# Exploring the Daily Routines and Pressure Levels of STAT423 Students in Winter 2023: A Survey-Based Study

Fysal Beauferris  30048436
Fangru Guo 30095243
Ruixi You 30104198

# Survey Design, Sampling Methodology, and Data Collection

## 1. Survey/Questionnaire Design

To gain insights into the daily life of students who registered STAT423 in Winter 2023, we carefully designed a survey to collect information on various aspects of their routines. The target population is all students enrolled in STAT423 and the sampling unit is a student. We began by identifying key areas of interest, such as academic work, gaming, physical exercise, socializing, and sleeping, and crafted questions to assess the amount of time students dedicate to each of these activities. Additionally, we were interested in understanding how the number of courses a student takes might affect their perceived pressure level, so we included questions on both the number of courses they are currently enrolled in and their reported pressure level.

We also included a question to ascertain the university year of each student, as we believed this could provide valuable context for our findings. Finally, we asked about how often students plan their daily lives, as we hypothesized that this might be an important factor in their overall well-being.

After drafting the questions, we reviewed and revised them several times to ensure clarity and accuracy. We also pilot-tested the survey with a small group of students to identify any potential issues or areas for improvement. Based on their feedback, we made further refinements and arrived at the final version of the survey that we used to collect our data.

## 2. Sampling Methodology

We decided to use simple random sampling, as it is a straightforward method that allows us to randomly select participants from a list of the entire population (N=53) using R studio. Although we considered other sampling techniques like stratified and post-stratified sampling, we were unable to use them because we lacked information about the population's strata and their weights. Additionally, because our population is small, we didn't require cluster sampling or re-sampling, but we could have used Bootstrap to estimate the population mean for average sleeping time.

We hand out a pilot survey first, then get 6 responses to calculate the largest minimum sample size among the parameter we wish to estimate through this formula using moe=1:

$$n = \frac{N\sigma^2}{(N-1) * \left(\frac{moe}{Z_{\frac{\alpha}{2}}}\right)^2 + \sigma^2}$$

Alternatively, we could have employed systematic sampling, but given our minimum n of 25 and the need to pick at least one out of two, we felt that simple random sampling was the most suitable method.

## 3. Data Collection

During the administration of our survey, we encountered a nonresponse rate of 32.5%, which is relatively high. To mitigate this, we implemented various strategies at the initial stage of survey administration, such as emphasizing the importance of participants' responses and using clear and concise language in our survey questions. However, despite our efforts, we still faced a considerable nonresponse rate.

In terms of potential response biases, we did observe some indication of social desirability bias. Specifically, some participants tended to overestimate the amount of time they spent socializing. This suggests that participants may have felt pressure to present themselves in a positive light.

To handle nonresponses in our dataset, we opted to use mean imputation as there were only one or two missing values per question. Unfortunately, we did not have enough information to determine whether the nonresponses were missing completely at random (MCAR), missing at random (MAR), or non-ignorable, as we lacked knowledge of the overall distribution of students' year of the program. However, based on our observations, we hypothesize that individuals with more courses and higher year of program may be likely to leave nonresponses.

In our study, we conducted a sample size calculation to determine the minimum sample size needed to estimate each parameter of interest. Based on our calculations, the largest minimum sample size required was 25. To account for potential non-response, we decided to send out 40 surveys, which resulted in a final sample size of n=27.

However, if we had unlimited resources, we could have chosen to use a sample size equal to the population size N=n=53. This would have allowed us to obtain more precise estimates of the parameters of interest and potentially increase the statistical power of our study. Nonetheless, we recognize that increasing the sample size beyond what is necessary to obtain accurate estimates can be wasteful and may not always be justified, especially if the resources could be better utilized in other aspects of the study.

# Data Analysis

We have a total of 11 survey questions in our dataset that will be used for data analysis. However, before proceeding with any estimation, we must address the issue of missing data. In order to do so, we use the imputation method to replace the missing values with the mean of each column. Presented below are the first six rows of our data frame prior to the imputation of missing values:

| Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 |
|---|---|---|---|---|---|---|---|---|---|
| 2nd | Rarely | 7 | 7 | 2.0 | 4.0 | 5.0 | 7.5 | 18 | 5 |
| 4th | Never | 2 | 7 | 5.0 | 2.0 | 1.0 | 9.0 | 10 | 3 |
| 5th or above | Never | 1 | 7 | 10.0 | 4.0 | 5.0 | 7.0 | 8 | 1 |
| 4th | A few times a week | 5 | 1 | 5.0 | 4.0 | 1.0 | 6.0 | 13 | 4 |
| 3rd | A few times a week | NA | NA | NA | NA | 2.5 | NA | 0 | 3 |
| 4th | A few times a week | NA | NA | NA | 0.0 | 0.0 | 8.0 | 0 | 3 |

After addressing the issue of missing data in our dataset by imputation method, we are now able to present the first six rows of our data frame as follows:

| Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 |
|---|---|---|---|---|---|---|---|---|---|
| 2nd | Rarely | 7.00 | 7.00 | 2.0 | 4.00 | 5.0 | 7.5 | 18 | 5 |
| 4th | Never | 2.00 | 7.00 | 5.0 | 2.00 | 1.0 | 9.0 | 10 | 3 |
| 5th or above | Never | 1.00 | 7.00 | 10.0 | 4.00 | 5.0 | 7.0 | 8 | 1 |
| 4th | A few times a week | 5.00 | 1.00 | 5.0 | 4.00 | 1.0 | 6.0 | 13 | 4 |
| 3rd | A few times a week | 4.72 | 3.64 | 2.9 | 3.18 | 2.5 | 6.9 | 0 | 3 |
| 4th | A few times a week | 4.72 | 3.64 | 2.9 | 0.00 | 0.0 | 6.9 | 0 | 3 |

During the data analysis, we employed various estimation methods to estimate the population proportion and mean, including SRS estimation, ratio estimation, and regression estimation. The formulas used for the calculations and confidence intervals are as follows:

SRS estimation of population mean: $\overline{X} \pm t_{\frac{\alpha}{2},n-1} * \sqrt{\frac{S^2}{n}\left(\frac{N-n}{N}\right)}$

SRS estimation of population proportion: $\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N}\right)}$

Ratio estimation of population mean: $\mu_X * \left(\widehat{R} \pm Z_{\frac{\alpha}{2}} \sqrt{\left(\frac{1}{\mu_X^2}\right)\left(\frac{N-n}{N}\right)\frac{S_R^2}{n}}\right)$

The regression model: $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ , where $\widehat{\beta}_1 = r\left(\frac{S_Y}{S_X}\right)$ , $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$

Regression estimation of the "Grand" mean of Y: $\overline{Y}_{Regression} \pm t_{\frac{\alpha}{2},n-2} \sqrt{\left(\frac{N-n}{N}\right)\frac{MSE}{n}}$

**1.**

Based on the data collected from the survey question regarding how often students plan and organize their daily activities, we aimed to estimate the proportion of students who do this every day. Out of the sample size, it was found that 7 students plan and organize their daily activities every day. Using SRS to estimate the population proportion, we calculated a 95% confidence interval. Our analysis indicates that we can be 95% confident that the proportion of students registered for STAT423 in Winter 2023 who plan and organize their daily activities every day

falls between 0.1413 and 0.3772. Our analysis suggests that a total of 7 to 20 students registered for STAT423 in Winter 2023 are likely to plan and organize their daily activities every day.

Then we employed the bootstrap method for this analysis. We conducted 2000 iterations and the results show that we can be 95% confident that the proportion of students registered STAT423 in Winter 2023 who plan and organize daily activities every day is between 0.1111 and 0.4444. In comparison to the SRS method, our analysis suggests that the bootstrap method does not provide a more precise estimate of the population proportion.

**2.**

An intriguing hypothesis to explore is whether there is a relationship between students' pressure of study and their course load or year of program. To investigate this, we utilized three survey questions that ask about students' course load, year of program, and the pressure of study. To begin, we established the null hypothesis and alternative hypothesis. The null hypothesis posits that there is no significant relationship between students' pressure of study and their course load or year of program. The alternative hypothesis, on the other hand, proposes that there is a significant relationship between students' pressure of study and their course load or year of program. With these hypotheses in place, we conducted a Chi-Square Test to examine the association between course load and students' pressure of study.

```
> chisq.test(survey$Q11, survey$Q12)

        Pearson's Chi-squared test

data:  survey$Q11 and survey$Q12
X-squared = 37.663, df = 12, p-value = 0.0001743
```

The resulting test statistic was 37.663 and the associated P-value was calculated to be 0.0001743, which is less than the significance level of 0.05. Thus, we reject the null hypothesis and conclude that there is a significant association between course load and students' pressure of study.

Next, we utilized the same model to test the association between the year of program and students' pressure of study.

```
> chisq.test(survey$Q2, survey$Q12)

        Pearson's Chi-squared test

data:  survey$Q2 and survey$Q12
X-squared = 14.64, df = 12, p-value = 0.2617
```

The resulting test statistic was 14.64 and the associated P-value was calculated to be 0.2617, which is greater than the significance level of 0.05. Hence, we fail to reject the null hypothesis and conclude that there is no significant association between the year of program and students' perceived pressure of study at the 0.05 significance level.

**3.**

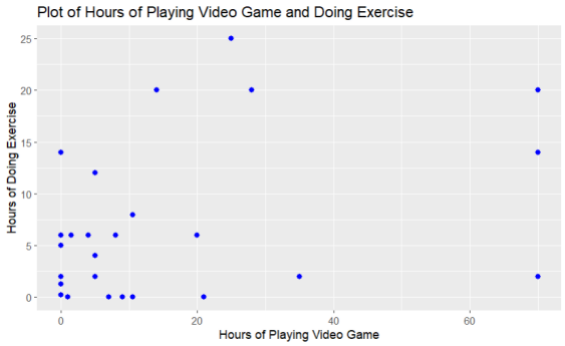
Our next goal is to estimate the average number of hours that students who registered for STAT423 in Winter 2023 dedicate to academic work each day throughout the semester. This includes attending classes and completing assignments. To accomplish this, we employed the SRS estimation of the mean. Based on our statistical analysis, we are 95% confident that the true average number of hours devoted to academic work each day during the semester, including classes and assignments, for students registered in STAT423 in Winter 2023, is between 3.5628 and 5.8772. Additionally, the standard deviation for this estimation is 2.9810. These findings provide us with valuable insights into the academic habits of students enrolled in this course and can inform future studies in this area.

**4.**

We have a survey question asking the days of playing video games in a week, and a question asking the hours of playing video games in a day. We calculate the hours of playing video games in a week by multiplying these two data. Using the SRS estimation of the mean, we can confidently estimate that the average number of hours spent playing video games per week for students registered in STAT423 during Winter 2023 is between 7.0770 and 24.0053 hours.

Furthermore, there are also two questions asking how many times take part in physical exercise or activity per week and how many hours take each time to take part in physical exercise or activity. We can also calculate the hours of taking part in physical exercise or activity in a week by multiplying these two data. We estimate that the average number of hours spent engaging in physical exercise or activities per week for these students is between 4.3434 and 10.1308 hours, also with a standard deviation of 7.4541.

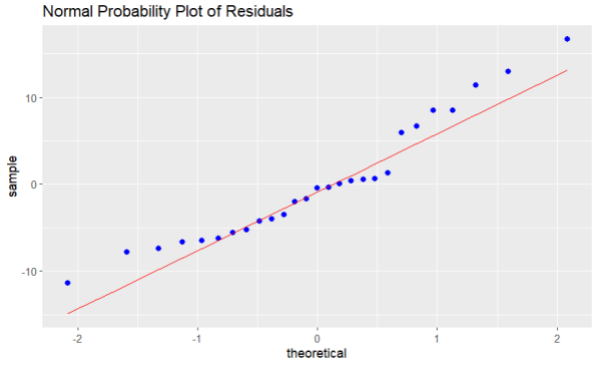
To improve the accuracy of our estimation, we attempted to use ratio and regression estimation. We found that the correlation between the hours of playing video games per week and the hours of engaging in physical exercise or activities per week was 0.3297, as shown in the scatter plot below:

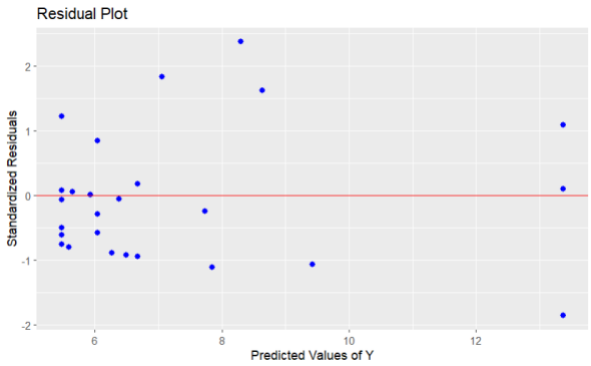



We use hours of playing video games per week as the subsidiary variable to estimate the mean hours of doing physical exercise or activities per week. Since we don't have the actual mean hours of playing video games per week, we assume it to be 16 hours, which is close to the

sample mean of 15.5412. With ratio estimation, we find that the 95% confidence interval of the mean hours spent in physical exercise or activities per week for students enrolled in STAT423 in winter 2023 is between 4.6956 and 10.2058.

To perform regression estimation, we used the model Y = 5.485496 + 0.1127078X, where Y represents the mean hours of engaging in physical exercise or activities per week and X represents the hours of playing video games per week. To check the conditions of regression, we used both the Normal Probability Plot of Residuals and the Residual Plot:



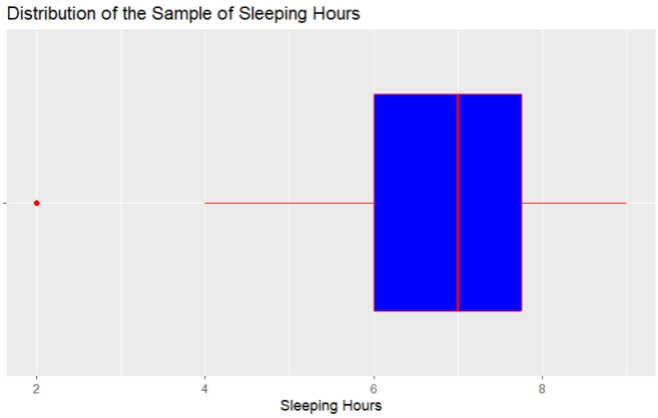

From the normal probability plot, it appears that the residuals follow a roughly linear distribution, indicating that the normality assumption is met. Additionally, the rectangular shape of the residual plot suggests that the assumption of homoscedasticity is also met.

Using regression estimation, the 95% confidence interval of the mean hours of doing physical exercise or activities per week for students registered STAT423 in winter 2023 is somewhere between 5.2964 and 9.2811.

When comparing the two estimation methods, we found that the MSE of the regression model was 51.5068, while the variance of the ratio estimation was 108.7533. This gives an empirical relative efficiency of 0.4736, indicating that there was more variation in the ratio estimation. Therefore, compared to the SRS estimation of mean, the regression estimation produced the most precise result.

**5.**

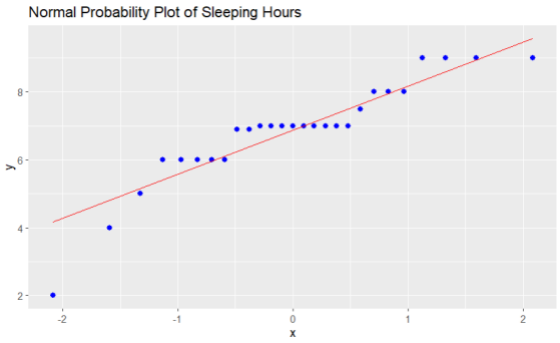
The last estimation is the students' average sleeping hours. We obtained numerical data from a survey question regarding the students' sleeping hours. Using the SRS estimation of mean, we estimate with 95% confidence that the average number of hours spent sleeping per day for students registered in STAT423 in winter 2023 is between 6.2649 and 7.4616, with a variance of 2.3755. To visualize the distribution, we created a box plot:

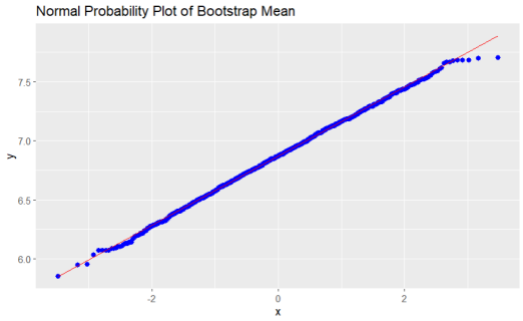



The box plot indicates that the distribution of sleeping hours is concentrated between 6 to 8 hours. Furthermore, the median of the sleeping hours is 7, which suggests that half of the students registered for STAT423 in winter 2023 slept for more than 7 hours per day on average, while the other half slept for less than 7 hours.

We used a bootstrap method with 2000 iterations to estimate the average sleeping hours of students registered in STAT423 during Winter 2023. To assess the validity of the method, we created two normal probability plots that depict the distribution of sleeping hours before and after applying the bootstrap.

The first plot shows the original distribution of sleeping hours. The second plot shows a nearly normal distribution of the resampled means obtained from the bootstrap method.




The resulting 95% confidence interval for the mean average hours spent sleeping in a day for students registered in STAT423 for Winter 2023 is between 6.2705 and 7.4225. It is worth noting that the bootstrap method provides a slightly more precise estimation than the SRS estimation.

# Conclusion

In conclusion, this report has presented a comprehensive analysis of the behavior and habits of students registered in STAT423 in Winter 2023. Through various statistical methods, we have estimated the proportion and mean of the time spent by the students on different activities, such as studying, playing video games, engaging in physical exercise or activities, and sleeping. Our findings suggest that the students tend to spend more time on studying and playing video games compared to other activities. Additionally, we have explored the relationship between the students' pressure of study and their course load or year of program, and our hypothesis test results indicate a significant association between the students' pressure of study and course load. Overall, the results of our analysis provide valuable insights into the behavior and habits of the students and can be used for further analysis and decision-making.