Case Study 2: Labs vs. Video Games

Question

Are students using video games as a means to better understand concepts they learn in the classroom? How do grades differ in statistics and probability classes for students who play video games in comparison to those who do not play video games?

Hypothesis

We hypothesize that labs that integrate video games would help with the students' understanding of statistical and probabilistic concepts

Literature Review

Many students face troubles in learning statistical and probabilistic concepts solely because the material is highly abstract and requires inter-relating logical reasoning, critical thinking, data analysis and interpretation skills (Boyle). The lack of application in the traditional teaching methods of statistics has drawn a lot of criticism in recent years. One way to counter the problems faced with the current pedagogical methods would be to use technology to target topics where students face difficulties. After all, most modern research points towards the use of technology for learning, stating that learning should be interactive, situational and problem-based (Boyle). In order to validate this relationship, we will analyze the performance of students in a statistics class in conjunction with the amount of video games they play.

Introduction of Data

The data we will be working with in this study (videodata.txt) consists of undergraduate students form Section 1 of a UC Berkeley lower division Statistics 2 course during the Fall of 1994. There were 314 students in the class, out of which 95 students were randomly selected to complete the survey. The data used is their response to the survey. Out of the 95 students selected, 91 completed the survey. The survey consisted of three open response questions to indicate the number of hours they played video games and the number of hours they worked in the week prior to the survey, The student’s age was also collected. There were a series of yes/no questions to get information about the students’ usage of computers, like if they owned a PC, had an email

etc. There were some personal yes/no questions as well, like the sex of the student, if they hated math etc. There were other categorical questions that they were asked to gauge the extent to which they play games and their preferences, like how often they play and where they like to play. The students were also asked what grade they expected in this class. If the student did not answer the question properly, or at all, the code 99 was set as their response. Students who did not like playing video games/ have never played video games were asked to skip a series of questions. There was a followup survey that covered whether students liked or disliked games and the reason for the same. The students were free to have more than one response for these questions and give reasons for some of the questions, like why they played the games they play.

Background

The class was held on Monday, Wednesday and Fridays from 1-2pm. There was 10 discussion sections of 30 students each that met on Tuesdays and Thursdays. The course is a requirement for students who intend to major in business and also satisfies the quantitative reasoning requirement. Since students majoring in business are more likely to take this class, our results may be skewed, in turn indicating how business majors learn statistical and probabilistic concepts using video games. The students were randomly sampled from the list of students who took the second exam of the semester. The survey was given to students a week after the exam, on the day they were supposed to pick up their exams from their respective discussion sections. The students who could not be found during discussion were located during lecture and asked to take the survey. Though there was complete anonymity in the survey 91 of the 95 students provided complete responses.

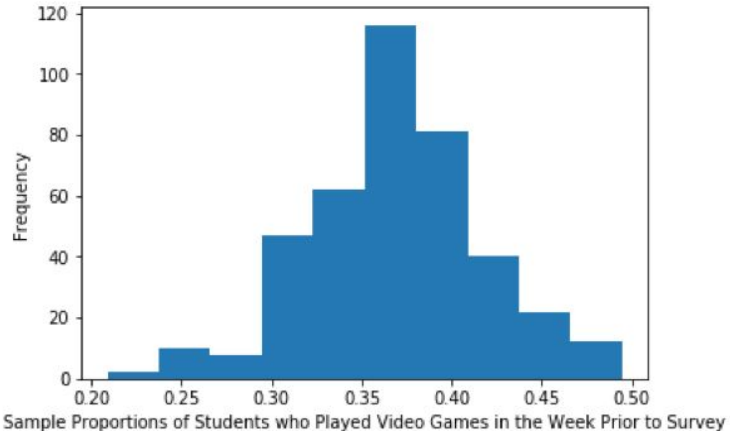
Basic Analysis

[Scenario 1]

An estimate of the fraction of students who played a video game in the week prior to the survey is: 0.374. Since the sample mean is an unbiased estimator of the population mean, our 0.374 estimate is legitimate.

Assuming the distribution of the sample averages is normal, a 95% confidence interval can be constructed by the interval:
[sample mean- 2* standard error, sample mean + 2* standard error]

To gain assurance that the distribution of the sample means is approximately normal, a bootstrap simulation was performed. A bootstrap population of 314 units was created based on the original sample. In each simulation, a sample of 91 units were selected from the bootstrap population and the sample's proportion was calculated. Four hundred simulations were run. The histogram below depicts the results of the four hundred simulations:



Visually, the distribution of sample proportions looks normal. This claim is validated by a kurtosis of -0.010 and a skew of 0.076. Since both these values are close to 0, the distribution is approximately normal, and the confidence interval described previously is valid. The calculated 95% confidence interval for the proportion of students who played video games in the week prior to the survey is:

[0.278, 0.469]

[Scenario 2]

Here is the table of descriptive statistics for students who reported that they play video games daily.

```
total hours:  40.0

count     9.000000
mean      4.444444
std       5.570258
min       0.000000
25%       1.000000
50%       2.000000
75%       4.000000
max      14.000000
```

Here is the table of descriptive statistics for students who reported that they play video games weekly.

```
total hours:  71.1

count    28.000000
mean      2.539286
std       5.499046
min       0.000000
25%       0.500000
50%       2.000000
75%       2.000000
max      30.000000
```

Here is the table of descriptive statistics for students who reported that they play video games monthly.

```
total hours:  1.0

count    18.000000
mean      0.055556
std       0.161690
min       0.000000
25%       0.000000
50%       0.000000
75%       0.000000
max       0.500000
```

Here is the table of descriptive statistics for students who reported that they play video games semesterly.

```
total hours:  1.0

count    23.000000
mean      0.043478
std       0.208514
min       0.000000
25%       0.000000
50%       0.000000
75%       0.000000
max       1.000000
```

Here is the table of descriptive statistics for students who did not answer or improperly answered the frequency question.

```
total hours:  1.0

count    23.000000
mean      0.043478
std       0.208514
min       0.000000
25%       0.000000
50%       0.000000
75%       0.000000
max       1.000000
```

As we can see from the descriptive statistics for each reported frequency of play, most people reportedly play video games weekly. In addition, the weekly frequency of play also has the highest number of hours played at 30 hours. Since there was an exam in the week prior to the survey, we would expect our previous estimates to be lower than we initially expected since students should spend their time studying for their exams instead of playing video games. As for the comparison of time spent playing video games in the week prior to the survey for each reported frequency of play, we would expect the total number of hours spent playing video games weekly would decrease since a total of 71.1 hours spent playing video games weekly seems like a lot given that students have their exams the same week. Upon further inspection, we see that one person who reported playing video games weekly recorded 30 hours spent playing video games weekly, so this person is the main cause for the unusually long amount of time spent playing video games weekly. However, even with that person who recorded 30 hours playing video games, the mean of people who played video games weekly is 2.539286, which makes sense given that students have exams the same week.

Below is my work to create a 95% confidence interval of the number of hours that students spent playing video games.

```
mean:  1.2428571428571429
std:  3.7770400773663693
std_err:  0.3959413840794146
mar_err:  0.7760451127956526
CI:  (0.4668120300614903, 2.0189022556527956)
```

Theory:

Confidence Interval: The distribution of sample means for samples of size n is illustrated as a normal distribution centered at μ. For any given value of z* the probability that a sample mean lies within z* standard deviations of the mean can be calculated using probability tables, and in this case, this probability is denoted as C. It is important to note that this probability does not bring any insight about the population mean, but rather only the sample means.

$$P\left(\mu - z^*\frac{\sigma}{\sqrt{n}} \le \bar{x} \le \mu + z^*\frac{\sigma}{\sqrt{n}}\right) = C$$

$$\mu - z^*\frac{\sigma}{\sqrt{n}} \le \bar{x} \le \mu + z^*\frac{\sigma}{\sqrt{n}}$$

$$-z^*\frac{\sigma}{\sqrt{n}} \le \bar{x} - \mu \le z^*\frac{\sigma}{\sqrt{n}}$$

$$-\bar{x} - z^*\frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{x} + z^*\frac{\sigma}{\sqrt{n}}$$

$$\bar{x} + z^*\frac{\sigma}{\sqrt{n}} \ge \mu \ge \bar{x} - z^*\frac{\sigma}{\sqrt{n}}$$

$$\bar{x} - z^*\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z^*\frac{\sigma}{\sqrt{n}}$$

$$P\left(\bar{x} - z^*\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z^*\frac{\sigma}{\sqrt{n}}\right) = C$$

Using this new form, C is called the confidence level, which determines the level of confidence that we have that the population mean lies within the indicated confidence interval. For example, if C = 0.95, then using probability tables, we find that z* = 1.96. With those values, we say that we are 95% confident that the population mean lies within the confidence interval. Here is the general inequality for a 95% confidence interval.

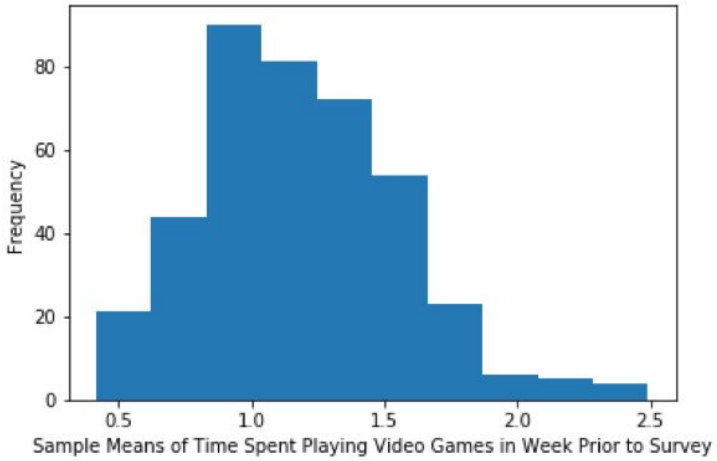
$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

[Scenario 3]

As described previously, assuming the distribution of the sample averages is normal, a 95% confidence interval can be constructed by the interval:

[sample mean- 2* standard error, sample mean + 2* standard error]

The sample mean of the amount of time the average student spent playing video games in the week prior to the survey is 1.24 hours. To ensure that the mean amount of time spent playing video games is approximately normal, a bootstrap simulation similar to the one described earlier was ran. The results of the simulation are depicted in the histogram below:



Visually, the distribution of sample proportions looks normal. This claim is validated by a kurtosis of 0.354 and a skew of 0.537. Since both these values are close to 0, the distribution is approximately normal and thus, the confidence interval described previously is appropriate. The calculated 95% confidence interval for the mean time spent playing video games in the week prior to the survey is:

[0.575, 1.910]

[Scenario 4]

In general, we think that most people do enjoy playing video games. To come to this conclusion, we first performed some data cleaning by consolidating the ‘like’ column from 5 categories to 2 categories, as a new column named ‘new_like’. We will denote 1 = never played, 4 = not really, and 5 = not at all, in the ‘like’ column as 0 = dislike in the new column ‘new_like’, and the rest as 1 = like in the 'new_like' column. Then, we took the proportion of values of 1 = like in the ‘new_like’ column and found that 0.7582417582417582, or about 75.82% of students who participated in this study, said that they like to play video games either somewhat or very much. Using the data from the follow-up survey, we performed some data analysis to create a list of the most important reasons why students like/dislike video games. We created a table showing the different reasons why students like playing video games. This table is shown below and contains the proportions of students who said that a certain column was the most important reason why they like to play video games. As we can see from the table, about 66% of students who participated in this survey said that they like to play video games because they are a form of relaxation for them. Students also like to play video games because of the feeling of mastery, the graphics and realism of video games, and video games provide a mental challenge for students.

relax	0.666667
master	0.287356
graphic	0.264368
challenge	0.241379
coord	0.045977

Similarly, we created a table showing the different reasons why students do not like playing video games. This table is shown below and contains the proportions of students who said that a certain column was the most important reason why they dislike to play video games. From the table, we see that the most important reason is that video games take up too much time out of students’ busy schedules. Between studying for classes, working jobs, and social activities, students have to manage their time effectively, which leaves very little free time to play video games. Another major reason for disliking video games is that they cost too much. Regardless of whether a student plays on a game console or PC, students will have to spend a lot of money to buy game consoles, games, or PC parts if they build their own PC. A third important reason is that students think that video games are pointless and provide no real value to them.

time	0.482759
cost	0.402299
point	0.333333
bored	0.275862
frust	0.264368
rules	0.195402
lonely	0.045977
friends	0.022989

In terms of any potential outliers or missing data, we found that in the ‘age’ column, there was a value of 33. Since the context of the study says that “All of the population studied were

undergraduates enrolled in Introductory Probability and Statistics, Section 1, during Fall 1994", the 'age' value of 33 initially implies that an undergraduate student who participated in this study was actually 33 years old at the time. Upon further inspection, as we see below, we see that this student put 40 as their value for 'work', suggesting that this student works full time (40 hours per week) and is attending school outside of work. Ultimately, we decided to keep this row of data since we rationalized that this particular row of data is feasible for this study.

	educ	sex	age	home	math	work	own	cdrom	grade	new_like
88	1	0	33	1	0	40	1	0	2	1

For missing data, we also considered those students who say that they have never played video games or do not at all like video games are asked to skip over some of these questions. Because students were asked to skip over some questions, these missing values are missing by design. In this case, since a 'like' value of 1 means that the student has never played video games, we looked at the videodata dataset where like = 1 (never played). Below is the resulting subset of the dataframe. From this dataframe below, we see that the student did not answer or improperly answered the questions corresponding to the 'where', 'freq', 'busy', and 'educ' columns, as indicated by the values of 99. Since we ended up using only the 'educ' column out of the four columns to implement the CART algorithm, we decided to keep this row of data and change the value in the 'educ' column to 0 to preserve the non-null data in this row and to use this row of data when creating the tree. Also, since this student spent 0 hours playing video games, it makes sense to replace 99 with 0 in the 'educ' column since this means that this student does not play educational games.

	time	like	where	freq	busy	educ	sex	age	home	math	work	own	cdrom	email	grade
81	0.0	1	99	99	99	99	1	19	1	1	10	1	0	0	4

In addition to looking at 'like' values of 1, we looked at 'like' values of 5 since a 'like' value of 5 means that the student does not like to play video games at all. So, we looked at the videodata dataset where like = 5 (not at all). Below is the resulting subset of the dataframe. As with the dataframe above, we also decided to keep these rows of data and change the values in the 'educ' column to 0 to preserve the non-null data in these rows and to use these rows of data when creating the tree. Also, since these students spent 0 hours playing video games, it makes sense to replace 99 with 0 in the 'educ' column since this means that these students do not play educational games.

	time	like	where	freq	busy	educ	sex	age	home	math	work	own	cdrom	email	grade
15	0.0	5	99	99	99	99	1	19	1	0	0	1	0	1	3
23	0.0	5	99	99	99	99	0	19	0	0	0	0	99	0	2
24	0.0	5	99	99	99	99	1	21	1	0	0	1	0	0	3
51	0.0	5	99	99	99	99	0	20	1	1	0	1	0	1	3
72	0.0	5	99	99	99	99	1	19	0	1	16	1	0	1	3
74	0.0	5	99	99	99	99	0	19	1	0	40	0	0	1	3
76	0.0	5	99	99	99	99	0	19	1	1	15	1	0	1	3

Below is the dataframe showing the descriptive statistics for each column, or feature, of the videomultiple dataset (the data from the follow-up survey). This dataframe is sorted by the mean in descending order. Next to the dataframe is the sum of the number of yes values for each column. As we can see from this dataframe, the column with the highest mean is the ‘relax’ column. This means that out of the 87 people who participated in the follow-up survey, most students (58 students said yes for ‘relax’ of the 87 total students who participated in the follow-up survey) said that relaxation is a reason why they like playing video games. This suggests that the main or most important reason that students like playing video games is because video games provide an outlet for relaxation for most students who participated in the follow-up survey.

	count	mean	std	min	25%	50%	75%	max
relax	87.0	0.666667	0.474137	0.0	0.0	1.0	1.0	1.0
strategy	87.0	0.632184	0.485006	0.0	0.0	1.0	1.0	1.0
action	87.0	0.517241	0.502599	0.0	0.0	1.0	1.0	1.0
time	87.0	0.482759	0.502599	0.0	0.0	0.0	1.0	1.0
cost	87.0	0.402299	0.493204	0.0	0.0	0.0	1.0	1.0
sport	87.0	0.390805	0.490759	0.0	0.0	0.0	1.0	1.0
point	87.0	0.333333	0.474137	0.0	0.0	0.0	1.0	1.0
master	87.0	0.287356	0.455153	0.0	0.0	0.0	1.0	1.0
adv	87.0	0.287356	0.455153	0.0	0.0	0.0	1.0	1.0
bored	87.0	0.275862	0.449539	0.0	0.0	0.0	1.0	1.0
graphic	87.0	0.264368	0.443553	0.0	0.0	0.0	1.0	1.0
frust	87.0	0.264368	0.443553	0.0	0.0	0.0	1.0	1.0
challenge	87.0	0.241379	0.430400	0.0	0.0	0.0	0.0	1.0
rules	87.0	0.195402	0.398809	0.0	0.0	0.0	0.0	1.0
sim	87.0	0.172414	0.379930	0.0	0.0	0.0	0.0	1.0
boring	87.0	0.160920	0.369587	0.0	0.0	0.0	0.0	1.0
coord	87.0	0.045977	0.210649	0.0	0.0	0.0	0.0	1.0
lonely	87.0	0.045977	0.210649	0.0	0.0	0.0	0.0	1.0
friends	87.0	0.022989	0.150736	0.0	0.0	0.0	0.0	1.0

relax	58
strategy	55
action	45
time	42
cost	35
sport	34
point	29
master	25
adv	25
bored	24
frust	23
graphic	23
challenge	21
rules	17
sim	15
boring	14
coord	4
lonely	4
friends	2

Theory:

After performing some descriptive statistics, we decided to create an algorithm to ultimately help us create a prediction tool to determine whether or not a student likes or dislikes video games. We used the CART algorithm from discussion to determine the most important reasons why students like and dislike video games. A classification and regression tree, or CART, is a machine-learning method for constructing prediction models from data. For classification purposes, we have a sample of n observations on a class variable Y that takes values 1, 2, ... , k, and p predictor variables, X_1,..., X_p. The goal is to find a model for predicting the values of Y (if Y is a continuous random variable) or classify Y (if Y is categorical) from new X values. The tree itself is formed by nodes and leaves, or terminal nodes. Each internal node is divided in two children nodes on the basis of a splitting rule. So, for each new observed value, the decision tree determines which leaves it belongs to according to the value of the explanatory variables.

With this in mind, let's apply these concepts to determine the most important reasons why students like and dislike video games.
Since our task only asks for like or dislike, we need to consolidate the 'like' column from 5 categories to 2 categories, as a new column named 'new_like'. We will denote 1 = never played, 4 = not really, and 5 = not at all, in the 'like' column as 0 = dislike in the new column 'new_like', and the rest as 1 = like in the 'new_like' column.
Below is the updated dataframe with the relevant columns, or features, from the original videodata dataset as well as the 'new_like' column to provide a binary representation of each student's preference over whether or not they like or dislike video games. This dataframe will help us implement the CART algorithm.

	educ	sex	age	home	math	work	own	cdrom	grade	new_like
0	1	0	19	1	0	10	1	0	4	1
1	0	0	18	1	1	0	1	1	2	1
2	0	1	19	1	0	0	1	0	3	1
3	1	0	19	1	0	0	1	0	3	1
4	1	0	19	1	1	0	0	0	3	1

Below is the resulting CART when we plot the decision tree with the 'new_like' column against the other columns, or features, listed in the dataframe above. To read this decision tree, each condition branches left for "true" and right for "false". When you end up at a value, the value array represents how many samples exist in each target value. So, value = [10, 3] mean there are 10 "dislike video games" and 3 "like video games" by the time we get to that point in the decision tree. In addition, value = [11, 22] means 11 "dislike video games" and 22 "like video games". To create this tree, we had to fill any values of 99 (meaning a question was not answered or improperly answered) with 0 to preserve any non-null values in the same row of the dataset and still use those rows of data.


```
{'educ': 0.16458170574526246,
 'sex': 0.07844058726673975,
 'age': 0.1018692333936571,
 'home': 0.04619923161361144,
 'math': 0.0831903158166613,
 'work': 0.23972892220454686,
 'own': 0.05696408339271145,
 'cdrom': 0.02069154774972558,
 'grade': 0.20833437281708414}
```

As we can see from the dictionary of the relevant videodata column names and their corresponding feature importances, we see that the number of hours that students work the week prior to the survey (the 'work' column) is the most important feature to determine whether a student likes or dislikes video games. This makes sense since if a student is working while in school, that student has to manage their time between studying for school and working their job,

which takes away from any free time that they may have to play video games, thereby influencing their opinion on whether or not they like video games. We also showed that the next two important features are the grade that the student expected to receive in the class (the 'grade' column) and whether or not the student plays educational games (the 'educ' column).

After creating the tree, we decided to create a random forest classifier to help predict if another student's values could determine if that student likes or dislikes video games. To do this, we created a random number generator to generate random numbers for each column. Then, we plugged those random numbers into the random forest classifier to see the predicted output. To see if the random forest classifier is working correctly, we used a row of data from the dataset where that row had a corresponding 'new_like' value of 0. Specifically, this student's values are the following: [0,1,20,1,1,10,1,0,3]. As expected, the random forest classifier gave an output of 0, which means that the classifier predicted that this particular student does not like video games. Since we know that this classifier is working properly, we decided to input random numbers for each relevant column, or feature, in the dataset. The following are the random numbers that were generated:

```
educ:   1
sex:    0
age:   19
home:   1
math:   0
work:  12
own:    1
cdrom:  1
grade:  3
```

When we used this set of random numbers as an input for the random forest classifier, we found that the classifier gave an output of 1, meaning that the classifier predicted that this student likes video games.

[Scenario 5]

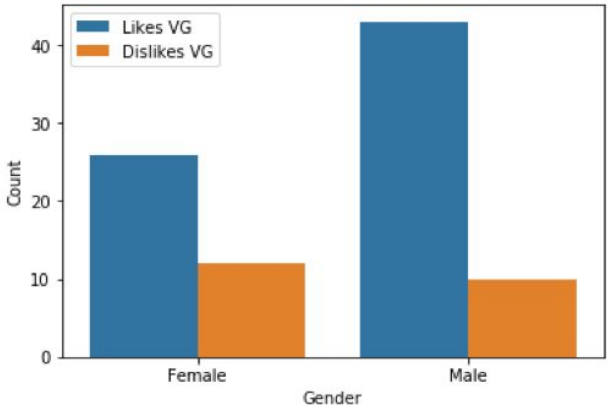
Not everyone likes to play video games. To find if differences between those who like to play video games and those who don't are correlated with gender, employment status and ownership of a computer, an analysis was performed.

Male vs. Female

To get an initial sense of the differences in how much each gender likes to play videogames, the proportion of those who like and dislike video games was calculated for each gender. Those who answered "like to play very much" or "like to play somewhat" were grouped into the "likes VG (videogames) " category; others were grouped in the "Dislike VG". The following table displays the results:

	Male	Female
Like VG	0.826923	0.684211
Dislike VG	0.173077	0.315789

The following histogram displays by gender the count of those who like and dislike video games:



From this initial analysis, it appears that while the majority of both genders like to play video games, a larger proportion of males like to play video games. To get a better sense of the differences among gender in liking video games, the proportions that members of each gender answered to the survey question of how much they like playing video games is displayed below:

Gender	Female	Male
Never Played	0.000	0.019
Very Much	0.132	0.346
Somewhat	0.553	0.481
Not Really	0.211	0.096
Not at All	0.105	0.058

Next, the time each gender spent playing video games in the week prior to the survey are shown below. Note, hours of playing were rounded up and those playing four or more hours of video games were aggregated into the "4+" group.

Gender	Female	Male
Hours Playing VGs		
0	0.763	0.528
1	0.105	0.132
2	0.053	0.245
3	0.026	0.038
4+	0.053	0.057

After analyzing these tables, our hypothesis that males like playing video games more than females is further validated. A higher proportion of females answered "not really" and "not at all" when asked how much they like playing videogames while a much higher proportion of males answered "very much". While a larger proportion of females answered, "somewhat", the proportions are similar. The second table displays that males played more hours of video games than females in the week prior to the survey. A larger proportion of males fell in each of the non-zero categories while, a larger proportion of females fell in the zero category. While the majority of each gender played no video games in the week prior to the survey, this observation is likely skewed. In the week prior to the survey, students in the survey were likely spending time studying for their upcoming stats exam and had little time to spend playing video games. If there was no exam during the week in question, the proportions of those playing no video games may have been lower.

To gain further insight, several additional statistics were calculated. For each gender, proportions that played video games in the week prior to the survey, proportions that like video games, average hours played in the week prior, the median frequency that video game players play, and proportions that play if busy were calculated. The results are shown below:

	Males	Females
Prop Played VGs	0.472	0.237
Prop Like to Play	0.827	0.684
Average Time Playing (in hours)	1.596	0.75
Median Freq VG Players Play	weekly	weekly
Prop Play if Busy	0.271	0.125

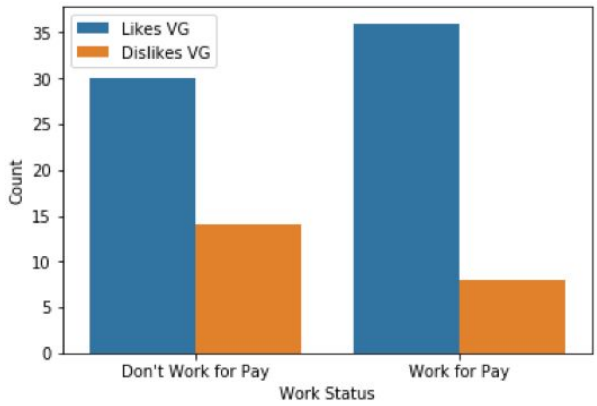
These results further validate our hypothesis that males enjoy video games more than females. In each of the statistics above, males have a higher proportion/average when compared to those of females. When considered with the previous findings and the population in question, namely the 314 students in Statistics 2, Section 1, during Fall 1994 at UC Berkeley, it is safe to conclude that a larger proportion of the male population like playing video games.

Work for Pay vs. Don't Work For Pay

To get an initial reading on whether a student's employment status [i.e., whether or not they work for pay] is correlated with their like/dislike of video games, the proportion of those who like and dislike video games was calculated. Those who answered "like to play very much" or "like to play somewhat" were grouped into the "likes VG (videogames)" category; others were grouped in the "Dislike VG". The following table displays the results:

	Work	Dont Work
Like VG	0.837	0.682
Dislike VG	0.163	0.318

The following histogram displays the count of those who like and dislike video games by employment status:



From this initial analysis, it appears that the while both the majority of those who work for pay and those who don't work for pay enjoy video games, a larger proportion of those who work for pay like to play video games. To get a better sense of the correlation of employment status and liking video games, the counts that members of each group answered when asked how much they like playing video games is displayed below:

Hours of Work	0	1-10	11-20	21+
Like Playing VGs				
Never Played	0	1	0	0
Very Much	9	6	5	3
Somewhat	21	9	13	0
Not Really	10	3	0	0
Not at All	4	0	2	1

A breakdown of the time spent playing versus the number of hours of work follows. Note that the hours of playing were rounded up and those playing four or more hours of video games were aggregated into the "4+" group.

Hours of Work Time Playing VGs	0	1-10	11-20	21+
0	30	10	13	3
1	4	1	4	0
2	7	5	3	0
3	1	2	0	0
4+	2	1	1	1

After analyzing these two tables, it appears that those who do and those who don't work for pay both like video games similarly and played similar hours video games in the week prior to the survey. As each of the above tables has 20 possible categories, these tables provide too much granularity to draw any conclusions.

To gain further insight into the favorability of video games when considering employment status, several less granular statistics were calculated. Proportions that played video games in the week prior to the survey, proportions that like video games, average hours played in the week prior, the median frequency video game players play, and the proportion that play if busy were calculated. The results are shown below:

	Workers	Non-Workers
Prop Played VGs	0.409091	0.318182
Prop Like to Play	0.837209	0.681818
Average Time Playing (in hours)	1.08182	1.45455
Median Frequency for VG Players	weekly	weekly
Prop Play if Busy	0.184211	0.230769

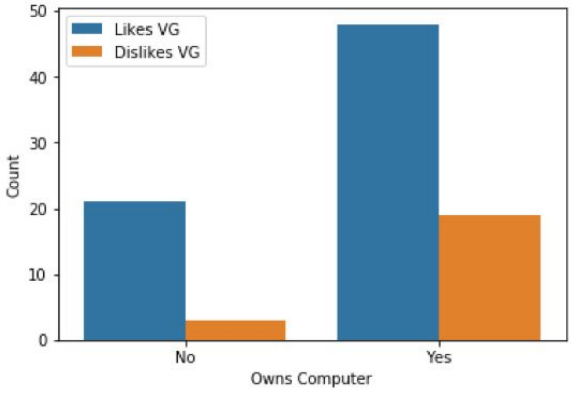
These results do not provide conclusive evidence that there is a difference in how much those who do and don't work for pay like video games. While those who work for pay had both a higher proportion who played video games in the week prior to the survey and a higher proportion of those who like to play video games, those who don't work for pay, on average, played more hours of video games and contain a greater proportion of students who will play if they are busy. Thus, there seems little predictive power in this distinction.

Owns Computer vs. Doesn't Own Computer

To get an initial reading on whether or not owning a computer is correlated to liking/disliking, the proportion of those who like and dislike video games was calculated for each of these groups. Those who answered "like to play very much" or "like to play somewhat" were grouped into the "likes VG" category; others were grouped in the "Dislike VG". The following table displays the results:

	Owners	Non-Owners
Like VG	0.727273	0.875
Dislike VG	0.272727	0.125

The histogram below displays the count of those who like and dislike video games for each category:



From this initial analysis, while it appears that the majorities of both groups like playing video games, those who don't own a computer have a higher observed proportion of liking video games. However, since only 24 of the 91 surveyed students didn't own a computer, slight perturbations to the data may drastically change each groups proportions. Thus these proportions alone cannot justify any conclusions on how owning a computer and liking video games are correlated. To get a better sense of the correlation between owning a computer and liking video games, the counts that members of each group answered when asked how much they like playing video games is displayed below:

Owns Computer	No	Yes
Like Playing VGs		
Never Played	0	1
Very Much	5	18
Somewhat	16	30
Not Really	1	12
Not at All	2	5

In addition, the counts of the hours each group member played video games in the week prior is shown below. Note that the hours of playing were rounded up and those playing four or more hours of video games were aggregated into the "4+" group.

Owns Computer	No	Yes
Time Playing VGs		
0	18	39
1	1	10
2	3	12
3	1	2
4+	1	4

After analyzing these tables, it appears that the initial proportions calculated may not have been representative of how much each group likes video games. While the counts of computer owners and non-owners are roughly proportional, the second table indicates that a larger proportion of computer owners played video games in the week prior to the survey. As the number of hours played in the week prior is an objective statistic, this may indicate that computer owners like video games more than those who don't own a computer. Thus, these results tell us that this distinction may not be an inform us about video game preferences.

To gain further insight into how owning a computer and whether liking video games are correlated, several additional statistics were calculated. Proportions that played video games in the week prior to the survey, proportions that like video games, average hours played in the week prior, the median frequency video game players play, and the proportion that play if busy were calculated. The results are shown below:

	Owner	Non-Owner
Prop Played VGs	0.41791	0.25
Prop Like to Play	0.727273	0.875
Average Time Playing	1.09851	1.64583
Median Frequency for VG Players	weekly	weekly
Prop Play if Busy	0.186441	0.285714

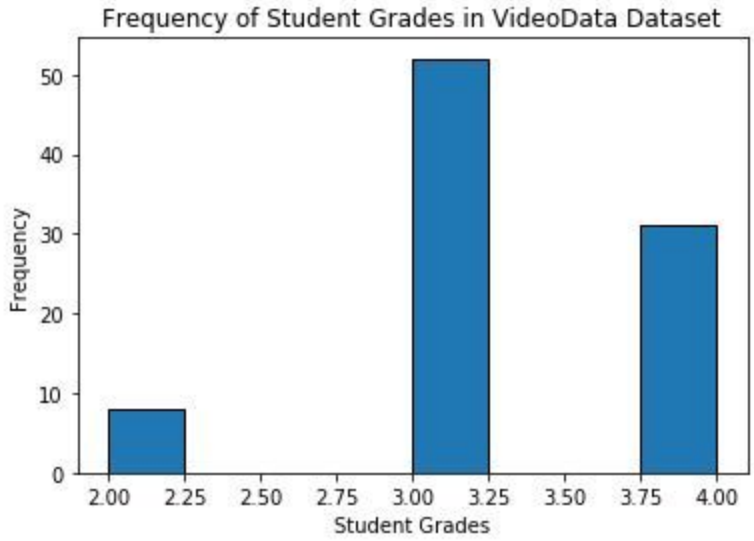
These results do not provide any conclusive evidence on whether owning a computer and liking to play video games are correlated. While a larger proportion of computer owners played video games in the week prior to the survey, non-owners on average played longer and are more likely to play video games when busy. The proportions that played video games in the week prior was likely swayed by the statistics midterm, and potentially underestimate the proportions of non-owners who play video games in a typical week. While there may be a statistically significant difference in the proportions of those in each group that like video games, the data produced by the survey don't enable us to conclude this.

[Scenario 6]

Below are the following items: a table showing the proportions of each grade value in the given videodata dataset, a table showing descriptive statistics of the grades in the given dataset, and a histogram binning the grade values automatically. As we can see from the first table, we see that about 34% of students who participated in the study said that they expected to receive an A, about 57% of students expected a B, and about 8% of students expected a C. This is very different (and therefore, does not match) the target distribution of 20% A's, 30% B's, 40% C's, and 10% D's or lower used in grading assignments. From the second table, we see that the mean expected grade is about 3.25, which translates to either a B- or B, which is a pretty high average when compared to the target average of a C.

	grade
3	0.571429
4	0.340659
2	0.087912

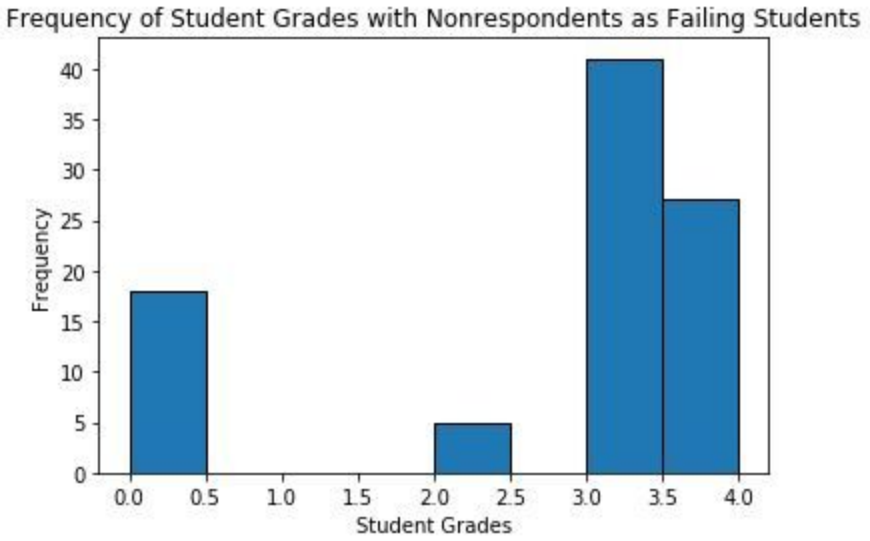
	count	mean	std	min	25%	50%	75%	max
grade	91.0	3.252747	0.607242	2.0	3.0	3.0	4.0	4.0

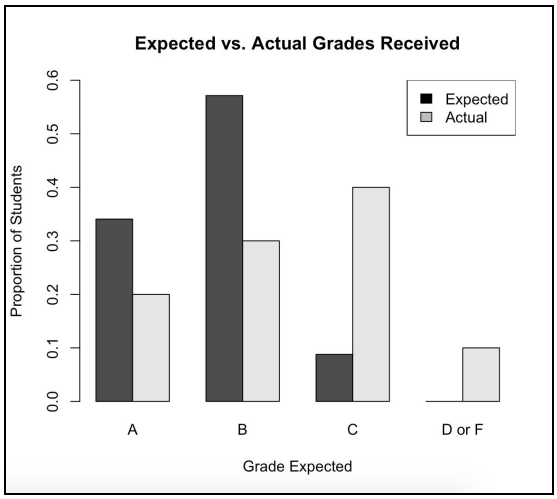


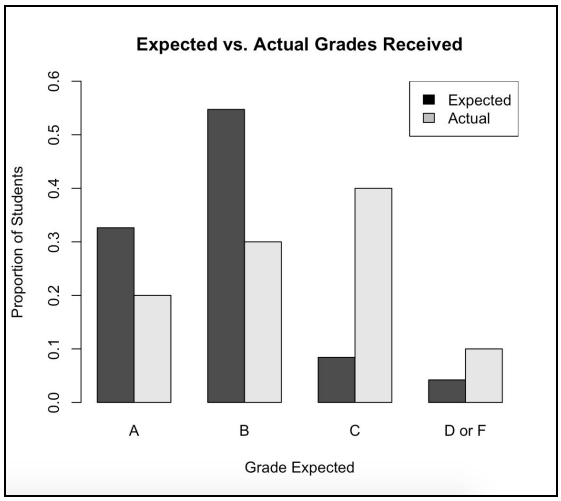
Below are the same items from above, but with nonrespondents represented as failing students. Specifically, we edited the grade values of nonrespondents to 0 to represent that these students expected to get an F. As we can see from the first table, we see that about 29% of students who participated in the study said that they expected to receive an A, about 45% of students expected a B, about 5% of students expected a C, and now about 19% of students (the nonrespondents) expected an F. This is still different (and therefore, does not match) the target distribution of 20% A's, 30% B's, 40% C's, and 10% D's or lower used in grading assignments. Specifically, we see a larger percentage of students expecting an F than those expecting a C. This is due to the fact that the grade values for the nonrespondents were changed to 0 to represent them as failing students, as per the question. From the second table, we see that the mean expected grade is about 2.64, which roughly translates to a C or C+, which is a closer average than the average from the given videodata dataset when compared to the target average of a C.

	grade
3	0.450549
4	0.296703
0	0.197802
2	0.054945

	count	mean	std	min	25%	50%	75%	max
grade	91.0	2.648352	1.424965	0.0	2.5	3.0	4.0	4.0







The image on the left compares the students' expected grades for the class and the target distribution presented in lecture as it appears in the data set. For the image on the right we have added the four non-respondents to the *D or F* bar and adjusted the other bars accordingly. You

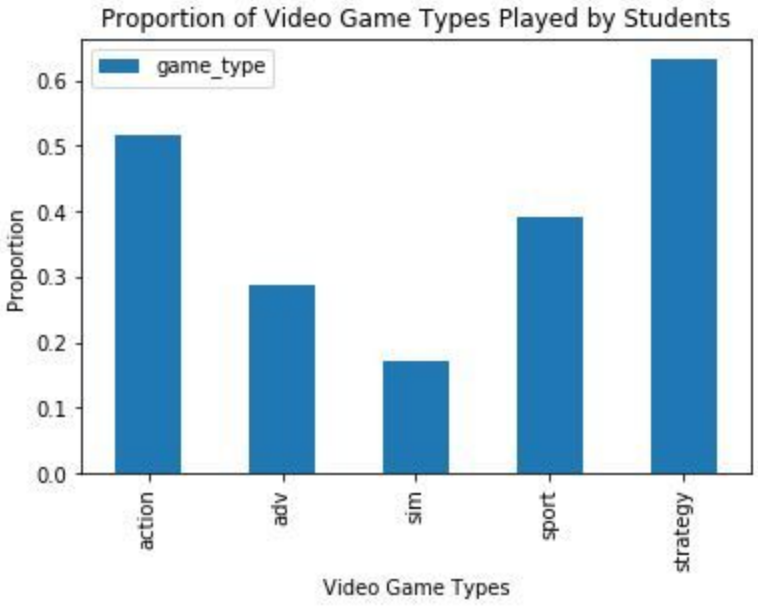
can see that the image changes, but only in the slightest, due to the fact that only 4.3% of the students selected for the survey neglected to respond. You can the similarity in proportions in the table below:

	A	B	C	D or F
Expected Grade	34.0%	57.1%	8.9%	0.0%
Adjusted Expected Grade	32.6%	54.7%	8.4%	4.3%
Target Distribution	20%	30%	40%	10%

We can see from the table that students have a tendency to overestimate their ability or the rigor of the class. The proportion of students who projected themselves to receive either an *A* or *B* was much higher than the target distribution whereas the opposite is true for grades of *C, D,* or *F*. Since the non-respondents were added to the *D or F* group (we consider failing to be having either a *D* or *F* due to the fact that this would require one to retake the class), the *A, B,* and *C* groups will decrease. The denominator that was used to find the original proportions will become larger with the addition of the four missing students, thus causing the proportion of the other three groups to decrease.

[Scenario 7]

Our additional question that we wanted to explore and answer was the following: Using the data from the follow-up survey, what was the type of video game that most students who participated in the study put as the type of video game that they play? So, we had to subset the given data to work with only the columns of data corresponding to each type of video game that was included in the follow-up survey. From there, we calculated the proportions of each type of video game and plotted those values in a plot, which is shown below.



As we can see from the plot, the type of video game that most students put as the type that they play is strategy games. This makes sense since strategy games are not as fast-paced as action games, so strategy games provide a form of relaxation that most students who play video games seek out, as evidenced by the fact that relaxation is the most important reason that students like playing video games, according to the given data. In addition, strategy games provide a mental challenge to students and a feeling of mastery when a student wins a game. Both the mental challenge and feeling of mastery are also some important reasons why students like playing video games.

[Theory]

The equation for the expected value of the sample mean is an unbiased estimator. Here is the proof for that equation.

$$E[\bar{x}] = E\left[\frac{1}{n}\sum_{j=1}^{n} x_{I(j)}\right]$$

$$= \frac{1}{n} E[x_{I(1)} + x_{I(2)} + \ldots + x_{I(n)}]$$

$$= \frac{1}{n} (E[x_{I(1)}] + E[x_{I(2)}] + \ldots + E[x_{I(n)}])$$

$$= \frac{n\mu}{n}$$

$$= \mu$$

Negative covariance means that one unit increases while the second unit decreases. Here is the proof for the variance equation.

$$E[x_{I(j)} - \mu]^2$$

$$= \sum_{i=1}^{N} (x_i - \mu)^2 \, P(I(j) = i)$$

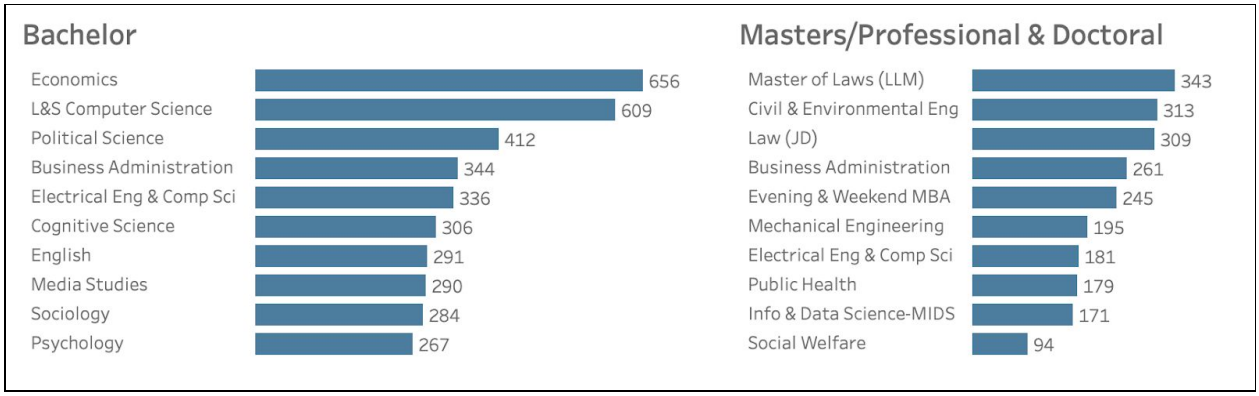
$$= \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$= \sigma^2$$

Advanced Analysis

Data Limitations

It is worth noting that there are some limitations to our data in the fact that the method from which our data was pulled is a bit biased. The survey that is administered should aim for a sample that serves as a representation of the total population of interest. Our population of interest is half of the 3,000 to 4,000 students that will be taking the introductory statistics course at UC Berkeley. The labs that the committee of faculty are trying to produce are meant to extend the traditional syllabus for a course by providing an interactive learning environment that offers students an alternative method for learning the concepts of statistics and probability. It is said that, "some have linked labs to video games" which is where we obtain our goal for this assessment. However, it may be a bit hasty to assume that a student's preference on video games translates to the classroom in a lab setting. Furthermore, the subjects in this survey/study are of a certain age range (as far as students go) as well as confined to certain types of majors. Below is the major breakdown for students who received degrees (both undergraduate and graduate) from the 2017-2018 class at UC Berkeley:



Another limiting factor of this survey was the fact that 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. As a result, we are left without some valuable information about these students that could've helped the faculty make a decision about how to better admisiter labs (possibly in a way that does not resemble video game play).

The simple random sample is the probability method for selecting the students in our case. This method is a probability model for assigning probabilities to all samples of size n from the entire population of size N. Each unique random sample (of size n units) has the same chance, $1/(NCn)$ of being chosen. Although, there is some dependence in our selections in that the second student chosen will depend on who the first student chosen was. Depending on the makeup of the class, this could have an effect on what the makeup is of those taking our survey.

Conclusion

The analysis performed on the given data sets strongly indicates the video games would be a good means to aid the teaching of statistical and probabilistic classes, accepting our initial hypothesis. We could see that the possible factors that affected students playing video included, the number of hours worked in the past week, the gender of the student, the ownership status of a PC and the reason the student played video games. From our analysis we saw that students mainly play video games for relaxation. Further, on performing chi square tests, we saw that we can not make the conclusion that disliking video games was associated with gender, or the ownership status of a PC.

Video games would certainly help with the learning process considering that most students would find the labs to be ‘relaxing’. Students will no longer see their homework/labs as an ordeal task, instead they would look forward to doing it. Since there was no association between gender and like/dislike of video games, we can safely presume that the new pedagogical methods are not favourable for one gender over the other. To add to this, we concluded that students who worked a lot, just didn't have enough time to spend on playing video games. Those same students are probably very stressed, working that many hours a week on top of their school work loads. If video games were integrated into the statistics curriculum, it would help them relax a little bit more, while also helping them understand concepts better.

Works Cited

https://www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf

https://www.sciencedirect.com/science/article/pii/S0360131514000141