Video Games and Computer Lab Design

Math 189 - Investigation #2

Professor Schwartzman

Spring 2020

Arthur Chang - A14410373

Raya Kavosh - A14826756

Siddharth Saha - A15572442

Contributions (edit later)
Raya - Introduction, Scenario III, Conclusion
Arthur - completes scenarios 1 and 4. For 1, I performed bootstraps to create a sampling distribution and utilized a kurtosis test to test for normality. Analyzed and came up with point estimates and interval estimates. For 4, I analyzed the major criterion for likeability.
Siddharth - Came up with and completed ideas/analysis of scenarios 2, 5 and 6 (which include graphical analysis, 2 sample proportion tests, chi squared goodness of fit tests and decision tree classifiers) for the videodata dataset as well as detailed on some of the decisions in their respective appendix section

Introduction

Every year at UC Berkeley, 3,000-4,000 students enroll in Statistics courses, half of which take introductory statistics courses to satisfy the quantitative reasoning requirement. As an additional instruction resource, a committee of faculty and students have designed a series of computer labs intended to provide an alternative, interactive learning environment for statistics and probability. Since some have linked labs to video games, students in advanced statistics courses conducted a survey of undergraduate students enrolled in a lower-division statistics course regarding the extent to which the students play video games and which aspects of video games they find most and least fun. The responses will be used to provide insight to the designers as to what features in the computer lab would be the most attractive and beneficial to students.

Data

Out of 314 students in Statistics 2, Section 1, during the Fall of 1994, 95 students were selected at random to participate in the survey. The list of all students who had taken the second exam of the semester was used to select the students to be surveyed. The exam was given a week prior to the survey. To limit the number of nonrespondents, data collectors visited both the Tuesday and Thursday meetings of the discussion sections in the week the survey was conducted. On Friday, those students who had not been reached during the discussion section were located during lecture. Out of those 95 students, 91 completed the questionnaire.

They were asked to identify how often they play video games and what they like and dislike about the games. The answers were then coded numerically as described below.

Variable	**Description**
Time	# of hours played in the week prior to survey
Like to play	1=never played, 2=very much, 3=somewhat, 4=not really, 5=not at all
Where play	1=arcade, 2=home system, 3=home computer, 4=arcade and either home computer of system, 5= home computer and system, 6=all three
How often	1=daily, 2=weekly, 3=monthly, 4=semesterly

Play if busy	1=yes, 0=no
Playing educational	1=yes, 0=no
Sex	1=male, 0=female
Age	age in years
Computer at home	1=yes, 0=no
Hate math	1=yes, 0=no
Work	# of hours worked the week prior to the survey
Own PC	1=yes, 0=no
PS has CD-Rom	1=yes, 0=no
Have email	1=yes, 0=no
Grade expected	4=A, 3=B, 2=C, 1=D, 0=F

If a question was not answered or improperly answered, then it was coded as 99. Those respondents who had never played a video game or who did not at all like playing video games were asked to skip many of the questions.

The students were then given a follow up survey in which more than one response may have been given. Out of the categories *Action, Adventure, Simulation, Sports, and Strategy*, they were asked to check all types that he or she plays. Those who have never played a video game or do not at all like to play video games were instructed to skip this question.

Students who did answer this question were also asked to choose up to 3 reasons why they play the games they do from the following: *Graphics/Realism, Relaxation, Eye/hand coordination, Mental Challenge, Feeling of mastery, Bored*.

Finally, all students were asked to select up to three reasons for not liking video games from the following: *Too much time, Frustrating, Lonely, Too many rules, Costs too much, Boring, Friend’s don’t play, It is pointless*.

Video Games could be classified according to the device on which they are played and the kinds of skills involved. Four categories were classified as follows:

Device: arcade, console, PC
Arcade games: fast and emphasize eye/hand coordination
Console games: action, adventure or strategy games
PC games: simulation and role-play exclusively and other types as well

Research Questions

The objective of this study is to investigate the responses of the participants in the study with the intention of providing useful information about the students to the designers of the new computer lab. We will explore the following features of the survey data:

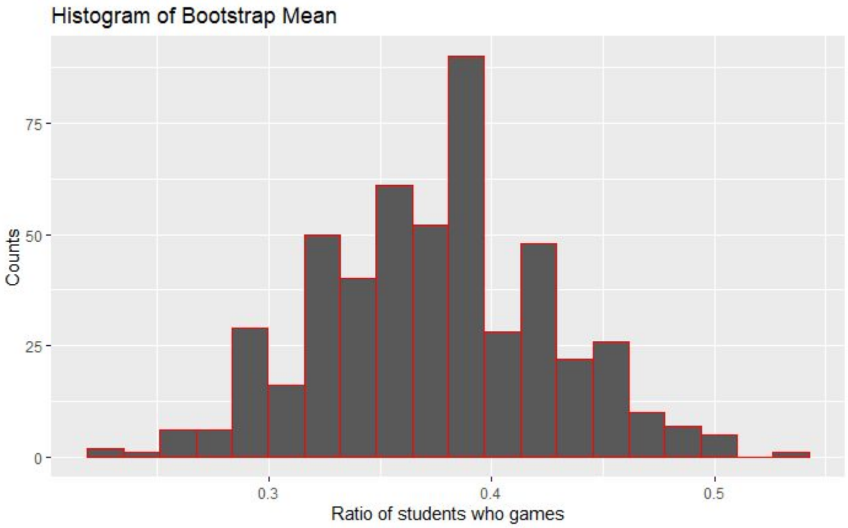
I. What fraction of students played video games in the week prior to the survey?
II. Does the fact that an exam was coming up affect the results of our survey?
III. Can the average time spent playing video games in the week prior to the survey be generalized to the entire population of statistics students who may utilize the new computer labs?
IV. What factors affect a student's likelihood of enjoying video games?
V. Are there any notable differences in the demographic of students who like to play games and students who don't like to play games?
VI. How close are the grades to the target distribution? What if all non respondents have already failed? Is the new distribution closer to the target distribution?

Analysis

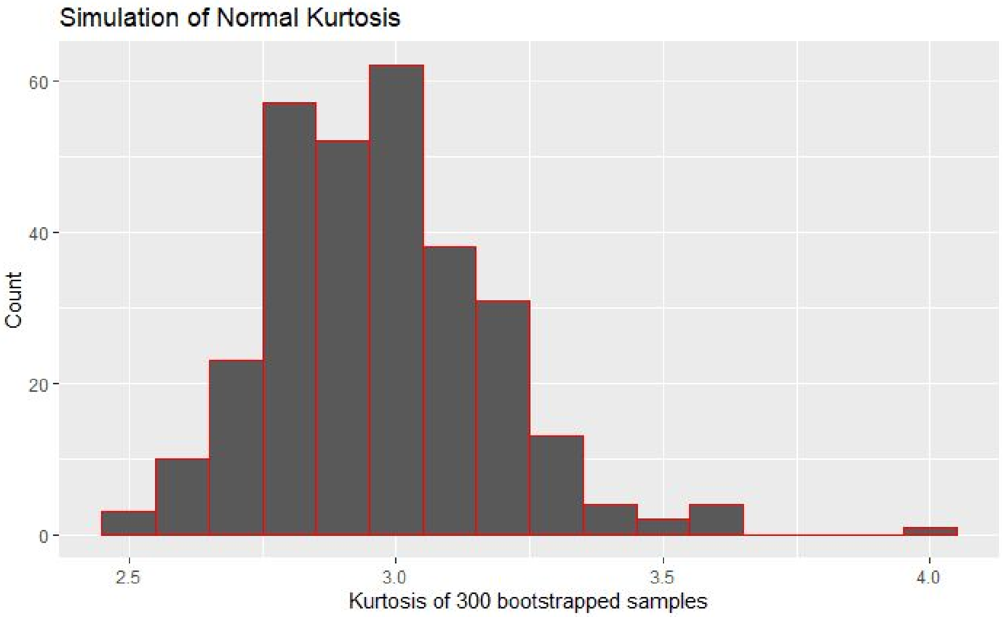
Scenario I

In order to get the fraction of students who played video games in the week prior to the survey, we first transformed the time played into dummy variables 0 and 1. After the calculation of sample mean within time feature we observed only 37.36 percent / 0.3736 of the 91 samples taken have played video games one week prior, this is also our point estimate.

The interval estimates are calculated with 95 percent confidence intervals over a normal distribution. In order to see if the data follows a normal distribution, we first run a bootstrap on the sample for 500 iterations and save the result to a variable.



We then performed the kurtosis test for 300 iterations as the outer loop for the sample mean bootstrap. The graph below showed the mode of kurtosis test to be right around 3, which means the samples are approximately normal.



After calculating for the 95 percent interval estimate over normal distribution we came to the result of: [0.2742318, 0.4730210].
Since n/N: 91/314 > 0.05, we decided to include population correction factor to get a better estimate of the interval. After population correction factor we get: [0.2654510, 0.4818017].

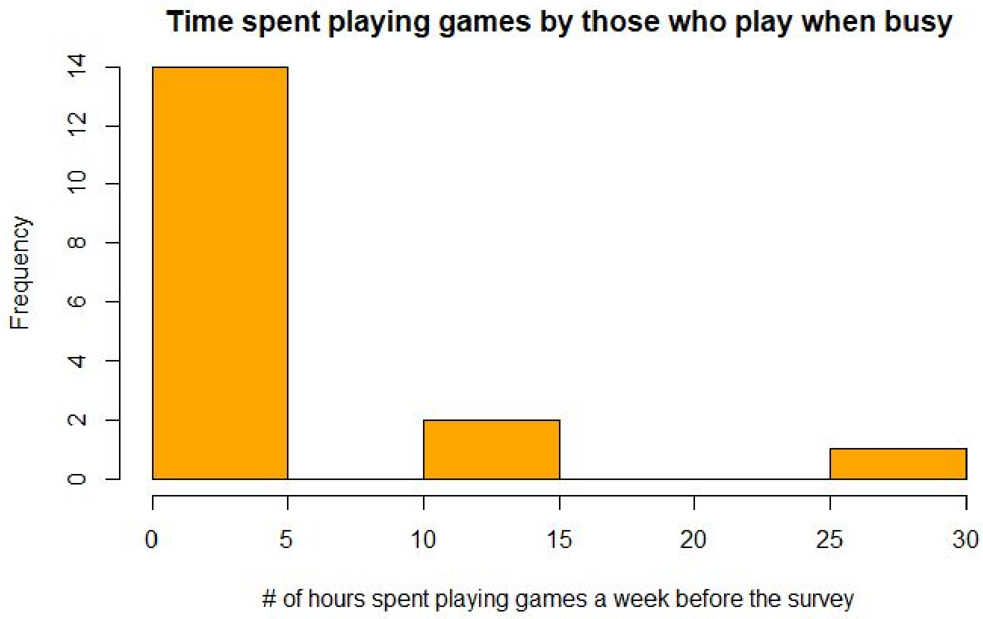
Scenario II

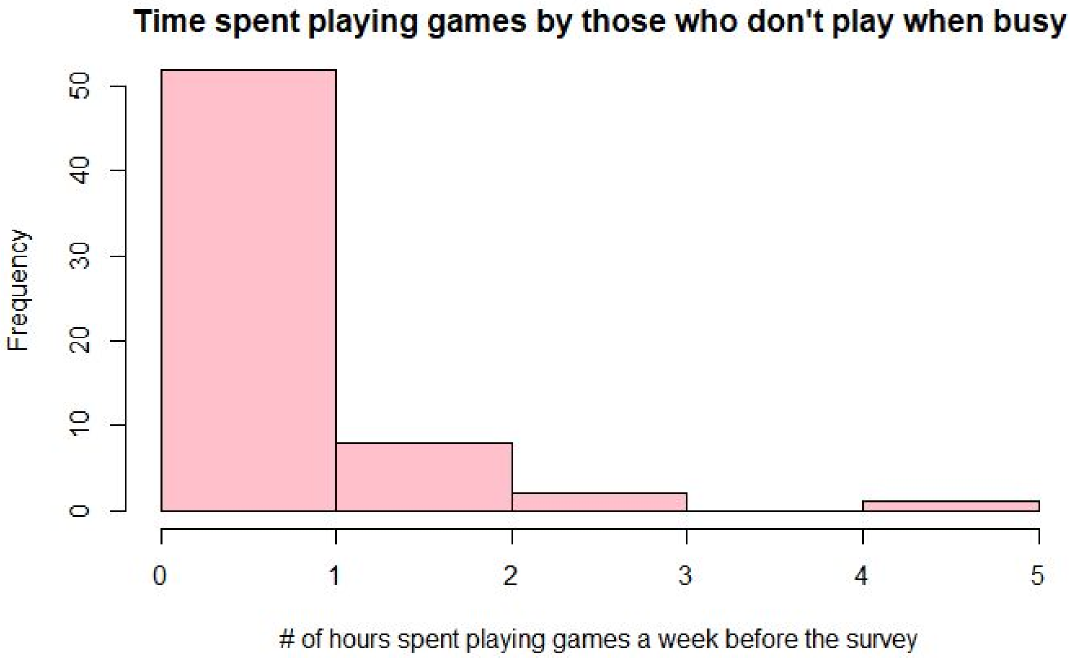
In this scenario we would like to compare the distributions of how frequently people play with the number of hours reported and see if the fact that an exam was coming up affects the results at all.

First let's see the number of people who don't play before exams and the number who do in our sample

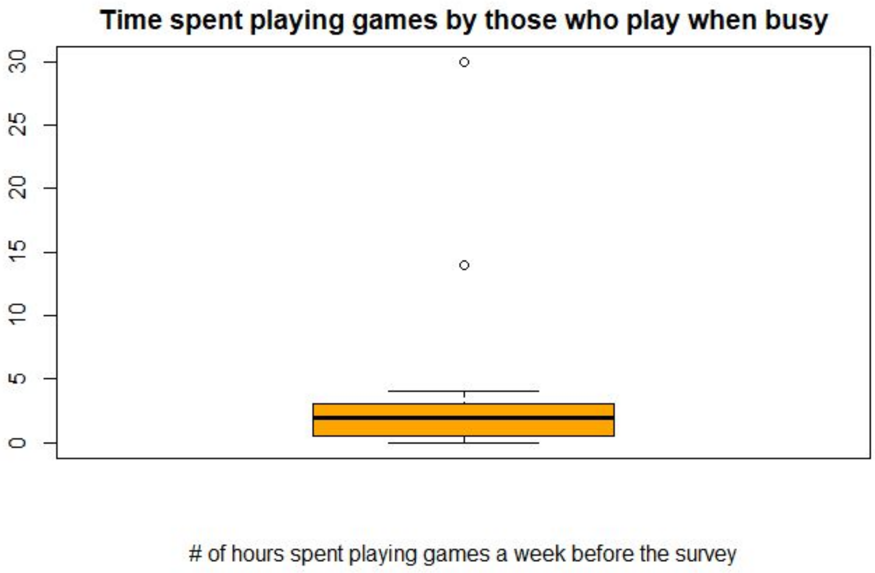
	Number in group
Don't play when busy	63
Play when busy	17

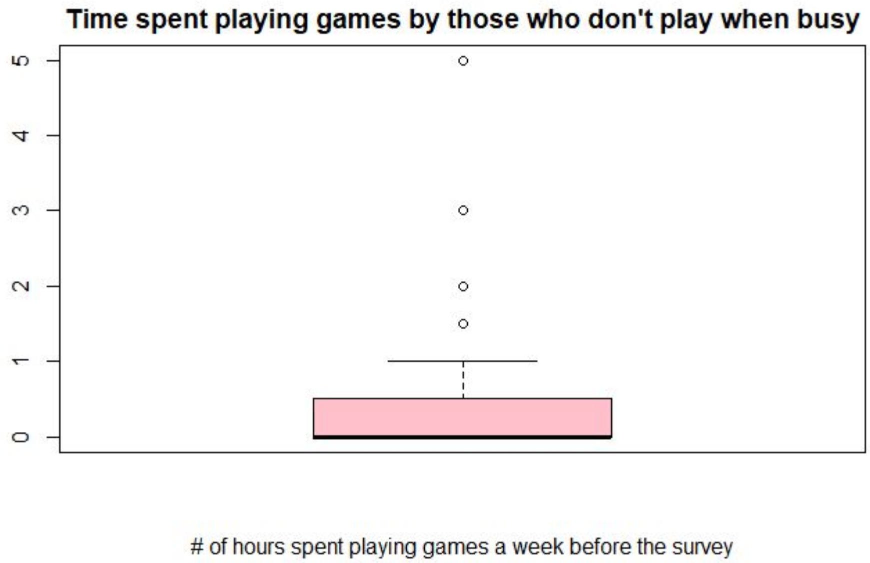
This tells us that we have a lot more people who prioritize their work over their games. This thus gives an indication that the fact an exam was coming up would affect the distribution of the amount of time they spent playing games in the week before the survey. Now let's analyze the time spent playing games by people who play when they are busy and people who play when they are not busy.



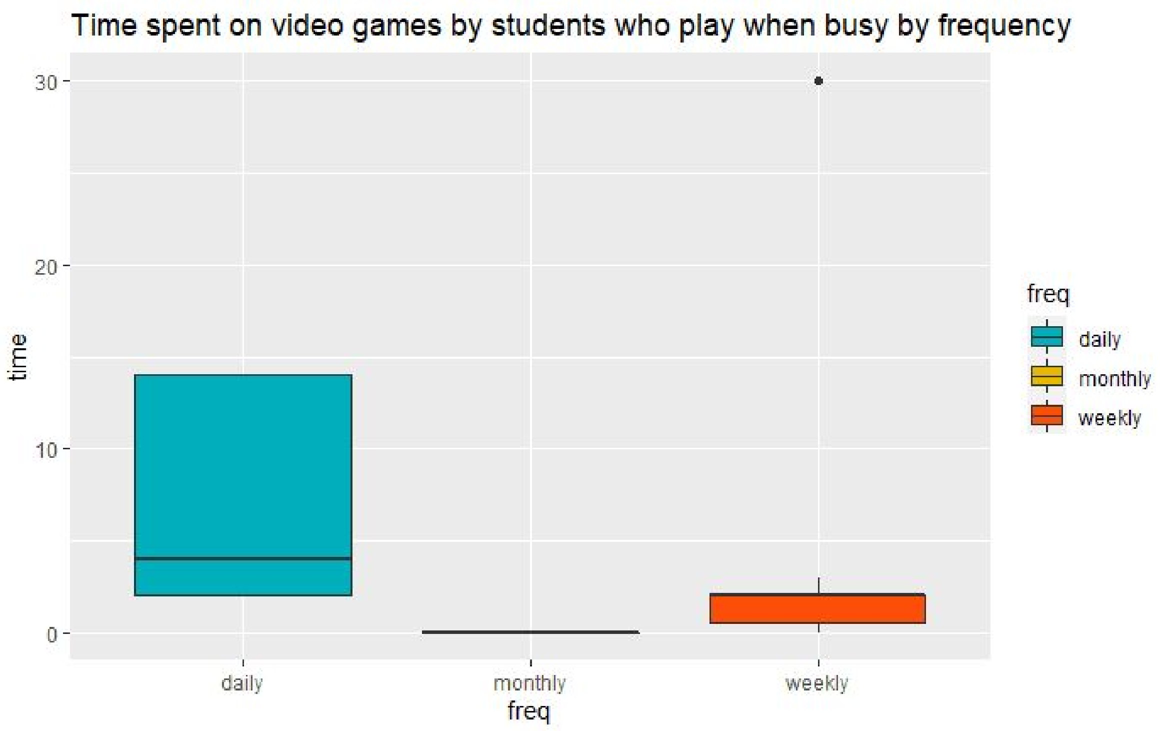


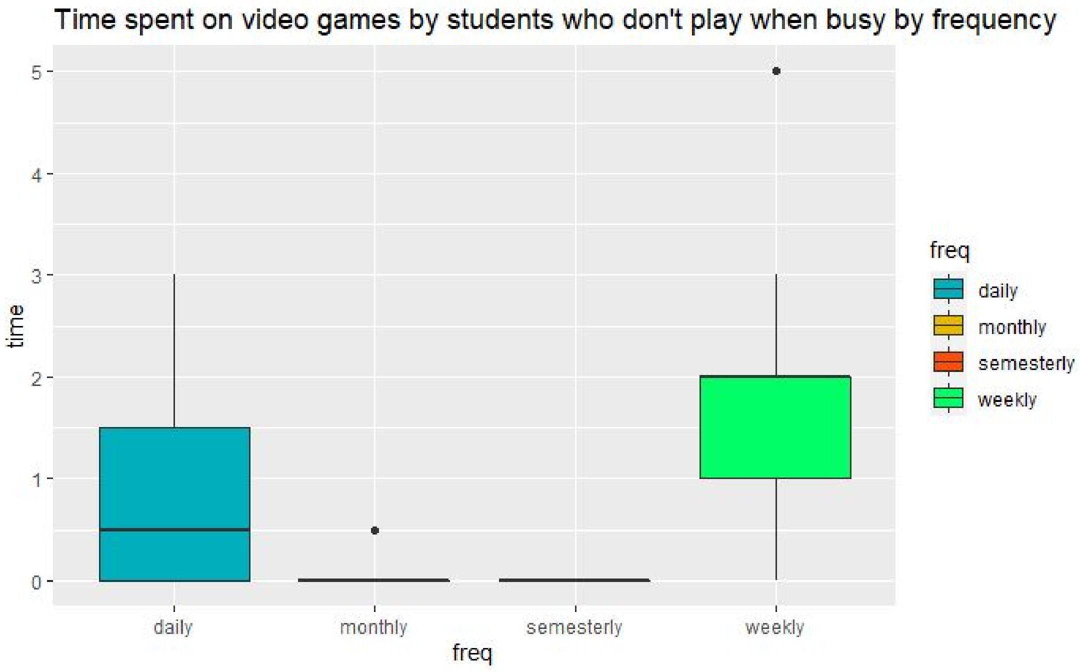
The histograms show that whether or not a student is busy can play a large factor in the number of hours played. The students who played games even when busy had logged in hours stretching up all the way to 30 while those students who did not play games only played up to 5 hours. A box plot would also help us confirm these findings through the median.





The boxplots show that the median number of hours spent playing games was higher among students who played games when busy were higher than the students who did not play games when busy. It also gives us an idea of outliers as for students who played games when busy any time above 5 hours was considered an outlier while for students who did not play games when busy anytime above 1 hour was considered an outlier. These significant differences matched with the number of students who don’t play when busy in our sample indicates that the fact that an exam was coming up would have seriously impacted our results. We could gain more in depth analysis by checking on the reported frequencies of playing.

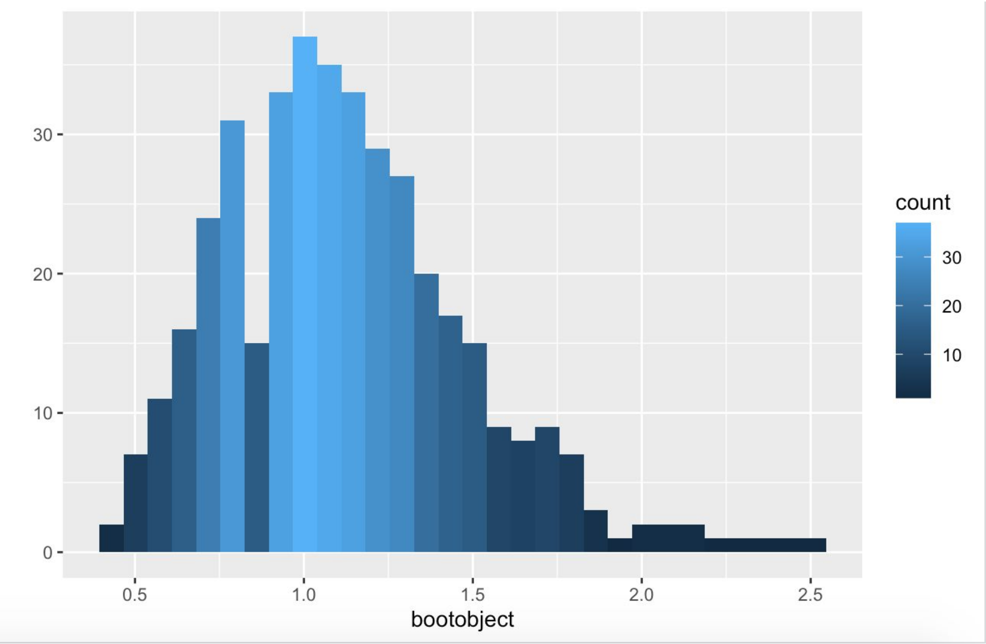




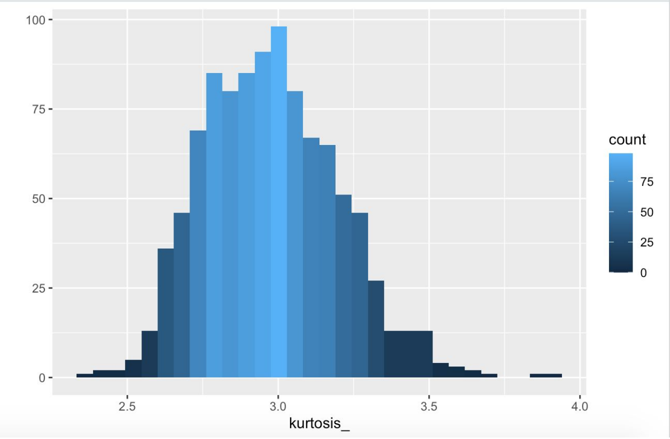
From these 2 boxplots we see that we have very few sample data points (in fact none for semesterly in students who play when busy) for students who play on a monthly or semesterly basis. However, this is not too big of a problem as the majority of the hours logged in our survey do come from the students who play on a daily or weekly basis. We do see that one sharp contrast is that the weekly students who don't play when busy have a much higher median than the ones who claim they play daily. This could either mean that the students who play on a weekly basis are either lying or that they spend the majority of their time in just a few days of the week (like a weekend). Being in college it is most likely the latter (especially because these are the students who say they don't play when they are busy) but we can't say for sure. For students who play even when busy we see the opposite trend i.e people who play weekly indeed do play lower than those who play daily signifying a slightly more hardcore gaming nature. Thus if there was indeed no exam coming up when this survey was taken our distributions would most likely be more skewed to the right indicating a significantly higher number of hours spent gaming in the week prior to the survey.

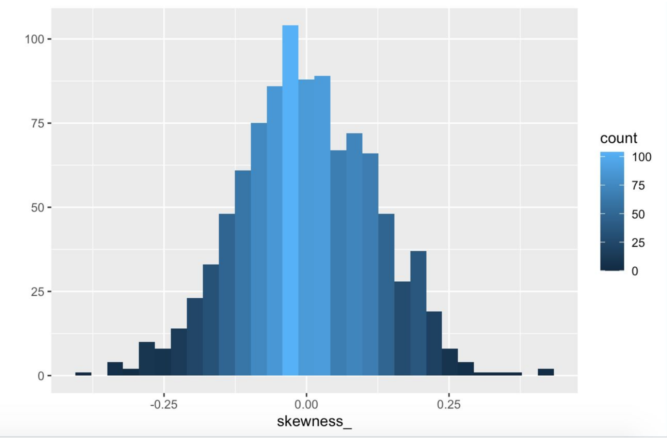
Scenario III

From our histograms above, we saw that the distribution of time students spent playing video games is very skewed. However, we don't have access to the entire population to take more samples. In order to determine whether the probability distribution of the sample average is normal, and therefore whether or not we can reasonably expect our sample mean to represent the true population mean, we used the bootstrap method. We created a bootstrap population of 314-the size of the original sample population- out of the 91 survey responses that were recorded in the original sample. That is, for each of the 91 students who responded, we created 3.45 units with the same time value, rounded to the nearest integer. Then from the bootstrap population, we selected a simple random sample of 91, with replacement- the bootstrap sample- and took its average. We repeated this process 400 times to simulate the probability distribution of the sample average, shown in the histogram below.



The average of all of the bootstrap sample means was 1.13, and the standard deviation was 0.36, meaning the standard error of the time spent video games was 0.36. This gives us a 95% confidence interval of [0.41, 1.85]. To justify this interval, we evaluated the normality of the distribution by comparing the kurtosis and skewness values from the bootstrap sample distribution to those of a normal distribution with sample size 400. Plots for normal kurtosis and skewness distributions appear below.





The average kurtosis value of a normal distribution, found through simple random sampling, was 2.98. A kurtosis value of 3 implies a lack of outliers. The average skewness value was 3.954293e-05. Skewness very close to 0 implies a symmetric distribution.

In comparison, the kurtosis value of our distribution was 3.86, meaning that our distribution is heavy-tailed, or has outliers. The outliers can be seen in the right tail of the histogram of sample means from the bootstrap population. Further, our skewness value was 0.76, meaning our data is positively skewed.

From the differences in these values from the values of a normal distribution, we can conclude that the time spent playing video games is not normally distributed. There are many high outliers from students who reported playing video games for over two hours, and therefore it is not appropriate to provide the confidence interval obtained from the bootstrap sample.

Scenario IV

In the following data displayed in scenario 5, it is observed that most of the students regardless of gender, working status etc mostly enjoy video games. In terms of what reasons students enjoy or not enjoy the games, we can analyze the features below.

In order to correctly analyze reasons why students like or dislike video games, we first removed the students who do not play video games (1) and those who don't enjoy them at all (5).

Graphic/Realism	26%
Relaxation	66%
Eye/hand coordination	5%
Mental challenge	24%

Feeling of master	28%
Bored	27%

Reasons given for those that like the video game

Too much time	48%
Frustrating	26%
Lonely	6%
Too many rules	19%
Costs too much	40%
Boring	17%
Friend's don't play	17%
It is pointless	33%

Reasons given for those that dislike the video game

Deduced from the above chart, we observe top 3 contributing reason for liking video games for students are:

1. Relaxation
2. Feeling of mastery
3. Boredom

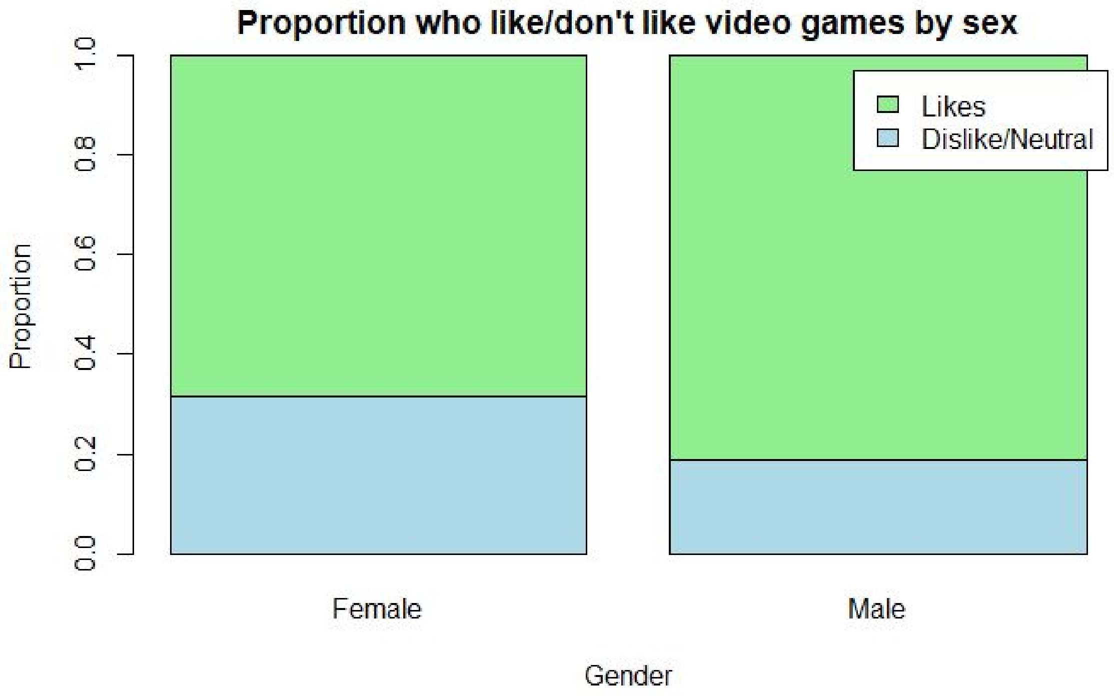
Deduced from the above chart, we observe top 3 contributing reason for disliking video games for students are:

1. Too much wasted time
2. Costs of the game
3. Pointlessness

Scenario V

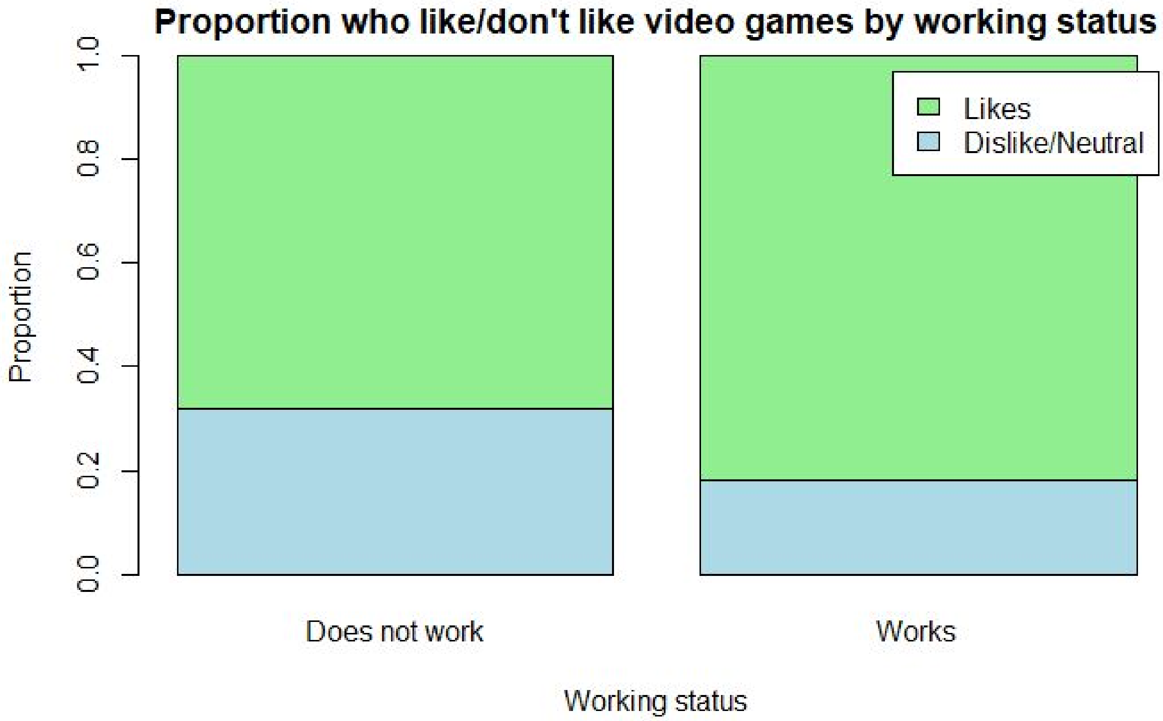
In this scenario we want to see if there are certain features of a person that make them like games. To do this we choose 3 features to test on: Sex, Working status and PC owning status. To carry out this test we group our "like" responses into 2 groups. They either like playing games if they answered very much or somewhat to our survey or they dislike/are neutral about playing games if they answered otherwise. We also remove any unknown values from like and the column we are analyzing in order to stop null values from affecting our analysis

First sex:



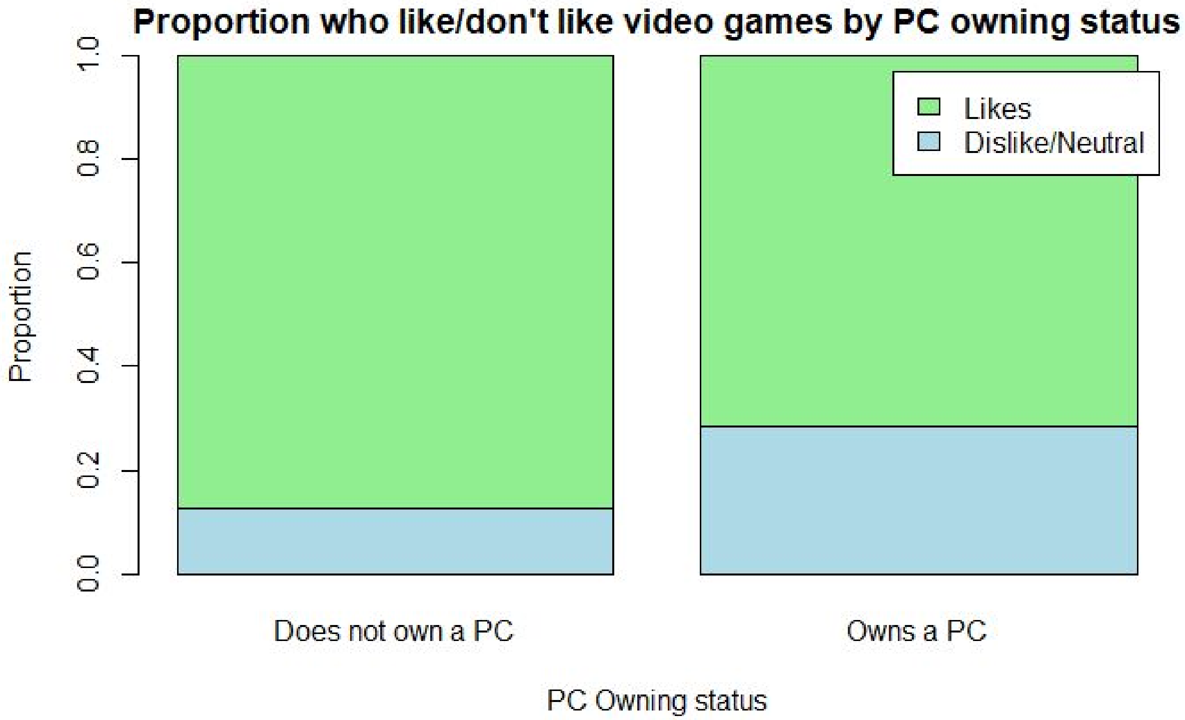
From the bar plot above we can see that generally males tend to like games more than females. We decided to run a 2 sample proportion test to verify this. Our null hypothesis is that the above difference is just by chance and the 2 groups are roughly similar while our alternative hypothesis is that the differences above are not just by chance but actually do indicate that males like playing games more than females. Our 2 sample proportion test results in a p value of .1254. While this is not an extremely high value it does indicate that the difference we see above or something more extreme can occur at least 12% of the time which means that the difference we see could just be because of chance. However this is not too reliable due to the extremely small number of samples we have and the fact that our sample is not lesser than 10% of the population(91/314 being roughly 29%) thus making normality assumption invalid.

Second working status:



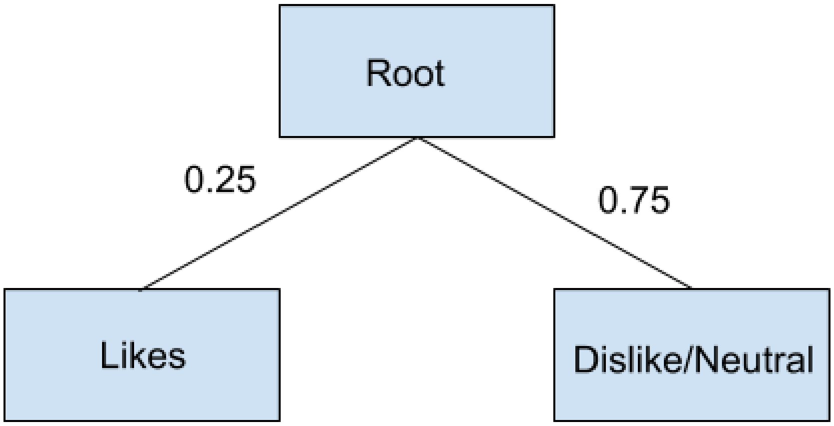
From the bar plot above we can see that generally working people tend to like games more than non working people. This could be because working people could afford more of them/favor it as a pastime. We decided to run a 2 sample proportion test to verify this. Our null hypothesis is that the above difference is just by chance and the 2 groups are roughly similar while our alternative hypothesis is that the differences above are not just by chance but actually do indicate that working people like playing games more than non working people. Our 2 sample proportion test results in a p value of .1092. While this is not an extremely high value it does indicate that the difference we see above or something more extreme can occur at least 11% of the time which means that the difference we see could just be because of chance. However this is not too reliable due to the extremely small number of samples we have and the fact that our sample is not lesser than 10% of the population(91/314 being roughly 29%) thus making normality assumption invalid. We also have only 8 working people who dislike/feel neutral about games meaning we don’t have 10 of each class at least which makes normality assumption invalid again.

Finally PC owning status:



From the bar plot above we can see that generally people who don't own a PC tend to like games more than people who do own them. This could be because of the presence of the consoles and some of the people reporting they don't own a PC simply because they prefer consoles (even though they may actually own one). We decided to run a 2 sample proportion test to verify this. Our null hypothesis is that the above difference is just by chance and the 2 groups are roughly similar while our alternative hypothesis is that the differences above are not just by chance but actually do indicate that people who don't own PCs like playing games more than people who do own PCs. Our 2 sample proportion test results in a p value of .1004. While this is not an extremely high value it does indicate that the difference we see above or something more extreme can occur at least 10% of the time which means that the difference we see could just be because of chance. However this is not too reliable due to the extremely small number of samples we have and the fact that our sample is not lesser than 10% of the population(91/314 being roughly 29%) thus making normality assumption invalid. We also have only 3 people who don't own PCs who dislike/feel neutral about games meaning we don't have 10 of each class at least which makes normality assumption invalid again.

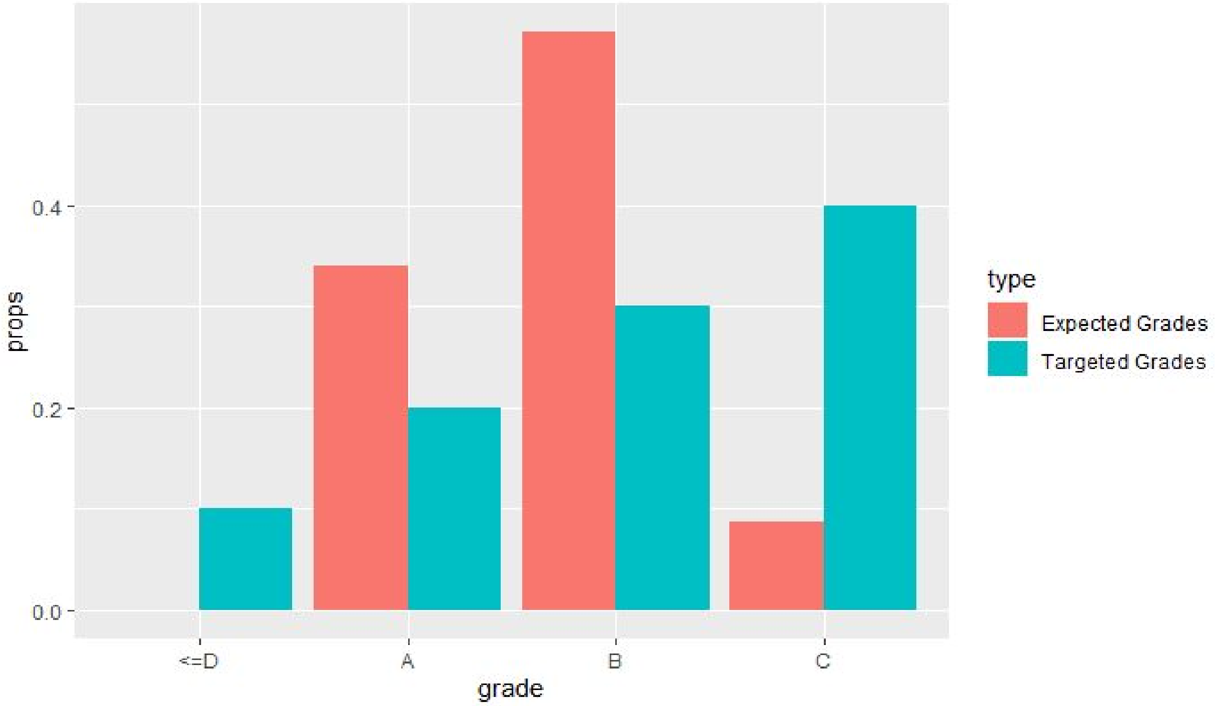
Finally we attempt to make a decision tree classifier using our dataset. This allows us to see which variables play an important factor in determining whether a person is most likely going to be a gamer or not. In accordance with our observations above we should find that a decision tree is not possible.



All our decision tree gets is that out of 88 students (after cleaning), only 22 like playing games while the remaining 66 are neutral/dislike games. It does not use any of the variables we analyzed above indicating that there may not be any such relationship between those attributes and whether or not a student likes playing games. While this follows our analysis it could again be a victim of our low sample size (despite attempting to overfit to at least find relationships in this sample).

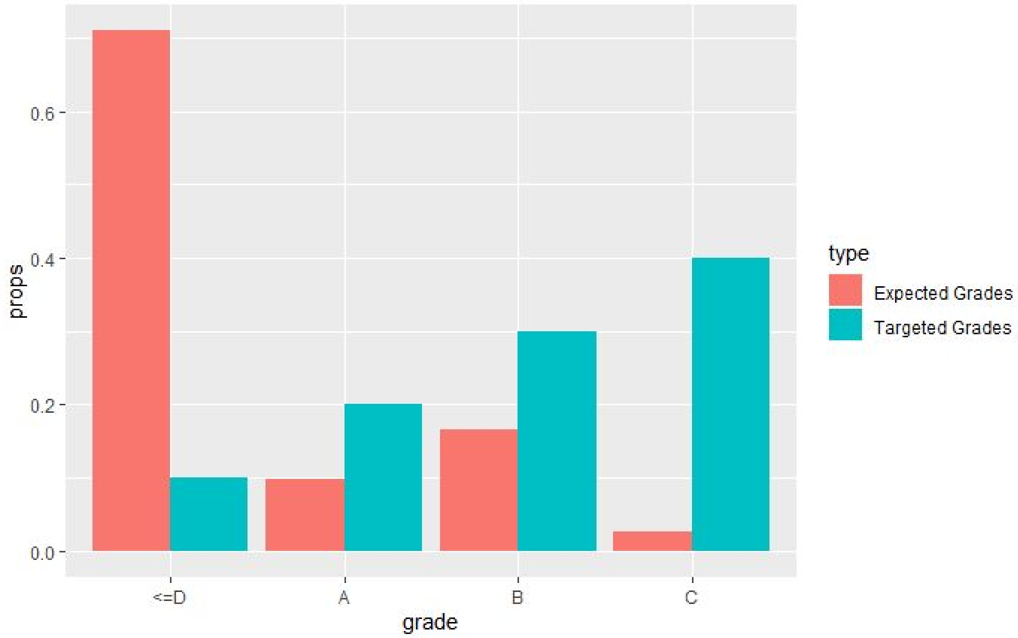
Scenario VI

In this scenario we analyze the expected grade distribution and see how close it is to the target distribution the faculty are trying to achieve.



From the above distribution we can see that the expected grades are extremely far off from the targeted distribution. In fact we don't have any students who are expecting a D. Students are overall expecting higher grades then the distribution targeted by the faculty. Running a chi-squared test gives us a p-value of $1.63 * 10^{-16}$

which is an extremely low p value indicating the 2 distributions are different. However if we know that the non respondents actually failed the class we get a whole new distribution as displayed below.



Running a chi-squared test on this distribution gives a much higher p-value of 2.2 * 10^{-16}, indicating that this distribution is still extremely far off the target distribution. Moreover, just from eyeballing it we can see that the distribution is still nowhere close. The proportion of students failing are much more than 10% while the proportion of students with Cs are way below the 40% goal. Overall, this change seems to give a completely opposite result and indicates that the students are expecting overall grades to be below than the target distribution.

Conclusion

Through our investigations of statistics student responses regarding their video game use and enjoyment, we were able to gain some insight into factors that influence students' experiences with and likelihood to use video games, which may be useful to designers of the new computer lab.

Using the population correction factor, we found that the proportion of students who played video games in the week prior to the study was between 27% and 48% using a 95% confidence interval. A point estimate gave us 37.36%. This proportion was affected by the fact that there was an exam approaching, confirmed through a comparison of the distributions of students who do and do not play video games when they are busy. Students who do not play video games when they are busy, even if they enjoy them, play less hours of video games- and at a lower frequency- than students who play regardless of whether they are busy. Students who play regardless of whether they are busy are outliers in our data, as they play many more hours almost daily, while the majority of students who play video games play weekly.

For both of these populations, the computer labs would be beneficial, and should be held weekly or biweekly to maximize attendance. Students who normally wouldn't play video games when they are busy would be more likely to participate since they would be learning statistics at the same time. Students who would play video games regardless would also likely participate since they could take advantage of the opportunity to learn while doing something they enjoy.

We found that the top two contributing reasons for liking video games were relaxation and feeling of mastery. That means that if the labs were designed to make course material less intimidating and more fun in their approach, they would appeal to the top reasons students like to play video games. Feeling like they mastered the material and experiencing less anxiety surrounding the subject would increase attendance and computer lab use, even with exams approaching.

In contrast, the top three contributing reasons for disliking video games for students were ‘too much wasted time’, associated costs, and pointlessness. Since the labs would be free of charge to students, and educational, these factors would not deter students from participating. However, the length of the labs should be limited to 30-90 minutes, in order to remain efficient and attract attendance, since the 95% confidence interval for the average amount of time played by students in our sample was about 24-111 minutes including some high outliers.

Appendix

Scenario I

There are two bootstraps nested inside each other due to the normality testing.
The outer loop records the number of times kurtosis is run over a different sample mean distribution.
The inner loop records the simulation of the sample mean distribution.
The loop for kurtosis is needed to check if most of the time the sample mean distribution is approximately normal.

Scenario II

Actual distribution of people who play when busy and not busy. Excluded Unknown in the body as it doesn’t add much information in understanding how an exam coming up may affect students

	Number in group
Don’t play when busy	63
Play when busy	17
Unknown	11

Scenario V

Along with plotting the number of students in each group we also run 2 sample proportion tests. This allows us to determine if there are any significant differences in each group. If there are any significant differences it would mean that a certain factor actually does influence whether a student likes gaming or not. All 2 sample proportion tests ran use a 1 sided greater alternative hypothesis. Which side lies on the greater side is determined from the graphs in the body

Distribution of liking games by sex

	Female	Male
Dislike/Neutral	12	10
Likes	26	43

P Value: .1254

Distribution of liking games by working status.

Note that only 8 working people dislike games which may make our data not normal/not independent

	Does not work	Works
Dislike/Neutral	14	8
Likes	30	36

P Value: .1092

Distribution of liking games by PC owning status

Note that only 3 people who don't own a PC dislike games here which may make our data not normal/not independent

	Does not own PC	Owns PC
Dislike/Neutral	3	19
Likes	21	48

P Value: 0.1004

Decision Tree output:

We run a decision tree to determine what conventional machine learning models think about what factors may affect whether or not a student is likely to like gaming. As seen below there is just the root which goes to like and dislike with probabilities of 0.25 and 0.75 respectively indicating that even the machine learning model does not feel there are significant factors. This could be just because of low data size though

Root 88 22 Likes (0.25, 0.75)

Scenario VI

Decided to run chi squared goodness of fit tests here to see how close the distribution is to the expected distribution. The chi squared test is perfect for such questions as it allows us to test how likely it is that an observed distribution has been made just because of chance. If there was a distribution made not because of chance then it would mean something significant caused the change in distribution and the faculty can't just rule them out as mere coincidences

Chi squared test before knowing non respondents failed

X-squared = 62.608, df = 3, p-value = 1.629e-13

Chi squared test after knowing non respondents failed

X-squared = 386.09, df = 3, p-value < 2.2e-16