The Relationship Between Workload and Living Quality of Undergraduates in UW Madison

STAT 302: Accelerated Introduction to Statistics, LifeExperienceMatters

Jiaer Zhang, Group Leader
Jiayi Gao
Hao Li

December 20th, 2019

Contents

Abstract

This study aims to find the relationship between workloads and life quality of Chinese international undergraduate students at the University of Wisconsin-Madison. The data is collected through an online survey. The data set includes the measure of learning intensity and the measure of life-quality of the students. Specifically, 257 cases are selected randomly to be the sample of this study. Additionally, four tests including randomization test, t-test, chi-square test, and linear regression, are performed to test the relationship between different variables. The explanatory variables used in these tests are ages of students, credits students take, hours of classes students take, the school year of each students, gender, major, and whether students are taking different levels of courses. The response variables used are the study hour, activity time, hours of sleeping, whether each student wakes up at night, and what kind of activity each student participates in. The general hypothesis for the whole study is that the high stresses affect the quality of life of each student. However, since different students have different personalities and lifestyles, further studies regarding professional psychology knowledge are needed.

Introduction

Background Information

Undergraduate students value the quality of their college life, which generally refers to physical health, comfort, and happiness experienced by the student. However, due to the high pressure and study workload of undergraduate education, many college students could not balance their college life in a satisfactory way.

One study showed that stress and workload can negatively influence the quality of life, which is the main element of well-being, of students (Weinstein & Laverghetta, 2009, p1). For instance, because of the high-level workload and stress, undergraduate students' mental health can be harmed. In some severe cases, students even have dangerous ideas such as suicide. One study showed that close to 20 percent of college students in China report suicide ideas and the range varies from 6%–39.2% in other countries (Zhang et al., 2012, p.2). As a result, it is significant to investigate the relationship between international undergraduate students' study workload and the quality of their life. Different amounts of undergraduates course workloads are determined by different majors. Specifically, the stress level from various majors can affect undergraduate students differently. Researchers have conducted one study based on undergraduate students randomly sampled from all U.S. universities, and the result indicated that "hard" science majors experienced significantly more perceived stress than "soft" (relatively easy) majors (May & Casazza, 2012, p. 1). Those who major in Engineering, Computer Science, and Nursing are the students who have a higher possibility of having more work and stress (Robinson, 2009, p.2). Moreover, Engineering and Computer Science department has the highest perceived stress at (15.4%) among students with Nursing coming in second at (14.1%) and Science and mathematics coming in third at (11.8%). High levels of workloads occupy undergraduate students a large amount of after-class time. To be specific, undergraduates have to give up some amount of their recreation time and even sleep time to complete high-leveled workloads. This indicates these undergraduates having lots of course workloads usually fail to achieve their expectations and goals about sleep hours and extracurricular life, perceiving a lower quality of life. In contrast, college students who have participated in extracurricular activities normally frequently join extra-curricular activities and meet their expectations often have a high quality of their college life.

Different from previous studies focusing on a broad view of the workloads and students' stress level, this study concentrates more on the stress level for different majors and tries to find whether recreational activities can help to reduce the perceived stress and improve students' life quality.

Broader Impact and Greater Significance

The target audience of this study is all Chinese international students. The results of this study could be helpful for many Chinese international undergraduates in the United States. The University of

Wisconsin-Madison is a typical American public research institute which offers various majors, and the academic requirements for the different major and learning environment of UW-Madison could at least provide a rough representation of other universities, Therefore, the target market for this study can be all other public American universities.

By finding meaningful relationships between workloads and the student's life quality, this study can provide pre-college students and undergraduates with undecided majors some insights about their major selections based on their learning abilities. For universities, the ideal educational outcome is to help the students find the most suitable major and prepare them for the future work field that they are interested in. From the students' perspective, they might want to choose a suitable major with manageable workloads to experience an ideal college life. The result of the research can be informative for undergraduate students to compare different majors and help them to balance academic studies and social life.

Statement of Purpose

Purpose

The purpose of the study is to find whether there is a meaningful association between the workloads of Chinese international undergraduate students and the quality of these students' life. Precisely, we want to find out whether high workloads are associated with negative quality of life for students.

Specific Aims and Objectives

The specific aim is to provide suggestions to Chinese international undergraduate students, whose college life quality does not meet their expectations, about the way to improve their life quality and provide some insights to Chinese international pre-college students on their major selection. The center challenges of this study are collecting data without too many biases.

Research Plan

The plane is to conduct an observational study. By randomly selecting samples using the methods of stratified sampling and cluster sampling, the researchers collect data from representative samples, then try to find meaningful associations between explanatory and response variables and identify possible biases for this study. Based on the association found in the study, researchers can generalize this result to the population of interest, which is all Chinese international students at the University of Wisconsin-Madison.

Methods

Data Collection

Population of Interest

The population of interest for this experiment is the Chinese international undergraduate students in the University of Wisconsin - Madison. The samples are expecting around 250 Chinese international undergraduate students who take our survey.

Type of Study

This is an **observational study** because we do not actively interact with the participants. We only collect data from them and perform data analysis. We will not derive causation from this study, but rather to identify an association between the variables. Because we want to collect information about students' living habits such as sleeping hours, studying hours and recreational time, it is very difficult to collect in an experimental setting. Even if we can collect them in an experimental setting, it is impractical to do since it is very expensive and time-consuming. Instead, an observational study can allow researchers to real-life and more accurate data compared with an experimental study.

Survey Protocol

Type of Survey

This survey is a cross-sectional survey because the fundamental features of cross-sectional survey are met. The survey is conducted at a single point in time and the interest is the experience of students at the present time. It is not aimed to keep track of a curtain population for a period of time. This survey also makes comparing and analyzing different variables simultaneously feasible. Inference could be drawn from the data of different variables collected. Therefore, this survey is classified as cross-sectional survey.

Location and Time

There are not specific restrictions on the location and time on this survey due to the nature of the data collection approach. The data is collected through internet questionnaires. The respondent could access to the survey anytime and at any location. The challenge the randomization procedure for sending the survey which will be discussed in the following section.

Survey Method

We used stratified sampling to randomly select our cases. Since Chinese international undergraduates, who widely use a social media named We-chat to communicate, have different chatting groups in We-chat based on different year-of-university, we have four stratum (year of university): freshmen, sophomore, senior, and junior We-chat chatting groups. In each strata, labeling Chinese international students of this strata from 1 to n (n is the number of individuals in this strata). Then, Using randInt(1,n) to generate m different individuals $m = samplesize \cdot \frac{stratasize}{populationsize}$ and select these m different individuals, and repeat this procedure for every stratum and add the selected cases to get a sample of size 4m.

Randomization Protocol

Confounding Variable

Some possible confounding variables include the individual's habits. For example, some students do not sleep as much as others and their average sleep hours may be less than the majority students no matter of their study workload

Season and weather in Madison could be another confounding variable. More specifically, in summer, the weather is nice and students are more likely to have outdoor activities such as hiking. This may increase extracurricular activity hours. In winter the weather is cold and the snows heavily, students might tend to stay in libraries or at home to study rather than go outside. This may decrease the extracurricular activity hours.

Representative of The Population

The researchers should balance the weight and proportion of each strata.

Possible Bias

The possible biases for this study include sampling bias and non-sampling bias.

The non-sampling biases includes: response bias, which occurs when the respondent that do not answer the questions according to their true situations. Question and wording bias, which is another non-sampling bias caused by the ambiguous words in the survey that might cause people giving unexpected answers.

Sampling bias is caused by insufficient randomization procedure. For example, in this study, sampling bias might be caused by: too many samples have the same sex, too many samples are from the same year of the school, too many samples are selected from people we are familiar with, too many samples from the same majors, samples are selected from the people who use "Wechat".

Reduce Bias

To reduce the response bias we have in this study, researchers could use contradictory questions in the survey. For example, researchers could ask both "Is your sleeping time above 7 hours" and "Is your sleeping hour below 7 hours" in the survey. To be more specific, the researchers could put the questions on different pages so that the sample who takes the survey has a higher possibility of telling the truth. By comparing the answers, the researchers can tell if the sample is lying or not.

To reduce the question and wording bias, the study asks Professor of Stat 302 to modify the wording in each question.

To reduce the sampling bias, this study can attempt to collect as much of the data as possible. Then the researchers could adjust the bias to some extent by stratification. For instance, in this survey, the researchers could stratify the samples by years of school such as freshmen, sophomore, junior and senior, and then according to the stratification to modify the weight of each group.

Sample Size Determination

For sample size determination at this stage, methods of a single mean is applied because this quantitative variable has the most variability. For the categorical variables, the margin of error of 0.2 and a confidence level of 95% is applied. The following is one example using variable "SleepHour" to estimate the desired sample size. The result is 96 in this case. (Code is attached in Appendix)

List of Variables of Interest

Our data include 17 variables: The specific variables are described below

We plan to collect our data through online surveys like "Qualtric". We will send out emails to Chinese international undergraduates and send links to the survey in different chatting groups. Moreover, We will also hand out paper surveys to Chinese international undergraduates at different locations on campus at different times.

Variable Table

Name of variable	Explanatory/Response	Categorical/Quantitative	Unit of Measurement
Age	Explanatory	Quantitative	year-old
Credits	Explanatory	Quantitative	credits
classHour	Explanatory	Quantitative	hours
StudyHour	Explanatory	Quantitative	hours
SchoolYear	Explanatory	Categorical	
Gender	Explanatory	Categorical	
Major	Explanatory	Categorical	
Elementary	Explanatory	Categorical	
Intermediate	Explanatory	Categorical	
Advanced	Explanatory	Categorical	
AcademicHour	Response	Quantitative	hours
ActivityTime	Response	Quantitative	hours
SleepTime	Response	Quantitative	hours
Fallasleeptime	Response	Categorical	
WakeupTime	Response	Categorical	
Activity	Response	Categorical	
ActivityType	Response	Categorical	

Data Dictionary

- **Age**: The age of each sample.
- **Credits**: the number of credits each sample.
- **classHour**: Hours of sample spend on attending classes on average per day.
- **StudyHour**: Hours of sample spend on studying on average per day.
- **SchoolYear**: The schoolyear of each sample (e.g Freshmen, Sophomore, Junior, Senior).
- **Gender**: The gender of each sample. (e.g, Male, Female)
- **Major**: The primary major that each sample takes: major category (e.g, STEM, Business, Social Science, Liberal Art, Agriculture, Education)
- **Elementary**: Whether each sample has taken any elementary coursework at UW–Madison: Yes or No
- **Intermediate**: Whether each sample has taken any intermediate coursework at UW–Madison: Yes or No
- **Advanced**: Whether each sample has taken any advanced coursework at UW–Madison: Yes or No
- **AcademicHour**: Hours of sample spends on academic life (both attending courses and studying) on average per day.
- **ActivityTime**: Hours that each sample spends on extracurricular activities on average per day.
- **SleepTime**: The number of hours of sleeping for each sample on average per day
- **Fallasleeptime**: Time that each sample takes to fall asleep (e.g less than 10 minutes; less than 20 minutes; less than 30 minutes; less than 1 hour; more than 1 hour)
- **WakeupTime**: Times of waking up at night for each sample (e.g 0;1;2;3; more than3)

- **Activity**: Whether each sample attends any extracurricular activities: Yes or No
- **ActivityType**: Types of extracurricular activities (e.g club, part time job, etc)

Statistical Method

This project focuses on two tests including a randomization test and a t-test to analyze these variables.

Firstly, a randomization test for difference in the mean is used to test whether there exists a significant difference between two means. In this study, the null hypothesis would be $H_0 : \mu_1 = \mu_2$ and the alternative hypothesis would be $H_a : \mu_1 \neq \mu 2$. To test if the null hypothesis is true, first, shift the sample means for each group such that the sample means are equal, then randomly sampled with replacement from the independent samples for each of the two groups and compare the randomization means. Repeating these steps 10000 times to build a symmetric and bell-shaped randomization distribution that centered at 0. In this case, the statistic samples are forced to be normal. After the randomization distribution of proportions is constructed, this study uses it to find the p-value by checking the proportion of the mean difference in randomization distribution that is more extreme than the original sample statistic $\mu_1 - \mu_2$. At last, if the p-value is less than the significance level α (usually 0.05), the study has enough evidence to reject the null hypothesis and conclude that the difference between the two means is greater than zero.

Secondly, to test if there is a difference in means one variable in different categories, a t-test for the difference between means is conducted. The null hypothesis for this test is $H_0 : \mu_1 - \mu_2 = 0$, and the alternative hypothesis is $H_a : \mu_1 - \mu_2 \neq 0$. However, before executing a t-test, this study has to check if the sample sizes for the variable in both categories are large enough, which means both the sample sizes have to be larger than 50. If the sample sizes are not large enough, the underlying distribution has to be approximately normal (no extreme outliers and skewness) to apply the Central Limit Theorem to estimate the sampling distribution. In order to find the p-value for the one-sided t-test, this study needs to calculate the area under the t-distribution to the right of the test statistics. Finally, if the p-value is less than the significance level α (usually 0.05), the study has enough evidence to reject the null hypothesis and conclude that the difference in the mean is greater than zero.

Thirdly, to test if there exists a significant association between two categorical variables, a chi-square test for association is performed. The null hypothesis for this test are: H_0 : there is no association between the two categorical variables. H_a : there is an association between two categorical variables. Before performing chi-square test, the study has to check if the expected count of each cell is greater than 5. After checking the expected count, the study can put the expected count into a two-way table that exactly matches the null hypothesis that there is no relationship. The remaining task is to calculate the chi-square statistic by using the formula $\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$. The degrees of freedom are calculated by using $(row - 1) \cdot (column - 1)$, This study can find the degree of freedom. The p-value for the test, which is the area that is greater than the test statistic under the χ^2 distribution with $(row - 1) \cdot (column - 1)$ degrees of freedom. Finally, if the p-value is less than the significance level α (usually 0.05), the study has fairly strong evidence that there is an association between the two variables.

Finally, a linear regression test for slope is conducted to test if there is a linear relationship between two quantitative variables. The t-test for slope is appropriate if the residuals are normally distributed, which is reflected in the *Normal Q-Q plot*. The *Residual V.S. Fitted plot* should show linearity. Also, the *Scale-Location plot* should show constant variance. If these conditions are met, the data is suitable to perform a t-test for slope. The hypothesis for this test is defined as $H_0 : \beta_1 = 0$ and $H_\alpha : \beta_1 \neq 0$. The t-test statistics: t and the confidence interval CI are calculated using the formula $t = \frac{b_1 - 0}{SE}$ and $CI = b_1 \pm t^* \cdot SE$ Since this is a two-tailed test, After obtaining the test statistics, the p-value is calculated by finding the area under the t-distribution curve with $n - 2$ degrees of freedom to the right of the t-test statistics($t > 0$) and to the left of the -t ($-t<0$). If the p-value is lower than the given significance level $\alpha = 0.05$ in this study, there is enough evidence to reject the null hypothesis and conclude the alternative hypothesis that the slope is not zero. There exists a linear relationship between the two quantitative variables.

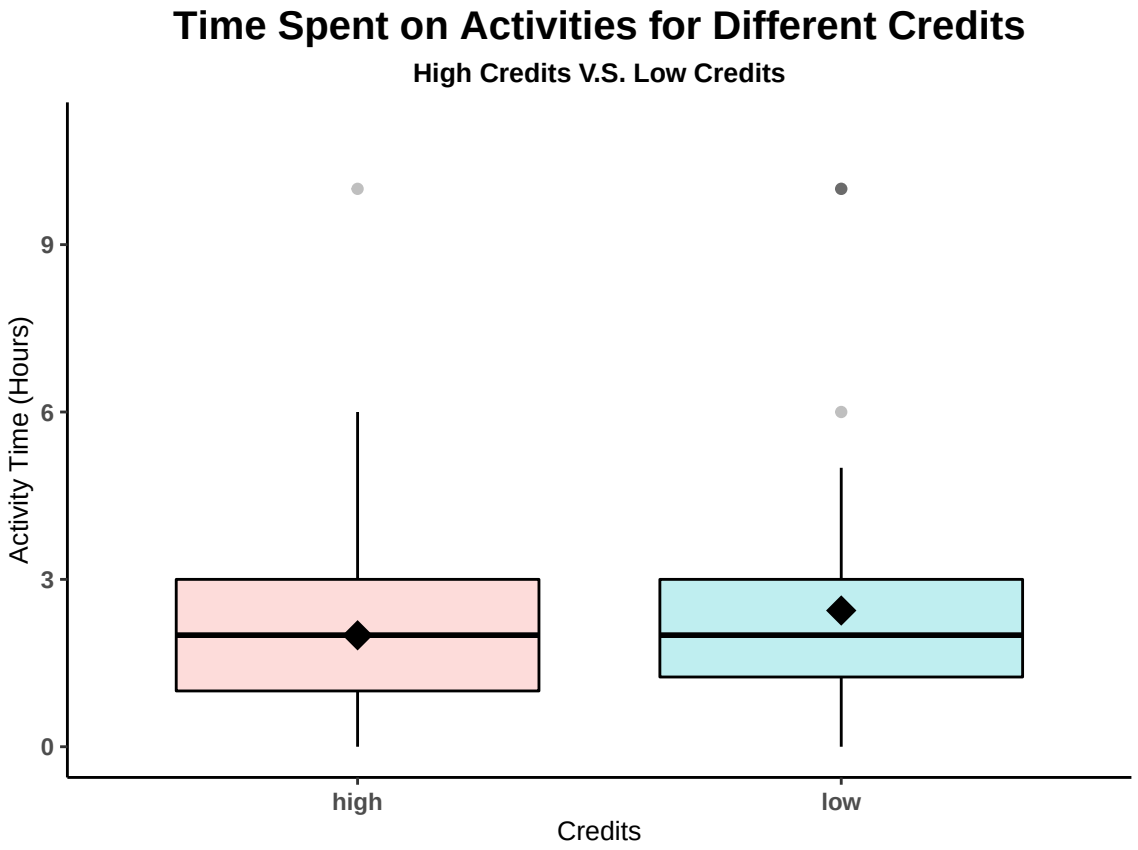
This study uses R studio as the technique to do the analyzing and graphing. Various types of packages in R studio are applied in this study including ggplot2, dplyr, knitr, ggfortify, tidyverse, reshape2,and xtable.

Results

Randomization Test

I.Difference in Mean: Credits and Extracurricular Activities

To determine whether credits taken will affect the quality of life for undergraduate students, this study examines is there an association between number of credits taken by the student and the time they spend on extra curricular activities per day. This study divide students into two groups: Low credits group (0-15 credits) and High credits group (16-23 credits).



The boxplots indicate that students take less than 16 credits spend a little more time than students take more than 16 credits on extracurricular activities. In order to check whether this difference is statistically significant. The following sections conducted a randomization tests for difference between mean extracurricular activity time of students in two groups.

The null and alternative hypotheses are defined as follows: where The mean extracurricular activity time of student in High credits group is defined as μ_h. The mean extracurricular activity time of student in Low credits group is defined as μ_l.

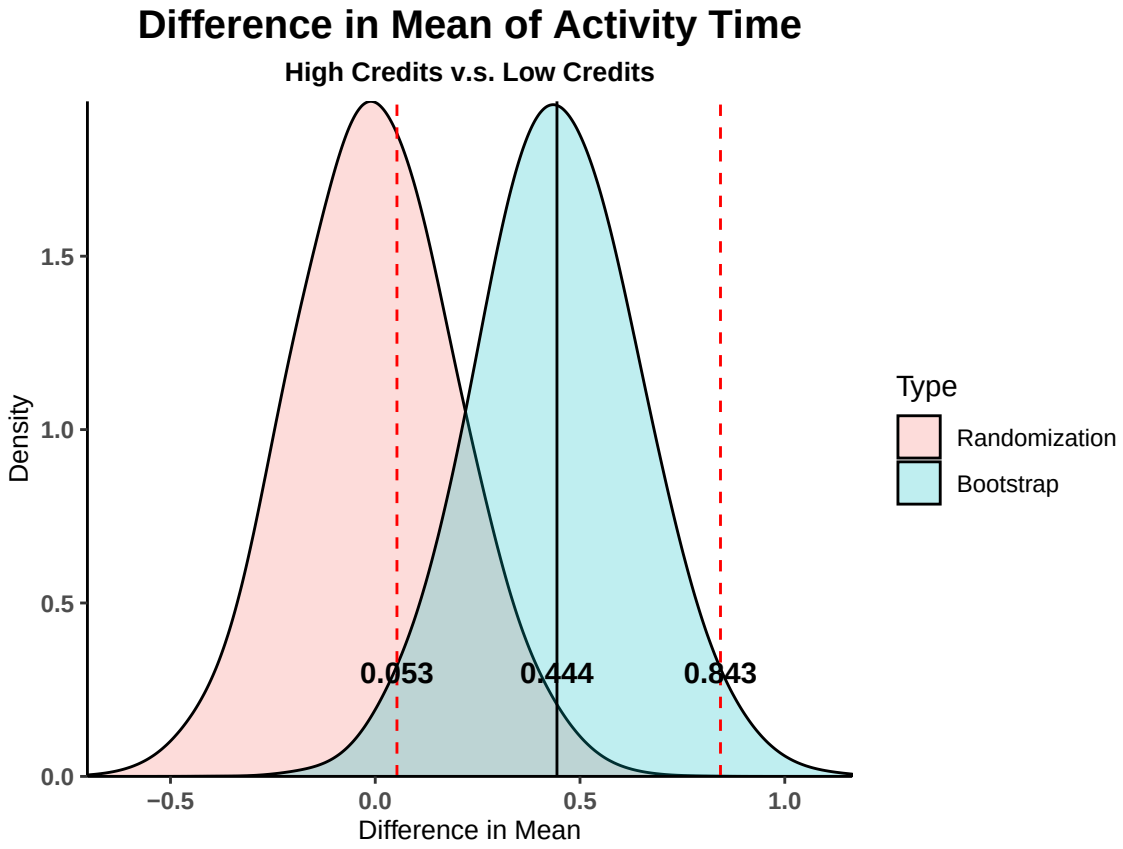
$$H_0 : \mu_l - \mu_h = 0$$
$$H_a : \mu_l - \mu_h > 0$$

Based on these hypotheses, the randomization distribution will be constructed assuming that the null hypothesis is true. This can be achieved by Shift the sample means for each group such that the sample

means are equal. Randomly sample with replacement from the independent samples for each of the two groups given the sample sizes and compare the randomization means. In this case, the sample size of Low credits group and High credits group are defined respectively as n_l and n_h.

In order to proceed analysis, the randomization distribution should be approximately symmetric and bell-shaped and centered at the null value 0. This study randomly samples 10000 times to create a relative symmetric and bell-shaped distribution.

Moreover, in order to create a confidence interval, a bootstrap distribution will be constructed. This can be achieved by sampling with replacement from the data of the two groups without shifting them. 10000 samples are drawn to make the bootstrap distribution relatively symmetric and bell-shaped and centered at the sample statistic.



The red distribution on the left is the randomization distribution, and the blue one on the right side is the bootstrap distribution. The randomization distribution is centered at 0 and is relatively symmetric and bell-shaped. The bootstrap distribution, it is centered at the sample statistic 0.4437022 (rounded to 0.444) and the distribution is approximately symmetric and bell-shaped.

The sample statistic, which is represented by the solid line on the graph, is $\overline{x}_l - \overline{x}_h \approx 0.444$. Here, $\overline{x}_l$ represents the sample mean sleep hours for *Low credits* group, and $\overline{x}_h$ represents the sample mean sleep hours for *High Credits* groups.

The red dashed line in the graph represents the 95% confidence interval. *One particular thing to notice is that even though this should be a one-tailed test based on the hypotheses, the two-tailed test is conducted here, since the confidence interval for two-tailed test is easier for interpretation.*

The 95% confidence interval for the difference in mean is calculated from the bootstrap distribution:

Table 2: 95% Confidence Interval

2.5%	0.053
97.5%	0.843

The **95% confidence interval** for this test is **between 0.053 and 0.843**.

Assuming the null hypothesis is true, the p-value measures how extreme the sample statistic is. On the graph, the p-value represents the area under the curve that is greater than 0.444. The p-value of this test calculated in R Studio is **0.0145**.

Because the p-value is 0.0145, which is smaller than 0.05 when $\alpha = 0.05$, there is enough evidence to reject null hypothesis and in favor of alternative hypothesis that students in Low credits group spend more time on extracurricular activities than students in High credits group. (one-tailed two means randomization test,$\overline{x}_h = 1.997761$, $\overline{x}_l = 2.441463$,$\alpha = 0.5$)

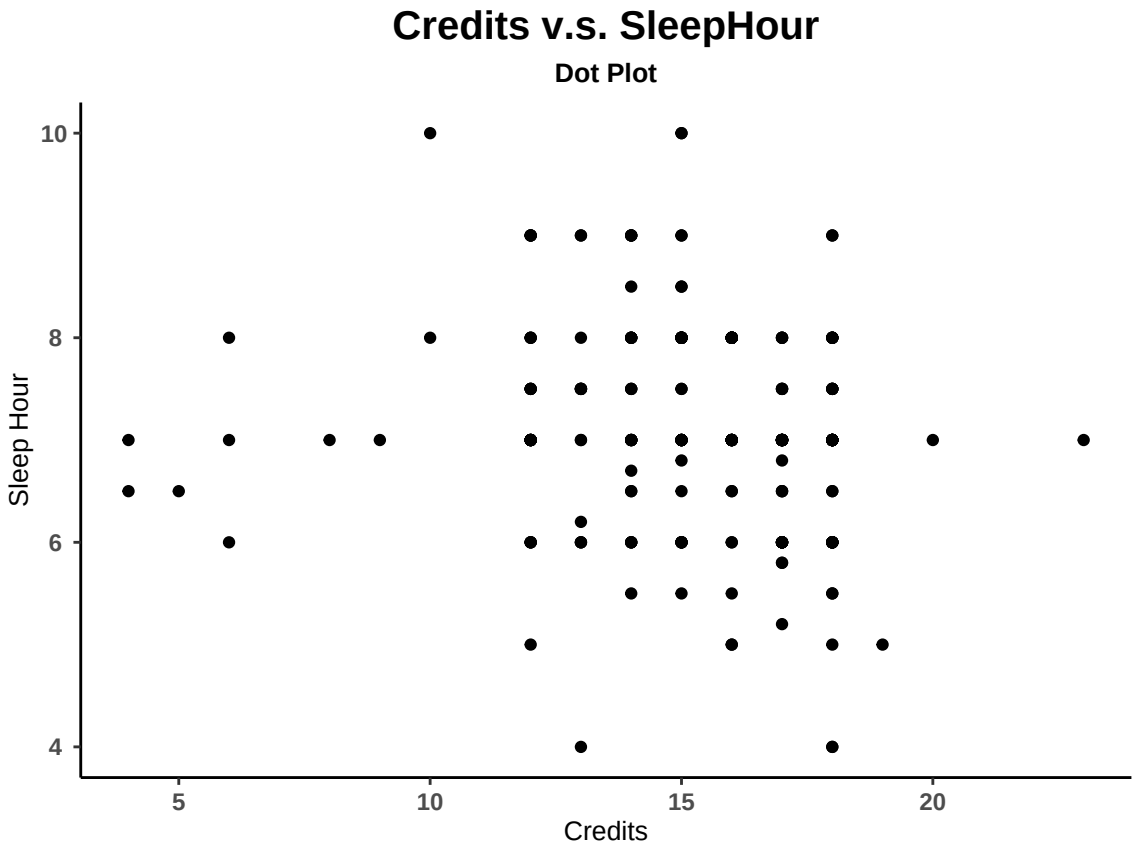
II.Randomization Test for Correlation: Credits Taken and Number of Sleep Hours

Hypothesis

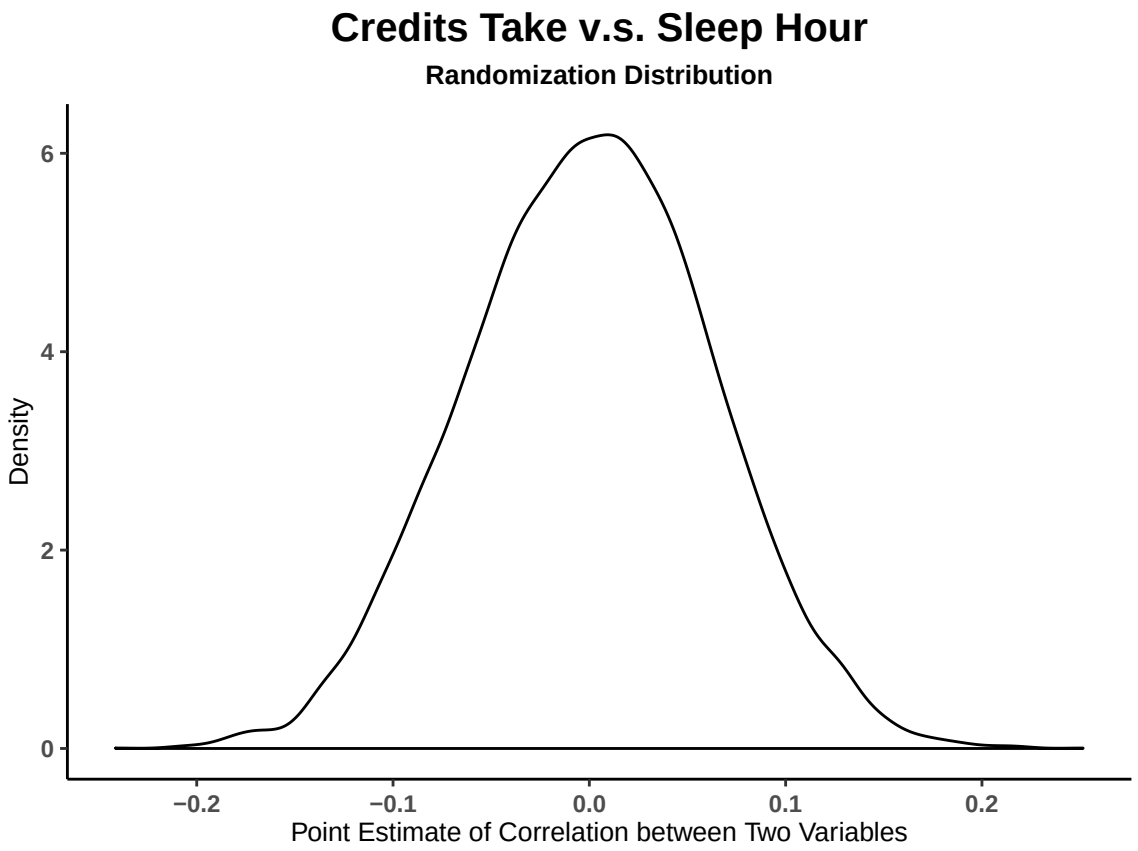
ρ is defined as the correlation between credits one student take for Fall 2019 semester and the average sleep hour for this student.

$$H_0 : \rho = 0$$
$$H_a : \rho < 0$$

Summary Figure



Generate Randomization Distribution



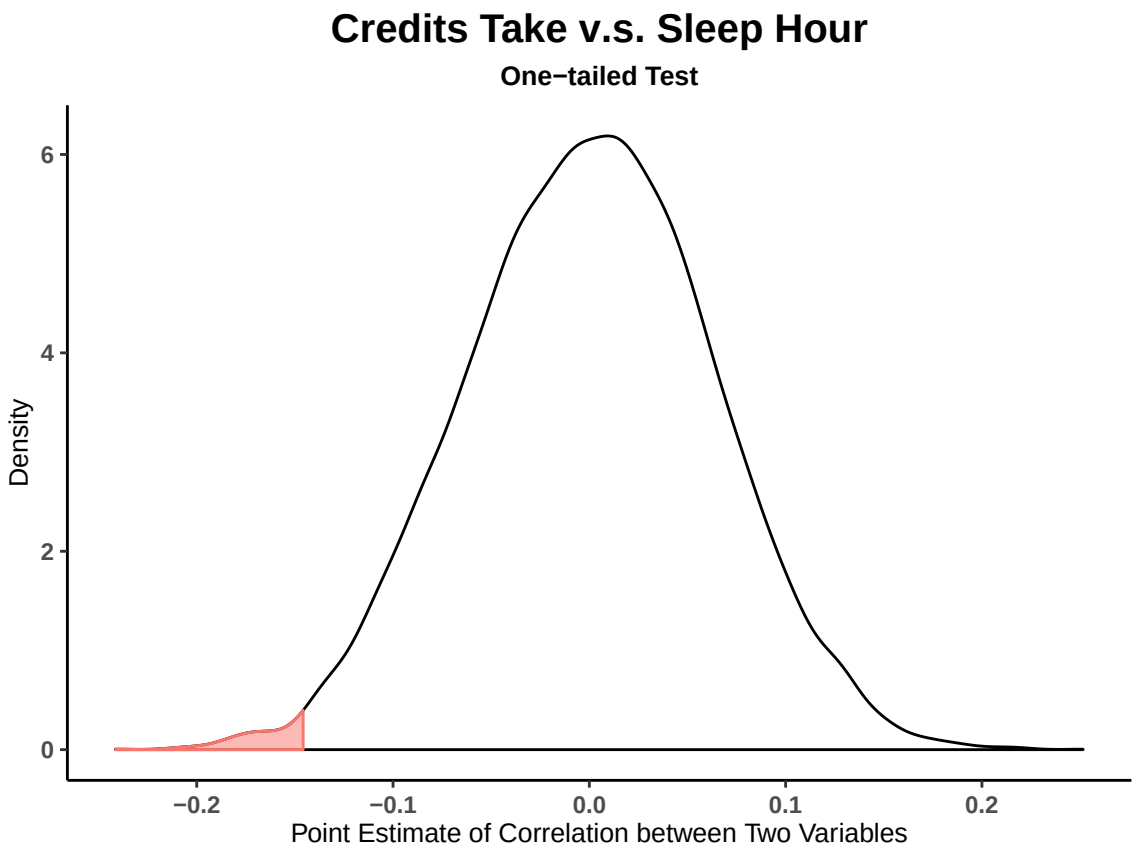
The randomization distribution is approximately symmetric and bell-shaped.

Calculate Test Statistic

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\overline{x})}{s_x}\frac{(y_i-\overline{y})}{s_y} = -0.1662545$$

where x_i is the credits took in Fall 2019 semester of i-th student and y_i is the sleep hours of i-th student. $\overline{x}, \overline{y}, s_x, s_y\$$ are sample mean and sample standard deviation of the corresponding variables.

Compute p-value



Since this is a one-tailed test, p-value is calculated by the portion of randomization distribution that assumes null hypothesis is true, which is the area less than -0.1662545. By using R, the p-value is calculated to be 0.0034. (The code for calculating p-value is included in the Appendix)

Interpretation

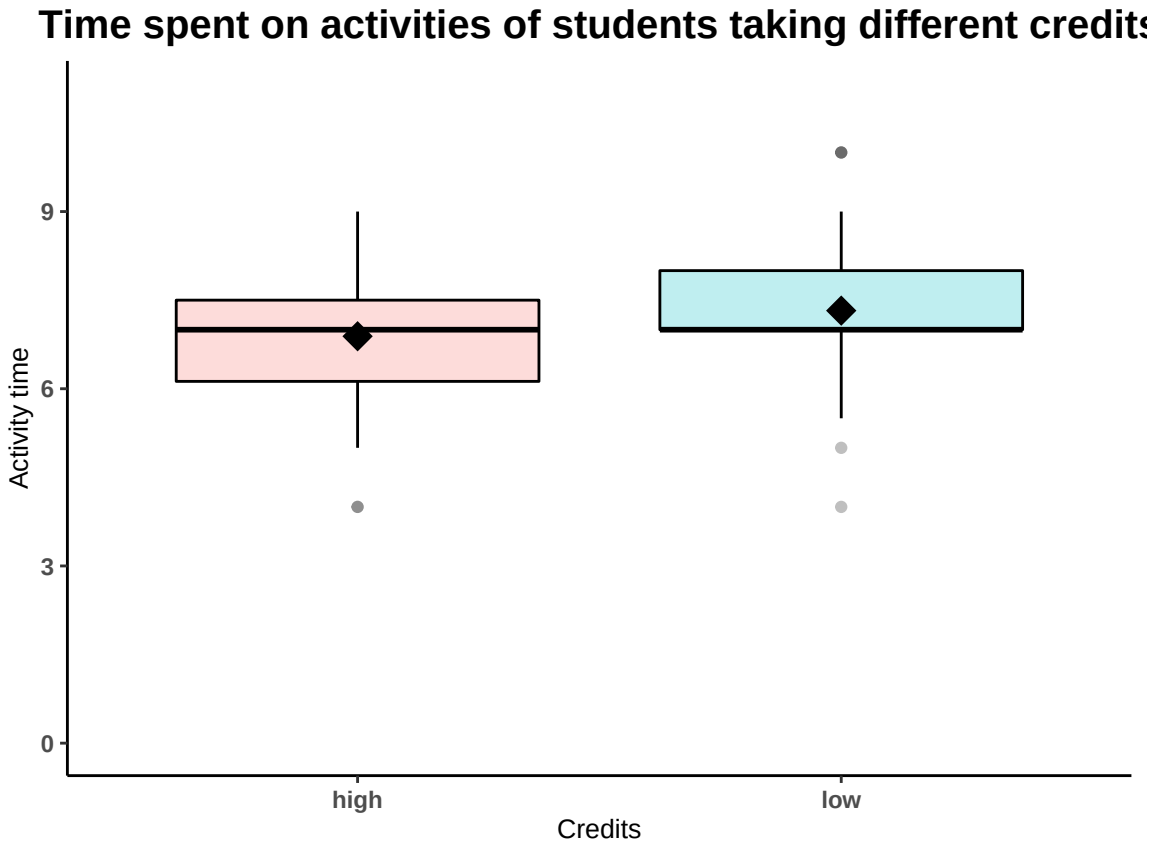
There is significant evidence showing that the credit took for one semester is negatively correlated with the sleeping hours that one student can get. (one-sided randomization test, $r = -0.1662545$, $p = 0.0034$, $\alpha = 0.05$).

Parametric Test

T-Test (Difference in Mean): Number of Credits and Sleep Hours

This study explores the association between number of credits taken in Fall 2019 semester and the average sleep hour per day. Because sleeping time is good indicator of the life quality, the test result could be useful for the the study.

For this test, the variable *Credits* is divided to two group, which is the same as the randomization test above. The *HighCredits* group is $credits \geq 16$, and the *LowCredits* group is $credits < 16$.



The box plots indicates that students in *Low credit* group sleep slightly more than students in *High credit* group. In order to check whether or not the difference is statistically significant, the following section conducts a parametric test of differences in means using the **t-distribution**.

Table 3: Summary of sleephours

Group	Size	Mean sleephour	SD sleephour
Highcredits	134	6.889	0.885
Lowcredit	123	7.323	1.052

Even though there are some outliers, a t-distribution can be used to conduct a two sample t-test, because

the assumptions for t-test are met: both samples are greater than 50 as $n_h = 134$ and $n_l = 123$, Where n_h is the sample size of *HighCredit* students and n_l is the sample size of *LowCredit* students. The mean sleep hours of *HighCredit* students is defined as μ_h. The mean sleep hours of *LowCredit* students is defined as μ_l. The standard error of sleep hours of *HighCredit* students is defined as s_h. The standard error of sleep hours of *LowCredit* students is defined as s_l.

The sampling distribution of difference in mean of sleep hours will follow a $T(\mu_l - \mu_h, \sqrt{\frac{s_l^2}{n_l} + \frac{s_h^2}{n_h}})$ distribution, where the degree of freedom will be $min(n_h - 1, n_l - 1) = 122$. Therefore, the 95% confidence interval will be calculated using the following formula. (Note: even the hypothesis tests performed are one tailed test, the confidence interval is a two-tailed interval). For calculating the confidence intervals, the critical value, t^*, would be the t-score at the 97.5 percentile with the degrees of freedom of 122, which is 1.98.

$$\begin{aligned} 95\%\ \text{CI} &= (\bar{x}_l - \bar{x}_h) \pm t^*_{122} \times \sqrt{\frac{s_l^2}{n_l} + \frac{s_h^2}{n_h}} \\ &= (\bar{x}_l - \bar{x}_h) \pm 1.98 \times \sqrt{\frac{s_l^2}{n_l} + \frac{s_h^2}{n_h}} \end{aligned}$$

Then, in order to conduct hypotheses tests, the t-score and the p-value will be calculated using the following formulas, depending on the direction of the alternative hypothesis.

$$t = \frac{\bar{x}_l - \bar{x}_h}{\sqrt{\frac{s_l^2}{n_l} + \frac{s_h^2}{n_h}}}$$

Hypothesis

It appears that students who take less credits tend to sleep more than students who take more credits. The hypotheses are following:

$$\begin{aligned} H_0 &: \mu_l - \mu_h = 0 \\ H_a &: \mu_l - \mu_h > 0 \end{aligned}$$

Table 4: Summary for t-Test

Mean.Lowcredits	Mean.Highcredits	SD.Lowcredits	SD.Highcredits	SE	t	df	p
6.889	7.323	0.885	1.052	0.122	3.562	239.411	2e-04

The study is 95% confident that there is a difference in the mean sleep hours between student in *HighCredits* and *LowCredits* group, $\mu_l - \mu_h$, falls within $(0.192, 0.675)$.

$$\begin{aligned} 95\%\ \text{CI} &= (\bar{x}_l - \bar{x}_h) \pm 1.98 \times \sqrt{\frac{s_l^2}{n_l} + \frac{s_h^2}{n_h}} \\ &= 0.434 \pm 1.98 \times 0.122 \\ &= (0.192, 0.675) \end{aligned}$$

By using the formula stated in previous part, the t-score is calculated to be:

$$\begin{aligned} t &= \frac{\bar{x}_l - \bar{x}_h}{\sqrt{\frac{s_l^2}{n_l} + \frac{s_h^2}{n_h}}} \\ &= \frac{7.323 - 6.889}{\sqrt{\frac{1.052^2}{123} + \frac{0.885^2}{134}}} \\ &= \frac{0.434}{0.122} \\ &= 3.56 \end{aligned}$$

Using this t-score, the p-value of this test is calculated.

$$\begin{aligned} T &\sim t_{122} \\ p - value &= P(T \geq 3.56) = 0.00026 \end{aligned}$$

This one tailed t-test for difference in mean number of sleep hours between students in *HighCredit* group and *LowCredit* group gives a p-value of 0.00026. Thus, this provides a strong evidence to reject the null hypothesis and in favor of the alternative that student who take more than 16 credits sleep less than the students who take less than 16 credits

Chi-Square Test: Gender and Major

The variable of the test are Gender, which is divided into two categories (Male and Female), and Major, which is divided into three categories (STEM, Business, and other majors). The question arises from the purpose of this research project whether gender will play a role in major selection. Therefore, the aim of the test is to test whether there is an association between gender and major choices.

Hypotheses

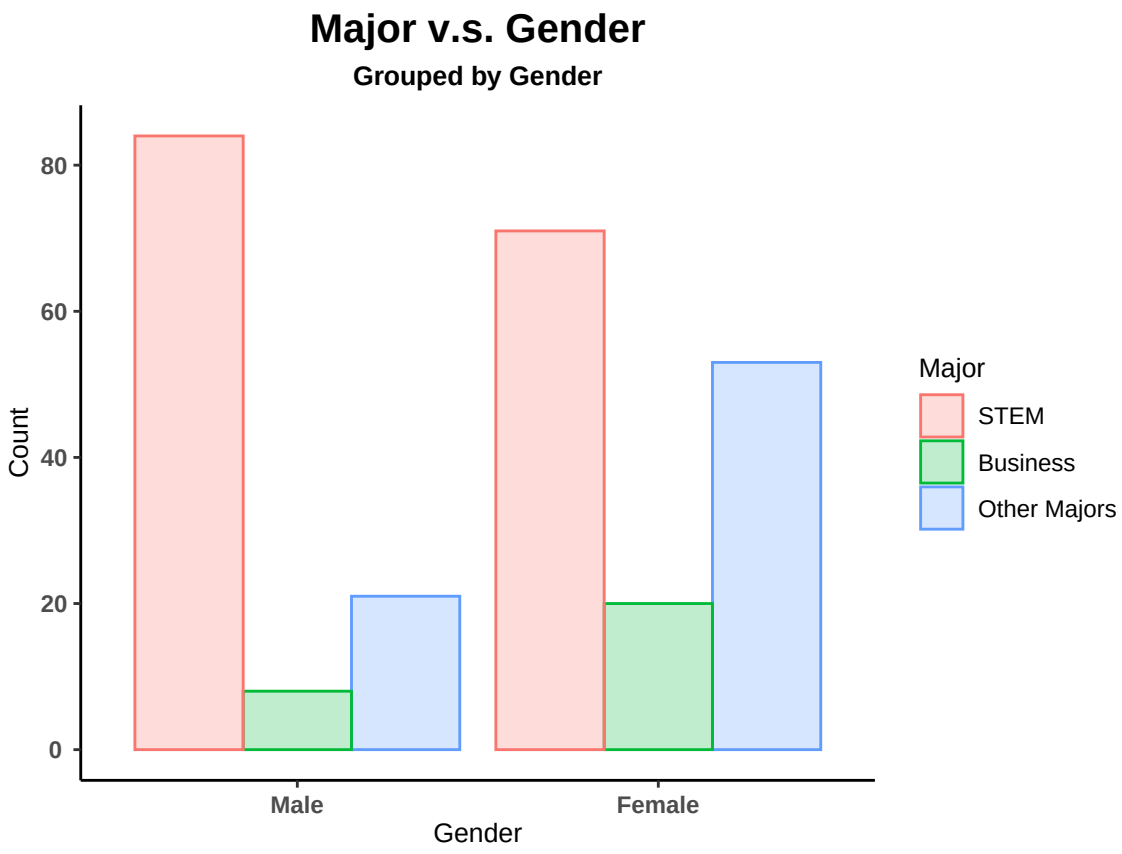
$$\begin{aligned} H_0 &: Major\ is\ not\ associated\ with\ Gender \\ H_a &: Major\ is\ associated\ with\ Gender \end{aligned}$$

Observed Count

Table 5: Observed Count Table (Major/Gender)

	Male	Female	Total
STEM	84	71	155
Business	8	20	28
Other Majors	21	53	74
Total	113	141	257

Summary Figure



Check for Assumptions

Table 6: Expected Count Table (Major/Gender)

	Male	Female
STEM	68.15175	86.84825
Business	12.31128	15.68872
Other Majors	32.53697	41.46303

From the table, all the expected counts are 5 or greater. Therefore, it is appropriate to use the χ^2-distribution with 2 degrees of freedom, $(r-1)\cdot(c-1)=2$.

Calculate Test Statistic

$$\begin{aligned}
X^2 =& \sum_{i=1}^{k} \frac{(x_i - np_{0i})^2}{np_{0i}} \\
=& \frac{(84 - 68.15175)^2}{68.15175} + \frac{(8 - 12.31128)^2}{12.31128} + \frac{(21 - 32.53697)^2}{32.53697} + \\
& \frac{(71 - 86.84825)^2}{86.84825} + \frac{(20 - 15.68872)^2}{15.68872} + \frac{(53 - 41.46303)^2}{41.46303} \\
&= 3.685408 + 1.509767 + 4.09078 + 2.89202 + 1.184748 + 3.210126 \\
&= 16.57285
\end{aligned}$$

Compute p-value

$$\begin{aligned}
X^2 &\sim \chi^2_{(r-1)(c-1)} \\
p - value &= P(\chi^2_{(r-1)(c-1)} \geq X^2) \\
&= P(\chi^2_2 \geq 16.57285) \\
&= 0.0002519135
\end{aligned}$$

Interpretation

There is significant evidence such that majors choices are associated with gender. (chi-square test for association, $\chi^2 = 16.57285$, $n = 257$, $df = 2$, $p = 0.0002519135$, $\alpha = 0.05$).

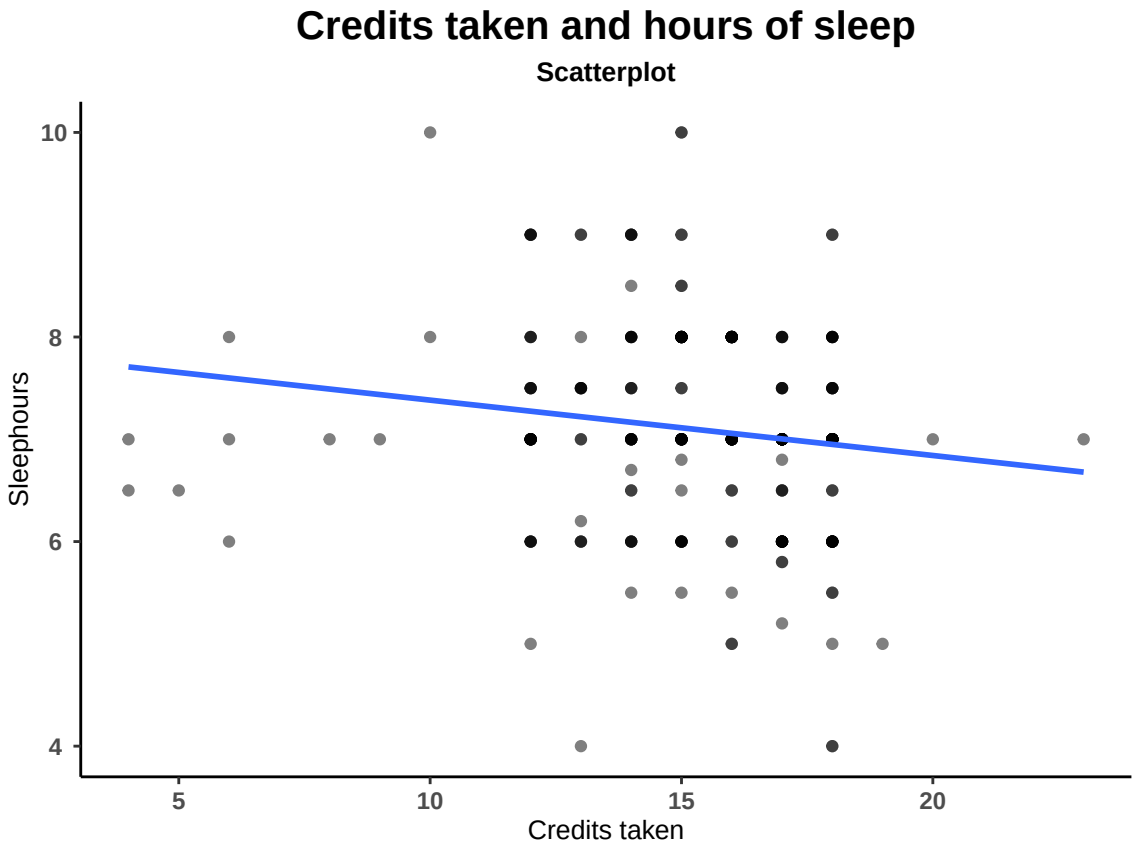
Linear Regression Test: Credits and Sleep Hours

Test Description

This test want find whether there are meaningful relationships between the credits taken by a students and his/her hour of sleep. To explore this idea, a linear regression test is applied.

Summary Plot

The following scatter plot describes the distribution of the credits taken and sleep hours of students.



Summary Tables

The following are the Anova Table and the summary table for the intercept and slope

Table 7: Anova Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Credits	1	5.311826	5.3118262	5.50672	0.0197094
Residuals	255	245.975022	0.9646079	NA	NA

Table 8: T-test Summary Table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.9241904	0.3579949	22.13493	0.0000000
Credits	-0.0541264	0.0230655	-2.34664	0.0197094

According to this scatter plot, the simple linear model seems to have a negative slope. In order to check if this negative slope is true for the population, a t-test for slope is conducted. The linear relationship is defined as follow $Y = \beta_0 + \beta_1 X + \epsilon$. The population slope between the number of credits and sleep hours is defined as β_1, and the hypotheses of this test are defined as follows.

Hypotheses

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

Compute Confidence Interval

In order to provide the reader a range that captures the true slope, this study construct 95 confidence interval for the slope.

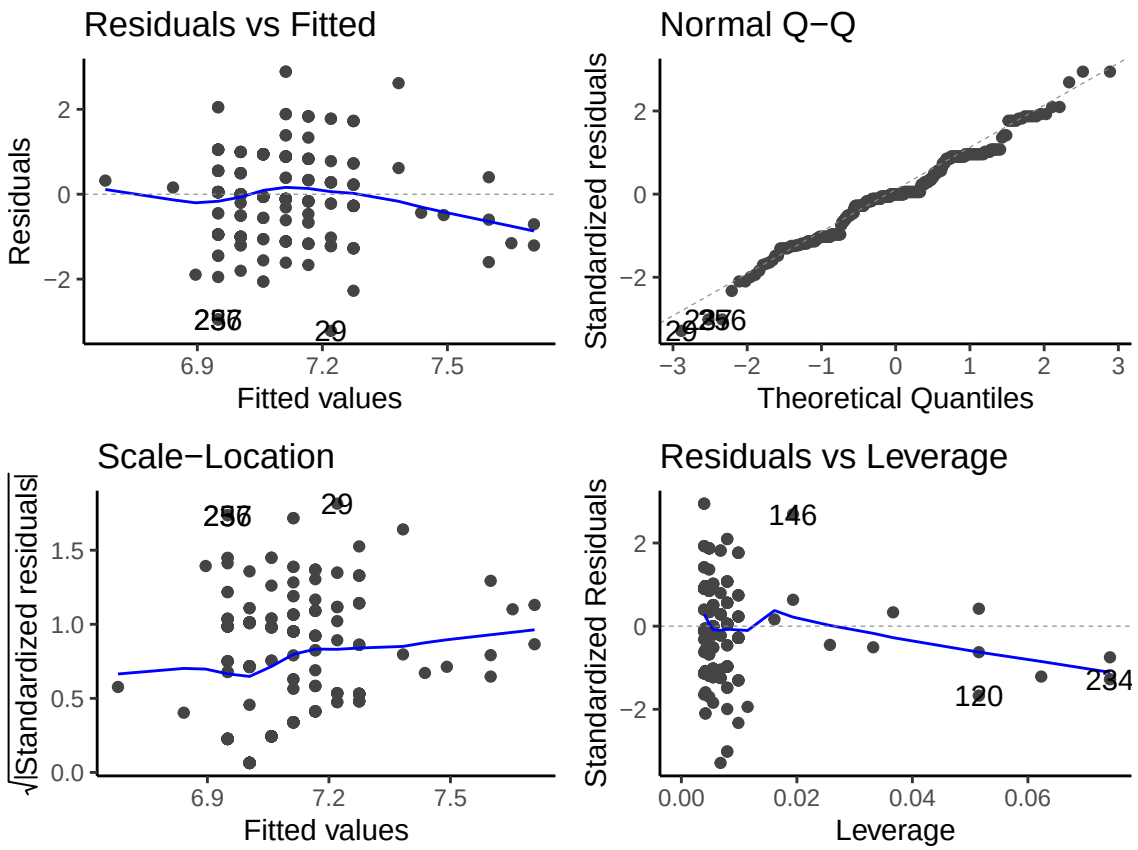
In the following equation t^* is the t-value for 95% confidence interval under the degree of freedom 255 b_1 is the slope of the regression line for this sample.

$$\begin{aligned} CI &= b_1 \pm t^* \cdot SE \\ &= -0.054 \pm 1.9693 \cdot 0.023 \\ &= (-0.0993, -0.0087) \end{aligned}$$

Interpretation

The study is 95% confidence that the true slope will fall within $(-0.0993, -0.0087)$

Check for Assumptions



In order to conduct a t-test for slope, the data in the sample should be linearly distributed, have constant variance, and the residuals should be normally distributed.

- Linearity Assumption: The Residuals vs Fitted plot could illustrate linearity, since the trend line is approximately horizontal at $y = 0$, and the dots are randomly distributed on the plot.
- Constant Variance Assumption: The Scale-Location plot shows that there is no fitted pattern for the standard residuals, and the dots are approximately horizontal. Therefore, the data has constant variance.
- Normality Assumption: The Normal Q-Q plot shows that the distribution of residuals are not perfectly normal since there are slight curves from the plot. However, the overall trend of these data is along the dashed line. The deviation is not unacceptable, therefore this data set fits a linear regression test.

Since the assumptions are met, the t-test statistic is computed using the following formula.

$$\begin{aligned} t &= \frac{b_1 - 0}{SE} \\ &= \frac{b_1}{SE} \\ &= \frac{-0.054}{0.023} \\ &= -2.35 \end{aligned}$$

The p-value is calculated by finding the area under the t-distribution with the degrees of freedom $(n-2) = 255$, to the left of -2.35, and to the right of 2.35.

$$\begin{aligned} T &\sim t_{255} \\ p-value &= P(|T| \geq |t|) \\ &= P(|T| \geq 2.35) \\ &= 0.019 \end{aligned}$$

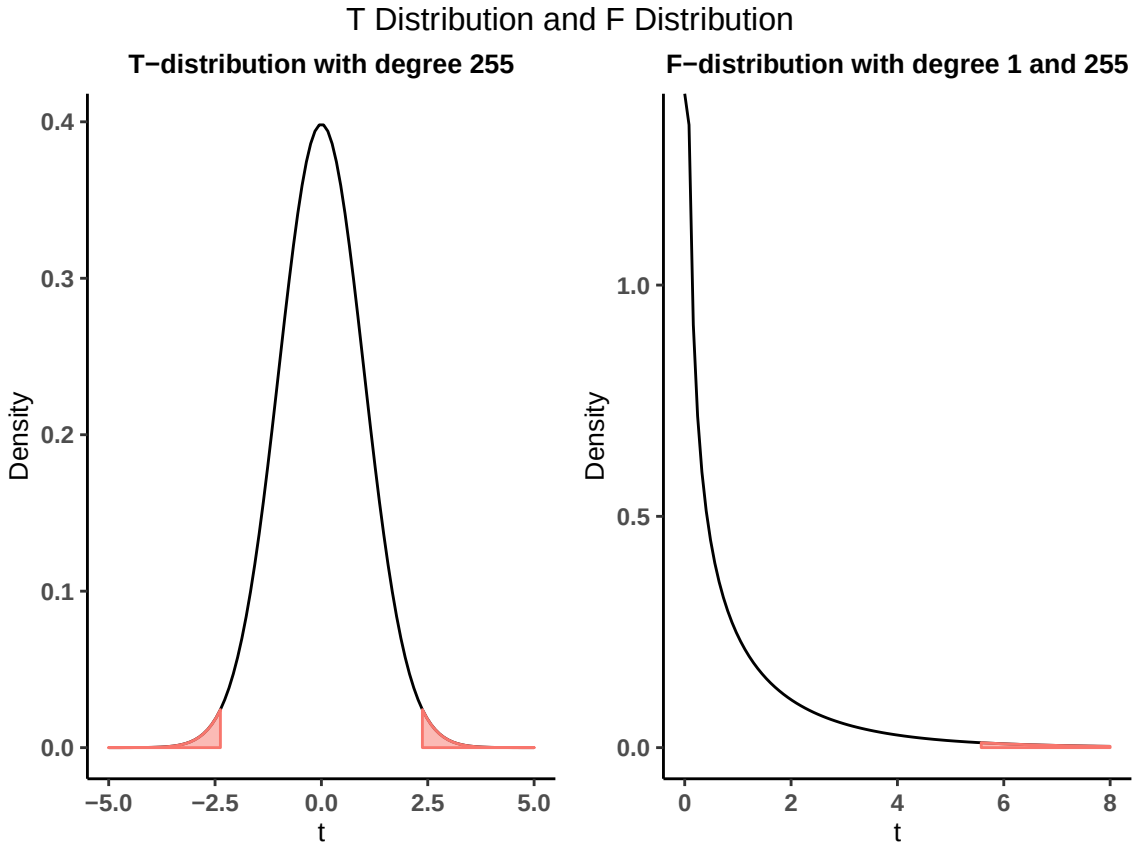
F-Test

F test should yield the same result as the T-test. Therefore, F test is also performed as a verification.

F-statistic is also computed from ANOVA table.

$$F = \frac{MSModel}{MSE} = \frac{5.3118262}{0.9646079} = 5.50672$$

$$\begin{aligned} F &\sim F_{255} \\ p-value &= P(|F| \geq |t|) \\ &= P(|T| \geq 5.50672) \\ &= 0.01970942 \end{aligned}$$



Conclusion

This p-value from this linear regression test is 0.019. Under $\alpha = 0.05$ significance level, there is significant evidence that the credits taken by a student is a good predictor for the hours of sleep of the student. There is a negative relationship between the number of credits taken and the sleep hours of a student. The more credits one student take, the less sleep he/she will get. (t-test for slope, $t = -2.35$, $df = 255$, $p = 0.019$).

Discussion

Objective

The objective of this study is to find out possible factors that are influencing the quality of life of college students, and the result of the study can provide some insights for Chinese undergraduate students, especially for those pre-college students, about major selections and providing approaches to alleviate their workloads.

Summary

First of all, a randomization test is performed to test the difference in mean time spent on activities between the *High Credit* group and the *Low Credit* group. Since extracurricular activities are a crucial part of college life, the mean time spent on extracurricular activities is a good indicator of the life quality of college students with different credits. In this test, the alternative hypothesis is that the mean time spent on extracurricular activities for *low credits* group is more than the *high credit* group. Based on the test statistic and p-value, the result of the study is statistically significant under the 95% confidence level. This result is also consistent with one previous study conducted by Lushington et al. (2015), whose study shows that there is a negative relationship between workload and time spent on extracurricular activities. Thus, the conclusion would be that there is significant evidence to support the alternative hypothesis that there are differences between these two groups. Specifically, students who took fewer credits spend more time on extracurricular activities.

Moreover, a randomization test for correlation between the number of credits taken by the students and their average sleep hour is conducted. Taking more credits means the student have more classes and homework to do, thus they need more time to study. Consequently, this might decrease their sleep time.The alternative hypothesis is that a negative correlation between the number of credits taken and their hours of sleep exists. The test yields significant results that there is strong evidence to indicate that the correlation is indeed negative. Therefore, taking more credits do decrease the hours of sleep student get.This result is also consistent with the result found by the T-test stated above.

Beyond that, a t-test is performed to further investigate the difference in average sleep hour between students who take 16 credits or more and students who take less than 16 credits. Specifically, the study is trying to find whether students who take fewer credits have more sleep hours than students who take more credits. The t-test yield significant results: Student who takes more than 16 credits sleep less than students who take less than 16 credits. This result can help the study to conclude that taking more credits will decrease the sleeping time for an undergraduate student.

In addition, a chi-square test is performed to find whether there is an association between *gender* and *major*. Here, major is divided into three groups, STEM (Science, Technology, Engineering & Mathematics), Business, and other majors. Gender is divided into two categories, Male and Female. Some people may hold a stereotype that male students are inclined to study STEM majors while female students are more likely to study Business related majors. Thus, the alternative hypothesis for this test is that there is an association between *Major* and *Gender*. The test statistic (χ^2) is 16.57285 and the degrees of freedom of the test is $(3-1)\cdot(2-1)=2$. Under a $\chi^2 - distribution$ with 2 degrees of freedom, the p-value is calculated to be 0.00025. Under a 95% confidence interval ($\alpha = 0.5$), the result is statistically significant. Therefore, there is significant evidence suggesting that there is an association between major and gender. In general, the workload for STEM majors is heavier than Business major or other majors. Major is a crucial indicator of the life quality of college students. Thus, students with different gender should have different expectations for their college life quality.

Lastly, linear regression is performed between two quantitative variables: credits taken and sleep hours. To check whether there is a linear relationship between these two quantitative variables, a t-test for slope is performed. The null hypothesis is that $H_0 : \beta = 0$The alternative hypothesis is that $H_\alpha : \beta \neq 0$. The p-value of this test is very small and this study yield significance evidence in favor of the alternative

hypothesis that there is a negative linear relationship between credits taken and sleep-hours. This result provides evidence for the intuitive interpretation that if a student takes more credits, his/her sleep hours will decrease. This could provide suggestions for Chinese international undergraduate students of how many credits they should take for every semester without sacrificing too many sleep hours.

Error Analysis

Even though this study yields some statistically significant results, possible biases may occur in the study, which can lead to errors.

One possible error is that data collected by this study might be slightly biased. The samples are collected through an online survey website *Qualtrics*. The link of the survey is distributed to Chinese international undergraduate students via we-chat, which is a popular social media software among Chinese students. Even though most Chinese international undergraduates are in the UW-Madison Chinese student we-chat group, it is not guaranteed that every Chinese student is included. Therefore, the sample might not be a fully accurate representation of the whole population.

Another sampling bias is that the data are collected only from students who are willing to share their information because some students might be reluctant to fill out the survey due to concerns about privacy issues. It is also possible that they did not fill the survey because they missed the message. These cases will lead to a non-representative sample.

Beyond those potential biases, confounding variables might also influence the results of the study. For example, variables like average sleep hours and average time spent on extracurricular activities of Chinese International undergraduates are investigated. These variables are subject to changes due to the different personalities of individuals, which is a possible confounding variable of this study. Another confounding variable is studying efficiency. Specifically, students with more effective learning habits and better time-management skills can spend less time to finish schoolwork and thus sleep more hours.

Future Studies

Future studies can generalize the population of interest to a larger group. Given the limitation of the samples, the result of the study can only be applied to Chinese international undergraduate students at the University of Wisconsin-Madison. If future researchers have enough funds and time, they can collect data from both domestic and international students of all U.S. universities. In that case, the result of the study could be generalized to all students in US universities.

Furthermore, the population of interest is limited to undergraduate students. Future studies could perform a study specifically for graduate students in universities. The reason is that the factors affecting the life quality of graduate students might be different from those of undergraduates. For example, graduate students normally take fewer credits than undergraduate students. However, the difficulty of graduate-level courses is higher than undergraduate-level courses. Besides these two differences, many other aspects should be taken into consideration for how to alleviate workloads for graduate students. Thus, it is necessary to conduct a separate study for undergraduate and graduate students.

References

May, R. W., & Casazza, S. P. (2012). Academic Major as a Perceived Stress Indicator: Extending Stress Management Intervention. College Student Journal, 46(2), *264–273*. Retrieved from http://search.ebscohost.com.ezproxy.library.wisc.edu/login.aspx?direct=true&AuthType=ip,uid&db=aph&AN=77698058&site=ehost-live&scope=site

Weinstein, L., & Laverghetta, A. (2009). College Student Stress and Satisfaction with Life. College Student Journal, 43(4), *1161–1162*. Retrieved from http://search.ebscohost.com.ezproxy.library.wisc.edu/login.aspx?direct=true&AuthType=ip,uid&db=aph&AN=55492493&site=ehost-live&scope=

Zhang, X., Wang, H., Xia, Y., Liu, X., & Jung, E. (2012). Stress, coping and suicide ideation in Chinese college students. Journal of Adolescence, 35(3), *683–690*. doi: 10.1016/j.adolescence.2011.10.003

Lushington, K., Wilson, A., Biggs, S., Dollman, J., Martin, J., & Kennedy, D. (2015). Culture,Extracurricular Activity, Sleep Habits, and Mental Health: A Comparison of Senior High School Asian-Australian and Caucasian-Australian Adolescents. International Journal of Mental Health, 44(1/2), 139–157.https://doi-org.ezproxy.library.wisc.edu/10.1080/00207411.2015.1009788

Appendix

Set Up

Sample Size Determimation

```
s = sd(Data$SleepHour)
t = qt(0.975,247)
((t*s)/0.2)^2
```

Randomization Test

I.Difference in Mean: Credits and Extracurricular Activities

```
NewData = Data %>%
  mutate(Stem=recode(Major,
                     `STEM (Science, Technology, Engineering, Mathematics)`="Yes",
                     .default = "No"),
         AcademicHour = ClassHour + StudyHour,
         CreditsHorL= recode(Credits,
                             '1' = "low", '2' = "low", '3' = "low",
                             '4' = "low", '5' = "low", '6' = "low",
                             '7' = "low", '8' = "low", '9' = "low",
                             '10' = "low", '11' = "low", '12' = "low",
                             '13' = "low", '14' = "low",'15' = "low", .default = "high")
         )
NewData%>%
  ggplot(aes(x = CreditsHorL, y = ActivityTime)) +
  geom_boxplot(color = "black", fill = c("#F8766D", "#00BFC4"), alpha = 0.25) +
  theme_classic() +
  labs(subtitle = "High Credits V.S. Low Credits",
       title = "Time Spent on Activities for Different Credits",
       x = "Credits",
       y = "Activity Time (Hours)") +
  theme(plot.title = element_text(size = 15, color = "black", hjust = 0.5, face = "bold"),
        plot.subtitle = element_text(size = 10, color = "black", hjust = 0.5, face = "bold"),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size = 10),
        axis.text.x = element_text(vjust = 0.6, face = "bold"),
        axis.text.y = element_text(hjust = 0.6, face = "bold")) +
  ylim(0, 11) +
  stat_summary(fun.y=mean, geom="point", shape=18, size=5, color="black", fill="black")
```

```
credit.low = filter(NewData,CreditsHorL== "low")
credit.high = filter(NewData,CreditsHorL== "high")

x.bar.low = mean(credit.low$ActivityTime)
s.low = sd(credit.low$ActivityTime)
n.low = length(credit.low$ActivityTime)
```

```r
x.bar.high = mean(credit.high$ActivityTime)
s.high = sd(credit.high$ActivityTime)
n.high = length(credit.high$ActivityTime)

se <- sqrt((s.low^2/n.low)+(s.high^2/n.high))

x.bar <- mean(NewData$ActivityTime)

shift.low <- x.bar.low - x.bar
shift.high <- x.bar.high - x.bar

low.0 <- credit.low$ActivityTime - shift.low
high.0 <- credit.high$ActivityTime - shift.high



set.seed(302)
B <- 10000


# randomization
mat.rand.low <- matrix(sample(low.0,B*n.low,replace=TRUE),
                       byrow = TRUE,
                       nrow = B,
                       ncol = n.low)
mat.rand.high <- matrix(sample(high.0,B*n.high,replace=TRUE),
                       byrow = TRUE,
                       nrow = B,
                       ncol = n.high)

rand.mean.low <- apply(mat.rand.low,1,mean)
rand.mean.high <- apply(mat.rand.high,1,mean)
rand.diff <- rand.mean.low - rand.mean.high

df.rand <- data.frame(Randomization=rand.diff)

# compute p-value
tol <- 1.0e-12
p.value2 <- mean(rand.diff >= x.bar.low-x.bar.high-tol)


# Bootstrap
mat.boot.low <- matrix(sample(credit.low$ActivityTime,B*n.low,replace=TRUE),
                       byrow = TRUE,
                       nrow = B,
                       ncol = n.low)
mat.boot.high <- matrix(sample(credit.high$ActivityTime,B*n.high,replace=TRUE),
                       byrow = TRUE,
                       nrow = B,
                       ncol = n.high)

boot.mean.low <- apply(mat.boot.low,1,mean)
boot.mean.high <- apply(mat.boot.high,1,mean)
```

```
boot.diff <- boot.mean.low - boot.mean.high

df.boot <- data.frame(Bootstrap=boot.diff)


# Randomization & Bootstrp Distribution
ci.pe <- round(quantile(boot.diff, probs = c(0.025, 0.975), type = 2),3)

df.melt <- melt(data.frame(Randomization = df.rand, Bootstrap = df.boot))

ggplot(df.melt, aes(x = value, fill = variable)) +
  geom_density(alpha = 0.25, adjust = 2) +
  theme_classic() +
  labs(subtitle = "High Credits v.s. Low Credits",
       title = "Difference in Mean of Activity Time",
       x = "Difference in Mean",
       y = "Density",
       fill = "Type") +
  theme(plot.title = element_text(size = 15, color = "black", hjust = 0.5, face = "bold"),
        plot.subtitle = element_text(size = 10, color = "black", hjust = 0.5, face = "bold"),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size = 10),
        axis.text.x = element_text(vjust = 0.6, face = "bold"),
        axis.text.y = element_text(hjust = 0.6, face = "bold")) +
  scale_y_continuous(expand = c(0,0)) +
  scale_x_continuous(expand = c(0,0)) +
  geom_vline(xintercept = ci.pe, color = "red", linetype = "dashed") +
  geom_vline(xintercept = x.bar.low-x.bar.high) +
  annotate("text",
           x = ci.pe,
           y = 0.3,
           label = round(ci.pe, 3),
           fontface = "bold") +
  annotate("text",
           x = x.bar.low-x.bar.high,
           y = 0.3,
           label = round(x.bar.low-x.bar.high, 3),
           fontface = "bold")
```

```
# print out the 95% confidence interval
df.ci <- data.frame(x=round(ci.pe,3))
df.ci %>%
  kable(caption = "95% Confidence Interval",col.names = "")
```

II.Randomization Test for Correlation: Credits Taken and Number of Sleep Hours

```
ggplot(Data, aes(x = Credits,y = SleepHour)) +
  geom_point() +
  labs(x = "Credits",
       y = "Sleep Hour",
       subtitle = "Dot Plot",
```

```
       title = "Credits v.s. SleepHour") +
  theme_classic() +
  theme(plot.title = element_text(size = 15,
                                  color = "black", hjust = 0.5, face = "bold"),
        plot.subtitle = element_text(size = 10, color = "black",
                                     hjust = 0.5, face = "bold"),
        axis.title.x =element_text(size = 10),
        axis.title.y = element_text(size = 10), legend.title = element_text(size = 10),
        axis.text.x = element_text(vjust = 0.6, face = "bold"),
        axis.text.y = element_text(hjust = 0.6, face = "bold"))
```

```
set.seed(302)

v1 = Data$Credits
v2 = Data$SleepHour

B = 10000
null.sample = numeric(B)

for (i in 1:B) {
  v1.permut = sample(v1, length(v1), replace = FALSE)
  v2.permut = sample(v2, length(v2), replace = FALSE)
  null.sample[i] = cor(v1.permut, v2.permut, use = 'complete.obs')
}
null.sample = data.frame(s = null.sample)

plot= ggplot(null.sample, aes(s)) +
  geom_density() +
  labs(x = "Point Estimate of Correlation between Two Variables",
       y = "Density",
       title = "Credits Take v.s. Sleep Hour",
       subtitle = "Randomization Distribution") +
  theme_classic() +
  theme(plot.title = element_text(size = 15,
                                  color = "black", hjust = 0.5, face = "bold"),
        plot.subtitle = element_text(size = 10, color = "black",
                                     hjust = 0.5, face = "bold"),
        axis.title.x =element_text(size = 10),
        axis.title.y = element_text(size = 10), legend.title = element_text(size = 10),
        axis.text.x = element_text(vjust = 0.6, face = "bold"),
        axis.text.y = element_text(hjust = 0.6, face = "bold"))


plot
```

```
#### Calculate Test Statistic
test.stat = cor(v1, v2, use = 'complete.obs')
sum(((v1 - mean(v1)) / sd(v1)) * ((v2 - mean(v2)) / sd(v2))) / (length(v1) - 1)

test.stat + quantile(null.sample$s, 0.025)
test.stat + quantile(null.sample$s, 0.975)
```

```
#### Compute p-value
p.value3 = mean(null.sample <= test.stat)

gg.data <- ggplot_build(plot)$data
temp <- gg.data[[1]] %>%
  filter(x <= test.stat)

ggplot(null.sample, aes(s)) +
  geom_density() +
  labs(x = "Point Estimate of Correlation between Two Variables",
       y = "Density",
       title = "Credits Take v.s. Sleep Hour",
       subtitle = "One-tailed Test") +
  theme_classic() +
  theme(plot.title = element_text(size = 15,
                                  color = "black", hjust = 0.5, face = "bold"),
        plot.subtitle = element_text(size = 10, color = "black",
                                     hjust = 0.5, face = "bold"),
        axis.title.x =element_text(size = 10),
        axis.title.y = element_text(size = 10), legend.title = element_text(size = 10),
        axis.text.x = element_text(vjust = 0.6, face = "bold"),
        axis.text.y = element_text(hjust = 0.6, face = "bold")) +
  geom_area(data = temp, aes(x=x, y=y, color=factor(1), fill=factor(1)), alpha = 0.5)+
  theme(legend.position="none")
```

Parametric Test

T-Test (Difference in Mean): Number of Credits and Sleep Hours

```
MyData = Data %>%
  mutate(Stem=recode(Major,`STEM (Science, Technology, Engineering,
                                  Mathematics)`="Yes",.default = "No"),
         AcademicHour = ClassHour + StudyHour,
         CreditsHorL = recode(Credits,
                              '16' = "high", '17' = "high",
                              '18' = "high",'19' = "high",
                              '20'="high", '21'="high",
                              '22'="high",'23'="high",
                              .default = "low")
         )

# Summary figure box-plot

MyData %>%
  ggplot(aes(x = CreditsHorL, y = SleepHour)) +
  geom_boxplot(color = "black", fill = c("#F8766D", "#00BFC4"), alpha = 0.25) +
  theme_classic() +
  labs(title = "Time spent on activities of students taking different credits",
       x = "Credits",
       y = "Activity time") +
  theme(plot.title = element_text(size = 15, color = "black", hjust = 0.5, face = "bold"),
```

```
        plot.subtitle = element_text(size = 10, color = "black", hjust = 0.5, face = "bold"),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size = 10),
        axis.text.x = element_text(vjust = 0.6, face = "bold"),
        axis.text.y = element_text(hjust = 0.6, face = "bold")) +
  ylim(0, 11) +
  stat_summary(fun.y=mean, geom="point", shape=18, size=5, color="black", fill="black")
```

```
Lowcredits = filter(MyData,CreditsHorL== "low")
Highcredits = filter(MyData,CreditsHorL== "high")
mean.Hsleep <- round(mean(Highcredits$SleepHour), 3)
mean.Lsleep <- round(mean(Lowcredits$SleepHour), 3)
sd.Hsleep <- round(sd(Highcredits$SleepHour), 3)
sd.Lsleep <- round(sd(Lowcredits$SleepHour), 3)

df.summary <- data.frame(
  Group = c("Highcredits", "Lowcredit"),
  Size = c(pull(count(Highcredits)), pull(count(Lowcredits))),
  Mean.sleephour = c(mean.Hsleep, mean.Lsleep),
  SD.sleephour = c(sd.Hsleep, sd.Lsleep)
)

kable(df.summary, caption = "Summary of sleephours",
      col.names = c("Group", "Size",
                    "Mean sleephour", "SD sleephour"))
```

```
qt(0.975,122)
```

```
se <- sqrt((sd.Lsleep^2/pull(count(Lowcredits)))+(sd.Hsleep^2/pull(count(Highcredits))))

ttest <- t.test(Lowcredits$SleepHour,
                Highcredits$SleepHour,
                alternative = "greater",
                conf.level = 0.95)

df.sum <- data.frame(
  x1 = mean.Hsleep,
  x2 = mean.Lsleep,
  s1 = sd.Hsleep,
  s2 = sd.Lsleep,
  SE = round(se,3),
  t = round(ttest$statistic,3),
  df = round(ttest$parameter,3),
  p = round(ttest$p.value,4)
)

kable(df.sum, col.names = c("Mean.Lowcredits","Mean.Highcredits",
                            "SD.Lowcredits", "SD.Highcredits",
                            "SE", "t", "df", "p"),
      caption = "Summary for t-Test",
      row.names = F)
```

Chi-square Test

```
# STEM Major
MStem <- Data %>%
  filter(Gender == 'Male') %>%
  count(Major == 'STEM (Science, Technology, Engineering, Mathematics)')

FStem <- Data %>%
  filter(Gender == 'Female') %>%
  count(Major == 'STEM (Science, Technology, Engineering, Mathematics)')

# Business
MBusiness <- Data %>%
  filter(Gender == 'Male') %>%
  count(Major == 'Business')

FBusiness <- Data %>%
  filter(Gender == 'Female') %>%
  count(Major == 'Business')

# Social Science
MSS <- Data %>%
  filter(Gender == 'Male') %>%
  count(Major == 'Social Science (Economics, Sociology, Psychology, etc.)')

FSS <- Data %>%
  filter(Gender == 'Female') %>%
  count(Major == 'Social Science (Economics, Sociology, Psychology, etc.)')

# Arts
MArts <- Data %>%
  filter(Gender == 'Male') %>%
  count(Major == 'Liberal Art (Common Arts, History, etc.)')

FArts <- Data %>%
  filter(Gender == 'Female') %>%
  count(Major == 'Liberal Art (Common Arts, History, etc.)')
# Education
MEducation <- Data %>%
  filter(Gender == 'Male') %>%
  count(Major == 'Education')

FEducation <- Data %>%
  filter(Gender == 'Female') %>%
  count(Major == 'Education')

# Agriculture
MAgri <- Data %>%
  filter(Gender == 'Male') %>%
  count(Major == 'Agriculture')

FAgri <- Data %>%
  filter(Gender == 'Female') %>%
```

```
  count(Major == 'Agriculture')
```

```
Chi.square.Table <- as.table(rbind(c(84,71,155), c(8,20,28),c(21,53,74),c(113,141,257)))
dimnames(Chi.square.Table) <- list(c("STEM", "Business","Other Majors","Total"),
                                   c("Male","Female","Total"))

kable(Chi.square.Table,caption = "Observed Count Table (Major/Gender)",
      col.names = c("Male", "Female","Total"))
```

```
Oberserved <- as.table(rbind(c(84,71), c(8,20), c(21,53)))
dimnames(Oberserved) <- list(Major = c("STEM", "Business","Other Majors"),
                             gender = c("Male","Female"))

Oberserved %>%
  as.data.frame() %>%
  rename(obs.counts = Freq) %>%
  ggplot(aes(x=gender,y=obs.counts,fill=Major)) +
  geom_bar(aes(color=Major),position="dodge",stat="identity",alpha = 0.25) +
  labs(x = "Gender",
       y = "Count",
       title = "Major v.s. Gender",
       subtitle = "Grouped by Gender",
       fill = "Major",
       color = "Major") +
  theme_classic() +
  theme(plot.title = element_text(size = 15,
                                  color = "black", hjust = 0.5, face = "bold"),
        plot.subtitle = element_text(size = 10, color = "black",
                                     hjust = 0.5, face = "bold"),
        axis.title.x =element_text(size = 10),
        axis.title.y = element_text(size = 10), legend.title = element_text(size = 10),
        axis.text.x = element_text(vjust = 0.6, face = "bold"),
        axis.text.y = element_text(hjust = 0.6, face = "bold"))
```

```
#### Check for Assumptions

obs.counts <- Oberserved

r <-nrow(obs.counts)
c <-ncol(obs.counts)

row.sums <-rowSums(obs.counts)
col.sums <-colSums(obs.counts)

n <-sum(obs.counts)

exp.counts <-outer(row.sums,col.sums,"*")/n

kable(exp.counts,caption = "Expected Count Table (Major/Gender)")
```

```
# Compute Chi-square Test Statistic
X.sq <- sum((obs.counts-exp.counts)^2/exp.counts)
```

```
# Two Methods Computing p-value
pchisq(X.sq,(r-1)*(c-1),lower.tail=FALSE)
1-pchisq(X.sq,(r-1)*(c-1))
```

Linear Regression Test

```
Data %>%
  ggplot(aes(x = Credits, y = SleepHour)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = F) +
  theme_classic() +
  labs(subtitle = "Scatterplot",
       title = "Credits taken and hours of sleep",
       x = "Credits taken",
       y = "Sleephours") +
  theme(plot.title = element_text(size = 15, color = "black", hjust = 0.5, face = "bold"),
        plot.subtitle = element_text(size = 10, color = "black", hjust = 0.5, face = "bold"),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size = 10),
        axis.text.x = element_text(vjust = 0.6, face = "bold"),
        axis.text.y = element_text(hjust = 0.6, face = "bold"))
```

```
lm.sleep <- lm(SleepHour ~ Credits, Data)
mod.sleep <- summary(lm.sleep)
anova = anova(lm.sleep)
xtable(anova) %>%
  kable(caption = "Anova Table")
xtable(mod.sleep) %>%
  kable(caption = "T-test Summary Table")
```

```
qt(0.975,255)
```

```
autoplot(lm.sleep) + theme_classic()
```

```
count(Data)
2*pt(-2.35,255,lower.tail = TRUE)
```

```
pf(5.50672, df1 = 1, df2 = 255, lower.tail = FALSE)
```

```
t = -2.35
x <- seq(-5, 5, length=100)
y <- dt(x, 255)
gg1 = ggplot(data.frame(x, y), aes(x = x, y = y)) + geom_line() +
labs(x = "t",
y = "Density",
title = "T-distribution with degree 255") +
theme_classic() +
theme(plot.title=element_text(size=10),
axis.title=element_text(size=10),
```

```
legend.title=element_text(size=10))

gg1.data <- ggplot_build(gg1)$data
temp1 <- gg1.data[[1]] %>% filter(x <= t)
temp2 <- gg1.data[[1]] %>% filter(x >= -t)

p1 = gg1 + geom_area(data = temp1, aes(x=x, y=y, color=factor(1), fill=factor(1)), alpha = 0.5) +
  geom_area(data = temp2,
            aes(x=x, y=y, color=factor(1), fill=factor(1)), alpha = 0.5)+
  theme(plot.title = element_text(size = 10 , color = "black", hjust = 0.5, face = "bold"),
        plot.subtitle = element_text(size = 10, color = "black", hjust = 0.5, face = "bold"),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size = 10),
        axis.text.x = element_text(vjust = 0.6, face = "bold"),
        axis.text.y = element_text(hjust = 0.6, face = "bold"))+
  theme(legend.position="none")

f = 5.50672
x <- seq(0, 8, length=100)
y <- df(x, df1 = 1, df2 = 255)
gg2 = ggplot(data.frame(x, y), aes(x = x, y = y)) + geom_line() +
labs(x = "t",
y = "Density",
title = "F-distribution with degree 1 and 255") +
theme_classic() +
theme(plot.title=element_text(size=10),
axis.title=element_text(size=10),
legend.title=element_text(size=10))

gg2.data <- ggplot_build(gg2)$data
temp <- gg2.data[[1]] %>% filter(x >= f)


p2 = gg2 + geom_area(data = temp,
                     aes(x=x, y=y, color=factor(1), fill=factor(1)), alpha = 0.5)+
  theme(plot.title = element_text(size = 10 , color = "black", hjust = 0.5, face = "bold"),
        plot.subtitle = element_text(size = 10, color = "black", hjust = 0.5, face = "bold"),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size = 10),
        axis.text.x = element_text(vjust = 0.6, face = "bold"),
        axis.text.y = element_text(hjust = 0.6, face = "bold"))+
  theme(legend.position="none")


grid.arrange(p1, p2, ncol = 2, nrow = 1, top = textGrob("T Distribution and F Distribution"))
```