Who Plays Video Games?
Yong Liu, Yunlin Tang, Jian Jiao
MATH 189, Spring 2020, HW2

## I. Introduction

Each year, there will be 3000-4000 students enrolling in statistics courses at UC Berkeley. To aid these students, a committee of faculty and graduate students have designed a series of computer labs. Therefore, to help the committee design the labs, a survey was conducted in order to determine the extent to which the students play video games and which aspects of video games they find most and least fun. The data were randomly collected 91 out of 314 students in Statistics 2, section 1, during Fall 1994. Since each person in the population is randomly chosen without replacement from the population and each person was equally likely to be chosen, the sample obtained is a simple random sample. There are 15 variables in the survey dataset and 21 in the followup survey dataset, and both of them contain 91 observations. In the following analysis section, we will investigate different scenarios in order to analyze the data.

## II. Analysis

### A. Scenario 1

Begin by providing an estimate for the fraction of students who played a video game in the week prior to the survey. We use both the point estimate and interval estimate from the sample to represent the population parameter, which is the proportion of students who played a video game in this particular week.

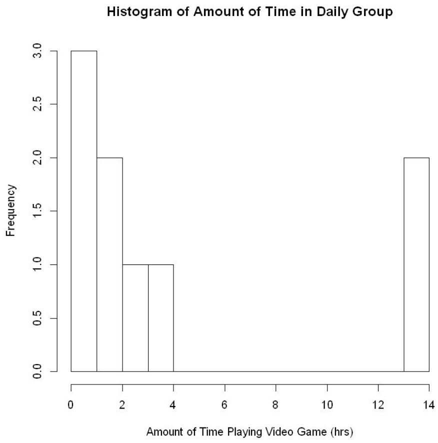
First, calculate the sample proportion (x bar) of students who played a video game this week by extracting information from the feature “time” which represents the number of hours played in the week prior to the survey. Then convert the number which is larger than 0 in “time” to 1 which represents the student had played, and convert the remaining to 0 which represents the student had not played this week. By calculating the mean of this converted vector we obtain the sample proportion, which is 0.3736. Then calculate the sample variance S^2 and the sample standard deviation S. By using the S to produce the approximate 68 confidence interval and 95 confidence interval (formula see Appendix A), found that they are (0.322, 0.424) and (0.271, 0.475) respectively.

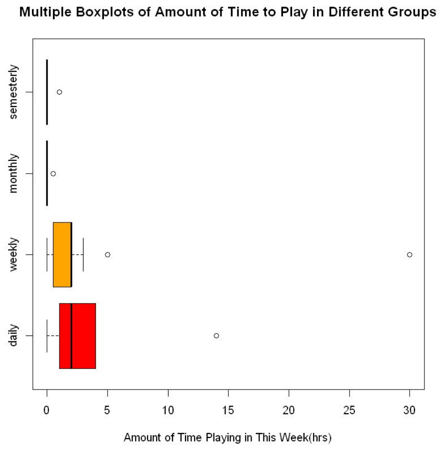
Since the sample is a simple random sample, the sample average (sample proportion in this case) is the sample statistic that can estimate the population parameter (population proportion in this case). In other words, this sample average 0.37 is an unbiased estimator of the population parameter. In addition, since the population variance is unknown, use the sample variance instead to produce the interval estimate (the approximate 68% and 95% confidence interval) for the population parameter which is the population fraction of students who played a video game in this particular week. It indicates that our sample proportion has a 68% chance to be within the interval (0.322, 0.424), and a 95% chance to be within the interval (0.271, 0.475).

## B. Scenario 2

In scenario 2, we check to see how the amount of time spent playing video games in the week prior to the survey compares to the reported frequency of play (daily, weekly, etc). By comparing these two questions, we want to see if there is any discrepancy between the reported amount of time and the usual frequency of playing video games. If so, we would further investigate how the fact that there was an exam in the week prior to the survey affect this comparison.

First, create a subset of data that contains people who answered '1' in the frequency question, which indicates they would usually play video games daily. Plot the histogram (left figure in below) of this group to check the distribution of the amount of time playing a video game in the week prior to the survey.

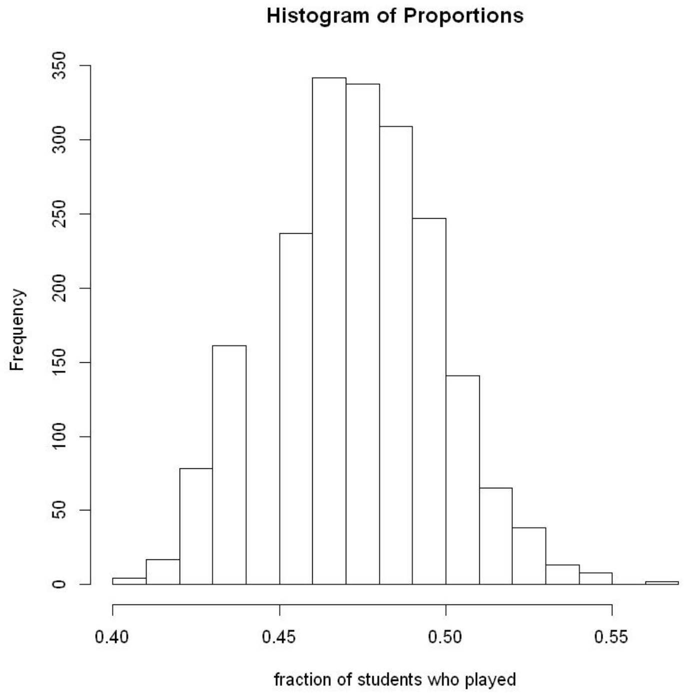




The histogram shows that the two most frequent amounts of time playing are in the [0, 0.5) interval. It is odd since we would expect people who usually play video games daily should at least play one hour per week. Then check how different the amount of time playing this week is between the daily and weekly groups. The boxplot (right figure in above) shows that the medians between these two groups are close which also indicates an odd event happened since we expect the daily people should spend more time playing.

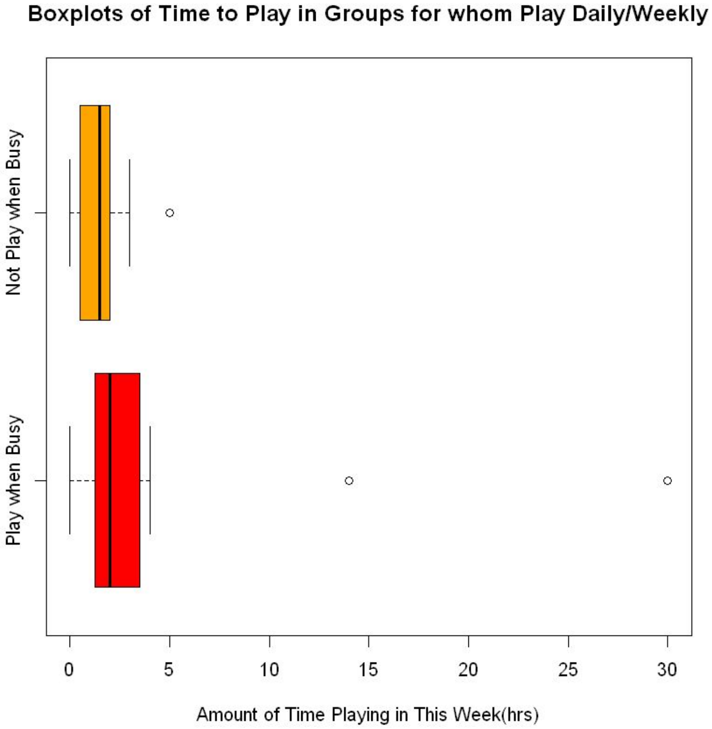
Then we decide to perform a simulation study by using the data in the sample which simulates the amount of time playing the video game during a usual week to further compare the data. Use 1 and 0 to represent they have played and did not play this week respectively. Based on their responses on the frequency question, assume the daily and weekly people will play this week, thus the probability of these people played is 1; then for monthly people, assume each of them has ¼ chance to play this week; for semesterly people, assume each of them has 1/15 chance to play since one semester roughly has 15 weeks; for those who did not answer or did not play video games at all, just assume that they have 0 probability to play. In each loop, randomly select 0 and 1 from the sample based on the probabilities given above and create a vector of size 91 which stores all the 0's and 1's. Then calculate means of each vector (2000 vectors in total since run for

2000 times), we obtain the fractions of students who played a video game in a usual week (see Appendix B for codes).



The histogram shown above reveals the distribution of fractions of students played in the usual week. Recalling the sample proportion is 0.3736 which indicates that there is only around 38% of people in the sample had played in the week prior to the survey. However, it is rare to happen in a usual week based on the histogram of simulation of the original sample. Thus we can suspect that there are some factors that influence the amount of time playing a video game in the week prior to the survey.

One potential factor might be the fact that there was an exam in the week prior to the survey. There is a question on the survey reveals that whether the responder will play video games when he/she is busy. By plotting the barplot of this feature, observe that there are significantly more people answering that they will not play (see appendix C). Then extracting the responders who answered that they will play weekly and daily in the frequency question, divide them into two groups: one who answered that they will play even when they are busy; one will not play. Plot the multiple boxplots to see the difference in the amount of time playing this week between these two groups.

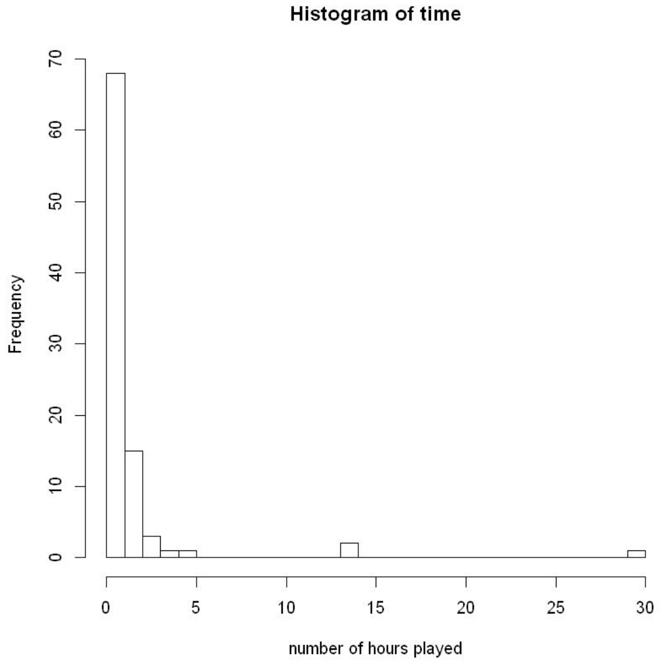


From the boxplot shown above, we see that the median in the group "play when busy" is slightly higher than the group "not play when busy". The 25th and 75th percentiles in the former group are also higher than the latter. It is obvious that there are numbers of people who play video games daily or weekly at usual will not play when they are busy. This observation might explain the discrepancy found above. Considering the fact that there was an exam in the week prior to the survey, it would be the potential factor that lowers the amount of time playing video games in this particular week.

## C. Scenario 3

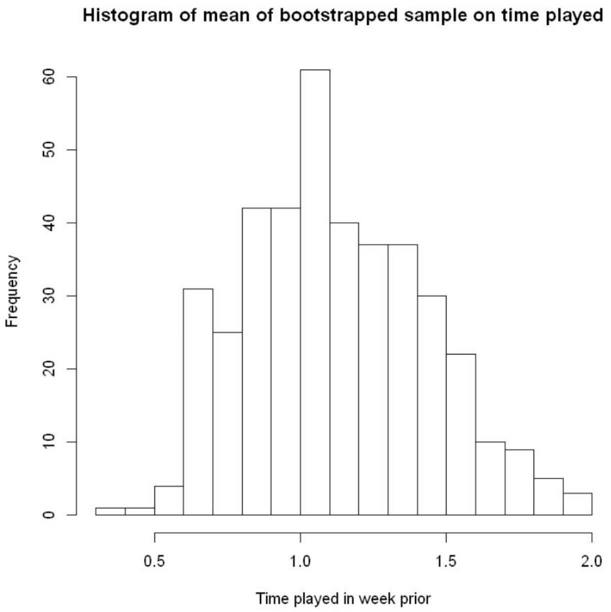
In this part, we will determine whether the mean of the amount of time spent playing video games in the week prior to the survey represents the real mean ( of time playing from the population by doing the interval estimate. The mean of the playing time from the sample set is 1.243. We will use the confidence interval of 95% to determine whether the mean of the sample is a good estimator by looking whether it falls in the interval [0.616,1.777]. Furthermore, we will also evaluate whether the distribution of the sample mean , which generates from bootstrap, is normal distribution or not, because if the distribution is not a normal curve, the confidence interval will not be helpful.

First, let's take a quick look of the distribution of the mean of playing time in sample:

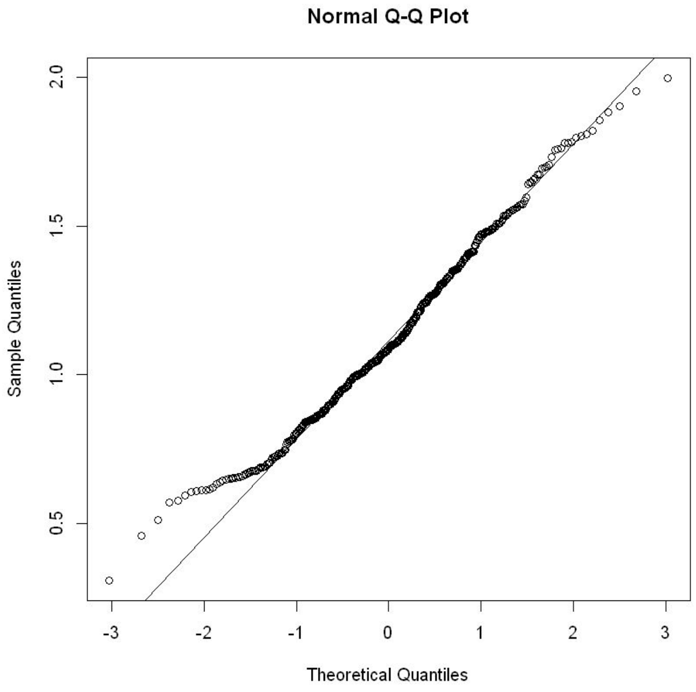


As you can see, it is obvious that the distribution of playing time in the sample set is skewed.

After performing the bootstrap without replacement, we get a figure of the distribution :



Based on the histogram , we can see that there is a belt-shape which is a normal distribution curve. However, we should not just make conclusions based on visuals in mathematical analysis. In order to test whether it is a normal distribution, we will use a quantile-quantile plot. See the figure below:



According to the qq plot above, we can notice that the actual data points approximately  fall on the theoretical line. However, the left tail of the data points is a little away from the line. So we need to do further analysis: we compare the kurtosis and skewness of a 500 bootstrap sample average with a normal s and k. Bootstrap skewness = 0.46 ; Bootstrap kurtosis = 3.33;  Normal skewness = 0.04 Normal kurtosis = 3.0: They are pretty close.

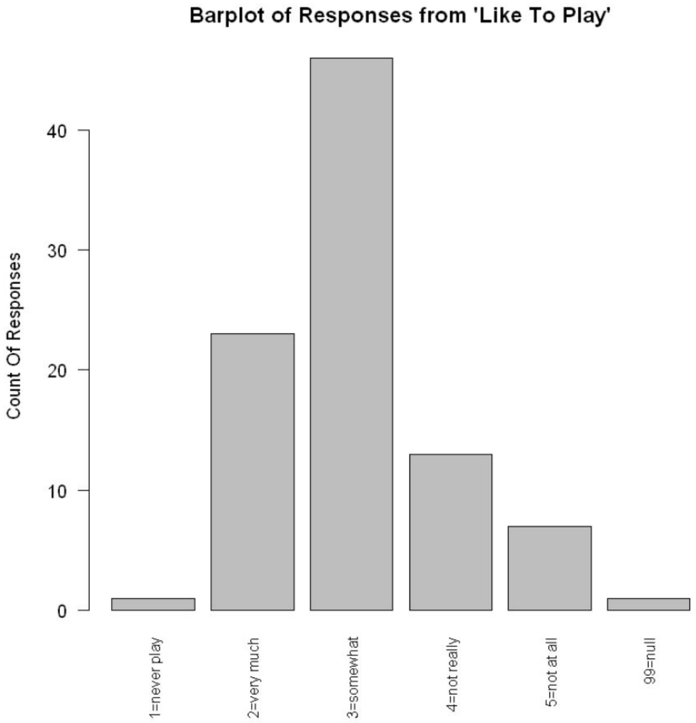
Therefore, we can come to the conclusion that the mean of time played from 91 students is good to represent the population and the distribution of the mean of bootstrap is normal distribution. The point estimate of the mean of bootstrap samples is 1.16.  falls in the interval [0.616,1.777].

## D. Scenario 4

In scenario 4, we check to see students' attitude towards video games.

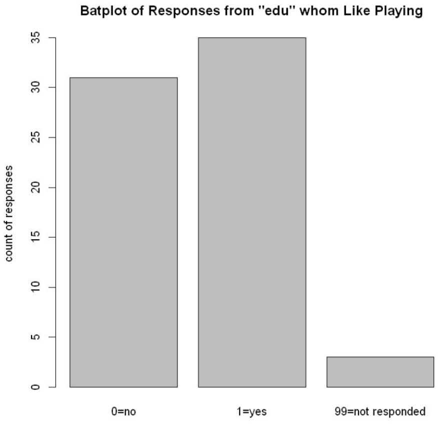
We first visualized the overall distribution of whether respondents like to play. Then we move on to the reasons that stimulate students to play games. We also discovered how students treat video games on an educational level by extracting "edu" data and visualizing. In addition, we figured out what students dislike about video games.

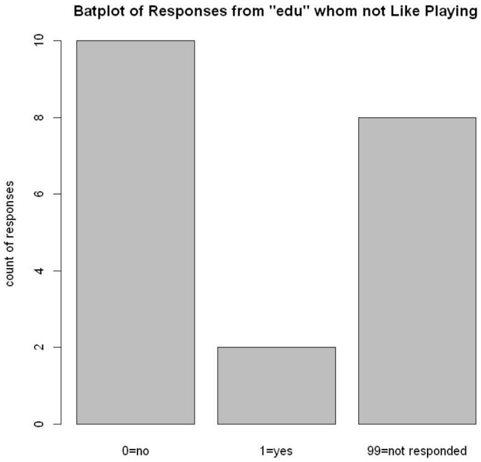
To show the distribution of whether respondents like to play, we created a bar chart below. From the graph, we can easily see that approximately 70% of students reported to like to play video games.



As the main quest of students is to study, we want to explore how educational proposes act as one of the reasons which makes students like to play video games. Based on the responses, we created bar charts to see whether students consider video games educational.

The graph shows that nearly 50% of students who like to play consider video games educational. While only 2 students of those who do not like to play think games have meaning of education.

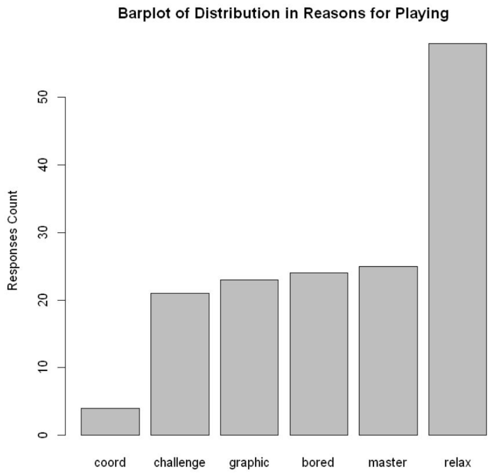


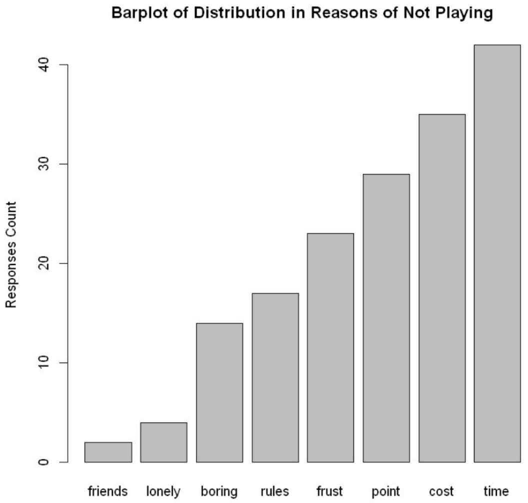


Since there are students who have never played video games or do not at all like video games, skip some questions, we dropped these null values and use the rest of data to construct two bar charts to show reasons why students like to play and dislike video games.

Results show that more than 50% of the respondents play games for relaxing, while sense of master, which is the second most selected reason, only takes about 25%. Which means that relaxation is the most important reason for students to play video games.

Besides that, we also discovered why students dislike video games. From the second graph below, we can see that more than 70% of students consider video games wasting time and costing much.



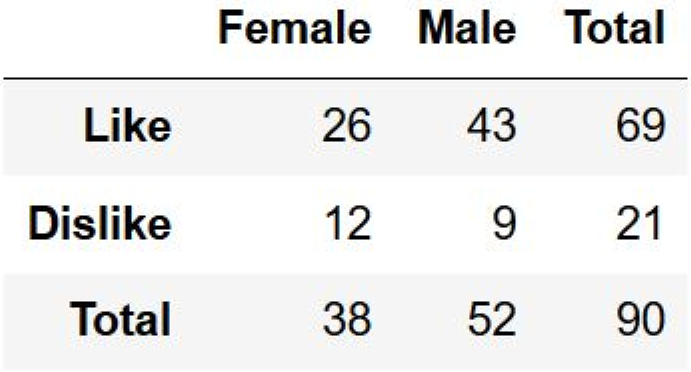


## E. Scenario 5

In scenario 5, we look for differences between students who like to play video games and those who do not by doing cross-tabulations.We created cross-tabulation tables to compare male and female students, students work for pay and those who do not and students who owned PC and those who do not owned PC.

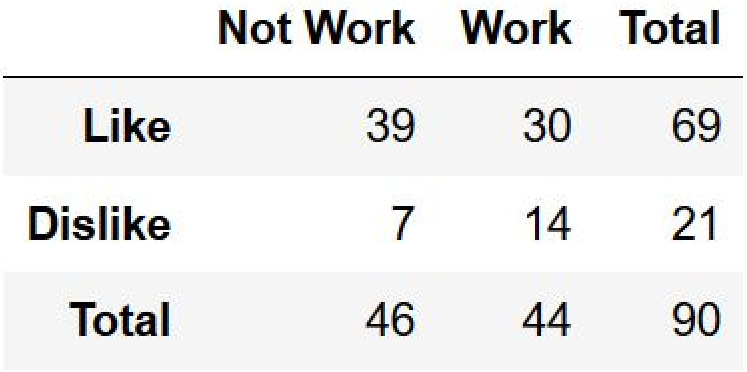
To make it easier to compare like and dislike, we first create a criteria to distinguish who likes who does not like:
Those who play video games very much (answered 2 for "like" question) and play video games somewhat (answered 3 for "like" question) are classified to "like" category and participants who answered 1 (Never), 4(Not really), 5(Not at all) are classified to dislike video games.

We first compare the difference in video game preferences between male and female survey participants by doing cross-tabulation.



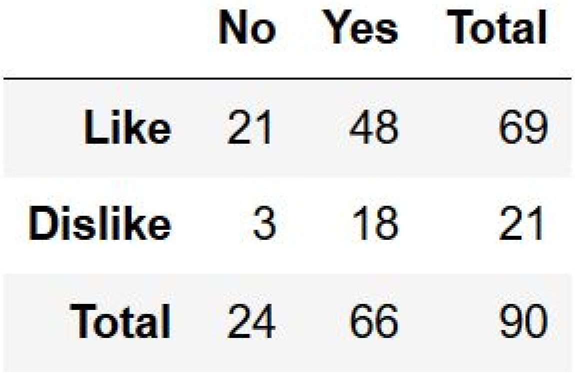
| | Female | Male | Total |
|---|---|---|---|
| **Like** | 26 | 43 | 69 |
| **Dislike** | 12 | 9 | 21 |
| **Total** | 38 | 52 | 90 |

We classified students who work 0 hours in a week as "Not Work" and students work more than 0 hours as "Work". After that, a cross-tabulation table was created.



| | Not Work | Work | Total |
|---|---|---|---|
| **Like** | 39 | 30 | 69 |
| **Dislike** | 7 | 14 | 21 |
| **Total** | 46 | 44 | 90 |

We then create a table for comparing the difference in video game preferences between survey participants who have owned a PC and those who have not.

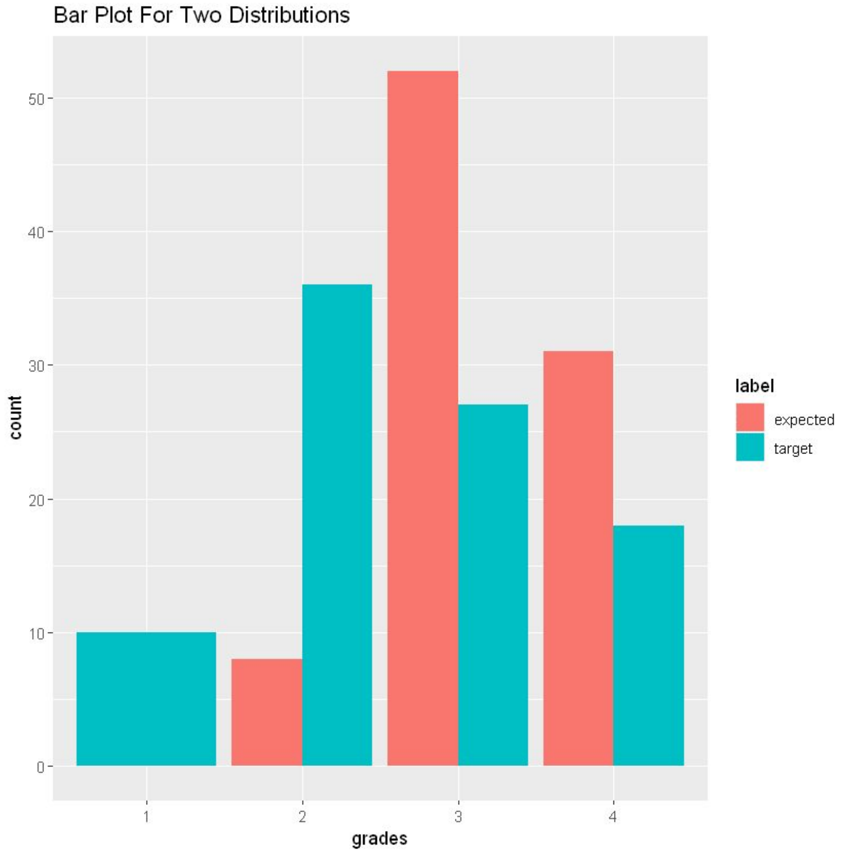


| | No | Yes | Total |
|---|---|---|---|
| **Like** | 21 | 48 | 69 |
| **Dislike** | 3 | 18 | 21 |
| **Total** | 24 | 66 | 90 |

The result from table 1 shows that males (82.7%) are more likely to enjoy video games than females (68.4%). Table 2 shows that 83.7% of participants who do not work like video games, however, only 68.2% of participants who work liked video games. Table 3 shows that 87.5% of participants who do not own a PC enjoy video games while 72.7% of participants who owned a PC enjoy video games.

## F. Scenario 6 (Advanced Analysis)

In this section, we have investigated the difference between the expected grade distribution and the target grade distribution. The expected grade is extracted from the responses of the 'grade' question which asked the respondents to answer their expected received grade in this class. The target grade distribution is used in grade assignment of 20% of A's, 30% of B's, 40% of C's, and 10% of D's. First, plot the barplot for each distribution side by side to visually inspect the difference.



The barplot in above shows the counts of grades in each distribution, where 1 represents D in grade, 2 for C, 3 for B, and 4 for A. It shows that the respondents in the sample are a little bit optimistic compared to the target grade distribution, for example the count of C's in target distribution is significantly higher than the expected distribution. We further use K-S test to confirm our guess that these two datasets do not come from one distribution. The null hypothesis is that these two datasets have the same distribution function. However, the small p-value carried out from the K-S test is 2.568e-07, which rejects the null hypothesis thus confirms our assumption that these two distributions are different. In addition, by counting those non-respondents (4 in total) as failing students who expect their grades as F, we again perform the K-S test then obtain a small p-value 2.379e-06, which is slightly larger than the p-value before.

## III. Importance and Conclusion

Our group successfully achieved our research goals by doing numerical analysis and graph visualizations and statistical estimates.

From Scenario 1, we have found that the fraction of students who played a video game in the week prior to the survey has 95% chance to lie in this interval (0.271, 0.475).

Scenario 2 shows that people who play video games daily or weekly as usual will not play when they are busy.

Scenario 3, we use interval estimates to come to the conclusion that the mean of time played from 91 students is good to represent the population and the distribution of the mean of bootstrap is normal distribution. The point estimate of the mean of bootstrap sample is 1.16.

Scenario 4 shows that relaxation is the most important reason for people who play video games and wasting time is the most important reason preventing students from liking video games.

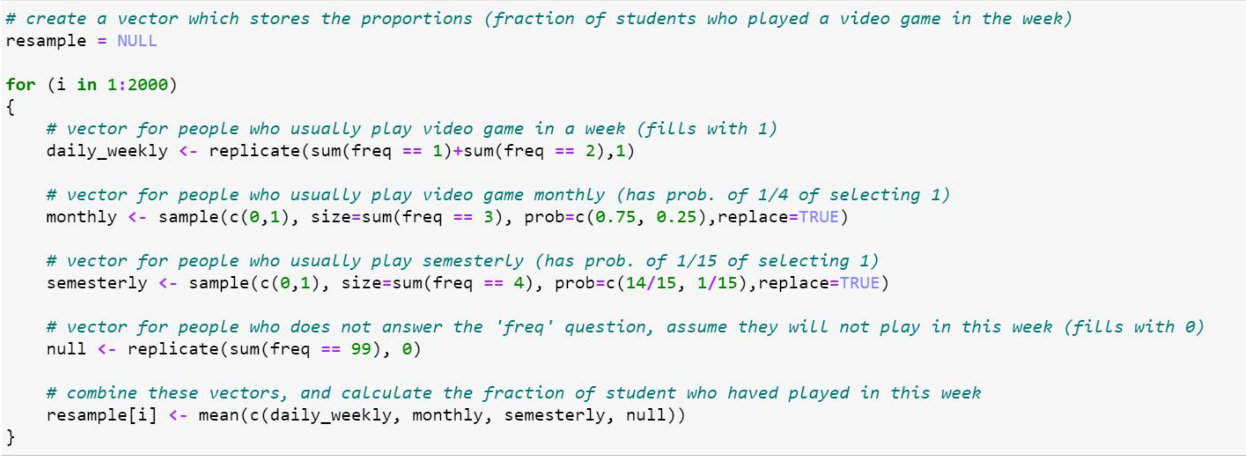
Scenario 5 shows that males tend to like to play video games more than females, people who do not work like to play games more than those who do not work, and people who do not own a PC like to play more than those who owned.

The limitation of our study is that our findings could not be generalized to the whole student population since the dataset is relatively small.

## IV. Appendix

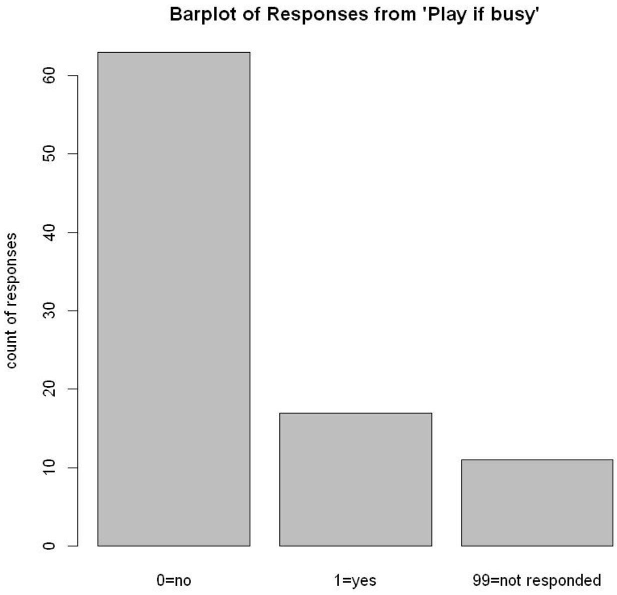
A. 68% Confidence Interval Estimate Formula: $(\bar{x} - \sigma/\sqrt{n}, \bar{x} + \sigma/\sqrt{n})$, and 95% Confidence Interval Estimate Formula: $(\bar{x} - 2\sigma/\sqrt{n}, \bar{x} + 2\sigma/\sqrt{n})$. When the population variance $\sigma$ is unknown, we use the sample variance s for substitution.

B. Code for Scenario 2:



```
# create a vector which stores the proportions (fraction of students who played a video game in the week)
resample = NULL

for (i in 1:2000)
{
    # vector for people who usually play video game in a week (fills with 1)
    daily_weekly <- replicate(sum(freq == 1)+sum(freq == 2),1)

    # vector for people who usually play video game monthly (has prob. of 1/4 of selecting 1)
    monthly <- sample(c(0,1), size=sum(freq == 3), prob=c(0.75, 0.25),replace=TRUE)

    # vector for people who usually play semesterly (has prob. of 1/15 of selecting 1)
    semesterly <- sample(c(0,1), size=sum(freq == 4), prob=c(14/15, 1/15),replace=TRUE)

    # vector for people who does not answer the 'freq' question, assume they will not play in this week (fills with 0)
    null <- replicate(sum(freq == 99), 0)

    # combine these vectors, and calculate the fraction of student who haved played in this week
    resample[i] <- mean(c(daily_weekly, monthly, semesterly, null))
}
```

C. Barplot of Responses from “Will you play video games when you are busy”:



## V. Contribution Statement

Yunlin worked on the introduction, scenario 1, scenario 2, partially scenario 4 and scenario 6.
Jian worked on scenario 5, partially scenario 4 and conclusions.
Yong worked on partially scenario 2, scenario 3, and edit/combine the codes.