Math 189 - Case Study #2

Labs vs. Video Games

February 16th, 2019

Name	PID
Erin Werner	A12612584
Emma Choi	A12635909
Talal Alqadi	A13816618
Ella Lucas	A13557332
Samantha De La Torre	A13300273

I. Introduction

UC Berkeley has 3,000 to 4,000 students enrolled in statistics courses every year, half of which take introductory level courses. A design committee of faculty members developed a series of computer labs to assist the students in introductory courses, by providing a more interactive learning curriculum. Some computer labs have been linked to video games. In order to optimize the use of video games in computer labs, a survey of undergraduate students was taken to determine how often students played video games and which aspects of video games they found most entertaining. The survey was conducted by a group of students in advanced statistics who developed a questionnaire, randomly selected participants, and collected the data. The goal of this study is to analyze the results of the sample survey and provide suggestions to the design committee, so they can create successful computer labs.

II. Data

From a population of 314 students enrolled in the Statistics 2 - Section 1 Fall '94 class, a random sample of 95 students was asked to complete the survey. Only 91 of the 95 took it, producing the data shown in Table 2.1. There will most likely be a large covariance and positive bias in the data as six of the fifteen features have the Bernoulli probability distribution. Also, any missing or improper responses were recorded with the value '99'. The students 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.

Table 2.1. Description of the variables and their importance in the study.

Variable	Description	Relevance in Statistical Lab
time	Number of hours played in the week prior to the survey.	Duration?
like to play	1 = never played, 2 = very much, 3 = somewhat, 4 = not really, 5 = not at all	Who would attend?
where to 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	Location?
frequency	1 = daily, 2 = weekly, 3 = monthly, 4 = semesterly	How often should it occur?
play if busy	1 = yes, 0 = no	Will people attend if they are busy?
playing is educational	1 = yes, 0 = no	Will people learn something?
sex	1 = male, 0 = female	Demographics.
age	Students age in years.	Demographics.

computer at home	1 = yes, 0 = no	Demographics.
hate math	1 = yes, 0 = no	Will people be successful?
work	Number of hours worked in the week prior to the survey.	Will people have time for it?
own pc	1 = yes, 0 = no	Demographics.
pc has cdrom	1 = yes, 0 = no	Demographics.
has email	1 = yes, 0 = no	Demographics.
grade expected	4 = A, 3 = B, 2 = C, 1 = D, 0 = F	Will people be successful?

The second part of the survey covered whether the student liked or disliked playing games and why. Because some students did not play or like video games, they were instructed to skip those questions. So, the responses only included students that played video games and liked them enough to answer questions about their preferences. These questions are different from the others as there are more than one response that may be given.

Table 2.2. Summary of the type of video games played.

Type	Percent
Action	50%
Adventure	28%
Simulation	17%
Sports	39%
Strategy	63%

Table 2.2 reflects the answers to the question: What types of games do you play? The student is asked to check all types that he or she played. At most three answers were allowed. 63% of students preferred strategic video games. Students were then asked to provide reasons why they play the games they do.

Table 2.3. Summary of the reasons for playing video games.

Why?	Percent
Graphics/Realism	26%
Relaxation	66%
Hand-Eye Coordination	5%
Mental Challenge	24%

Feeling of Mastery	28%
Bored	27%

Table 2.3 reveals reasons for playing the different types of video games. Again, at most three answers were allowed. Many students, about 66% of them, view video games as a relaxation method.

Table 2.4. Summary of students disliked about video games.

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

Finally, Table 2.4, contains a summary of what the students do not like about video games. All students, regardless of video game preferences, were asked to answer this question. Once again, they were asked to select up to three reasons for not liking video games. Many students replied that video games take up "too much time" (48%), which can have an impact on whether or not people would want to attend the lab. Additionally, responses like "frustrating" (26%) and "boring" (17%) also reveal different, important reasons as to whether or not people would attend the lab.

The third part of the survey collected general information about the students such as age, sex, etc. This helped to build the demographics about the students who responded to the survey.

Table 2.5. Numerical summary statistics of the survey responses.

time	like	where	freq	busy
Min. : 0.000	Min. : 1.000	Min. : 1.00	Min. : 1.00	Min. : 0.00
1st Qu.: 0.000	1st Qu.: 2.000	1st Qu.: 3.00	1st Qu.: 2.00	1st Qu.: 0.00
Median : 0.000	Median : 3.000	Median : 3.00	Median : 3.00	Median : 0.00
Mean : 1.243	Mean : 4.077	Mean :21.97	Mean :16.46	Mean :12.15
3rd Qu.: 1.250	3rd Qu.: 3.000	3rd Qu.: 5.00	3rd Qu.: 4.00	3rd Qu.: 1.00
Max. :30.000	Max. :99.000	Max. :99.00	Max. :99.00	Max. :99.00
educ	**sex**	**age**	**home**	**math**
Min. : 0.00	Min. :0.0000	Min. :18.00	Min. :0.0000	Min. : 0.000
1st Qu.: 0.00	1st Qu.:0.0000	1st Qu.:19.00	1st Qu.:1.0000	1st Qu.: 0.000
Median : 1.00	Median :1.0000	Median :19.00	Median :1.0000	Median : 0.000
Mean :14.55	Mean :0.5824	Mean :19.52	Mean :0.7582	Mean : 1.407
3rd Qu.: 1.00	3rd Qu.:1.0000	3rd Qu.:20.00	3rd Qu.:1.0000	3rd Qu.: 1.000
Max. :99.00	Max. :1.0000	Max. :33.00	Max. :1.0000	Max. :99.000
work	**own**	**cdrom**	**email**	**grade**
Min. : 0.00	Min. :0.0000	Min. : 0.000	Min. :0.0000	Min. :2.000
1st Qu.: 0.00	1st Qu.:0.0000	1st Qu.: 0.000	1st Qu.:1.0000	1st Qu.:3.000
Median : 5.00	Median :1.0000	Median : 0.000	Median :1.0000	Median :3.000
Mean :10.37	Mean :0.7363	Mean : 5.604	Mean :0.7912	Mean :3.253
3rd Qu.:14.50	3rd Qu.:1.0000	3rd Qu.: 0.000	3rd Qu.:1.0000	3rd Qu.:4.000
Max. :99.00	Max. :1.0000	Max. :99.000	Max. :1.0000	Max. :4.000

Table 2.5 shows all the summary statistics of the responses for each of the parts of the survey. Many features have a maximum value of '99' as they didn't play video games and were thus instructed to skip certain questions. We can see that many of the students spent less than one hour playing video games the week prior to taking the survey. This could have an impact on the results of this study and reflect whether or not people would want to attend the lab.

III. Background

The 95 participants in the survey were randomly selected out of the 314 students enrolled in Statistics 2, Section 1, during Fall 1994. However, only 91 of the 95 surveys were complete. The class is a lower division prerequisite for students intending to major in business. During the fall, the class met on Mondays, Wednesdays, and Fridays from 1-2PM in a large lecture hall that seats 400 people. In addition to lectures, students attended small, one hour, discussion sections on Tuesday or Thursday. For the 314 students, there were 10 discussion sections with 30 students in each section.

The students were selected from a list of students who submitted the second exam, which was taken 1 week before the survey. Each student was assigned a number from 1 to 314, and then a pseudo random number generator was used to select 95 numbers from 1 to 314. The student's anonymity was maintained to encourage honest responses.

A 3 stage system of data collection was utilized to limit the occurrence of non-respondents. Since the exam was taken the previous week, it was being passed back in the Tuesday and Thursday discussion sections the week the survey was being given, as previously mentioned. Therefore, data collectors attended the Tuesday and Thursday discussion sections the

week the survey was being given. During the Friday lecture of the survey week, the data collectors attempted to locate any of the selected students who were not reached during the discussion sections earlier that week. This system of data collection resulted in 91 completed surveys out of the 95 students selected. The data collectors briefly informed the students of the purpose of the study and informed them of their anonymity to promote accuracy in the survey.

Video games can be classified by the type of device used to play the game and the skills required to play the game. The main types of devices are console games, PC games, and arcade games. Arcade games are fast and require hand-eye coordination. Console games are usually classified as action, adventure, or strategy games, whereas PC games are usually classified as simulation or role play games. The skills required to play Console or PC games in each category are defined in Table 3.1.

Table 3.1. Summarizes the attributes typically found in each video game category.

	Eye/Hand	Puzzle	Plot	Strategy	Rules
Action	**X**				
Adventure		**X**	**X**		
Simulation				**X**	**X**
Strategy				**X**	**X**
Role-play		**X**	**X**		**X**

IV. Investigation

The objective of this study is to investigate the responses of the participants in the survey with the intention of providing useful information to the designers of the new computer lab. The information can then help the designers construct a computer lab that people would actually attend and benefit from.

A. Scenario 1

To provide an estimate for the fraction of students that played a video game, we look at the time variable, and equalize everything greater than 0 to 1, indicating that the person has played a video game. While, all the people who haven't played will remain 0. By calculating the average of the new data sample, we can get the point estimate. The sample mean is calculated to be 0.3736.

Furthermore, the interval estimate may be calculated in different ways. Using the central limit theorem, the 95% confidence intervals can be calculated. The central limit theorem allows

us to assume that the sample average of the population distribution has an approximate normal distribution, with an appropriate sample size and sample proportion. Using the normal distribution confidence intervals, the intervals are calculated as [0.2716395, 0.4756132]. However, these may not be accepted as true intervals due to the fact that n/N equates to 0.28, not a very small sampling proportion. We can also recalculate the confidence intervals with the addition of the population correction factor to get a little more accurate range. The new approximate confidence interval gives us [0.2893981, 0.4578547]. However, since the sample size and n/N are not favorable, these confidence intervals may not be accurate. A further bootstrap simulation will allow us to acquire more accurate confidence intervals and see if the data follows a normal distribution.

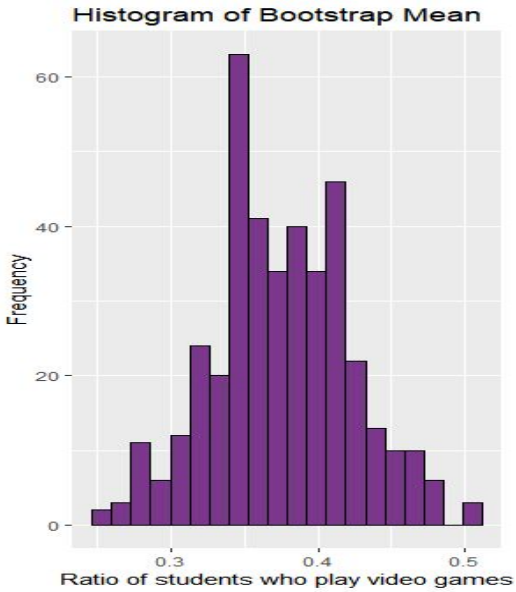

Figure 4.1.1. Distribution of bootstrap means.

Figure 4.1.1 illustrates the distribution of 400 sample means with sample size 91. Using the new bootstrap samples, we can construct 2.5% and 97.5% confidence intervals: [0.274, 0.461]. However, in order to trust these intervals we must first confirm whether the distribution approaches that of a normal one. The kurtosis and skewness values of the sample means are respectively, 2.9181 and 0.490, showing evidence that this is a normal distribution.

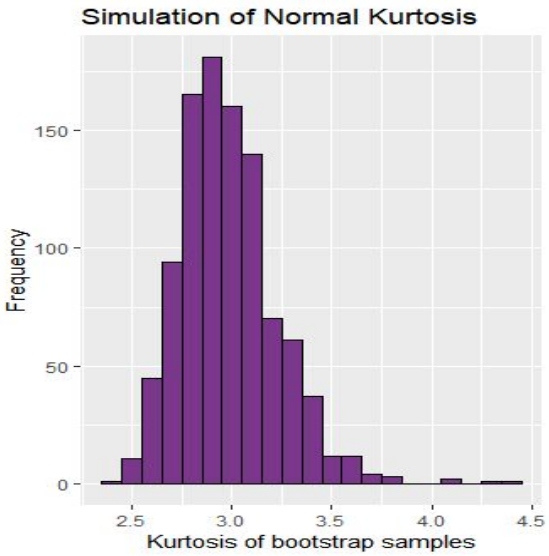

Figure 4.1.2. Distribution of monte-carlo simulation of kurtosis.

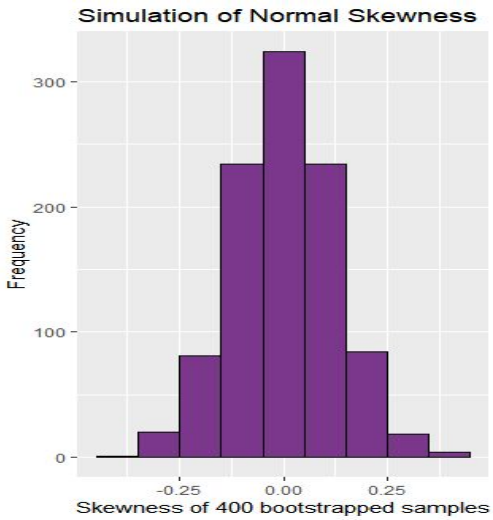

Figure 4.1.3. Distribution of monte-carlo simulation of skewness.

Figure 4.1.2 and 4.1.3 show the distribution of kurtosis and skewness of 1000 monte-carlo simulations with 400 samples. The kurtosis distribution seems normally distributed around 3. While the skewness distribution is normally distributed at 0. Thus, we can conclude that the bootstrapped sample means distribution is normally distributed. Therefore, we can accept both the confidence interval previously calculated as [0.274, 0.461], and the interval calculated earlier with the CLT and population correction factor as [0.289, 0.458]. A smaller range may be preferred to be chosen, so the CLT may be a better pick for the interval. This allows us to know that there is a 95% chance that the true mean lies within the confidence interval.

B. Scenario 2

1. Time and Frequency Playing Games:

In order to compare time playing video games the week before the survey and the frequency of play reported, we plot students who spent no time playing games the week prior to the survey in comparison to students who spent some time, against their respective frequency. These results are shown in Histograms 4.2.1. and 4.2.2. below.

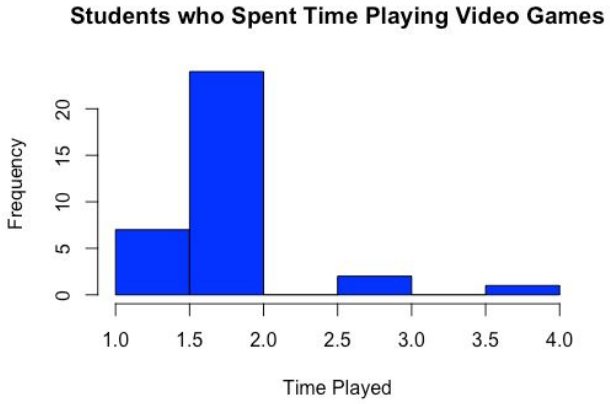

Figure 4.2.1. Distribution of students who played video games in the week prior to the study

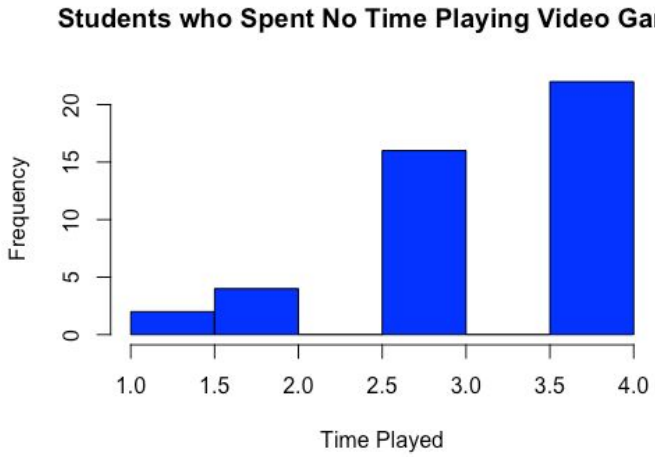

Figure 4.2.2. Distribution of students who didn't play video games the week prior to the survey.

These distributions show us that students who spent some time playing games the week prior to the survey play daily or weekly. In comparison, the negative skewness of students who spent no playing video games shows that the majority of these same students reported playing monthly or semesterly. These results make sense since people who spent time playing games or not, should be reflected in the frequency that they play. Next, we want to look at these distributions displayed on a box plot, which will show us different statistics of the two groups, such as the minimum, median, and maximum.

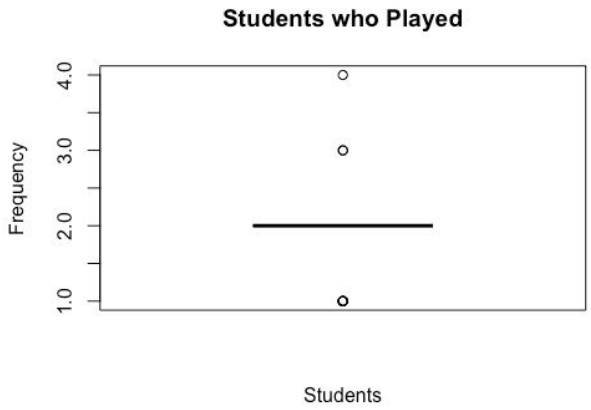

Figure 4.2.3. Boxplot of students who played the week prior and their distribution of frequencies.

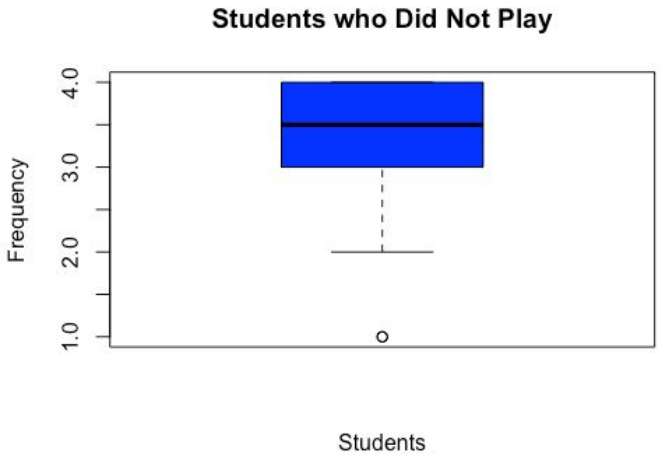

Figure 4.2.4. Boxplot of students who did not play the week prior and their distribution of frequencies.

Figure 4.2.3. tells us that 75% of students who said they play daily or weekly (1 or 2 frequency) did spend some time playing games the week before the survey, even though there was a test. The line for monthly and semesterly frequency means that the size of each bucket is very small, dense, or there is little variation. All students in the 3 and 4 groups spent little to no time playing video games. Figure 4.2.3. illustrates what we predicted would have been shown: students who did not play within the week prior to the test have a tendency to play only monthly or semesterly. However, there are still around 25% of students who report playing weekly or monthly. This lower quartile could be due to busyness of students since there was a test the week that this data was taken. This observation will be explored further in the next section.

2. Busyness Effect on Playing Games:

A comparison between the actual amount of time spent on video games and the reported frequency of playing is presented in the last section, and in this section, the effect of an exam, a cause for busyness, in the week prior to the survey will be estimated. In real life, if there was an exam prior to the survey, most students would prioritize studying for the exam over playing video games. As a consequence, the actual amount of time spent on video games would decrease, and the relationship between time playing and reported frequency would be weakened. Whether the result from the survey data set actually matches the assumption will be determined through a sign test.

In the survey dataset, the binary variable "busy" indicates whether a student plays video games when he or she is busy; to explain further, if busy equals to 1, the amount of time playing video games of the agent remains unchanged, but if busy equals to 0, the actual time playing changes to 0. Through Figure 4.2.5 below, we show that more students prefer to restrain from playing video games in an exam week.

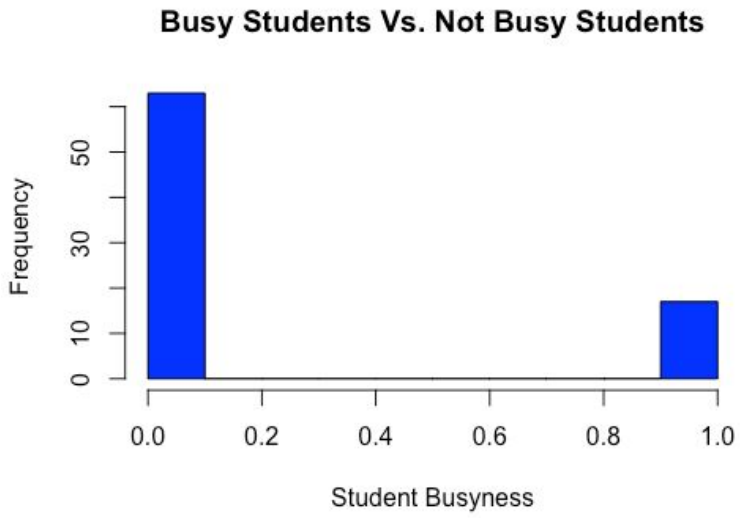

Figure 4.2.5 Frequencies of students who were busy vs not busy the week prior.

In addition, a box plot of reported frequency versus the variable "busy" would be helpful in displaying the difference in how often students play for students who have different answers to "busy". Below are two boxplots that show the distribution of frequencies for two different categories: students who identified as busy and students who identified as not busy.

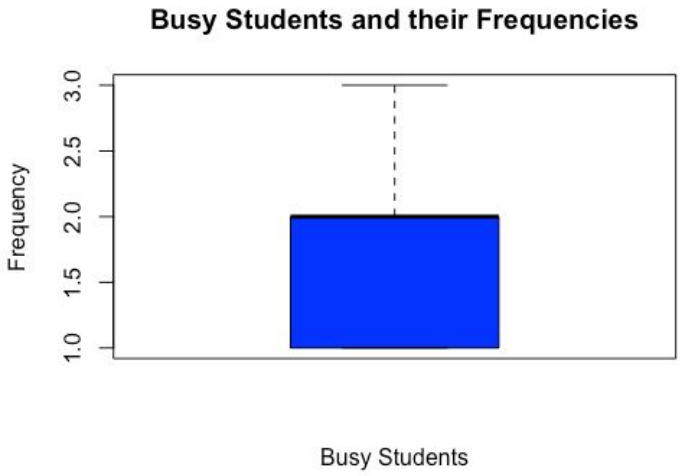

Figure 4.2.6 "Busy" students and their relative frequency of playing video games.

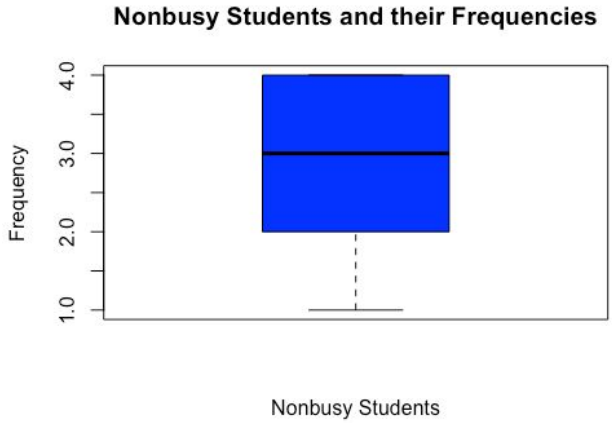

Figure 4.2.7 "Non-Busy" students and their relative frequency of playing video game.

These boxplots above illustrate that students who do not play video games when busy actually have higher reported frequency of playing video games. This observation leads to the inference that the effect of an exam is strong on whether a student will play video games, since the reduction in time spent on video games would be significant, given that the reported frequency and actual time playing have a positive correlation. In order to test this hypothesis, we will use a sign test which will indicate whether the claim that tests have an impact on the amount of video game time is significant. Our null hypothesis will be that the median of "time" and our new calculated variable "time.w.exam" should be equal if the presence of a test does not alter the amount of time played. Our alternate hypothesis will be that the medians will not be the same, since we believe that the presence of an exam will affect the amount of time playing video games and it would not be the same as time played when there is no exam.

To start this calculation, we create a new data series that is directly based off of the data column “busy”. If the value of “busy” is equal to 0, the corresponding value in our new series “time.w.exam” will be set to 0, otherwise it keeps the same value as the series “time” for that corresponding student. With this new series, we can run the sign test in order to obtain a p-value which will indicate whether we can reject or fail to reject the null hypothesis. A two-sided test with a 95% confidence level gives us a p-value of around 0.0000045. Since this p-value is less than our significance level of .05, we can reject our null hypothesis for the alternate hypothesis. Therefore, it is reasonable to claim that an exam exerts a huge influence on the amount of time that students spent on playing video games in the prior week.

C. Scenario 3

In the week prior to the survey, the students who actually took the survey reportedly spent 1.243 hours, on average, playing video games. But, this was a statistic taken from a sample of students and, thus, doesn’t necessarily fully represent the population’s true mean (μ). A confidence interval of 95% can help us to deduce whether or not it is fair to assume that the sample mean is a good representative of the entire population. The confidence interval provides a range of values where there is a 95% chance that μ is within those bounds. If μ is outside of the interval, we can reject the null hypothesis and can deduce that the mean is not a good representation of people’s behavior beyond the sample. The simple approximate confidence interval, derived by using the central limit theorem, for the sample is [0.45, 2.03]. Yet, this doesn’t reflect that the population is finite and was sampled from without replacement. So, it is more accurate to use the finite population correction factor. As a result, the 95% confidence interval then becomes [0.59, 1.90]. It is a slightly smaller range, making it more precise.

Another crucial factor for whether or not the sample mean is an accurate representation of the population is the sample size. The ratio of the number of samples to the size of the entire population (n/N) must be close to 0. In this case, 91/314 = 0.29. However, it is difficult to know if this value is close enough to 0. So, we will also use the bootstrap method to correct for this.

Yet, the distribution of the data can cause an issue in relation to building confidence intervals. The data needs to follow a normal distribution, otherwise the confidence intervals will not be useful for determining μ.

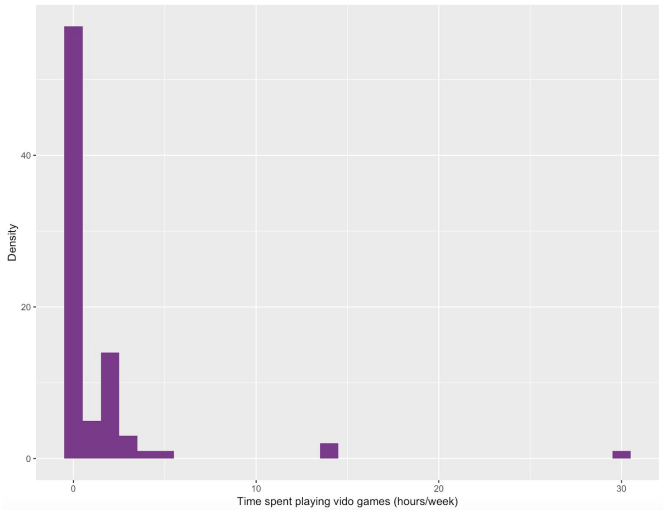

Figure 4.3.1. Distribution of hours per week spent playing video games.

It is not obvious if the distribution in Figure 4.3.1 follows a normal curve without. Therefore, employing the bootstrap method by using simple random samples without replacement provides a more revealing distribution, shown in Figure 4.3.2.

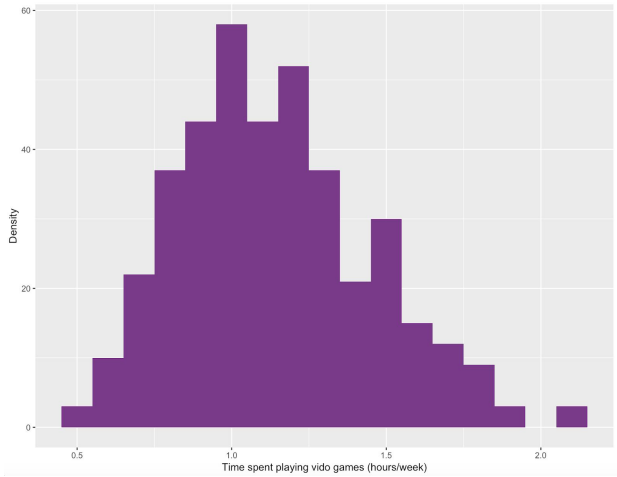

Figure 4.3.2. Bootstrap distribution of hours per week spent playing video games.

A quantile-quantile plot is useful for comparing the shapes of a distribution. More specifically, the qq-plot helps to test whether the data sequence of time spent playing video games is normally distributed, as demonstrated in Figure 4.3.3. The plot compares the bootstrap sample mean distribution to its theoretical normal distribution. The data looks mostly symmetric with the addition of heavy tails on both sides, indicating the presence of extreme values. The data is mostly linear but the intercepts do not equal zero, implying that the distribution is not necessarily normal.

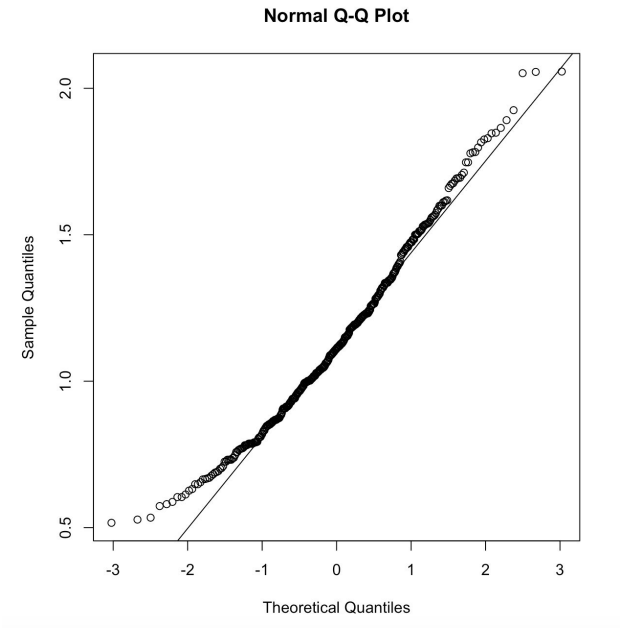

Figure 4.3.3. Normal quantile-quantile plot for the hours per week spent playing video games.

Once we run our bootstrap method, we take the 2.5% and 97.5% quantiles from the distribution to form a confidence interval for the population mean. This range contains 95% of the sample means, so the actual population mean would most likely be within the bounds of the interval. The bounds, as a result of our bootstrap implementation, are [0.63, 1.82]. Yet, the distribution of the sample means must follow the normal distribution in order to determine if this confidence interval is appropriate for generalization. Thus, we ran a Monte-carlo simulation of normal kurtosis and skewness to verify the type of distribution. The simulation we ran used a sample size of 1,000.

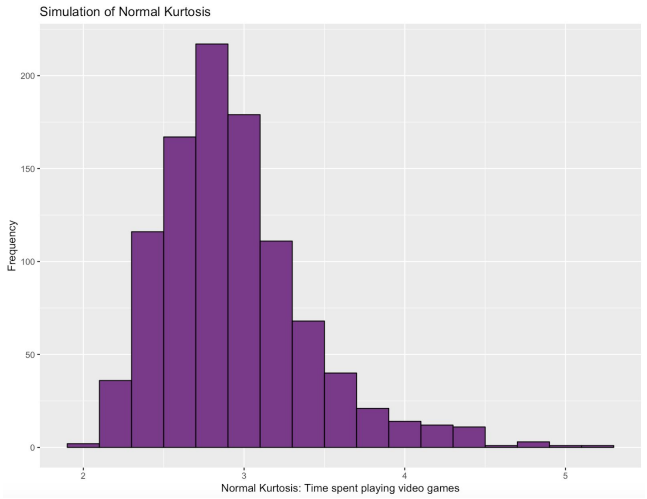

Figure 4.3.4. Distribution of Monte-carlo simulation of normal kurtosis.

The resulting distribution from the Monte-carlo simulation on kurtosis does not follow the normal curve, as seen in Figure 4.3.4. This is because the kurtosis is 2.81, which is a highly unlikely value in the normal distribution. Additionally, the skewness is of the distribution of sample means is 0.46. The distribution of Monte-carlo simulations on skewness (Figure 4.3.5) further confirms the non-normal distribution.

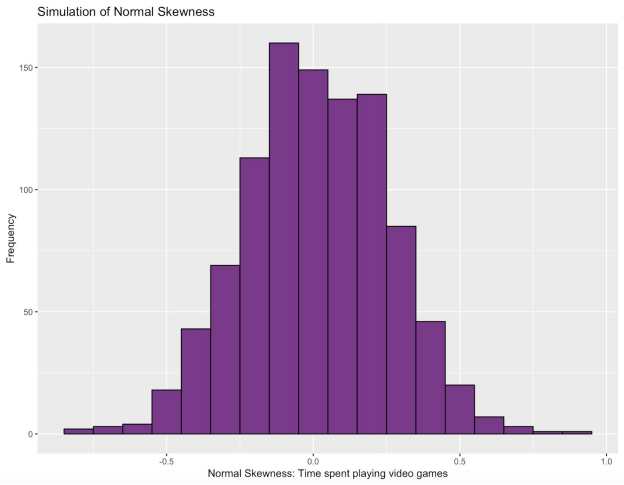

Figure 4.3.5. Distribution of Monte-carlo simulation of normal skewness.

A value of 0.46 for skewness is unlikely for a normal distribution. So, it is fair to assume that the resulting samples means from the bootstrap method do not follow the normal curve. Thus, it is not appropriate to use the confidence interval from that distribution, making the confidence interval derived from the central limit theorem with the finite population factor more accurate. Again, this interval is [0.59, 1.90]

By applying the results of the survey to the research question, the designers of the lab should consider that students would prefer to spend between 35 to 114 minutes a week in the statistics lab. These values were formulated by multiplying the confidence interval bounds by 60. So, the lab would be most effective if it held one session per week, lasting for 35-114 minutes. This conclusion is based on the consideration that most students who responded to the survey reportedly play video games weekly.

D. Scenario 4

In consideration of students' attitude towards video games, we utilized the survey results and ran them into algorithms (Decision Tree) that output a quantifiable judgement of students' opinions and preferences. Please reference Table 2.2, 2.3, and 2.4 from the Data section to see the survey results for type of video games played and reasons for playing. 63% of respondents play strategy games, and 66% of respondents play to relax. Other popular reasons for playing was to combat boredom (27%) and to experience the graphics/realism (26%). Table 2.4 indicates reasons why students dislike playing video games. This information is an important indication of

attitude because it provides insight on preventable reasons why students may have adverse reactions to a video-game based lab. Most students (48%) dislike how much time video games occupy, however other concerns were financial cost (40%) and that it is “pointless” (33%). It is important to note the implications of non-respondents here, who add bias.

To gather more insight (leading to more accurate assumptions) on the general attitude of students towards video games, we ran a Decision Tree model on the cleaned (removed null values of ‘99’) data. The Decision Tree classification model outputs a diagram that visualizes a statistical probability analysis, where each branch of the Decision Tree represents a potential outcome. This makes it a great choice to represent factors, driven from Tables 2.2 and 2.3, that influence preferences. The Decision Tree algorithm predicts that the educational value attributed to a game would most impact whether the students like or dislike the program in the given context. This makes sense considering that from the survey results, students are concerned about video games being a pointless, time-consuming activity - adding educational value would alleviate these concerns.

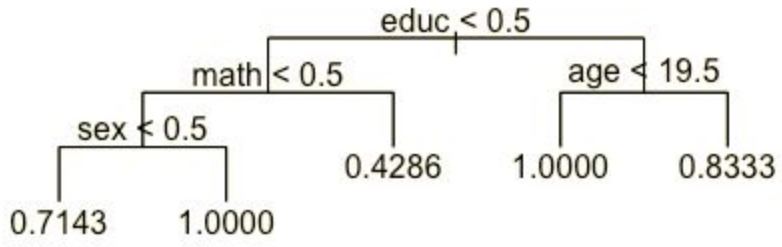

Figure 4.4.6. Decision Tree Model of factors influencing student gaming preferences.

E. Scenario 5

In order to highlight differences between students who like to play video games and those who do not, the original “like” column from the survey was reduced to 2 categories, like and dislike. The original 5 degrees of “like” were redefined into the like and dislike columns by the following criteria: 1 (never played), 4 (not really), and 5 (not at all) were counted in the dislike column, and 2 (very much) and 3 (somewhat) were counted in the like column. Once the new like and dislike columns were defined, a Chi-Squared Test was used to create a cross-tabulation table (Table 4.5.1) in order to compare the difference in video game enjoyment between the male and female students that were surveyed.

Table 4.5.1. Cross-tabulation for males and females.

Cross-Tabulation	Like	Dislike	Total
Female	26	12	38
Male	43	9	52
Total	69	21	90

Next, we wanted to compare video game enjoyment between the students who work and students who do not work. Students who worked more than 0 hours the week before the survey was taken were classified as students who work, and students who worked 0 hours the week before the survey was taken were classified as students who don't work. Another chi-squared test was used and the following cross-tabulation was made.

Table 4.5.2 Cross-tabulation for students who work those and those who do not work.

Cross-Tabulation	Like	Dislike	Total
Work	36	7	43
Don't Work	30	14	44
Total	66	21	87

We then wanted to compare video game enjoyment between the students who owned a PC and the students who did not own a PC. Another cross-tabulation, shown below, was made using a chi-squared test as well.

Table 4.5.3. Cross-tabulation for students who own a PC and those who do not.

Cross-Tabulation	Like	Dislike	Total
Own	48	18	66
Don't Own	21	3	24
Total	69	21	90

From the data above, more than half of the students who liked video games were male. Of the 52 male survey responses, 82.7% of them liked video games, whereas, only 68.4% of the female responses showed that they liked to play video games. Additionally, 83.7% of students who work liked video games, but only 68.2% of students who don't work liked video games. Table 4.5.3 shows that 12.5% of students who don't own a PC dislike video games and of the students who liked video games 30.4% of them don't own a PC.

F. Scenario 6

Another important feature to consider from this survey is the comparison between the target grade distribution (20% A's, 30%B's,40% C's and 10%D's and F's) and the expected grade distribution. Based on the raw data, the grade expectation is generally more optimistic than the target, so it is fair to conclude that the two distributions will be quite different. This is supported by the results of a Kolmogorov-Smirnov test. From the expected grades of the 91 students who responded to the survey and the simulated target grades with a sample size of 90, the small p-value (4.235e-07) implies that we can reject the null hypothesis and conclude that the two distributions are different. The conclusion that the two distributions are distinct is further reinforced by Figure 4.6.1 and Table 4.6.1.

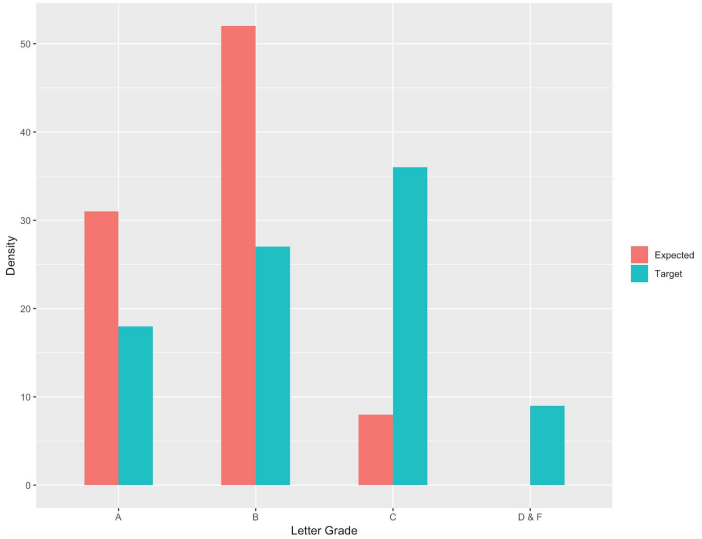

Figure 4.6.1. Distribution of expected and target grades of the responding students.

Table 4.6.1. Contingency table for the target and expected grades.

	D & F	**C**	**B**	**A**
Target	0.100	0.400	0.300	0.200
Expected	0.000	0.088	0.571	0.341

However, if we regard the students that didn't respond to this question as failing, the sample size and the distribution of expected grades will be different. The Kolmogorov-Smirnov test then yields a p-value of 4.96e-06, which is larger than the previous test. This means that there is a smaller difference between the distribution of expected and target grades. These new results are supported by Figure 4.6.2 and Table 4.6.2.

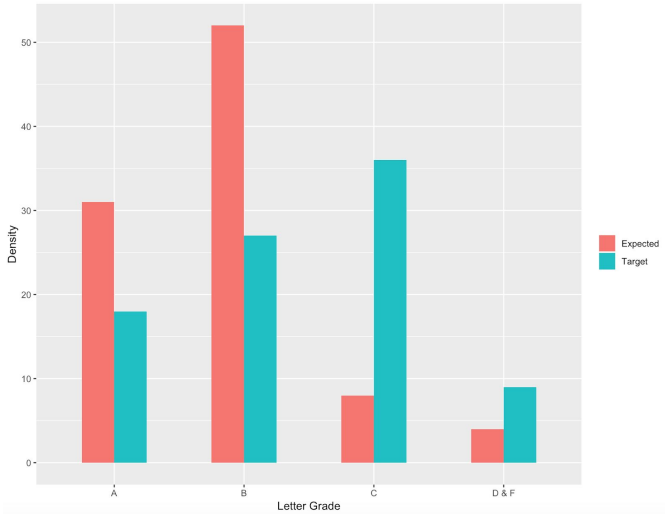

Figure 4.6.2. Distribution of expected and target grades of all the students from the sample.

Table 4.6.2. Contingency table for the target and expected grades, including non-respondents.

	D & F	**C**	**B**	**A**
Target	0.100	0.400	0.300	0.200
Expected	0.042	0.084	0.547	0.326

V. Theory

In this section, we will examine the problem of estimating the average amount of time students in the class spent playing video games in the week prior to the survey. To determine the exact amount of time for the entire class we would need to interview over 3000 students. Instead, a subset of them can be interviewed, and the information collected from this subset could provide an approximation to the full group. We will do this by using a simple random sample as our probability method for selecting students.

A. Terminology

- Population units make up the group that we want to know more about. In this lab, the units are the students enrolled in the 1994 Fall semester class of Introductory Probability and Statistics.
- Population size, usually denoted by N, is the total number of units in the population. For a very large population, the exact size of the population is often not known. Here, we have 314 students in the class.

- Unit characteristic is a particular piece of information about each member of the population. The characteristic that interests us in our example is the amount of time the student played video games in the week prior to the survey.
- Population parameter is a summary of the characteristic for all units in the population, such as the average value of the characteristic. The population parameter of interest to us here is the average amount of time students in the class spent playing video games in the week prior to the survey.
- Sample units are those members of the population selected for the sample.
- Sample size usually denoted by n, is the number of units chosen for the sample. We will use 91 for our sample size, and ignore the four who did not respond.
- Sample statistic is a numerical summary of the characteristics of the units sampled. The statistic estimated the population parameter. Since the population parameter in our example is the average time spent playing video games by all students in the class in the week prior to the survey, a reasonable sample statistic is the average time spent playing video games by 11 students in the sample.

B. Probability Model

The probability model fit to analyze survey data is the simple random sample. From probability theory, there are certain characteristics of the simple random sample that can be deduced and will be elaborated on in this section. A simple random sample entails selecting a certain number of individuals from a larger population. The model assigns probability to all samples of size n from a population of size N. Notably, survey data over a finite population is not an independent and identically distributed sample (i.i.d.) because individuals do not act as random variables, you can estimate expectation from a sample mean but won't know it's exact value - you can determine the population mean if n = N. By rule, each of the C(N n) samples are equally likely to be selected. Each unique sample of n has the same chance, roughly 1/ C(N n) of being selected. From the probability stated, we can implement the simple random sample model by:

- Assigning each unit a unique number from 1 to N.
- Recording each number, putting all records in a box to be mixed together.
- Drawing n tickets one at a time from the box without replacement.

C. Dependence

Conditional probability is derived from the formula P(A | B) = P(A intersection B) / P(B). Due to its definition, the probability that any two units in a population are chosen for the sample is n(n-1) / N(N-1), where P(one unit chosen for the sample) - n / N and P(another unit is chosen | one unit is chosen) = (n-1) / (N-1). So, in our study, the probability that any two students from the population were chosen to complete the survey and form our sample is (91*90) / (313*314), which equals 0.08.

D. Sample Statistics

1. Mean

Let $X_1, X_2, ..., X_N$ be the values of time for population unit i, where N = 314, then the population mean is:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Let $X_1, X_2, ..., X_N$ be the values of time for sample unit i, where N = 91, then the sample mean is:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_{I(j)}$$

Since, $x_{I(j)}$ represent the value of time spent playing video games by the sampled units, then the sample average is an unbiased estimator of the population parameter

$$\mathbb{E}\left(x_{I(j)}\right) = \sum_{i=1}^{N} x_i \mathbb{P}\left(I(j) = i\right) = \sum_{i=1}^{N} x_i \frac{1}{N} = \mu$$

2. Variance and Standard Deviation

To find the standard deviation of $\bar{x}$, we first find the variance of $x_{I(j)}$

$$\mathrm{Var}(x_{I(j)}) = \mathbb{E}\left(x_{I(j)} - \mu\right)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \sigma^2$$

Note: σ^2 represents the population variance

Then, we find the variance of $\bar{x}$

$$\begin{aligned}\mathrm{Var}(\bar{x}) &= \frac{1}{n^2}\mathrm{Var}\left(\sum_{j=1}^{n} x_{I(j)}\right) \\ &= \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{Var}\left(x_{I(j)}\right) + \frac{1}{n^2}\sum_{j=1, j\neq k}^{n}\mathrm{Cov}\left(x_{I(j)}, x_{I(k)}\right) \\ &= \frac{1}{n}\sigma^2 + \frac{n-1}{n}\mathrm{Cov}\left(x_{I(1)}, x_{I(2)}\right)\end{aligned}$$

Note: The previous equality is gotten by knowing that both $x_{I(j)}$ and $x_{I(k)}$ are Identically distributed. In addition, the covariance between the two samples units is not 0 but:

$$\mathrm{Cov}\left(x_{I(1)}, x_{I(2)}\right) = -\frac{\sigma^2}{N-1}$$

Thus, the Var($\bar{x}$) and SD($\bar{x}$) will be:

$$\mathrm{Var}(\bar{x}) = \frac{1}{n}\sigma^2\frac{N-n}{N-1}, \qquad \mathrm{SD}(\bar{x}) = \frac{1}{\sqrt{n}}\sigma\sqrt{\frac{N-n}{N-1}}$$

Looking at the above, it's important to note the inclusion of the finite population correction factor (1 - sampling ratio), which is impossible to ignore in our case because of the values of n and N.

Standard deviations for estimators (Standard Errors):

If σ^2 is unknown, then an estimator for population variance is:

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{I(j)} - \bar{x})^2$$

Using the new population variance, we can calculate the sample variance using:

$$\frac{s^2}{n}\frac{N-n}{N-1}$$

Note: we use s^2 to estimate and replace the unknown σ^2

A similar unbiased estimator of var(x) is:

$$\frac{s^2}{n}\frac{N-n}{N}$$

3. Totals and Percentages

When population parameters is requesting a proportion of students who play video games in the past week, x_i can now be a bernoulli trial, with i = 1, … , 314

$$x_i = \begin{cases} 1 & \text{if the ith student in the population owns a PC} \\ 0 & \text{otherwise} \end{cases}$$

Let $r = \sum x_i$, and $\pi = \frac{1}{N} * r$, where π is the proportion of students in the population of 314, who own PC. With this new population average, xbar still remains an unbiased estimate of the population average. Thus, new Var($\bar{x}$) and new standard error estimator formulas can be used:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \pi)^2 = \pi(1-\pi). \qquad \hat{SE}(\bar{x}) = \frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{n-1}}\frac{\sqrt{N-n}}{\sqrt{N}}$$

All the above can be found in a more compact way in the table below:

	Average	Proportion	Total
Parameter	μ	π	τ
Estimator	$\bar{x}$	$\bar{x}$	$N\bar{x}$
Expectation	μ	π	τ
Standard Error	$\frac{\sigma}{\sqrt{n}}\frac{\sqrt{N-n}}{\sqrt{N-1}}$	$\frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$	$N\frac{\sigma}{\sqrt{n}}\frac{\sqrt{N-n}}{\sqrt{N-1}}$
Estimator of SE	$\frac{s}{\sqrt{n}}\frac{\sqrt{N-n}}{\sqrt{N}}$	$\frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{n-1}}\sqrt{\frac{N-n}{N}}$	$N\frac{s}{\sqrt{n}}\frac{\sqrt{N-n}}{\sqrt{N}}$

Table: Properties of sample statistics

4. Confidence Intervals

a) Central Limit Theorem

We calculate the 95% approximate confidence intervals using the central limit theorem and the following formula:

$$\left(\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right)$$

where $\bar{x}$ is the sample mean, σ is the standard deviation, and n is the sample size, given a large enough sample size, and a sample sampling ratio.

b) Population Correction Factor

To get a more accurate confidence intervals, we apply the population correction factor to the finite population that we are using. Using the standard error we calculated earlier:

S.E = sqrt($\bar{x}$*(1 - $\bar{x}$)*(N-n)/((n-1)*N))

($\bar{x}$ − 1.96 * S.E., $\bar{x}$+ 1.96 * S.E.)

c) Bootstrap

To start a bootstrap, we create a new population that is of size 314 using the sample, this is called the bootstrap population. Technically every unit in the sample is repeated 314/91 times or 3.45 times. We create a histogram of 400 bootstrap sample means. If the new distribution is a normal distribution, we can calculate the 95% confidence interval by looking at the 0.975 and 0.025 quantiles of the bootstrap distribution of sample means.

E. Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test is a nonparametric tool to test equality of continuous, one-dimensional probability distributions for the purpose of comparing a sample and a referenced probability distribution. It may also be used to compare two sample populations. From this definition, this test is equipped to determine if a given sample may be attributed to a

particular probability distribution. We utilized the Kolmogorov-Smirnov Test to power our approach to Scenario 6. Please see the concluding distributions (Figures/Tables 4.6.1 & 4.6.2). Here, the K.S. test allowed us to analyze the expected and target grades of respondents.

Below is a brief elaboration on the K.S. statistic and associated distribution.

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[-\infty,x]}(X_i)$$

This function is characteristic of an empirical distribution. The associated statistic for a cumulative distribution function F(x) is as follows:

$$D_n = \sup_x |F_n(x) - F(x)|$$

"Sup" is representative of the supremum of the distance between Fn(x) and F(x).

F. Sign Test

The Sign Test is a non-parametric tool in statistics, meaning it can be used on data (such as the survey data we analyzed) that does not follow a normal distribution. This test is run on the null hypothesis and seeks to determine if the median of a distribution is equal to some value. Essentially, it tests the differences between pairs of observations. The sign test can be used either in place of a one-sample t-test, in place of a paired t-test, or for ordered categorical data. We implemented the Sign Test in our approach to Scenario 2 in order to better quantify the "busyness effect" on playing games. To do so, we made the assumption that students would prioritize studying before an exam, which would consequently decrease the amount of time they had to play video games. The Sign Test is equipped to test this assumption. To understand how the test works, consider the following assumptions:

- Data should come from two samples. The population may differ between the two.
- Dependent samples should be joined.

Procedure:

- Calculate the +/- sign for given distribution.
- Denote the total number of signs by 'n', and the number of less frequent signs by 'S'.
- Obtain the critical value (K) at .05 of the significance level with this formula:

$$K = \frac{n-1}{2} - 0.98\sqrt{n}$$

- Compare the value of 'S' with the critical value (K). If S > K, accept the null hypothesis. If S < the critical value of K, accept the null hypothesis.

G. Additional Question

From the survey data provided, there is ample opportunity for additional considerations that could inform researchers on how to effectively implement video games in statistics labs. The survey data also exists as an excellent standard of comparison with other survey data related to video game demographics or user optimization that have been previously and independently conducted. This digression will answer the following question: What attributes drive video game sales in the standard consumer? This question is significant to lab designers because the better that they know the market, the more informed decisions they can make when it comes to a well-received design in the lab setting.

We found an additional dataset that features sales data from over 16,500 games. From Table 2.2, we can see that from the sample of students surveyed from the original study, they were prompted to select the types of games played. From our additional dataset, one of the columns specifies genre which can be evaluated with the survey results for type preferences in mind. Comparing and contrasting the survey results to a huge sales dataset will be valuable insight for researchers to gather understanding of the market population while possessing knowledge of the preferences of their sample population.

When analyzing the new data frame indicating historic video game sales, we were particularly interested in the rank, name, year, and genre of the video games to answer our question of factors driving sales. Since our study is in America (UC Berkeley), we removed columns displaying sales in other countries (See Table 4.5.4).

Table 5.7.1. Top 10 video games as of 10/26/2016.

Rank	Name	Platform	Year	Genre	Publisher	NA_Sales
1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49
2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08
3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85
4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75
5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27
6	Tetris	GB	1989.0	Puzzle	Nintendo	23.20
7	New Super Mario Bros.	DS	2006.0	Platform	Nintendo	11.38
8	Wii Play	Wii	2006.0	Misc	Nintendo	14.03
9	New Super Mario Bros. Wii	Wii	2009.0	Platform	Nintendo	14.59
10	Duck Hunt	NES	1984.0	Shooter	Nintendo	26.93

To understand more about which genres are most popular in the American video game market (quantified by sales), we created a barplot to depict the data. Importantly, the market data possesses every category that was quantified in the survey data: Action, Adventure, Simulation, Sports, and Strategy. This makes for a statistically significant comparison between student-rated game types played and market-valued genres. As is depicted in Figure 4.6.3, the market value of games that were on the student survey and measured for profitability are ranked as follows in descending order: Sports, Action, Simulation, Strategy, and Adventure. Interestingly, students surveyed yielded different results and ranked (descending): Strategy, Action, Sports, Adventure, Simulation. A possible confound could be that the two studies were examining genre/type based on different metrics. The survey data asked the question of which games do you play while the market data ranked based on profitability.

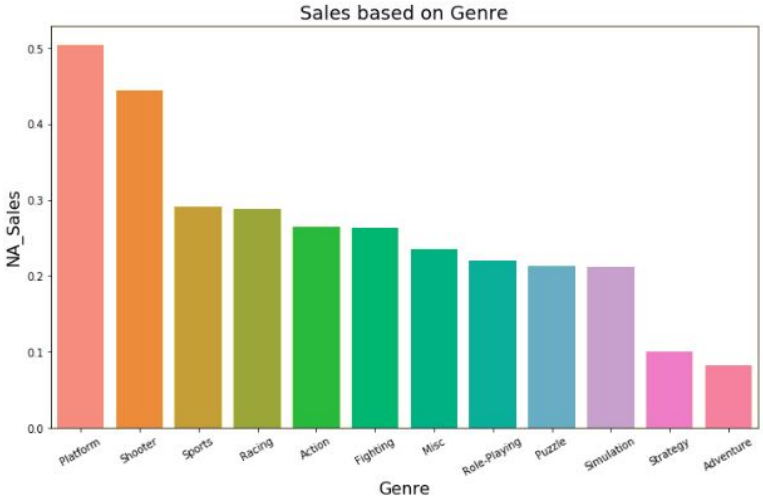

Figure 5.7.1 Top selling video game genres.

To answer the question of which attributes drive video game sales in the American market, we examined data on sales classified by genre because this metric established a commonality between our survey data and the market data. The significance of asking and answering this question is to provide UC Berkeley researchers with enhanced contextual information on what drives a successful video game experience. While the survey data provides information from a relevant sample population, it is our hope that market data will show what is historically viable. We found based on the types of video games on both the survey data and market data, it would be the researchers safest bet to design either a Sports or an Action game due to those types being both widely played by the sample population and viable in the historic market.

V. Conclusion

Through various numerical statistics and forms of graphical presentations, we are able to showcase numerous findings on this dataset of students and their video game play compared to their exams. We found the majority of students play strategy and action games, and that playing once a week is the optimal frequency of playtime. Yet, when there is an exam coming up, we found that students who report they play video games frequently are less likely to play when an exam gets close, proving that exams have a direct impact on video game play for students. We also discovered that length matters the most when students determine which video games to play, as the majority of reported reasons for disliking video games is due to "too much time". It is reported that most students prefer to play games that require a lot of work/computation, as this helps them relax more. Furthermore, the expected grade distribution and the target grade distribution have a significant difference, and the expected grade distribution is at a higher level since students receive a lower grade than they originally expected. As always, a limitation of our study is any of the claims we made cannot be generalized to students everywhere, but only to the population that we pulled our data from.

VI. Suggestions for Making a Better Computer Lab

After thorough review of the survey data and investigation of outside literature and data sources relevant to the topic of video games, our recommendation to lab designers at UC Berkeley program of statistics are as follows:

- Design games in the genre of Action or Sports because they are already played by a majority of the sample population and are historically profitable. (Data & Additional Question)
- Labs should be held once per week lasting between 35 and 114 minutes (Scenario 3)
    - However, during busy exam weeks, it is recommended to lower the frequency of labs
- Ensure that games have a clear, specific purpose relevant to the educational objectives of students enrolled in the course. (Scenario 4)
- Adjust the target grade distribution so it is closer to the expected grade distribution. (Scenario 6)
    - Make the grading of the lab more flexible and more dependent on how hard the students work

Works Cited

Bradic, Jelena. Math 189 Winter 2019 Slides
Smith, Gregory. “Video Game Sales.” Kaggle ,www .kaggle.com/gregorut/videogamesales/home.