Extent of Tendencies in Playing Video Games for Students

Author Contribution

James Lu: wrote the analysis, r code, and plots for the basic analysis sections 1 - 4. I also wrote the conclusion for my sections and the discussion section.

Matin Ghaffari: Intro, analysis, r code and plots for the basic analysis sections 5 - 6. I also wrote the advanced analysis and wrote the sections of the conclusion and discussion for the sections that I contributed and the limitations of the data.

Introduction

This study examines the relationship between College Students and their tendencies in playing video games and how potential factors influence the responses of the participants in the data we are analyzing. The dataset we are using is from a study at UC Berkeley in Fall 1994 that was aimed towards determining the extent to which students played video games and their tendencies regarding video games. Specifically the data was obtained through 91 students who completed surveys out of 95 randomly selected students from the 314 students in UC Berkeley Statistics 2, Section 1 of Fall 1994. This survey had multiple parts, where it was first given to 95 pseudo random students and was later followed up by the 91 students in the discussion section the week after they had taken their second exam. The 91 students who participated in the study completed surveys that collected data on information such as aspects of video games they find most and least fun and what they like and dislike about the games. The survey also collected information about why they dislike or like playing video games in addition to other information about themselves regarding their lifestyle and gaming tendencies. Furthermore the questionnaire survey also collects other numerical and categorical data that is specific to the personal characteristics of participants and information about general gaming features which we ultimately used in our analysis in order to provide useful information to designers of a new computer lab. We consider the dataset's shortcoming of only surveying one particular statistics class which may not be generalizable or representative to the population of all computer lab users, since our target population in the data is the 3,000 − 4,000 students in statistics courses at UC Berkeley, and therefore introducing dependencies in our data that we accounted for. While on the other hand, the dataset practiced good surveying methodologies which provided the advantage of having the participants randomly sampled and them being kept anonymous to encourage honest independent responses.

We will provide designers with this useful information by estimating the proportion of students who played a video game in the week prior to the survey, which we do in order to consider and analyze the potential confounds of the exam being given a week prior to the survey. By analyzing the amount of time students spent playing video games in the week prior by comparing the time spent playing video games daily, hourly, and weekly which we can use to help determine the effect of the exam upon previous estimates and these comparisons in time intervals. Furthermore with this dataset we perform a simulation study to determine our appropriate point estimate and an interval estimate for the average amount of time spent playing video games in the week prior to the survey. Additionally, we consider the other potential dependencies and relationships among the data such as the attitude and potential influences in the questions themselves in addition to analyzing personal characteristics and tendencies of the participants. We do this by performing numerical and graphical analyses to make comparisons amongst features and attributes of the data which allows us to analyze differences between those who like to play video games and those who don't. Lastly we look deeper into the analysis to investigate the relationship between this data and the grade that students expect in the course and how it matches the target distribution used in grade assignment.

2. Basic Analysis

2.1) Proportion of Students Who Played a Video Games in The Week Prior to the Survey

To begin our analysis, we estimate the fraction of students who played a video game in the week prior to the survey in the form of a point based estimate as well as an interval based estimate. Our point based estimate is derived by taking the number of students who played a video game in the week prior, which had a count of 34, and dividing this number by the total number of students, which had a count of 91. After performing this basic calculation we got a proportion of 0.3736, in other words ~ 37.36% of sampled students played video games in the week prior to the survey. Using this proportion we also generated an interval estimate in two forms: a standard confidence interval and a finite population corrected confidence interval. For both confidence intervals we used a 95% confidence level and a standard deviation derived from a normally approximated bernoulli distribution.

For our standard confidence interval we used the formula:

$$(\hat{p} - z_{\alpha/2} * \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2} * \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}})$$

Using the values $\hat{p}$ = 0.3736, n = 91 for our sample size, we calculated the standard confidence interval to be: (0.274, 0.473)

Although this confidence interval gives us an idea of the range of our estimate, we use a finite population correction factor to account for sampling without replacement when our population size of 314 (N) is not very large in comparison to our sample size of 91 (n). The formula for the corrected confidence interval is as follows:

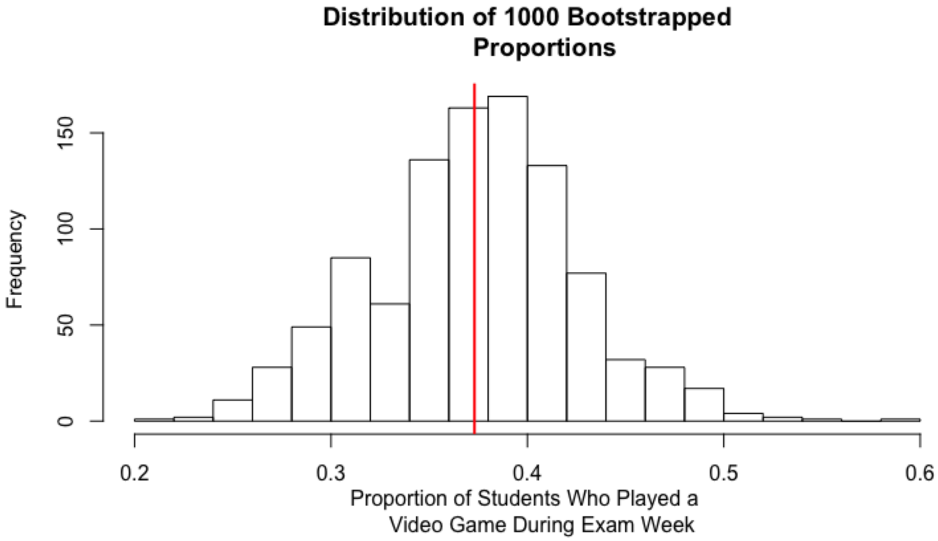
$$me = z_{\alpha/2} * \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}}$$

$$(\hat{p} - me, \hat{p} + me)$$

Using the values $\hat{p}$ = 0.3736, n = 91, N = 314 we calculated the corrected confidence interval to be: (0.289, 0.458)

Then to test the robustness of our estimate, we conducted 1000 bootstrap simulations to test the validity of our calculations as shown below in Figure 1.

Figure 1: Histogram of 1000 bootstrapped proportions with the red line representing the mean of the distribution



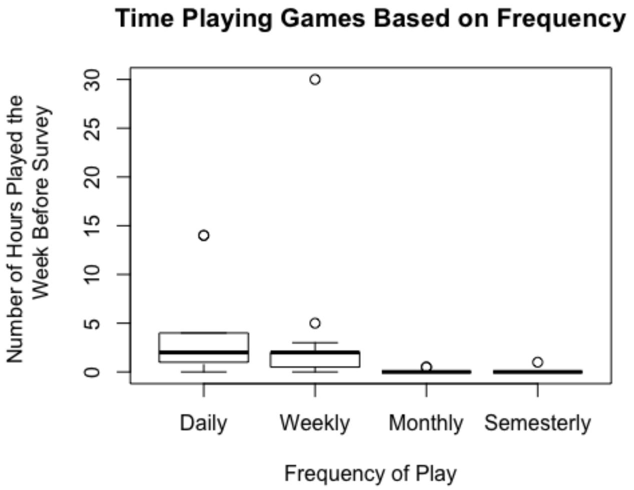
After running our bootstrap simulation we found the mean proportion of the 1000 runs was 0.362, with the new 95% confidence intervals being:
Standard CI: (0.264, 0.461)
Population Corrected CI: (0.279, 0.447)
We estimate that 36.2% of students played video games the week prior to the survey and we can say with 95% confidence that the estimated proportion is between 27.9% and 44.7%.

2.2) Analysis of The Time Spent Playing Video Games in The Week Prior to the Survey Compared to the Reported Frequencies.

After finding basic statistics regarding the number of students that play video games, we now analyze the relationship between the amount of time spent playing in the week prior to the survey against the students' reported frequency of play.

To begin our analysis, we generated a boxplot to examine the general distribution of the number of hours played for each category of frequency. In this analysis we are comparing amongst daily, weekly, monthly, and semesterly in order to best see how the data is distributed.

Figure 2: Boxplot visualizing the number of hours played in the week prior to the exam against typical frequency of play



After generating our plot, we found that students who reported playing daily had the widest range of hours played, with the range tapering off as the intervals between play got longer. However we examined some outliers in the weekly category that fell outside of the interquartile range and one student that claimed they usually play weekly but played ~30 hours the week before the exam. Excluding some outliers, we found from the boxplot that as the intervals between play increased, the range, median, and IQR for the number of hours played decreased.

Another method we used to analyze the data was to find the proportion of players in each frequency category who did or did not play when busy and the average time played. We generated a table to try and find a relationship between the average time played, the reported frequency of play, and the proportion of students who play when busy.

Table 3: Proportion of students who play when busy and the average time played the week prior to the survey for each frequency

Frequency	Average Time Played Week Prior to Survey (Hours)	Proportion Who Play When Not Busy	Proportion Who Play When Busy
1: Daily	4.44	0.44	0.56
2: Weekly	2.54	0.60	0.40
3: Monthly	0.06	0.94	0.06
4: Semesterly	0.04	0.96	N/A

In our analysis we found that the average time played decreases as the frequency of play decreases, and the proportion of students who play when busy decreases when the frequency of play decreases as well. This trend appears logical as those who play video games less frequently will have a smaller average time played and will be even more unlikely to play when they are busy.

2.3) Analysis of the average amount of time spent playing video games in the week prior to the survey.

To begin our analysis of the average amount of time spent playing video games in the week prior to the survey, we construct a point based and interval based estimation by taking the mean of the time variable and constructing a 95% confidence interval from it.

We found the average amount of time playing video games prior to the survey to be 1.243 hours. Using this estimate as our mean, we constructed a confidence interval using the confidence interval formula:

$$me = z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} = .776$$

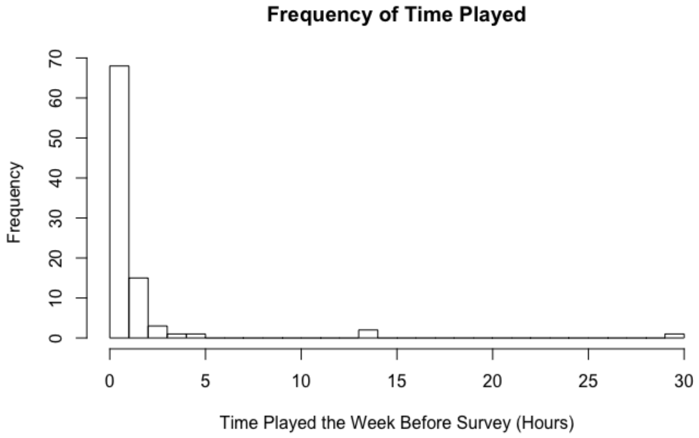
95% confidence interval = $(\overline{x} - me, \overline{x} + me)$ = (0.467, 2.019)

We also constructed another 95% confidence interval using a finite correction factor that also accounts for sampling without replacement when our population size of 314 (N) is not very large in comparison to our sample size of 91 (n).

$$me = z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = 0.655$$

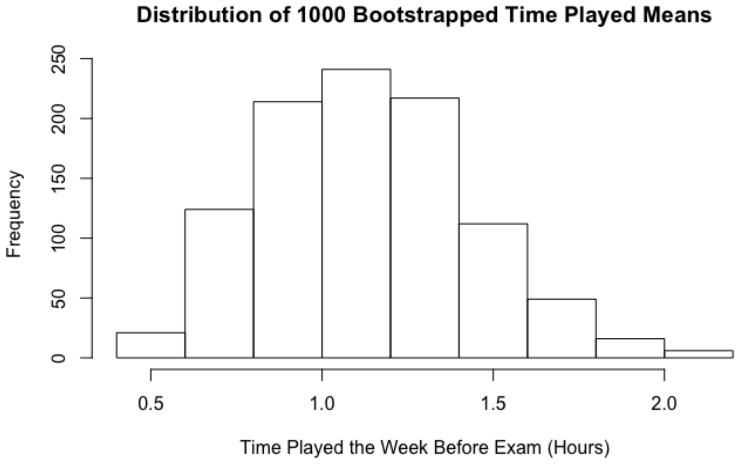
95% confidence interval = $(\overline{x} - me, \overline{x} + me)$ = (0.588, 1.898)

Figure 3: Histogram representing the frequency of time played



However, these estimates hold little weight because of the shape of the distribution, which is non-normal and highly skewed to the right as depicted above by Figure 3. To attempt to normalize our data and find more accurate point based and interval based estimates we performed 1000 bootstrap simulations of the sample mean as well as 1000 simulations of the empirical normal distribution in order to compare for normality.

Figure 4: Histogram visualizing the frequency of bootstrapped means



After conducting our bootstrap simulations we generated the histogram for this bootstrapped distribution in Figure 4 above. Although the distribution looks fairly normal, we wanted to test the kurtosis and skewness against a randomly sampled normal distribution of similar sample size.

Figure 5: Histogram representing the frequency of 1000 bootstrapped sample skewness of the empirical normal distribution

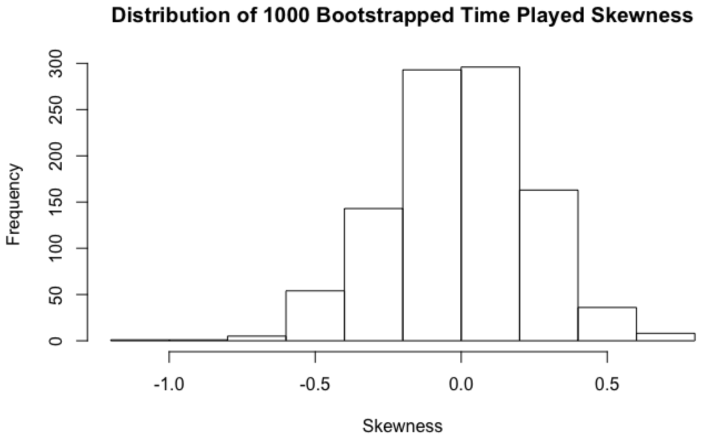
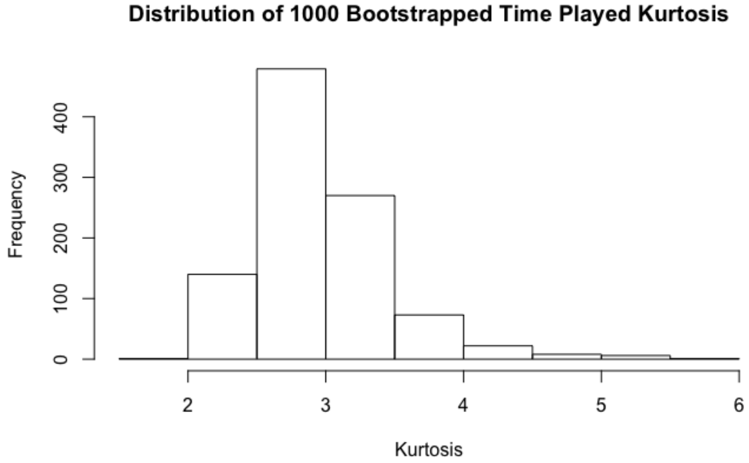


Figure 6: Histogram representing the frequency of 1000 bootstrapped sample kurtosis of the empirical normal distribution



After generating our bootstrap samples for the kurtosis and skewness of the normal distribution we created a table to compare our bootstrapped mean distribution to the standard normal distribution as shown on the next page on Table 4.

Table 4: Table containing kurtosis and skewness calculations from both bootstrapped samples

	Bootstrapped Standard Normal (Mean)	Bootstrapped Sample Mean
Kurtosis	2.938	3.3
Skewness	-0.003	0.5

Comparing the kurtosis and skewness of both bootstrapped samples we can say that our distribution of sample means is approximately normal as the kurtosis and skewness are close to that of a standard normal with the same sample size (91). After performing our bootstrapping simulation we calculated our new mean and 95% confidence intervals using our approximately normal data. Our new calculated mean was 1.136 hours, and our new confidence intervals are as follows:

$$me = Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} = .0618$$

95% confidence interval = $(\overline{x} - me, \overline{x} + me)$ = (1.08, 1.2)

$$me = z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = .052$$

95% confidence interval =$(\overline{x} - me, \overline{x} + me)$ = (1.09, 1.19)

2.4) Analyzing "Attitude" Questions to Determine in General if Students Enjoy Playing Video games?

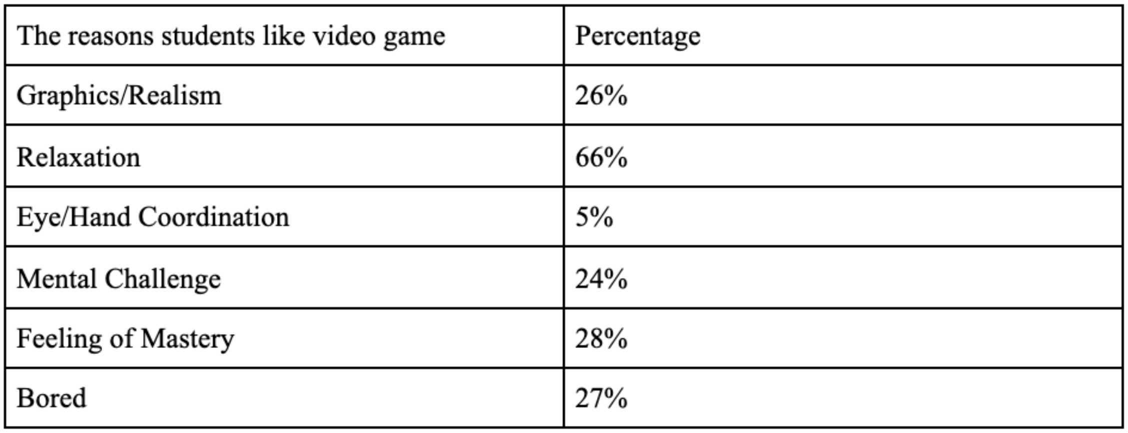
To quantify whether or not students enjoy playing video games, we estimate the proportion of the sample that like and dislike games. Students were grouped in the "like" category if they marked 2 (very much) or 3 (somewhat) in the like section of the survey. Conversely students were grouped in the dislike category if they answered 4 (not really) or 5 (not at all) in the like section of the survey. There were a small percentage (2.3%) of students who did not answer this question for one reason or another, however ~97% of the sample did.

Table 5: A table outlining the percentage of the sample that like and dislike video games.

	Percentage of Sample
Like Video Games	75.8%
Dislike Video Games	21.9%
Did Not Answer	2.3%

From this rudimentary analysis we found that a majority of the sample (75.8%) did like playing video games, and as for the reasons that students liked or disliked video games we look at the data collected in the second survey.

Table 6: A table outlining the reasons students like playing video games.



The reasons students like video game	Percentage
Graphics/Realism	26%
Relaxation	66%
Eye/Hand Coordination	5%
Mental Challenge	24%
Feeling of Mastery	28%
Bored	27%

Analyzing the data in the second survey we find that most students (66%) like video games because it relaxes them, while most students dislike video games because they have limited time (48%) or because the hobby costs too much (40%).

Using both Table 5 and Table 6 we find that the most important reason students like video games is because they can be used for relaxation, competition, escape, or even just to pass the time. Similarly the most important reasons students don't like video games is because they take up too much time, they are an expensive hobby, they might not provide any value to the person, and they can be frustrating. In general, we found that students did like playing video games with the majority of those students citing relaxation as a major reason for their enjoyment.

2.5) Analysis of Differences Between Those Who Like to Play Video Games and Those Who Don’t

Here in our analysis we categorize those who classify themselves as liking video games as well as those who don’t like video games in order to observe the trends and features that these two categories may differ or relate in. We accomplish this by using the categorical ordinal variable of “like” which has level 1-5 to indicate how much the individual likes or dislikes video games which we convert to a factor in order to group these categories into 2 genaralized categories for any liking and any disliking. These 2 general categories exclude the one observation that indicated “Never Played” since we would like to focus on those who indicated a level of liking or disliking and also because one single observation would fail to provide us with meaningful and generalizable results for that group. Next as shown below, we use these two categories to then run an analysis on their shared attributes that are categorical and nominal with binary values indicating their gender, if they own a PC, and if they work for pay or not. Additionally, several of the categories had the value 99, which we removed from this analysis, since these values indicate improperly or not answered questions which may introduce potential erroneous information from our analysis.

Figure 7: Here we have barplots for the two grouped categories of observations who indicate liking or disliking video games, which we use in this barplot to visualize the differences in these groups amongst males and females.

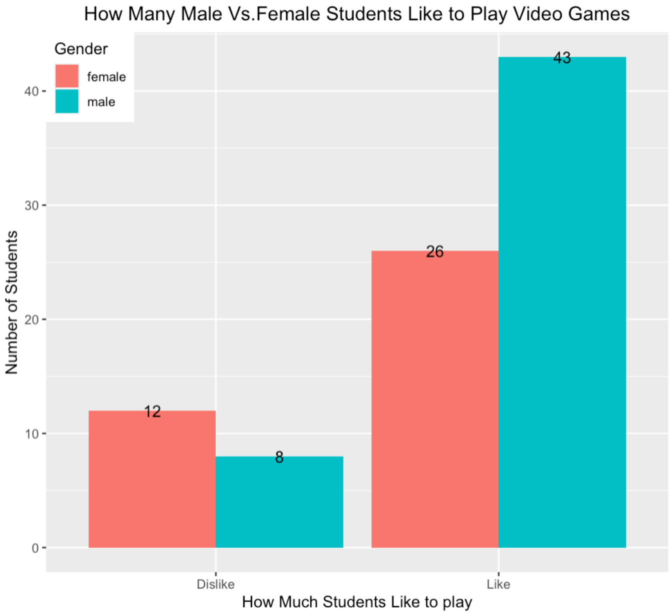


Figure 8: Here we have barplots for the two grouped categories of observations who indicate liking or disliking video games, which we use in this barplot to visualize the differences in these groups amongst those who own a PC and those who do not own a PC.

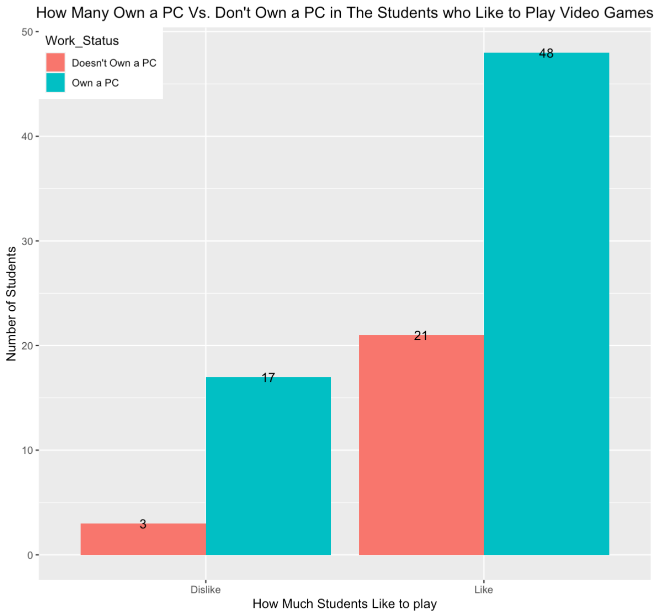


Figure 9: Here we have barplots for the two grouped categories of observations who indicate liking or disliking video games, which we use in this barplot to visualize the differences in these groups amongst those who work for pay and those who don't.

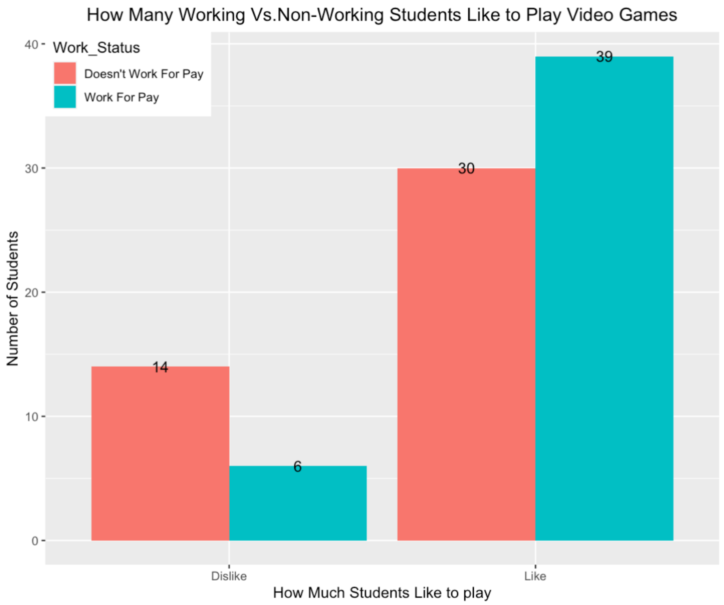
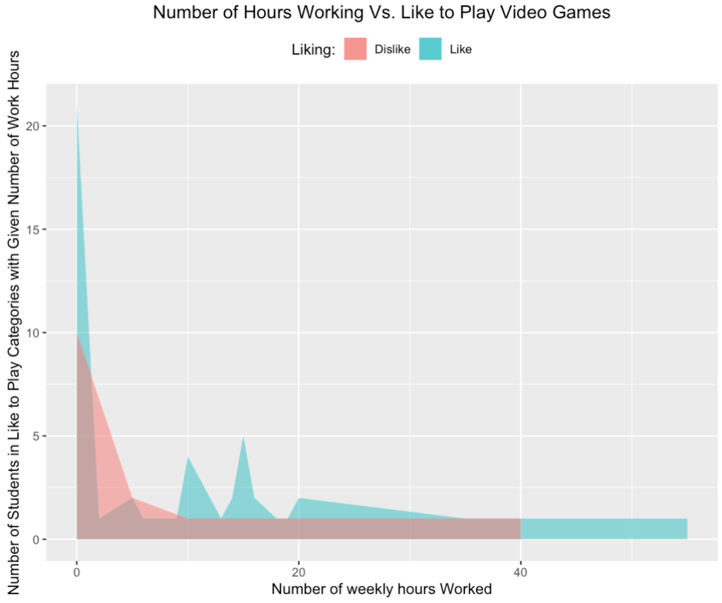


Figure 10: Here we have a stacked area chart in order to see the differences between the two grouped categories of observations for those indicating liking or disliking video games, who we analyze by observing the count in each category for different amounts of hours worked in a week.



With the analysis above, we were able to notice quite a few notable differences amongst these attributes of people who categorized themselves as liking or disliking video games. To begin with in figure 7 we see that there are almost twice as many boys who like video games than females, which is also supported in the dislike category since there is more than 30% more females who dislike video games. Another relationship that may be seen is between owning a PC and liking video games, since we can see in figure 8 how there is much more than double for those who like video games and own a PC, which is additionally supported by the dislike category which has significantly less for those who own a PC. Lastly we analyze these groupings' relationship to a binary categorical variable we created to compare the counts in those who work and don't work in figure 9, as well as the counts for those who like and dislike in relation to the discrete variable for the hours worked in a week as displayed in figure 10. In figure 9, the data suggests that those who work for pay tend to like video games more, which is also supported by figure 10 since the area plot shows this trend even for when those who like video games are working at a greater number of hours per week. Although these relationships may be considered to be justified through simple logic thinking, because of the limitations in our data regarding its size and generalizability, as will be discussed after, we can't say these analyses are enough to make definitive conclusions.

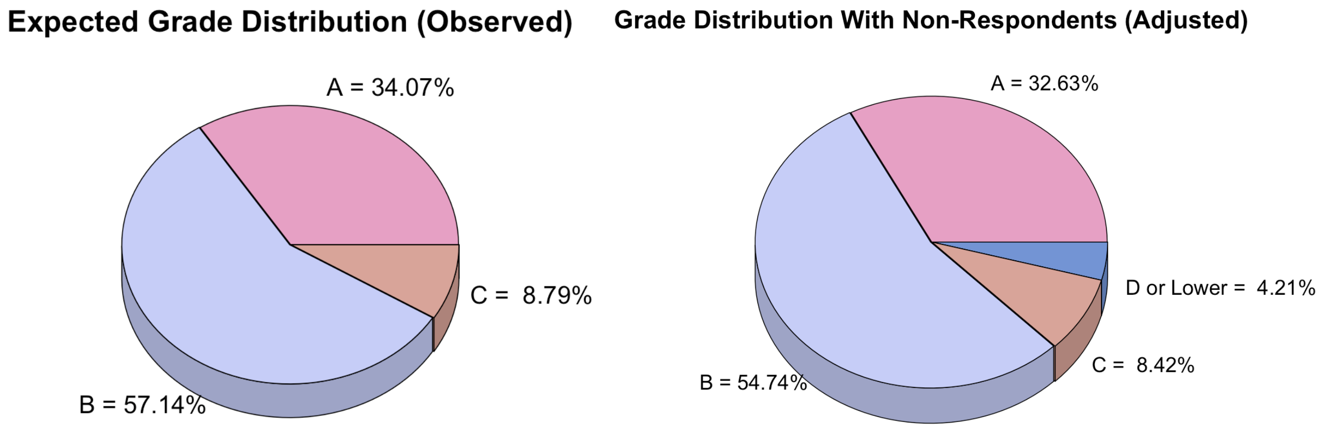
2.6) Analysis of the Distribution of Observed Expected Grades Vs. Target Distribution

When examining the data, we looked at the ordinal categorical variable of "grade" which represents the letter grades A, B, C, D, and F with the respective digits 4, 3, 2, 1, and 0. Using this information we grouped all the observations in the data set by this categorical variable in order to calculate the total number of observations that fell under each one of these grading categories. Afterwards, we used these count values to then compute the proportions for each category in order for us to analyze and determine the distribution for these grading categories and how they differ from the target distribution of 20% A's, 30% B's, 40% C's, and 10% D's or lower. Furthermore, we also observe the potential effect of the 4 non-respondents by adding them to our dataset and considering them as failing students who no longer bothered to come to discussion.

Table 10: Percentages of A's, B's, C's, and D's or lower for the data's student grade expectations, target distribution, and adjusted distribution if considering non-respondents as failing students.

	A	**B**	**C**	**D or lower**
Grade Distribution of Expectations (%) ***(Observed)***	34.07%	57.14%	8.79%	0%
Target Grade Distribution (%) ***(Target)***	20%	30%	40%	10%
Grade Distribution if non-respondents were failing students (%) ***(Adjusted)***	32.63%	54.74%	8.42%	4.21 %

Figure 11: Pie Charts to visualize the change in distribution for the students expected grade distribution (left-hand chart) from the observed data versus the chart to the right which shows the new adjusted distribution after considering non-respondents as failing students.



As the analysis above demonstrates, we see that both the observed expected grade distribution and the adjusted grade distribution both vary significantly from the target distribution in each of the categories. We see that the percentage of students who get an A is over 10 % from the target and the percentage of B students is significantly over the target distribution by over 20 %. As a result the percentage of those who expect to get a C and D are tremendously differing from the target with over a 30% difference in C and with no students who expect to get a D or lower as opposed to the target of 10%. This alone based on the data shows how different the expected distribution we observe is from the target due to the large differences in every category. Furthermore, when the data set and analysis is adjusted to include the non-respondents as failed students, we see that this very minimally shifts the distribution towards the target since there is too little of an amount of 4 non-respondents compared to the large number of students who expect A's and B's. With this being said, we can suggest from this data that the non-respondents isn't the factor the that is causing the observed expected grade distribution to vary so greatly from the target, however to make a more reasonable suggestion we would need a larger sample to have less miniscule values for potential factors such as these non-respondents.

Advanced Analysis

In a further analysis we would like to maker a better determination for what we did above in problem 6 by using the chi-squared framework in order to better understand how the observed grade expectation distribution matches the target distribution of 20% A's, 30% B's, 40% C's, and 10% D's or lower. Since this analysis question is a problem of 3 or more proportions and we are interested in determining if the our observed data counts match with the given expectation values, we can better approach this problem by using the chi-squared goodness-of-fit test to better explore if our given set of proportions matches the stated values for the target.

In order to use the chi-squared framework, we must first must make sure that we meet the following assumptions needed about the data in order to ensure that we can properly use the X^2 stat for our problem. First we set up a parameter studied as well as a null and alternative hypothesis, and afterwards we will satisfy the 3 needed assumptions in order for us to run the chi-squared goodness-of-fit test.

1. **Parameter**: Let p_A , p_B , p_C , p_D be the percentages for the grade distributions of all the A, B, C, and D or lower students (respectively).
2. **Null Hypothesis (H_0)**: $p_A = 20\%$, $p_B = 30\%$, $p_C = 40\%$, $p_{DF} = 10\%$. In other words, the given percentages for the students expected grade distribution are a good fit for the true proportions (target distribution).
3. **Alternative Hypothesis (H_1)**: The given percentages for the students expected grade distribution are NOT a good fit for the true proportions (target distribution).

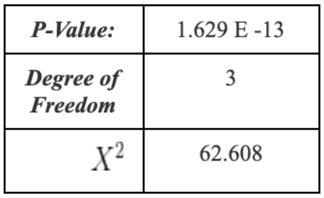
The 3 assumptions we must meet to use the chi-squared goodness-of-fit test ensure that the test stat is approximately X^2_{k-1} in order for us to use the sampling distribution: $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$ on χ^2_{k-1}
These assumptions are:

1. We have observed counts → We satisfy this assumption as we saw above in Table 3 and Figure 6.
2. The data comprising the counts is independent → We satisfy this since we know from the study that the students were randomly sampled with pseudo randomization without replacement which allows us to assume the data comprising the counts is independent.
3. We expect to see at least 5 in each category. Thus assuming H_0, $E_i \geq 5$ for all i. → We satisfy this rule since we have a sample size of 91 and are rarest outcome is 10% for D, thus 91 * 0.1 = 9.1 > 5, therefore because our lowest expected count is ≥ 5 then we know that all expected counts are large enough to meet this assumption.

Therefore since, we met all the necessary assumptions we can conduct the chi-squared goodness-of-fit test either with R using chisq.test() function or by hand by computing the test statistic and using the sampling distribution above and get the following p-value, x-squared, and degree of freedom values.

TABLE 10: Observed count figures on the table to the left and the table to the right contains the results from the chi-squared goodness-of-fit test using R (chisq.test()).

	A	**B**	**C**	**D or Lower**
Observed Counts	31	52	8	0



P-Value:	1.629 E -13
Degree of Freedom	3
X^2	62.608

Therefore since we have a p-value of 1.629 E -13 = 0.00000000001629 ≈ 0 < 0.05 (using 0.05, conventional significance level) it is so small that we can reject the null hypothesis in favor of the alternative hypothesis that the given percentages for the students observed expected grade distribution are NOT a good fit for the true proportions (target distribution).

Conclusion and Discussion

From our findings in question 1, we came to the conclusion that the expected percentage of students that will utilize the new computer lab to play video games to be roughly 36.2%. We also found that we can say with 95% confidence that the percentage falls between 27.9% and 44.7%. These findings imply that there is a sizable subset of students that may utilize the new computer lab to play video games. From our findings in question 2, we came to the conclusion that as frequency of play decreases, the likelihood of a student playing video games when busy also decreases. The significance of this finding in relation to the building of a new computer lab would be that the computer lab may see less usage for video games during busy weeks, ie. midterms or finals. From our findings in question 3, we estimated the expected hours played the week prior to the survey to be 1.136 hours. We can also say with 95% confidence that this expectation will fall between 1.09 hours and 1.19 hours. The significance of this finding with respect to the new computer lab is the fact that during busy weeks we should expect ~1.1 hours of usage for video games. From our findings in question 4, we came to the conclusion that most students enjoy video games. We estimate that ~75% of students at least somewhat like video games and many of them (~66%) cite relaxation as the reason for enjoying playing. With this information we can be confident that the majority of students can benefit from its creation. In question 5 we see that variables for working or not working, sex, and if they own PC or not does have a relationship liking of video games but we cannot make a definitive conclusion because of our limited sample. We also see this in question 6, we find that the distribution is far from the expected by comparing the portion of students in each grading category. Furthermore, we saw from the data that we can suggest that the non-respondents isn't the factor that is causing the observed expected grade distribution to vary so greatly from the target, however to make a more reasonable suggestion we would need a larger sample to have less miniscule values for potential factors such as these non-respondents.

This leads us into our discussion where I will first explain the limitations of this data set which largely concern the small sample size of the data and the way the data was collected. This dataset's limitation are that it only is surveying one particular statistics class which may not be generalizable or representative to the population of all computer lab users, since our target population in the data is the 3,000 − 4,000 students in statistics courses at UC Berkeley, and therefore introducing dependencies in our data that we accounted for. While on the other hand, the dataset practiced good surveying methodologies which provided the advantage of having the participants randomly sampled and them being kept anonymous to encourage honest independent responses. Additionally, from our investigations of this data set we revealed many interesting intricacies from the study, however the generalization of our findings might not be feasible due to the circumstances under which the data was collected. For example, we wanted to test on average how many hours did students gaming time decrease due to the presence of an exam but we did not have each student's average gaming time from a standard week. Using another study conducted by LimeLight Network, they found that an average gamer plays for ~6 hours a week (LimeLight Network). If we had a baseline for a college student under more normal circumstances we could compare the difference between their hours played during an exam week compared to a normal week. Our data was limited to a small subset of the population since our sample came directly from a college campus it cannot not be generalized to a wider population. Some other investigations that could be pursued might lie in the effect that gaming has on expected grades and actual grades, or the relationship between gaming and sleeping.

Works Cited

"THE STATE OF ONLINE GAMING – 2020." *The State of Online Gaming – 2020*, www.limelight.com/resources/white-paper/state-of-online-gaming-2020/.

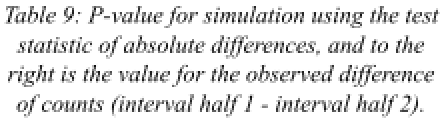


Table 9: P-value for simulation using the test statistic of absolute differences, and to the right is the value for the observed difference of counts (interval half 1 - interval half 2).