Video Game Survey

Austin Clark, Timothy Lee, Andy Schleicher, Tom Tang, Aiden Yoon, Benson Wu

February 16, 2019

Abstract

Method

This statistical analysis was conducted using a survey data set of 91 students from a lower division statistics course at UC Berkeley. This survey was intended to help the academic committee design more interactive labs to facilitate an interactive learning environment for students to better understand statistics and probability. We hypothesized that attributes related to visual appeal and mental stimulation, and that the lab duration of the lab should be around 1 hour to retain students' attention. Methods used in this analysis include bootstrap resampling for confidence intervals, Chi-square goodness of fit test, classification tree, and logistic regression. Data was analyzed and visualized with R software.

Result

Initially, we found out that there are approximately 37% of people that played games a week prior to taking the survey. However, before the students filled the survey there was a test they needed to take. We assume that because of this test that people played fewer hours the week prior to the survey to study for the test. After investigating the data, we discovered that there was indeed a drop in hours played compared to their usual number of hours played. Due to the nature of the data we can't conclude this fact but it will be worthwhile to explore it further through a second survey. Then, we found that people usually play around an hour of game per week based on the number of hours they played a week prior to the survey. So this result might be skewed left as there is potential that students play more a week on a regular basis. Using this knowledge, it might benefit students to construct the lab for about an hour or slightly more to maintain their full attention. To determine which students will enjoy playing games, we found that education level, owning a computer at home, age, sex, grade in statistics class, affinity towards math, and number of hours they work all play an important role in determining whether a student likes playing games. From our data, we discovered that play who have played in arcades like playing games more than those that don't. We believe that this is due to the time period when gaming culture didn't revolve around computer games as much as the arcade or console.

Conclusion

Around 37% percent of students enjoy video games, and that overall for this class we expect a lab of about an hour in length that focuses on the graphic aspect of gaming would be most beneficial to the students. Furthermore, we do not want to entirely base the new lab on video gaming because there are subsets of the population that do not enjoy gaming and we do not want to alienate that portion of the population.

Background

Over the last 30 years video games have seen a huge rise in popularity and the amount of time played. Video games have become integrated into mainstream culture and today's youth are spending hundreds of hours playing, mastering, and competing in video games. The fact that video games are so much more in depth and interactive than many other sources of media have caused people to question what specifically it is about video games that inspires such focus and motivation to play. Papers such as Endogenous fantasy and learning in digital games looks into specifically the motivation factor that drives people to play video games with a psychological approach. Other papers investigate the competition and learning components that drive people to video game addiction and which has made video games into a competitive E-sport. Furthermore, institutions such as UC Berkeley have begun looking into whether some of these motivating tools from video games can be used in a classroom setting to engage students in newer and better ways. Students who were enrolled in advanced statistics courses conducted the survey by creating a questionnaire and selected 95 out of 314 students in an introductory statistics course in Fall 1994 to participate in the survey; only 91 complete surveys were received.

What proportion of people actually play video games?

The goal is to calculate the proportion of students who played video games the week prior to the exam as well as construct a confidence interval and a point estimate for that proportion. The variable time is the number of hours played by students the week prior to the exam. Due to the fact that we could not survey the 3000 plus student body, we need to survey a portion or subspace of the population. The simple random sample techniques will help us do this by giving us a probability method for surveying the students. In this survey our population size will be 314 ($= N$) and our sample size will be 95 but we will use 91 ($= n$) as four students did not respond. So we calculate how many students had a nonzero response in reference to the time variable and then divide that by the total population, and resulted in $\hat{p} = 0.037$. We continued by creating an approximate confidence interval using a standard error as our standard deviation with the following formula.

$$S.E(\hat{p}) = \frac{\sqrt{\hat{p} * (1 - \hat{p})} * \sqrt{(N - n)}}{\sqrt{n - 1} * \sqrt{(N)}}$$

Then, we would create a confidence interval around p-hat by multiplying by 2 times the standard error divided by the square root of 91. With this, we get an approximate 95% confidence interval of (0.3646, 0.3826). However, because n isn't too large and n/N isn't too small, we will use bootstrap method to construct a confidence interval since we aren't sure whether or not the data is normally distributed.

First, to use the bootstrap method, we populate the data by roughly multiplying the count by 3.45 to increase the population to 314. Then, we calculated the mean and standard deviation of proportion of people that played games prior to the survey and compare to the bootstrapped mean and standard deviation by sampling without replacement from the populated data. We use the t-test statistics to construct a confidence interval of the proportion of people who played games prior to taking the survey by averaging over 1000 bootstrap samples. Below, the table shows the 95% confidence interval constructed through bootstrap.

Table 1: Bootstrap CI of Proportion of People Who Played a Week Prior to Survey

2.5%	97.5%
0.2918	0.4589

Did the students reported frequency of play match up with the actual time they spent playing video games in the last week?

Chi-Square Test: Expected Vs. Observed game playing before the exam

Our objective in this scenario is to find out whether the exam that was given affected the student's video game playing time. In order to do this, we examine the observed data, the amount of time that the students spent playing video games the week prior, compared to the expected data which is how frequently the students play video games. The frequency of the expected data is separated into four categories: daily, monthly, weekly, and semesterly. We also denoted the observed data as the variable denoted time. Comparing these two sets of data might allow us to explain if the exam had any affect on the student's video game playing time the week prior. In order to answer this question we utilize the chi-squared test to assess whether our observed data is significantly different from our expected data. Our hypothesis for this test are as follows:

$$T = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

H_0 =The expected and observed video game time played comes from the same distribution

H_1 = The expected and observed video game time played comes from different distributions

As we stated earlier, the chi-squared test can help us determine whether our observed data and expected data are significantly different which leads us to determine whether both data sets come from the same distribution. In scenario 6 we explore why the test statistic has a chi-squared distribution. The first objective was to set up a specific range for each bin (category) that we organize our observed data. The survey data for our time category was denoted as Daily = 1, Weekly = 2, Monthly = 3 and Semesterly/Never = 4. Note that the decision was made to include never with the semesterly category as there was thirteen N/A responses which corresponded to the 0.0 response in our observed data. We are not sure that the thirteen values coincide with the student never playing video games which could lead to a confounding variable which we will cover later. We used half an hour increments as our standard unit of time. For each result from our time data set we arranged the data as follows:

Number of hours played (x)	How often the student played
<3.5 hrs	Daily (1)
0.5 - 3.5 hrs	Weekly (2)
0 - 0.5 hrs	Monthly (3)
0 hrs	Semesterly/Never (4)

	observed	expected
Daily	5	9
Weekly	2	3 28
Monthly	6	18
Semesterly/Never	57	36

After our bins have been set we ran the chi-squared ( ) test. The test resulted in a chi-squared test statistic of 12.375, degrees of freedom value of 3, and a p-value of 0.006203.

The last thing to consider would be confounding variables. The most notable confounding variables would be the thirteen N/A responses from our observed data and the arbitrary nature of the selection bins. The N/A responses can be interpreted as students who never play video games and did not play video games the week prior. This might be a naive assumption as we do not know whether these students completed the survey out correctly or were being completely honest while taking the survey. To assume these thirteen students

do not play video games might produce inaccurate results which is why these responses can be considered confounding variables. The next confounding variable is the arbitrary way in which the bins can be set up. We chose to have the categories be measured in half an hour increments with the range of the categories in the table above. This process is completely arbitrary with no correct way of assigning these categories. Thus, this process is considered a confounding variable.

How much time do students spend playing video games?

Estimating an interval for average time spent playing video games through bootstrap resampling

We were interested in finding an average amount of time that students spend playing video games so we computed both and estimate based on our sample data and then simulated multiple samples through bootstrap to see how appropriate an interval we have. From the histogram above we see that our data is clearly skewed right with a large number of people in the 0 time spent playing video games in the week before and several severe outlier cases therefore we cannot proceed as if our data was normally distributed. We calculated our sample mean using (slide 28) $\bar{x} = 1.242857$. Since we are dealing with survey data which is not necessarily i.i.d and is skewed we used (slide 32 sample variance) and then calculated the unbiased estimator (slide 33) s=3.771021. By the central limit theorem we assume with 95% confidence that the true mean is within one or two standard errors of xbar(slide 39) which means from our data we assume the mean to be in the interval (0.4522, 2.0335). Since our data appears to be very skewed we ran bootstrap on our sample to come up with an interval estimate to see how appropriate ours is. Using the same process described in scenario 1. We yield a 90% confidence interval (0.1134, 1.6727). If we compare the two intervals we see that the bootstrap interval is shifted to the left and since our data is so heavily skewed we believe this to be the better interval to estimate the true mean for the amount of time people play video games. We believe with 90% certainty that the true mean for the amount of time that people enjoy playing video games is between 0.113 and 1.67 hours. A confounding variable to note from this estimate is that in section 2 we showed that there is statistical evidence that the fact that there was a test the week before may have had an effect on the amount of time that people play video games so that our interval may not be an accurate representation of a typical week of play. Also, another statistic that may be useful is to just look at the population of people that enjoy playing video games to see the amount of time that they spend on it in a given week. Through bootstrap we estimate the mean amount of time with 90% certainty to be (0.5120, 2.499). Overall we believe that if we were to design a lab which is meant to engage students in a way similar to that of video games the amount of time that students would be interested in such an activity would be somewhere in the 0.11 to 1.7 hour range.

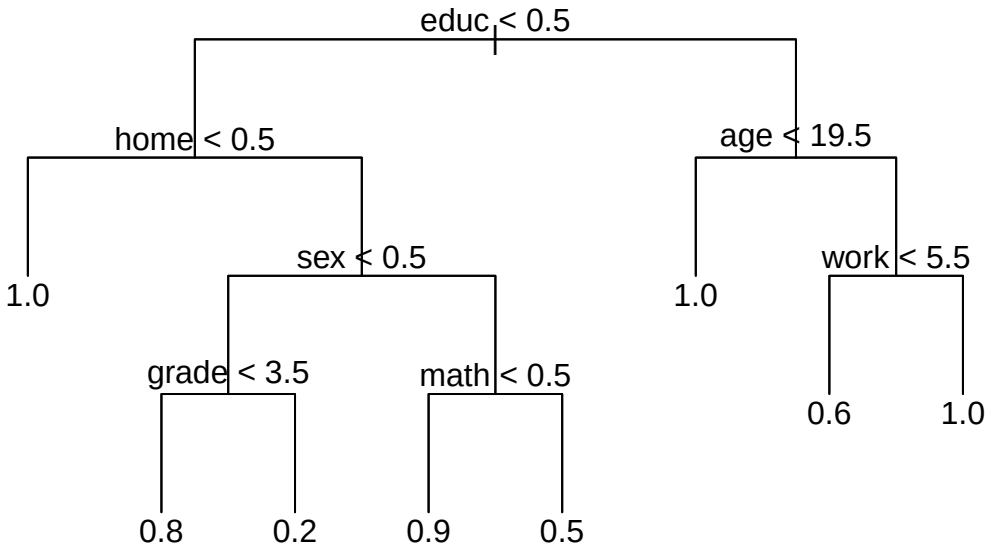
Liking Video Games

In order for the tree to make accurate regressions, our first step was to preprocess the data by changing the one row with an education value of 99 to NA. We established a categorical variable "likesgames" that was given either a 0 or 1 based on what information the student filled out for the "like" response on the survey.

With the resulting variable established, we proceeded by setting the formula for the regression tree as follows:

likesgames educ+sex+age+home+math+work+own+cdrom+grade

The results are shown below:



It is apparent that some of the input variables that were given were not actually used in the formula. The tree can be read in the following manner: The data is grouped based on every split, and following down each branch and it's conditions will yield a value between 0 and 1. The value is between 0 and 1 because our resulting variables are either 0 or 1 (conditional on whether or not a student likes playing games). The numerical value at the end represents the average of individuals who like playing games based on the conditions for the branches that come before it. For example, all individuals in our sample who have an education value of 0 and a home value of 0 like playing video games. Of all individuals in our sample who have education value of 0, home value of 1, sex value of 0, and grade values greater than 3, the average amount of them who like playing video games is 0.2. This result will always yield a value between 0 and 1, with a value less than 0.5 representing more individuals who don't like games.

$$\frac{Number\ of\ Individuals\ in\ Branch\ who\ like\ games}{Total\ Number\ of\ Individuals\ in\ Branch}$$

An individual believing video games are not educational, have a computer at home, are female, and expect an A in the course (value of 0.2). An individual believing video games are not educational, have a computer at home, are male, and hate math (value of 0.5). If we look at the amount of the number of students who fall into the likes and doesn't like categories we find that 69 students enjoy playing video games and 21 students do not enjoy video games. With these along with the regression tree that we established, it appears to be that more students enjoy playing video games.

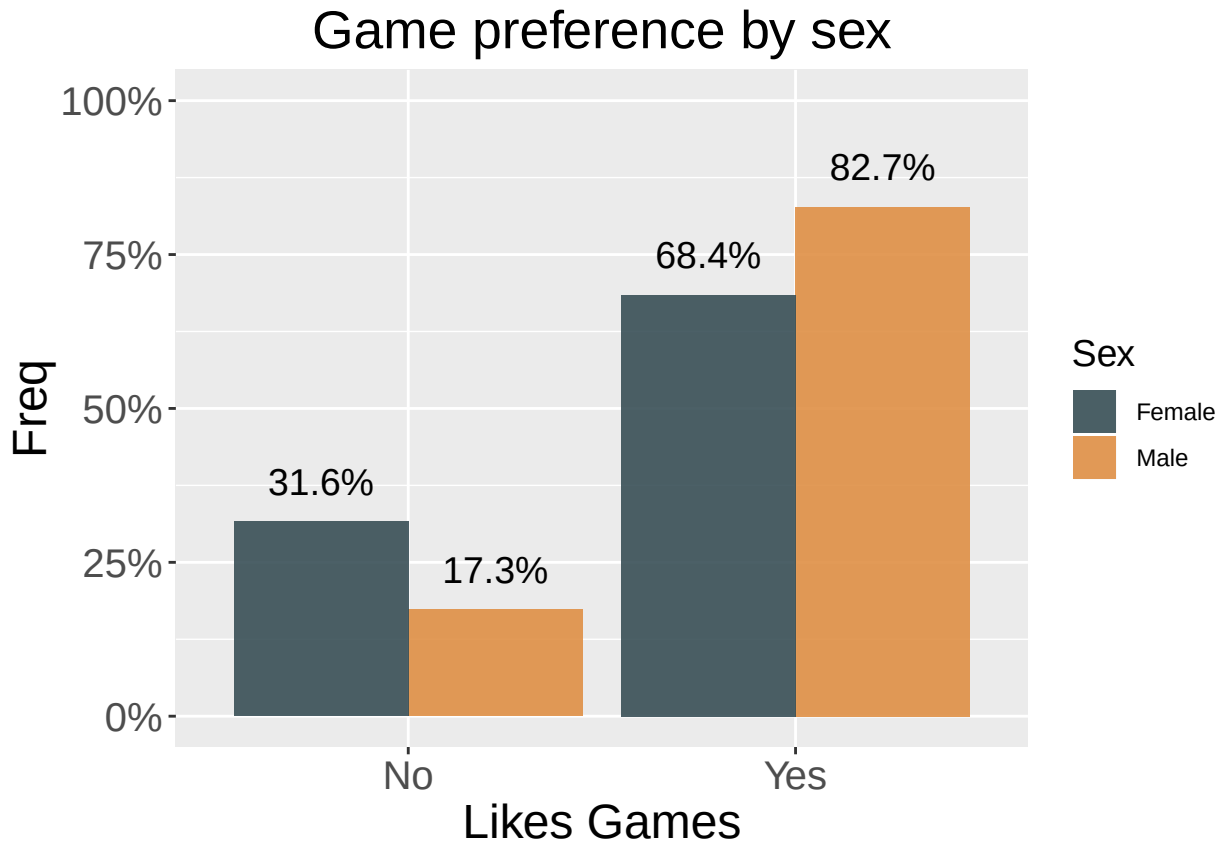
We also ran cross tabulations to analyze the relationship between video game liking against some other binary categorical variables of choice. We grouped people who like games "very much" and "somewhat" into a category of people who "like games" and everyone else into a category of people who "don't like games" for this analysis. Prevalence is calculated with the following formula such that Y represents the person's affinity for games (likes games or does not like games) and X represents the variables that we are tabulating game preference against; 0 and 1 represent the different categories in each variable.

$$Prevalence = \frac{\#(Y = i \cap X = j)}{\#\sum_{j=0}^{1}(Y = j \cap X = i)} \quad for\ i, j = 0, 1$$

First, we wanted to look at how gaming preference splits between males and females. We can see from the frequency plots that there is a greater frequency of males who like games, and a greater frequency of females who do not like games.

Table 4: Sex (n = 90)

	Female	Male
Doesn't like Games	12	9
Likes Games	26	43

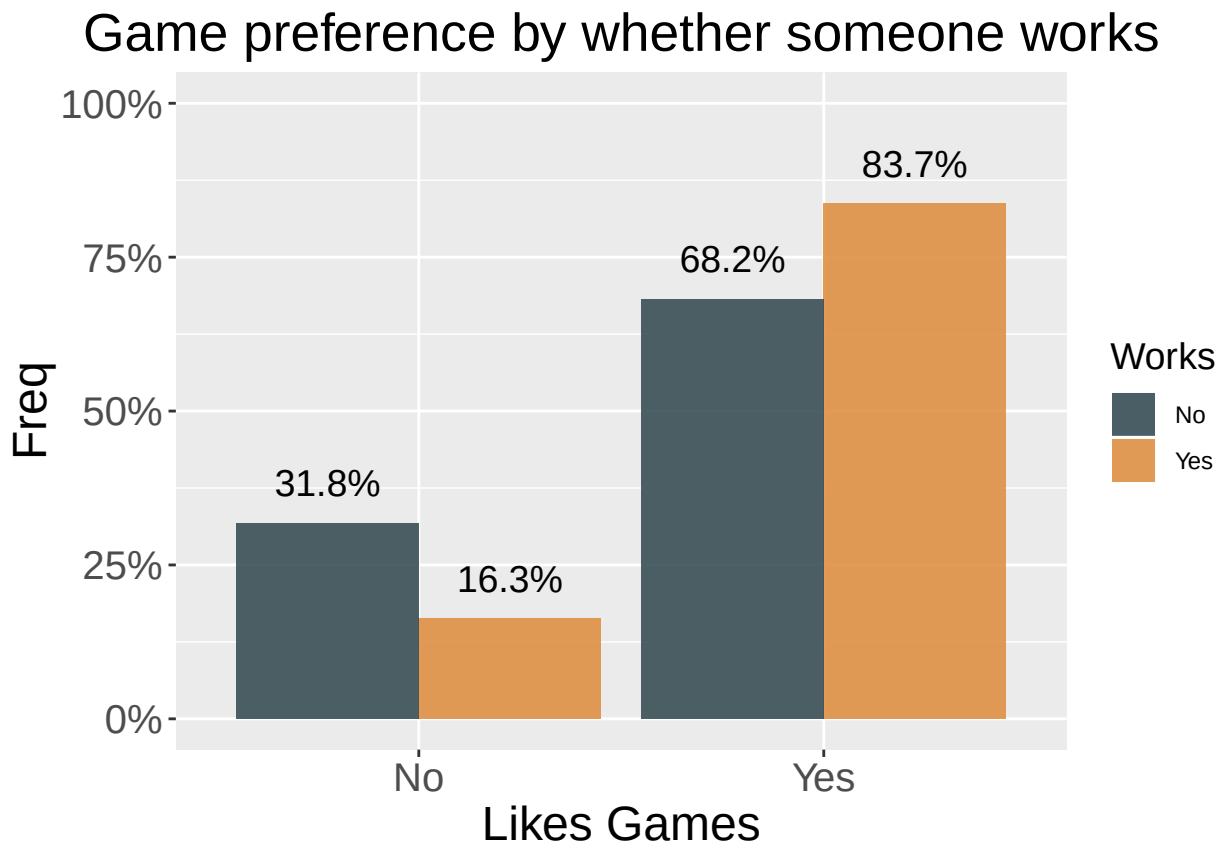


Next, we compare how the gaming preference splits between people who work and people who do not work. Anyone who logged any work hours in this survey were classified as people who work, and otherwise for people who logged 0 hours. The highest prevalence in this cross tabulation seems to be the number of people who work and like games, which is surprising because we expect that people who work might not have a particular affinity for playing games because of them giving their time to work.

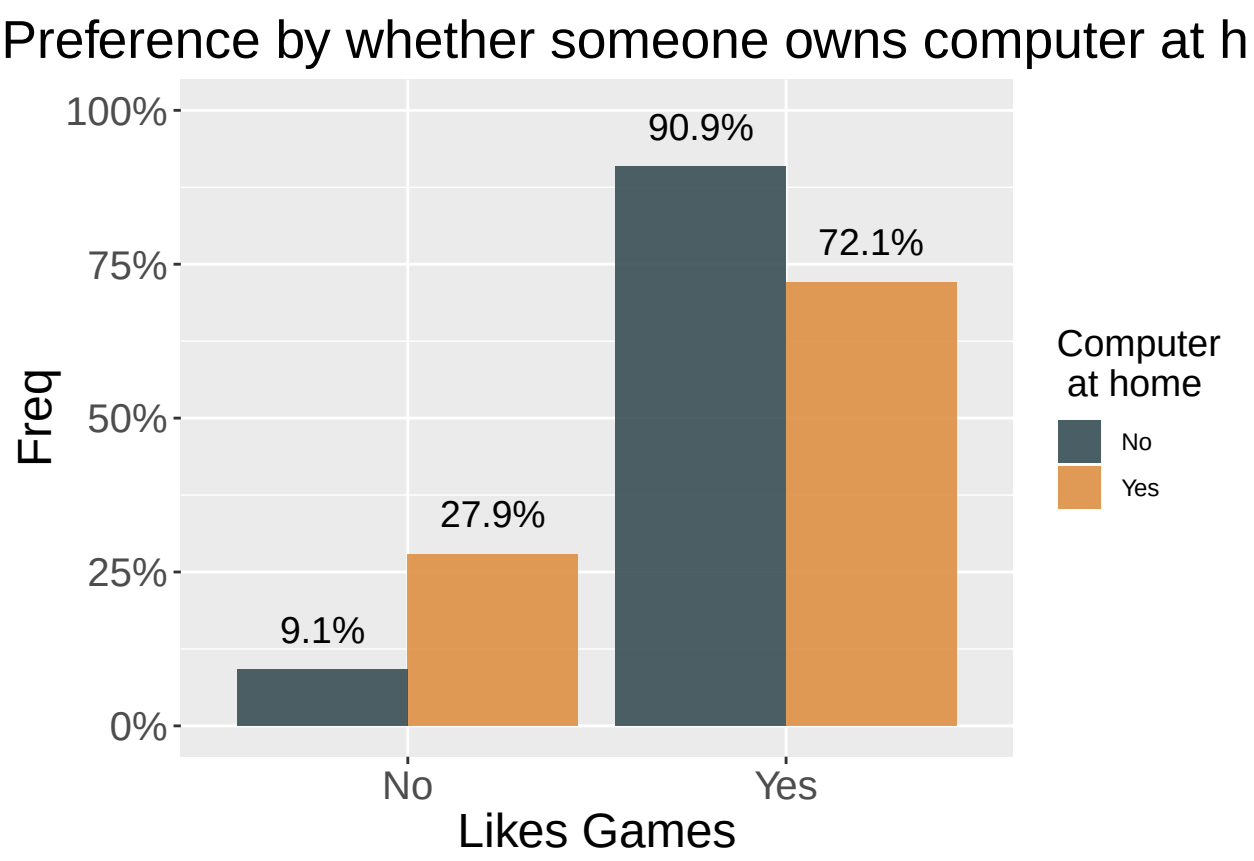
Table 5: Work (n = 87)

	Doesn't Work	Work
Doesn't like Games	14	7
Likes Games	30	36

	Doesn't Work	Work



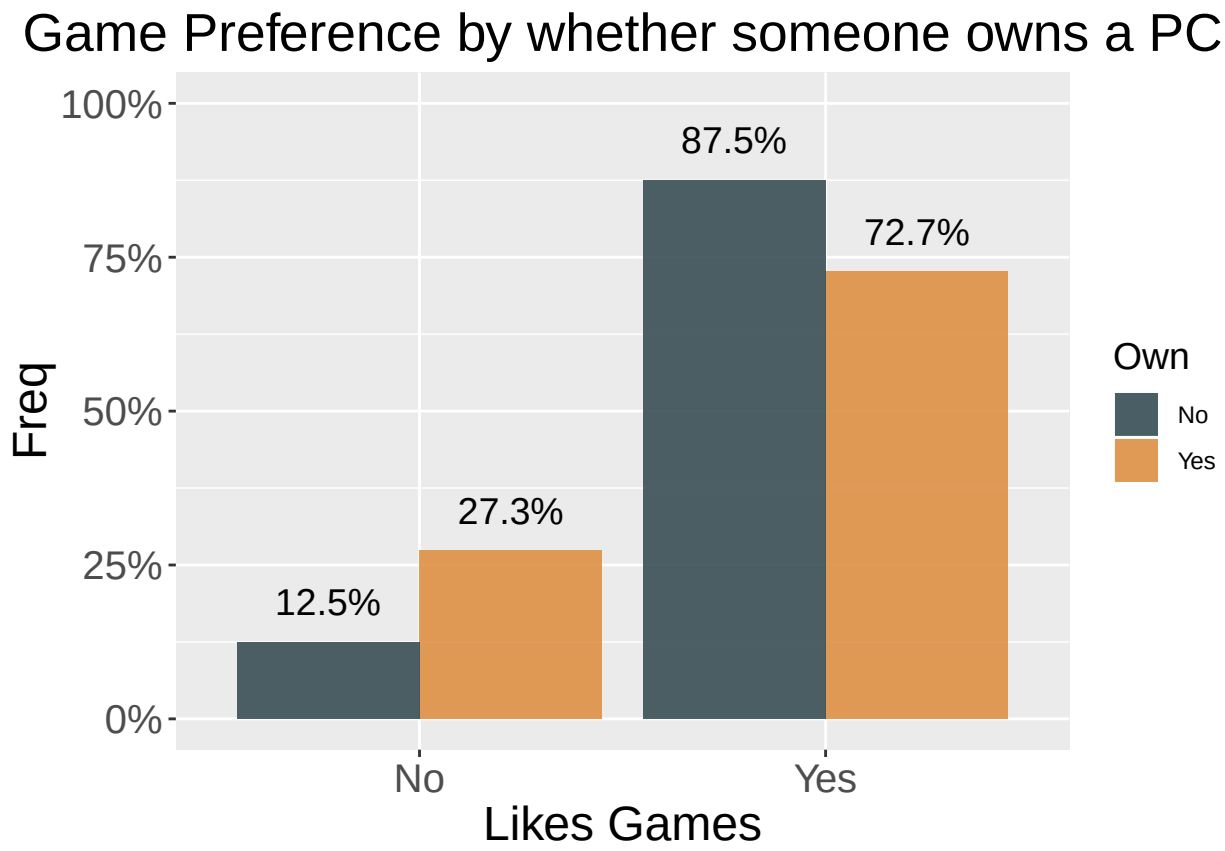
We also wanted to look at how ownership status of a computer at home and ownership of a PC (personal computer) might also have implications of people who like to play games to play games at an Arcade. We can see that the highest prevalence lies in the group of people who like to play games but don't have a computer at home. | | No | Yes | |:——————|:—:|:—:|
|Doesn't like Games| 2| 19| |Likes Games | 20|49| Table: Whether there is a computer at home (n = 90)



However, interestingly enough when we compare gaming preference to whether or not someone has a PC, the highest prevalence lies in the group of people who likes games, but do not own a PC. Because of that, it was in our interest to look at the relationship between whether someone liked games and their preferred location/console for playing games.

Table 6: Owns a PC (n = 66)

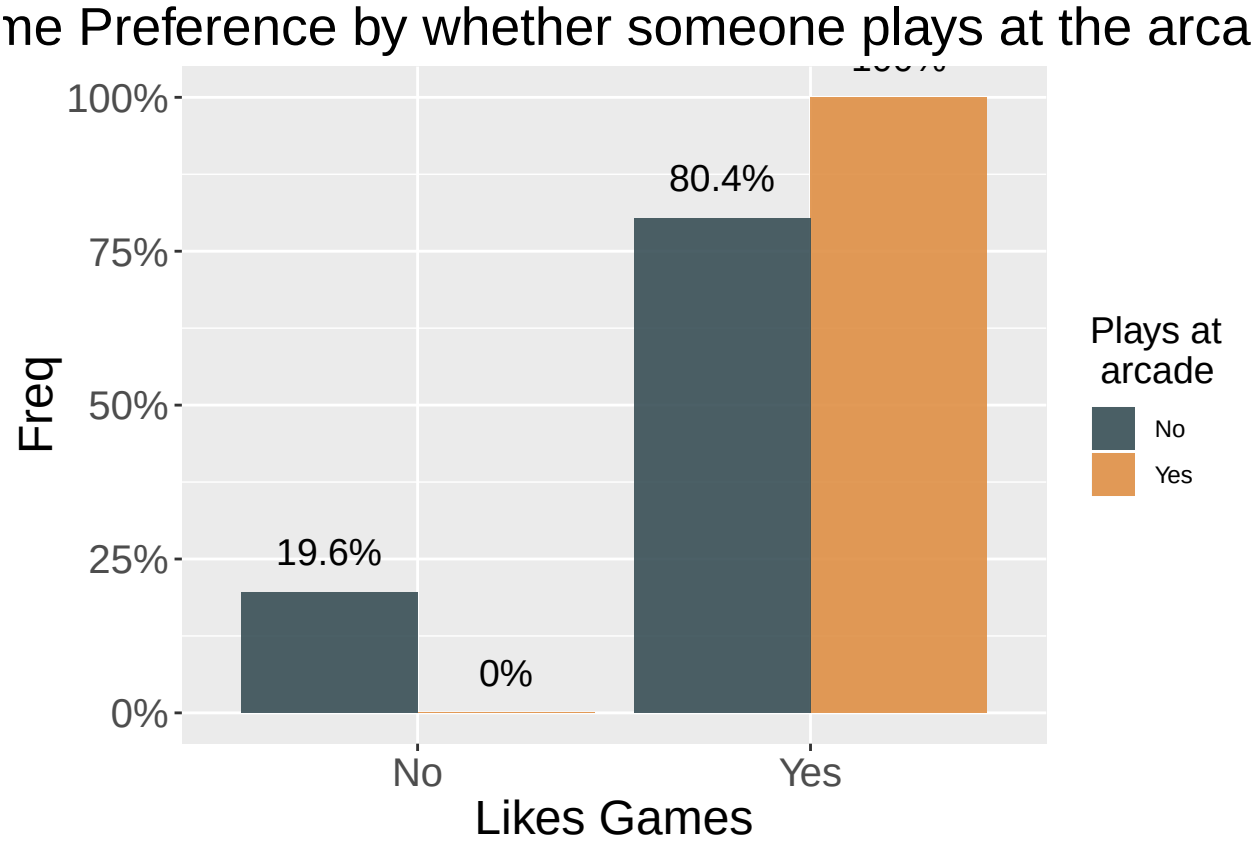
	No	Yes
Doesn't like Games	3	18
Likes Games	21	24



It turns out that the highest prevalence lies in the group of people who like games and go to the arcade. We hypothesized that this could have been a result of the availability of gaming consoles and computer games, as arcades were still a popular source of ways to play games back in the 1990's.

Table 7: Plays at arcade (n = 73)

	No	Yes
Doesn't like Games	10	0
Likes Games	41	22



Expected Grade vs. Targeted Grade Distribution

To investigate whether the grade distribution of students' expectation matches the discrete target distribution, we first use chi-square Goodness of Fit test. The reason that we believe this test is appropriate for this task is that we are comparing an empirical distribution with a target distribution, with which we can divide the population range into bins and compute the corresponding observations and expectations for each bin. The test statistics for Goodness of Fit test is given in eq (1):

$$T = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

In this scenario, we are comparing the empirical grade distribution from the survey to the target distribution Pwhere

$$P(X)$$

is given in eq (2):

$$P(X) = 20\% \ \ (for\ X = A)$$
$$P(X) = 30\% \ \ (for\ X = B)$$
$$P(X) = 40\% \ \ (for\ X = C)$$
$$P(X) = 10\% \ \ (forX = D\ or\ X = F) \ \ (2)$$

We use Goodness of fit test for the following null (H_0) and alternative (H_1) hypothesis:

H_0 = The expected and observed video game time played comes from the same distribution

H_1 = The expected and observed video game time played comes from different distributions

Based on eq (1) $T = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$ (1), where i=4 for A, i = 3 for B, i = 2 for C, i = 1 for D or F.

Assuming the 4 non-respondents expect grades that are random, for the 91 data collected, $E_i > 5$ for i = 1,2,3,4. The test yields $T = 62.608 for *df* = 3 and p-value is 1.629e-13. For a cut off value of 5%, we reject H_0 and conclude that the empirical distribution does not follow P

Suppose the 4 non-respondents all got F, then the new data to be tested against H_0 is the 91 samples amended with 4 Fs. Again, $E_i > 5$ for i = 1,2,3,4, and a goodness of fit test is appropriate. The test yields $T = 53.825$ for df = 3, and p-value is 1.223e-11. For a cut off value of 5%, we reject H~0' and conclude that the empirical distribution does not follow P

We therefore conclude, based on the results of the goodness of fit test, that the distribution of the expected grades does not match the target grade distribution.

Bootstrap

Assuming the 4 non-respondents expect grades that are random, following the sampling method as described in Scenario 1, we draw 91 samples in each iteration. These 91 samples form an expected grade distributions between A, B, C, and D/F, which gives a percentage for A, B, C, D/F respectively. We collect these percentages in corresponding vectors, and repeat this sampling for 100 iterations. Below are the critical quantiles for the empirical sampling distributions.

Grade	Min	2.5%	5%	25%	50%	75%	95%	97.5%	Max
A	0.253	0.264	0.274	0.319	0.341	0.374	0.407	0.418	0.440
B	0.484	0.495	0.505	0.549	0.571	0.604	0.659	0.665	0.670
C	0.033	0.038	0.044	0.055	0.077	0.099	0.110	0.121	0.143
D/F	0	0	0	0	0	0	0	0	0

Based on the above table, it is clear that we conclude that the expected grade distribution is does not match the distribution of the grades to be assigned, with an empirical p-value of 5%.

Assuming the 4 non-respondents expect F grades, we first amend the 91 expected grade samples with 4 Fs. Then following the sampling method as described in Scenario 1, we draw 95 samples in each iteration. These 95 samples form an expected grade distributions between A, B, C, and D/F, which gives a percentage for A, B, C, D/F respectively. We collect these percentages in corresponding vectors, and repeat this sampling for 100 iterations. Below are the critical quantiles for the empirical sampling distributions.

Grade	Min	2.5%	5%	25%	50%	75%	95%	97.5%	Max	Target
A	0.242	0.253	0.263	0.295	0.326	0.358	0.389	0.389	0.411	0.2
B	0.432	0.463	0.474	0.526	0.558	0.579	0.611	0.621	0.632	0.3
C	0.032	0.042	0.053	0.063	0.084	0.095	0.116	0.126	0.189	0.4
D/F	0	0.011	0.011	0.029	0.042	0.053	0.064	0.074	0.084	0.1

Based on the above table, it is clear that we conclude that the expected grade distribution is does not match the distribution of the grades to be assigned, with an empirical p-value of 5%.

Alternative Hypothesis

H_0 = The expected grades that students get are independent from how much they like to play computer games.

H_1 = the expected grades that students get are not independent from how much they like to play computer games.

In order to test for dependence between two variables, we use chi-square independence test. The test statistic is given in eq (3): $T = \sum_{r,c}^{4} \frac{(O_{r,c,} - E_{r,c})^2}{E_{r,c}}$ (1)

For similar reason as Goodness of Fit test explained in Scenario 6, T ??? ??(r-1)(c-1). The grade distribution table is given below:

| Like Computer Games | A | B | C | Total | |——— ————|—|—|—|——-| | Yes| 24 | 38 | 7 | 69 | | No | 7 | 14 | 1 | 22 | | Total | 31 | 52 | 8 | 91|

Combining the B and C bins, the test yields $T=$0.11475 and p-value=0.99. Thus, we do not reject H_0 with a cut off value of 5%. We therefore conclude that the test result is not significant enough to conclude that preferences on playing computer games is not independent from the expected grades.

Regression Tree Theory

A regression tree is used within this paper in order to determine what factors could potentially affected whether a student likes or dislikes video games. Since our data can be split between students who like and don't like video games, we can make our data set carry a categorical variable that represents this binary. The general version of a decision tree is based off predicting some response or class Y based on given inputs $X_1..X_n$. Each node of the tree is represented by a binary split between an input Xi. The process is continued until a leaf node is reached where an evaluation about the dataset can be made.

With the "tree" package that was used in this report, the response variable and input variables are determined based on the formula passed into the parameters. NuWmerical variables are split such that $X > a$ and $X < a$ represent the two groups, whereas the levels of factors are split into two non-empty groups. The goal of decision trees is to optimize the nodes and splits such that impurity is minimized. Impurity is measured in a manner such that the end nodes are as homogenous as possible (higher proportion of similar classes per each region that is created by the splits). A more concrete explanation of impurity can be seen in the results section.

Logistic Regression

To investigate the reasons, i.e. most significant factors, that attract students to play a type of game, we ran a logistic regression on the attributes of video games against students' preferences on the type of video game. Following is a table of the coefficients corresponding to the types of video games and the video game attributes.

Bootstrap Theory

Since n isn't too big and n/N isn't small we can't be sure that the data will be normally distributed by the law of large numbers. So we use bootstrap to confirm this. First, we need to increase sample size as 91 is too small and might cause the data to be dependent if we create a sample because we are sampling with replacement. To avoid this, we would use bootstrap population. To increase size to 314, we multiply out count by 3.45(= 314/91). Then we sample by taking 91 data points from bootstrap population without replacement. Then, I continued on using t-test statistics to get a confidence interval.

To interpret this table, we focus on the most significant attributes for each type of video games. We see that the most significant factor that attracts students to action games is their graphics, the most significant factor that attracts students to simulation games is their hand-eye coordination. Graphics draw students to sports games, and relaxation draws students to strategy games. For both sports and strategy games, students' preferences inversely relate the hand-eye coordination aspect of the games.

Conclusion

In conclusion, if we were to apply the results of this study into designing a lab which specifically incorporates elements of video games to engage students and promote more active learning we expect this type of lab to benefit around 37% of students based on the proportion of students that played video games in the week prior

Table 10:

	Dependent variable:				
	action	adv	sim	sport	strategy
	(1)	(2)	(3)	(4)	(5)
relax	0.470	0.916	−0.040	−0.161	1.684**
	(0.638)	(0.731)	(0.768)	(0.634)	(0.761)
coord	−0.064	−17.411	1.631	−1.478	−2.173*
	(1.295)	(1,935.940)	(1.130)	(1.281)	(1.280)
master	0.661	−0.056	−0.005	0.496	0.203
	(0.582)	(0.578)	(0.649)	(0.549)	(0.667)
bored	−0.124	0.608	0.507	0.609	0.669
	(0.591)	(0.592)	(0.670)	(0.583)	(0.757)
graphic	1.209*	0.858	−0.031	1.395**	−1.359*
	(0.667)	(0.596)	(0.696)	(0.611)	(0.708)
Constant	−0.273	−1.714**	−1.566**	−0.754	0.285
	(0.648)	(0.766)	(0.785)	(0.653)	(0.690)
Observations	66	66	66	66	66
Log Likelihood	−39.580	−38.058	−32.940	−41.836	−31.199
Akaike Inf. Crit.	91.161	88.117	77.880	95.671	74.398
Note:					*p<0.1; **p<0.05; ***p<0.01

to the survey. Based on our interval estimate for the amount of time students spent playing video games we believe that the ideal length of time for this lab to be around an hour. It is worth noting that the fact there was a test the week prior to the survey had some effect on the length of time people played so that it is not consistent with their reported frequency of play. So, we believe further investigation into the amount of time people spend playing will may increase the amount of hours people play in a week. Also, we found that access to computers, sex, attitude toward video games, and affinity to math are all factors that effect whether or not people enjoy video games. It is important to note that incorporating too many aspects of video games may not engage everyone equally as we found earlier that only 37% may enjoy video games and the rest not so much. Furthermore, we concluded that graphics are the most engaging part of video games and should be the focus of a lab built to incorporate gaming and the lab should try its best to stay away from requiring too much hand eye coordination.