Psychological and Emotional Effects of Social Media on Students

The Matrix

Shaun Campbell (Team Leader)
Nan Yang; Ziwei Zhao; Shuqing Zhao
STAT 302: Accelerated Introduction to Statistical Methods

20 December 2018

Contents

Abstract

The objective of this study is to uncover whether social media use is associated with positive, negative, or no psychological and emotional effects on students. In particular, the population of interest is undergraduate students at the University of Wisconsin - Madison. Data was collected through a survey sent out electronically to students and consists of 41 responses. The primary explanatory variable of interest is average social media use time per day. Gender is a potential confounding variable of interest and is investigated. Response variables of interest that are tested are stress level and GPA. Results of the test imply that social media use time is not associated with stress level or GPA. However, due to issues with sampling and violation of statistical assumptions, no conclusion can be drawn from this study.

Introduction

Social media has a wide-reaching presence in modern-day society. Usage of social media seems particularly prevalent among young people, such as college students. Platforms such as Instagram, Facebook, Snapchat, and Twitter provide entertainment as well as a way for students to communicate. A problem presents itself in the potential effects of over-exposure to social media. These potential effects may or may not include psychological and emotional effects on a person, which can be measured by an effect on real-life relationships, an effect on academic performance, and effects on sleep, happiness, and stress. Finding whether social media use is associated with effects on these psychological and emotional measures serves as the motivation for this study.

The effects of social media on people's lives are not entirely know, as social media is fairly new with popularity growing over the last decade. In 2013 Kross et al published a study that found that an increase in Facebook usage predicted a negative shift in how people feel moment-to-moment and how satisfied they are with their lives[1]. In 2012, Pantic I, Damjanovic A, Todorovic J, et al published a study that found that online social networking has a statistically significant positive correlation with depression scores in high school students[2].

Since these studies were published, social media use has risen greatly. For example, Facebook use has risen from around 57% of US adults using the platform in 2013 to 68% in 2018, and Instagram use has risen from around 17% in 2013 to 35% in 2018[3]. This study aims to take a new look on the effects of social media use specifically in college students at UW-Madison. The study will use different measures of specific emotional and psychological areas to generate a broader understanding of the effects of social media on emotional and psychological factors. Due to the growing amounts of time people are spending on social media platforms, it is important to study whether this has positive or negative effects. Researching this will hopefully reveal if social media use has a negative effect on students' wellness and should be moderated, or if it has a positive effect and should be encouraged. The objective of this study is to find whether social media is associated with positive, negative, or no psychological and emotional effects on students.

To examine the effects of social media on the lives of students, several measures will be investigated. The psychological and emotional effects that are measured include the following: time spent face-to-face with friends; the student's self-rating of their stress level, happiness level, and level of satisfaction with real-life relationships; GPA; level of sleep. These measures will give a sense of the psychological and emotional state of the students.

Methods

The population of interest for this study is all undergraduate college students at UW-Madison. The interest in effects on students and the nature of testing resources led to this being the best population to study. A survey was conducted in order to find associations within the data that can be applied to the population of interest. This study is an observational study because it involves no manipulation of variables.

Due to limitations on obtaining email lists to distribute the survey, data was collected by sending a survey to colleagues through messaging applications. The survey link reached approximately 60 students and 41 responded. One outlier which reported 21 hours of social media use per day is believed to be erroneous and is removed from the data resulting in a data set with 40 cases. It is possible that there is both response bias and non-response bias. Students who felt particularly strongly about the subject may have chosen to respond while students who felt less strongly may have chosen not to. The sample is not random because the recipients of the survey were not selected randomly. This could have a significant influence on the interpretation of results and will be dealt with further when the results are discussed. Ideally, the survey would have been sent to a group of students randomly selected from a list of all undergraduate students at UW-Madison. This would solve the issue of the selection of survey respondents not being random. Requiring students to respond to the survey would fix response and non-responses biases.

For the sample to be representative, the proportions of the sample demographics would need to be like the demographics for the university as a whole. The proportions for the population are 48.6% male and 51.4% female, and 7.6% Asian/Pacific Islander and 67.2% white[4]. The sample was 46.3% male and within reason for being representative of the population. The sample was 100% Asian/Pacific Islander and not within reason for being representative of the population. This will have important ramifications for how results can be interpreted. Expanding survey distribution to a more diverse group of students would be a step towards fixing this problem.

To be appropriate for statistical tests and analysis, the sample size needs to be large enough to be approximately normally distributed for quantitative variables, so ideally $n_i > 30$. Some tests do not satisfy this assumption and will be discussed further. Categorical variables must also satsitfy $n_i p_i \geq 10$ and $n_i(1 - p_i) \geq 10$ for each level of the categorical variables.

Potential Confounding Variables

Variable Name	Variable Type	Description	Variable Level
Year	Categorical	Class Standing	Freshman; Sophomore; Junior; Senior; 5th Year+
Major	Categorical	Major of the student	Write-in
Gender	Categorical	Which gender the student is	Male; Female; Non-Binary; Gender Fluid
Ethnicity	Categorical	Ethnic origin of the student	White; Hispanic/Latino; Black/African American; Native American/American Indian; Asian/Pacific Islander; Other
Extroverted	Categorical	Whether the student describes themselves as extroverted or introverted	Extrovert; Introvert

Explanatory Variables: Social Media

Variable Name	Variable Type	Description	Variable Level
MostUsed	Categorical	The app the student spends the most time using	Snapchat; Instagram; Facebook; Other
TimeUse	Categorical	The time of day that the student most commonly uses social media	Morning; afternoon; evening; night
PrimaryUse	Categorical	The primary reason the student uses social media	News; Entertainment; Communication; ; Study; Killing Time; Other
TimeOfDay	Quantitative	Hours per day the student spends on social media	
LikesPerDay	Quantitative	Number of likes/shares a student sends on Facebook per day	
Post	Numerical	How many social media posts the student makes per week	
Like	Numerical	How many followers the student has on Instagram	

Response Variables

Variable Name	Variable Type	Description	Variable Level
FacetoFace	Categorical	Whether the student prefers to talk to their friends in person or online	In person; electronic devices
StressLevel	Categorical	The student's rating of their stress level	1-10
HappinessLevel	Categorical	The student's rating of their happiness	1-10
Satisfaction	Categorical	How the student rates their satisfaction with their real-life relationships	Very Dissatisfied; Somewhat Dissatisfied; Neutral; Satisfied; Very Satisfied
GPA	Quantitative	The student's GPA	
Sleep	Quantitative	Hours the student sleeps on average	
TimeWithFriends	Quantitative	Hours spent with friends in real-life per week	
EffectOnVacation	Categorical	If the student spends more time on social media on vacation	Definitely not; probably not; might or might not; Probably; Definitely
EffectOnExam	Categorical	If the student spends more time on social media in exam weeks	Definitely not; probably not; might or might not; Probably; Definitely
EffectOnAcademic Performance	Categorical	If the student believes social media affects academic performance	Yes; no
EffectPositiveOr Negative	Categorical	The effect the student believes social media has on them	Extremely positive; Moderately positive; slightly positive; neither positive nor negative; slightly negative; moderately negative; extremely negative

Statistical Tests

The explanatory variables are split into groups to test for difference of means for different groups of the explanatory variables. Groupings that are analyzed include low social media use and high social media use, low stress and high stress, and confounding variables. It is hypothesized that spending lots of time on social media may be associated with an increased stress level and gender may be a confounding variable. Therefore, it is appropriate to test for a difference of means between the groups. Statistical tests such as randomization and t-test are used to test for statistical significance. Linear regression is also performed to test for a relationship between hours of social media use time per day and GPA. R and RStudio with dplyr and ggplot2 packages are used to clean, visualize, and analyze the data. The results of the statistic tests performed with R are used to generate confidence intervals for the relevant test statistics which are then interpreted.

Results

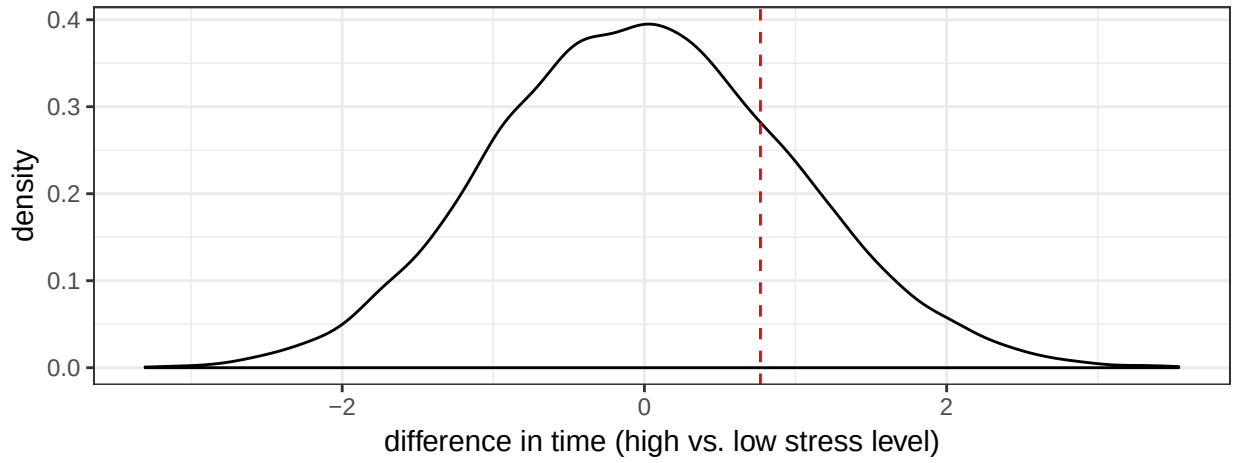
Social Media Use Time Per Day Depending on Stress Level

This test aims to find an association between stress level and time spent on social media per day. The variable "StressLevel" from the data is used to split that data into high stress (> 5) and low stress (≤ 5) groups. The high stress group has 28 observations and the low stress group has 12 observations. Average social media use time per day is calculated using the "TimeOfDay" variable. It is hypothesized that students with low stress spend a different amount of time on social media than students with high stress. The null and alternative hypotheses are as follows:

$H_0 : \mu_{\text{low}} = \mu_{\text{high}}$
$H_a : \mu_{\text{low}} \neq \mu_{\text{high}}$

The following box-plot visualizes the hours spent on social media per day for the high stress and low stress groups:

R is used to calculate the test statistic and p-value using a randomization test (Appendix, 2). The means of the high stress group and low stress group are shifted in order to do the randomization test. Bootstrap samples are created for each group using sampling with replacement for 10,000 samples. Means for each bootstrap sample are calculated and the difference of means for the bootstrap samples are calculated. The p-value is the proportion of bootstrap difference of means that are more extreme than the difference of means from the original sample. Below is a graph visualizing the randomization distribution with the difference of means from the original sample marked on the distribution.



The following summarizes the test statistic and the p-value:

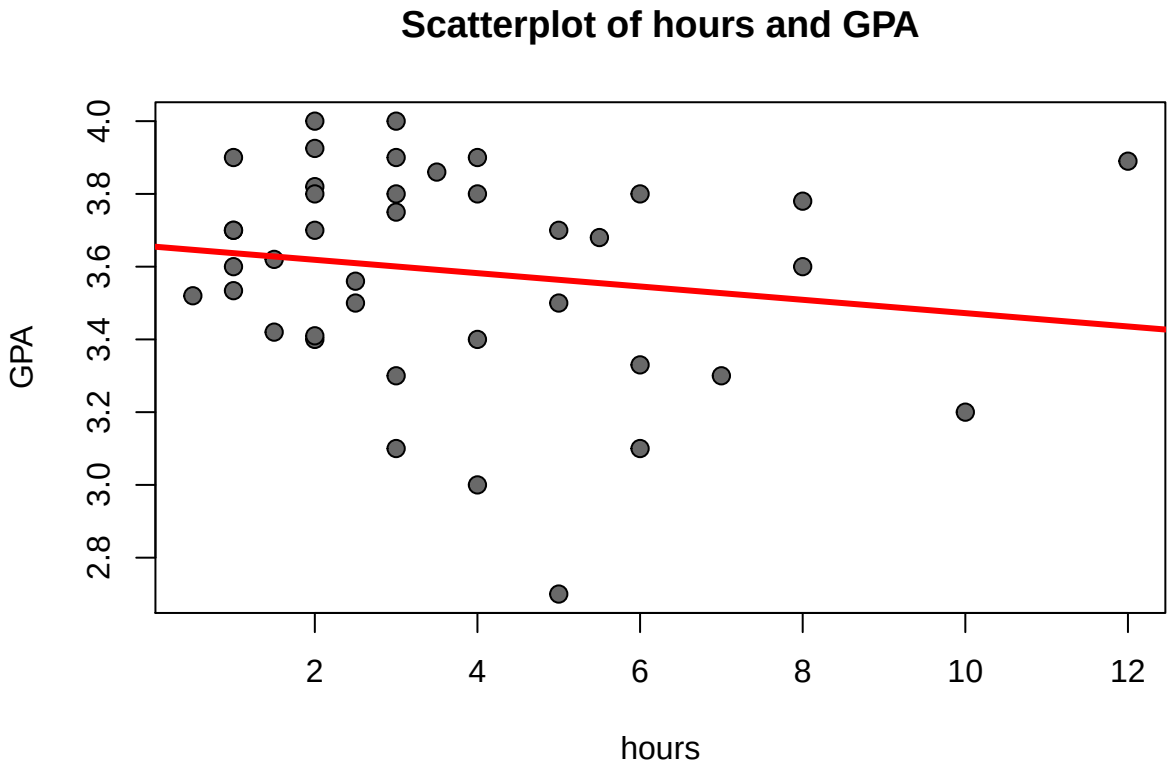
Statistic	Value
$\bar{X}_{\text{low}}$ - $\bar{X}_{\text{high}}$	`0.7678571`
t-statistic	`0.7526952`
P-Value	`0.445`

The 95% confidence interval using a t-distribution with 11 degrees of freedom and the standard error of the bootstrap difference in means is [`-1.0047613`, `2.5404756`]. The p-value is quite large, so it cannot be said that students with high stress levels (> 5) have a different social media use time per day than students with low stress levels (≤ 5).

Hours of Social Media Per Day Versus GPA

This test aims to test whether the hours per day a student spends on social media is associated with the student's GPA. The variable "TimeOfDay" is used to attempt to build a linear regression model to predict GPA. It is hypothesized that time spent on social media is associated with a higher or lower GPA. Therefore the null and alternative hypotheses for this test are:
$H_0 : \beta = 0$
$H_a : \beta \neq 0$

The following graph displays a scatterplot of the data with a linear regression line (Appendix, 3).



The following summarizes the test statistic and p-value for the simple linear regression (Appendix, 4):

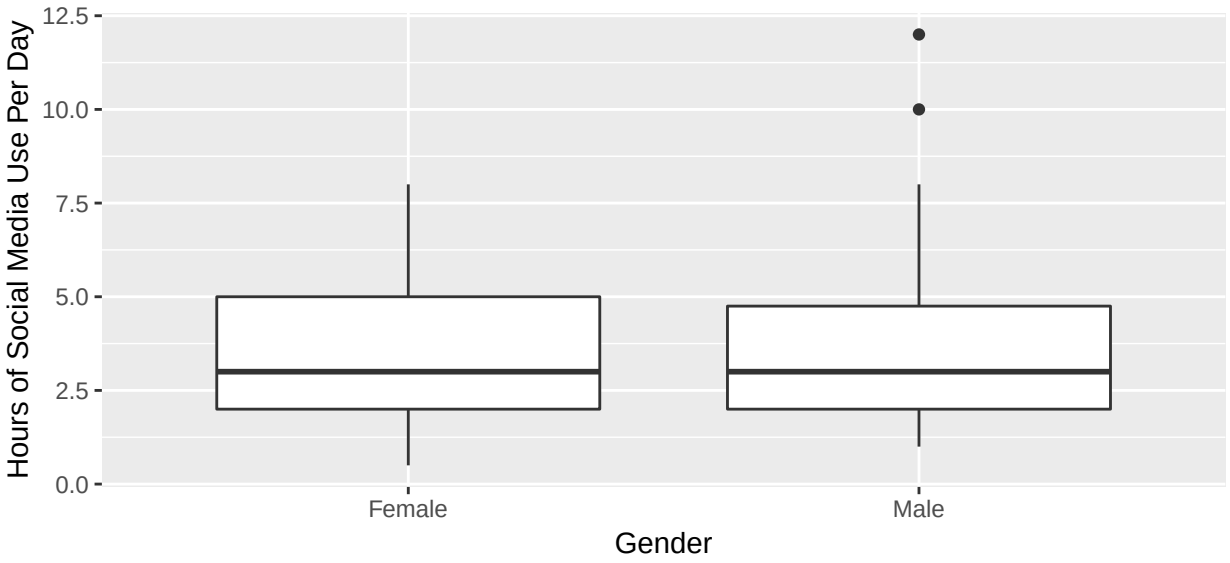
t-statistic	`-1.0048844`
P-Value	`0.321`

The critical t-statistic for statistical significance at the 5% level with 38 degrees of freedom is 2.023. The 95% confidence interval for the slope of the regression line using this critical t-statistic and the standard error of the slope coefficient is [`-0.0436037`, `0.0069656`]. There is a slight downward slope of the regression line. However, the t-statistic calculated in the test, the confidence interval, and the p-value imply that there is no significant linear relationship between hours of social media use per day and GPA.

Confounding Variable Gender Does Not Explain Stress

This test aims to test a potential confounding variable of gender on time spent on social media. Time spent on social media is an important measure for this study and the removal of a confounding variable may help with interpretation of results. The variable "Gender" is used to split the data into male and female groups. It is hypothesized that the mean use time is the same for both genders. Therefore, the null and alternative hypotheses for this test are:
H_0: $\mu_{\text{male}} = \mu_{\text{female}}$
H_a: $\mu_{\text{male}} \neq \mu_{\text{female}}$

The following box-plot shows the distrubitions of use time for the two genders:



R is used to perform a two-sided t-test to analyze whether the average use time is different for males and females. The following summarizes the test statistic and the p-value (Appendix, 1)

t-statistic	`-0.5448795`
P-Value	`0.5900573`

The difference in sample means between female and male use time is `-0.472`. The 95% confidence interval for the difference in means using a t-distribution with 28.622 degrees of freedom and the relevant t-statistic is [`-2.2457482`, `1.3013037`]. The t-statistic of t = `-0.545` implies that there is not a statistically significant relationship between gender and average hours of social media use per day at the 5% level. This result can be used to remove gender as a confounding variable for hours of social media use per day which may be helpful for analysis of other results.

Discussion of Results

The objective of this study is to find whether social media use is associated with positive, negative, or no psychological and emotional effects on students at UW-Madison. In original analysis of the data, it appeared that there may be significant effects on students' self-reported stress level and GPA in particular. The analysis of the data thus focuses on associations between time spent on social media per day and stress level/GPA.

First it is relevant to understand whether gender is a confounding variable on social media use time per day. A two-sided t-test uncovers that there is no statistically significant relationship between gender and average social media use time per day. Gender can then be said to have no significant effect on social media use time, so if stress is found to have an association, gender can be discounted as a confounding variable. The sample size for each gender is not > 30, so the central limit theorem cannot be applied. A t-distribution is instead used to calculate the p-value and confidence interval, which is more conservative and accounts for the sample size being smaller than it should be ideally. This issue could be resolved with a larger sample size.

Next it is tested to see whether social media use time is associated with stress level. The data is split by stress level into groups of a stress level > 5 and ≤ 5. The difference of mean average social media use time per day is calculated and a randomization test results in no statistically significant relationship being found. Again, the size for each group is not > 30, so a t-distribution is used to account for this. It can be seen by the density plot of the randomization distribution that the distribution is normal so no assumptions are violated. Since no statistically significant association is found, the results imply that social media use time per day has neither a positive nor negative effect on stress level.

The data is also analyzed to see if there is any significant relationship between hours spent on social media per day and GPA. The variable TimeOfDay is used as the explanatory variable and the variable GPA is used as the response variable for a simple linear regression. The regression shows no significant linear relationship between the two variables. Looking at the scatterplot of the two variables, there appears to be no relationship. There appears to be fanning in the data, which means the equal variance of residuals assumption of linear regression is violated. The violation of equal variance implies that a simple linear regression is not an appropriate model for the data. Because of this, the test does not help answer the research question.

The method of data collection resulted in the sample being neither random nor representative of the population. The way in which recipients were selected for the survey was not random. The sample was 100% Asian/Pacific Islander, and the population proportion is 7.6%[4], which means the sample is not representative. Due to these issues, the results of the tests must be thrown out.

Given more time and resources, the assumptions that are violated could be resolved. A more random sample with a much larger sample size could be obtained by randomly selecting a sample of students from a list of all undergraduate students at UW-Madison and requiring each one of them to take the survey. The resulting sample should then be random, representative, normal, resilient to outliers, and able to have the central limit theorem applied for tests. Any results could then be applied to the population of interest.

Ideally, a different form of study other than a survey would be performed. The survey, which asked how much time a student spends on social media per day, is based on having the student estimate the figure. A study could instead be performed that more closely monitors the exact amount time the student spends on social media. Instead of having the student rate their stress level on a scale, the student could take a test such as the BDI-II-II test taken in the study performed by Pantic I, Damjanovic A, Todorovic J, et al[2]. The data collected in place of a self-report of stress level would be more accurate and much less subjective. The results would then be more easily applied to the population.

The data collected in this test did not have a large enough sample size and was not random or representative and violations of statistical assumptions were an issue. Due to the small sample size, standard errors for variables were large and thus p-values were large and did not indicate statistically significant results. The results of this study's statistical tests imply that social media use has neither positive nor negative psychological and emotional effect but must be thrown out because the sample was not random or representative. Given a larger sample size and a more random method of collecting data, it is believed many of these issues would be resolved. However, the results of this study are inconclusive.

References

1. Kross E, Verduyn P, Demiralp E, et al. Facebook use predicts declines in subjective well-being in young adults. PloS One 2013; 8:e69841
2. Pantic I, Damjanovic A, Todorovic J, et al. Association between online social networking and depression in high school students: behavioral physiology viewpoint. Psychiatria Danubina 2012; 24:90-93
3. Smith, A., Anderson, M. (2018, March 1) Social Media Use in 2018. Retrieved from http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/
4. Enrollment Report Fall 2018-2019. Retrieved from https://d1hpan0pemm9te.cloudfront.net/wp-content/uploads/sites/36/2018/09/report-enrollment-2018fall.pdf

Appendix

1. T-test for Gender and TimeOfDay

```
raw_file = read.csv("C:/Users/shaun/Documents/Book1.csv")
data = subset.data.frame(raw_file,select = c(Gender,TimeOfDay))
male =data$TimeOfDay[data$Gender=="Male"][-19]

p.value = t.test(x = data$TimeOfDay[data$Gender=="Female"], y = male)$p.value
mean_diff  = mean(data$TimeOfDay[data$Gender=="Female"]) - mean(male)
CI.lower = t.test(x = data$TimeOfDay[data$Gender=="Female"], y = male)$conf.int[1]
CI.upper = t.test(x = data$TimeOfDay[data$Gender=="Female"], y = male)$conf.int[2]
sd.f = sd(data$TimeOfDay[data$Gender=="Female"])
n.f = length(data$TimeOfDay[data$Gender=="Female"])
sd.m = sd(male)
n.m = length(male)
t.stat = mean_diff / sqrt((sd.f^2/n.f)+(sd.m^2/n.m))
```

2. Randomization test for difference in means, TimeOfDay and StressLevel

```
library(ggplot2)
MyData <- read.csv(file="C:/Users/shaun/Documents/Book1.csv", header=TRUE, sep=",")
NewData <- subset(MyData,
select=c(StressLevel, TimeOfDay))
low = NewData$TimeOfDay[NewData$StressLevel<=5]
high = NewData$TimeOfDay[NewData$StressLevel>5][-29]
mlow = mean(low)
mhigh = mean(high)
mdiff = mlow - mhigh
sdlow = sd(low)
sdhigh = sd(high)
nlow = length(low)
nhigh = length(high)
t.score = mdiff/sqrt((sdlow^2/nlow)+(sdhigh^2/nhigh))

low = NewData$TimeOfDay[NewData$StressLevel<=5]
high = NewData$TimeOfDay[NewData$StressLevel>5][-29]
mean.l = mean(low)
mean.h = mean(high)
mean.all = mean(c(low,high))
## null samples with equal means
low0 = low - mean.l + mean.all
```

```
high0 = high - mean.h + mean.all
## Randomization
B = 10000
n.l = length(low)
mat.l = matrix(sample(low0,size=B*n.l,replace=TRUE), B, n.l)
n.h = length(high)
mat.h = matrix(sample(high0,size=B*n.h,replace=TRUE), B, n.h)
rand.l = apply(mat.l,1,mean)
rand.h = apply(mat.h,1,mean)
rand.diff = rand.l - rand.h
SE = sd(rand.diff)
CI = mdiff + c(-1,1)*1.796*SE
## p-value
p.value = mean( abs(rand.diff) >= abs(mean.l - mean.h) )
## Graph the randomization distribution
df = data.frame(rand.diff)



ggplot(df, aes(x=rand.diff)) +
  geom_density() +
  geom_vline(xintercept=mean.l-mean.h, color="red",linetype="dashed") +
  xlab('difference in time (high vs. low stress level)') +
  theme_bw()
```

3. Regression of GPA on TimeOfDay

```
gpa <- c(3, 3.5, 3.86, 3.9, 3.1, 3.7, 3.68, 3.8, 3.8, 3.82, 4, 3.89, 3.8, 3.33, 3.56, 3.4, 3.925, 4, 3.9
hours <- c(4, 2.5, 3.5, 3, 3, 2, 5.5, 3, 6, 2, 3, 12, 2, 6, 2.5, 2, 2, 2, 4, 4, 5, 1.5, 1.5, 1, 1, 10, 3
model <- lm(gpa ~ hours)
plot(hours, gpa, pch=21, bg="dimgray",cex=1.2,
     xlab="hours",ylab="GPA",main="Scatterplot of hours and GPA")
abline(model, lwd=3, col="red")
b1.hat = coef(model)[2]
dof=40-2
t = b1.hat/0.01823
t.crit = qt(.975,dof)
CI.slope = b1.hat + c(-1,1)*t.crit*.012490
p.val = 0.321
```