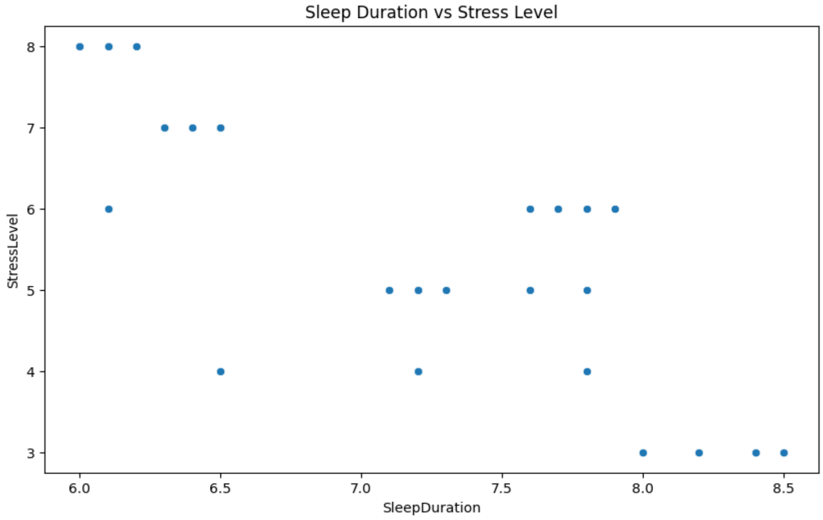
  The 95% confidence interval for the difference in proportions between males and females is (-0.077, 0.277), suggests that there is a difference between the proportion of males with Stress Level $>$ 6 and the proportion of females with Stress Level $>$ 6.

4.3 Regression analysis.

Sample data

Person ID	Sleep Duration	Stress Level
329	8.5	3
33	6.1	8
15	6.0	8
325	8.5	3
57	6.0	8
239	6.4	7
76	6.0	8
119	7.2	4
332	8.4	3
126	7.3	5
233	6.3	7
39	7.6	6
153	7.2	5
55	6.0	8
155	7.2	5
278	6.1	8
0	6.1	6
231	6.3	7
336	8.4	3
101	7.2	4
9	7.8	6
180	7.8	5
72	6.1	8
237	6.5	4
255	6.5	4
137	7.1	5
225	6.3	7
194	6.5	7
193	6.5	7
370	8.0	3
25	7.9	6
108	7.8	4
42	7.7	6
361	8.2	3
288	6.0	8
114	7.2	4
63	6.2	8
22	7.7	6
152	7.2	5
176	7.6	5

Scatter Plot



Sample Correlation Coefficient

The sample correlation coefficient is calculated as follows:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Using the provided data, the sample correlation coefficient is:

$$r \approx -0.819$$

Estimated Regression Line

The estimated regression line is found using the linear regression model:

$$Y = \beta_0 + \beta_1 X$$

Where:

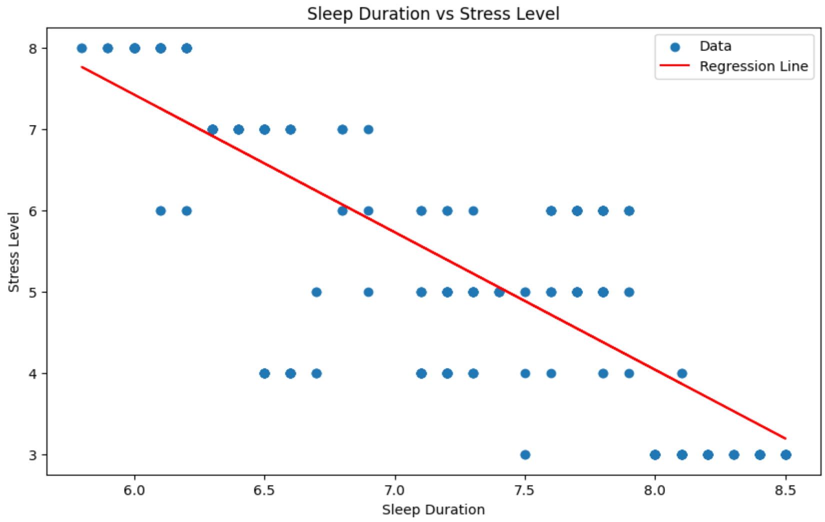
$$\beta_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Using the provided data, the estimated regression line is:

$$\hat{Y} = -1.692X + 17.579$$

Best Fit Regression Line



5. Conclusion

Hypothesis Testing

- There is insufficient evidence to conclude that the average sleep duration is less than 7 hours.

- The proportion of people with sleep duration greater than 7 hours is significantly higher than 50%.

Differences Between Genders

- Significant difference in sleep duration between males and females with stress levels greater than 6.

- No significant difference in the proportion of males and females with stress levels greater than 6.

Regression Analysis

- A strong negative correlation (0.819) between sleep duration and stress levels.

- The estimated regression line: $\hat{Y} = 1.692X + 17.579$