Case Study 2: Who Plays Video Games

April 27, 2018

Chengyu Chen (A14051607), 2nd Year Applied Mathematics, MATH 189
Chenyue Fang (A13686794), 2nd Year Probability and Statistics, MATH 189
Daniel Lee (A13726312), 2nd Year Probability and Statistics, MATH 189
Xinran Wang (A13564644), 2nd Year Probability and Statistics, MATH 189
Yuqi Wang (A13532155), 2nd Year Applied Mathematics, MATH 189
Ning Xu (A92061610), 3rd Year Probability and Statistics; Economics, MATH 189

Introduction

Every year, three to four thousand students enroll in statistics courses at UC Berkeley, half of which take introductory statistics courses to satisfy their quantitative reasoning requirement. To aid the instruction of these students, an innovative committee of faculty and students designed a series of computer labs, with the goal of extending traditional methods of teaching statistics by providing an interactive learning environment. Because the characteristics of labs have been linked to parallel characteristics in video games, the committee conducted an survey to better understand which aspects of video games students find the most and least fun. Students enrolled in advanced statistics courses developed the study and selected the students to be sampled.

This paper presents an analysis of the sample survey data with advice and proposals to the design committee to best design labs in statistics courses at UC Berkeley.

Since this study revolves around a survey, throughout this paper, special attention will be given to survey methodology and survey sampling, a broad field of statistics. A survey follows scientific methodology to collect data from individuals, typically sampled from a large population, in hopes of describing, exploring, and/or explaining characteristics of the population as a whole. Good survey research must be quantitative, careful, replicable, impartial, representative (of the population), and theory-based. In general, individuals must have an equal chance of being selected in the sample and the sample must be generalizable to the total population of interest. Survey sampling must be used only when appropriate. For example, if all members of a population were identical, survey sampling would be redundant.

Data

Out of 314 students in a lower division UC Berkeley statistics course in Fall 1994, 95 were selected to participate in the study's survey via a simple random sample. The survey was

anonymous in order to minimize response bias and for student's privacy. Furthermore, there was minimal non-response (hence minimal non-response bias) in the survey, with 91 completing the survey out of the selected 95 students. Non-response was minimized via a three stage system of data collection. First, data collectors visited the discussion on both Tuesday and Thursday sections. Second, examinations were returned to students on the week of the survey. Since a large proportion of the students come to collect their examinations, researchers hoped this would minimize non-response bias. Third, students who were not able to be reached during sections were given the survey during lecture (Nolan et al.).

Data from these 91 students is used throughout this paper's analysis. The following data variables regarding students' video game habits were collected:

Table 1. Description of Survey Variables

Variable	Data Type	Description
time	Numerical	Number of hours student played in the week prior to survey
like	Categorical (Ordinal)	Measure of student's sentiment toward playing 1 = never played 2 = very much 3 = somewhat 4 = not really 5 = not at all
where	Categorical	Where student plays 1 = arcade 2 = home system 3 = home computer 4 = arcade and either home computer or system 5 = home computer and system 6 = all three
freq	Categorical (Ordinal)	How often student plays 1 = daily 2 = weekly 3 = monthly 4 = semesterly
busy	Categorical	Whether student plays when busy 1 = yes, 0 = no
educ	Categorical	Whether student believes video games are educational 1 = yes, 0 = no
sex	Categorical	1 = male, 0 = female
age	Numerical	Student's age (years)
home	Categorical	Whether student has computer at home 1 = yes, 0 = no
math	Categorical	Whether student hates math 1 = yes, 0 = no

work	Numerical	Number of hours student worked the week prior to survey
own	Categorical	Whether student owns PC 1 = yes, 0 = no
cdrom	Categorical	Whether student's PC has CD-Rom 1 = yes, 0 = no
email	Categorical	Whether student has email 1 = yes, 0 = no
grade	Categorical (Ordinal)	Grade expected by student 4 = A 3 = B 2 = C 1 = D 0 = F

Since the survey sample size is relatively small and comes from a single section of a single course during one particular term, note that each individual's data may have a small degree of dependence and correlation with other individuals' data. This complication may be mitigated with a larger survey sample. However, for the purposes of this analysis, we assume that each individual's data is independent and identically distributed.

If a question was not answered or improperly answered by the student, the data value for that specific variable is coded as 99. Respondents who either never played a video game or did not like video games (at all) were asked to skip many of the questions.

In addition to the original survey above, a second part of the survey was conducted to determine whether the student likes or dislikes playing games and why. These questions are different from the original survey in that more than one response may be given. The following tables summarize the results of the additional survey questions:

Table 2: "What types of games do you play? (at most three answers)"

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

Table 3: "Why do you play the games you checked above? (at most three answers)"

Why?	Percent
Graphics/Realism	26%
Relaxation	66%
Eye/Hand Coordination	5%
Mental Challenge	24%
Feeling of Mastery	28%
Bored	27%

Table 4: "What don't you like about video game playing? (at most three answers)"

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%

Background

Prior to analysis of the dataset, literature review was conducted on other survey studies and methodologies.

In "Physical Activity, TV viewing, and Weight in U.S. Youth: 1999 Youth Risk Behavior Survey" by Eisenmann et al., researchers used survey data to determine the relationship between TV watching, physical activity, and weight in fourteen to eighteen year old teenagers in the United States. In order to give a representative sample, three-stage cluster sampling was used in the survey (countries, schools and classes). Students from randomly selected classes in sampled

schools recorded their responses on answer sheets to complete the survey. A total of 15,349 questionnaires were received from 144 schools with a student response rate of 86%. Three variables including TV watching (TV), moderate physical activity (MPA) and vigorous physical activity (VPA) were measured by days attending the specific activity in the last seven days. MPA and VPA were distinguished by whether the physical activity made students sweat or breathe heavily. BMI was calculated from the height and weight of students. Analysis of covariance was used to compare BMI across TV, MPA and VPA. The relationship between TV watching, physical activity and weight statues was measured by logistic regression models with odds ratio and 95% confidence level. The results showed that 44.6% and 64.7% of the students engaged in MPA and VPA three or more days in the last week with a higher rate in boys than in girls. 24.7% of the students watched more than four hours per day. The least-squares mean BMI showed a significant difference in BMI for each level of TV watching in both boys and girls. Furthermore, the regression analysis showed a significant relationship between decreased level of MPA and VPA and a higher risk of overweight status in both boys and girls (Eisenmann et al.).

In "New York State Case Manager Survey: Urban and Rural Differences in Job Activities, Job Stress, and Job Satisfaction" by Gellis et al., 421 samples were randomly selected from the New York States Mental Health Case Management and Coalition membership list, and 42% of them submitted the survey by mail. These participants were classified into two groups based on their work area: urban or rural. 30 variables were used to measure job stress from four aspects: stress frequency, stress intensity, job pressure and inadequate support. Each item was rated by participants on an ordinal nine point scale. Varimax rotation was used to relate items to job stressors. Cross tabulation, frequency distribution, confidence intervals, and Pearson correlation coefficients were used to examine group differences. The result yielded that total stress frequency was significantly positively related to total stress intensity with $r = 0.44$ and $P < 0.001$. The small and inverse relationship between age and lack of support was supported with $r = -0.20$ and $P < 0.001$. Female employers had higher overall scores the survey than male employers. Employers in urban areas had higher scores in the inadequate support and stress frequency sections than their rural counterparts did with $r = 0.25$ and $P < 0.001$, and $r = 0.25$ and $P < 0.001$ respectively (Gellis et al.).

In order to better understand the dataset during analysis, literature review was also conducted on studies specifically related to video games.

In "The Development of Attention Skills in Action Video Game Players" by Green, the Attentional Network Test measured three major components of visual attention: alerting (utility of cues to allocate a target), orienting (the ability to pay attention to imminent stimulus) and executive control (the ability to filter out distractors) across a control group and treatment group (who played action games) consisting of 7 to 22 year-olds. The study concluded that playing

video games caused the treatment group to have better visual attention and faster and more accurate responses to stimuli compared to the control group (Green).

In "Effect of Playing a Video Game on a Measure of Spatial Visualization" by Dorval, 70 students were randomly assigned into a control and a treatment group. The treatment consisted of eight sessions of video game playing. The study concluded that playing video games caused the treatment group to have significant improvements in spatial visualization test scores compared to the control group (Dorval).

In "Differences in Eye-hand Motor Coordination of Video-game Users and Non-users" by Griffith, 62 subjects were equally randomly assigned into a control group and a treatment group (who played video games). The study concluded that playing video games caused the treatment group to have better eye-hand motor coordination compared to the control group (Griffith).

Therefore, studies have shown that playing video games may improve attention, spatial visualization, and eye-hand motor coordination when compared to not playing video games.

Investigation

Scenario 1.

Point estimate is a single value predicting a parameter of a population. In this case, the fraction of students who played a video game in the week prior to the survey from the survey data is used as the point estimate for the population in the UC Berkeley's statistics class. A confidence interval is a defined range of values with specified probability containing the population estimate.

1). Simple confidence interval via CLT:

The sample size is 314, which is greater than thirty. By rule of Thumb of the Central Limit Theorem, the sampling distribution of the sampling means approaches normal when sample size gets larger. Though the given survey data are not independent and identically distributed random variables, the estimate generated by CLT can still provide a perspective of how the population data is distributed. By appending a new column called "indicator" containing the status of whether people played video games in the week prior to survey (1 indicates yes, and 0 indicates no), we made an estimate on the proportion of people played video games and built a confidence interval around the estimate.

Point Estimate: $\overline{X} = 0.374$

Formula: $(\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$
Confidence Interval: (0.274, 0.473)

By applying Central Limit Theorem, we estimate that 37.4% of students played video game in the prior week, and the the confidence interval is (0.274, 0.473), reflecting that the probability that fraction of students who played video game in the last week falls in this interval is 0.95.

2). Confidence interval via CLT with finite sample population correction:

However, central limit theorem is based off of the assumption that the sample size goes to infinity. Since the survey data collected is not large enough, with the same point estimate, a finite sample corrector is used to adjust the existing confidence interval.

Point Estimate: $\overline{X} = 0.374$
Formula: $(\overline{X} - Z_{\alpha/2}\frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{n-1}}\sqrt{\frac{N-n}{N}}, \overline{X} + Z_{\alpha/2}\frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{n-1}}\sqrt{\frac{N-n}{N}})$
Corrector: 0.843
Standard Error: 0.043
Confidence Interval: (0.289, 0.458)

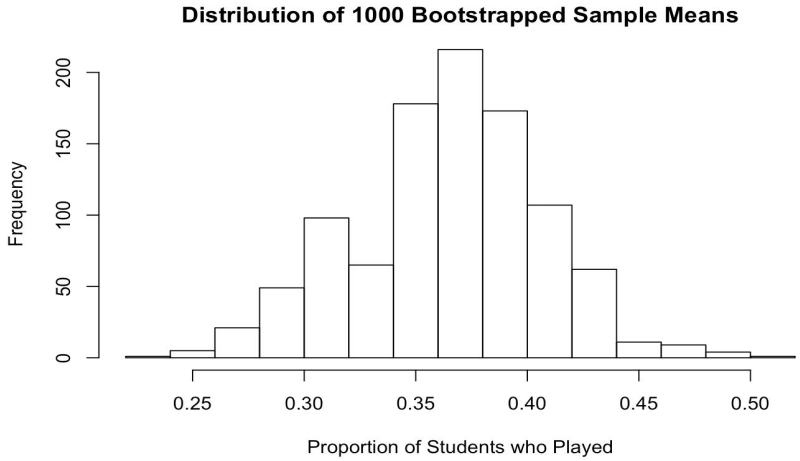
Using Finite Sample Corrector, the output reflects that the probability of the fraction of students who played a video game in the week prior to the survey lies within 0.289 and 0.458 is 0.95.

3) Confidence interval via bootstrap

In the case when sample size is finite, bootstrap resampling method can also be used to make the estimate and to construct the confidence interval. After checking the mean of each set of simulated data being approximately normally distributed, we computed the expected value of the bootstrap population's mean and set that as the estimate for the population. Then a confidence interval is constructed by using the values in the specified 2.5% and 97.5% percentile of the mean value.

Point Estimate: $\overline{X} = 0.365$
Confidence Interval: (0.275, 0.440)

Figure 1: Distribution of 1000 Bootstrapped Sample Means on proportion of student video game players



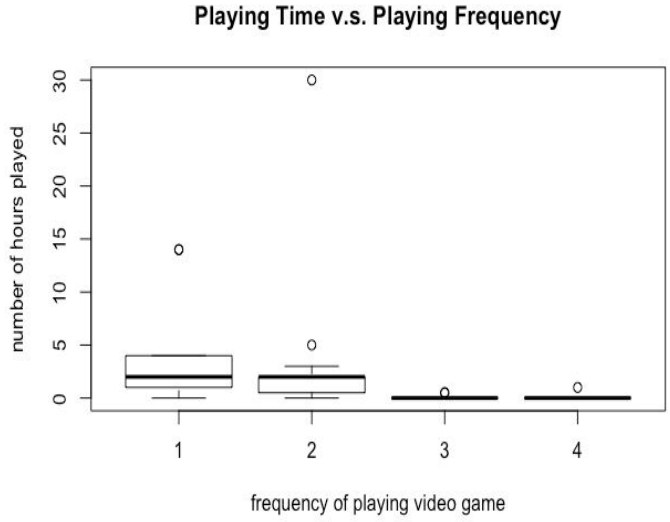
Bootstrap sampling method produced an estimated fraction of students who played video game in the prior week to be 36.5%, which is slightly lower than the previous estimate. The probability that the fraction of students who played video game in the prior week lies between 0.275 and 0.440 is 0.95.

Confidence Interval Comparison and Discussion:

The confidence interval generated by Central Limit Theorem is the widest, then comes the one with finite sample population correction and bootstrap. Therefore, the confidence interval computed by bootstrap resampling method is the most precise and accurate among all three.

Scenario 2.

Figure 2: Boxplot: Playing Time v.s. Playing Frequency



The four boxplots above compare and visualize the amount of time students spent on playing games prior to the survey based on four different playing frequencies: daily(1), weekly(2), monthly(3), and semesterly(4). The median of the first two groups are higher than the latter two, which matches the intuition that students who play video games frequently tend to spend longer time than those who play less frequently.

By comparing the first two boxplots, it is evident that the first boxplot has a higher third quartile value than that of the second, which indicates that more than 75% of people playing video games daily tend to spend more time playing than those who played weekly. Therefore, a positive association between the frequency of playing games and the amount of time spending on video games playing prior to the survey is claimed.

Table 5: Average of time spent on video games based on the frequency of playing

Frequency of playing video games	Average of time spent (hr)
1	4.44
2	2.54
3	0.056
4	0.04

Table 6: Association between hours played and busy

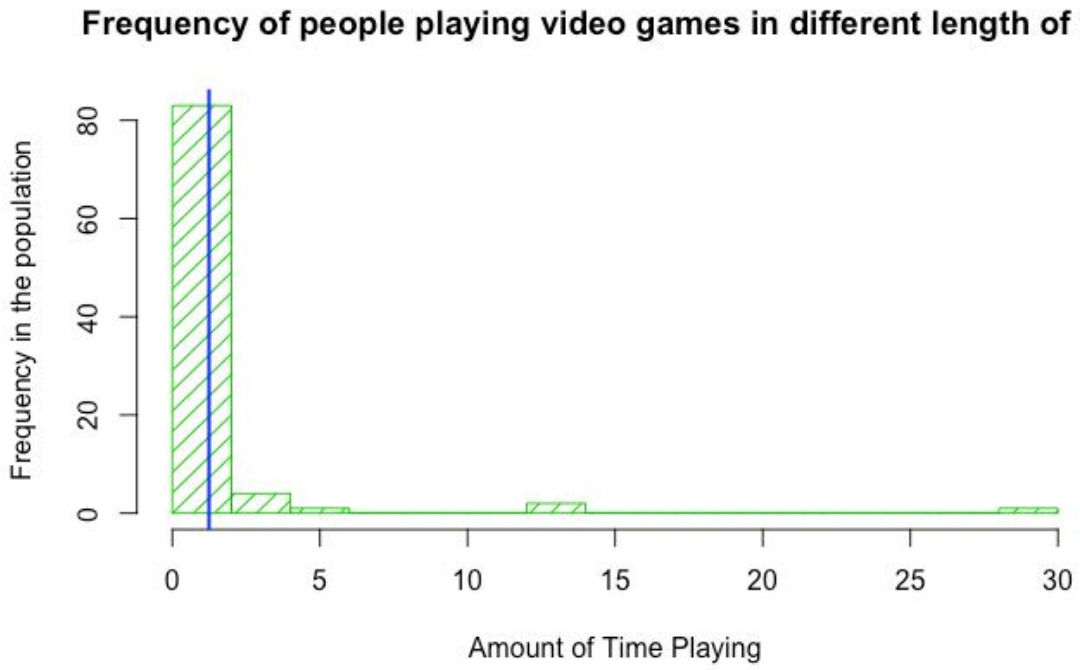
Frequency Group	# Hours Played by Students who Play Even when Busy	# Students played even busy	# Hours Played by Students who do not Play when Busy	# Students do not play when busy	Proportion of students play when busy
Daily	7.2	5	1	4	0.56
Weekly	4	11	1.594	17	0.39
Monthly	0	1	0.0588	17	0.06
Semesterly	NA	0	0	22	0

Table 5 lists the average number of hours students spent on playing video games for four groups of students who reported to play in different frequencies. Among all, the group of students who play video games daily has the highest average time of playing time.

In conclusion, for those who play video games while busy, the average time spent playing video games is 4.705882 hours, and for those who do not play while busy, the average time spent playing video games is 0.5095238 hour. In order to investigate how would having an exam in the week prior to the survey affect the previous estimates, we looked into the average number of hours students spend on playing when they were busy and the fraction of students who would play when they were busy. We discovered that the proportion of students playing even when busy is higher for people who played video games more frequently. Also, according to the data, students who play daily or weekly are the most affected by the presence of an exam, while those who play monthly or semesterly are not, as there is a high chance they were not initially planning to play in the prior week. Therefore, we claim that there exists a positive correlation between the frequency of playing, and the level of business the students encounter. Thus, the previous estimates could be higher than the true mean data if given the fact that an exam was given the week prior to the survey. In other words, the presence of an exam prior to the survey may have skewed the data, and special attention must be given to this fact.

Scenario 3

Figure 3: Histogram: Frequency of people playing video games in different length of time



The histogram above reflects the amount of time students usually spend on playing video games. X-axis of the histogram reflects the amount of time and the Y-axis represents the number of times the pattern of people spending certain that is discovered. From the graph, most people

spend between 0 and 2 hours playing video games, and the mean time of playing video games lies in this interval.

Confidence Interval:

1) Central Limit Theorem

Point Estimate: $\bar{x} = 1.243$

Confidence Interval: (0.467, 2.019)

2) Finite Sample Corrector:

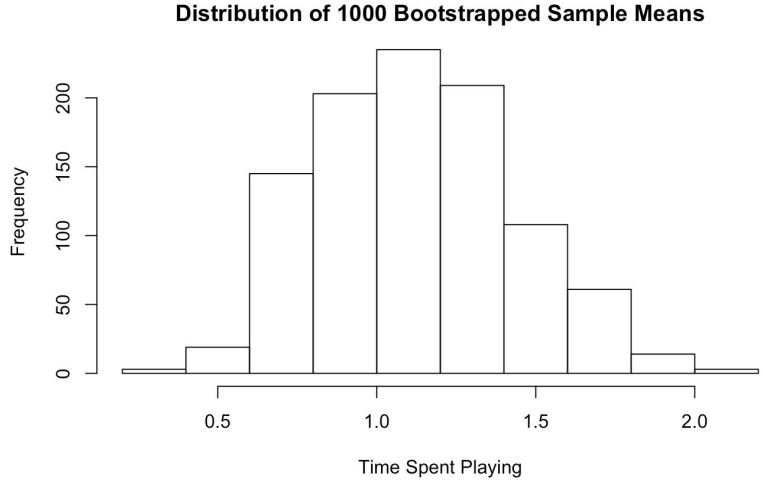
Point Estimate: $\bar{x} = 1.243$

Confidence Interval: (0.589, 1.897)

3) Bootstrap

Simulation:

Figure 4: Distribution of 1000 Bootstrapped Sample Means



The graph above reveals that the distribution of mean time playing video games is approximately normal with mean 1.14039. The density in the population gets larger as getting closer to the mean.

Point Estimate: $E[\overline{X}] = 1.124$ hr played prior to the week the survey conducted

Confidence Interval: (0.605, 1.780)

Figure 5: Distribution of 1000 Bootstrapped Sample Skewness

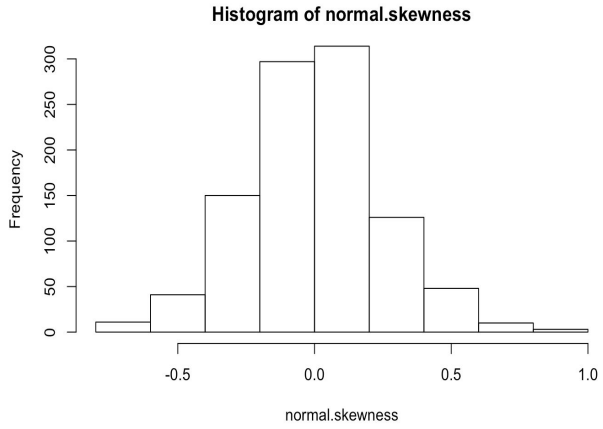
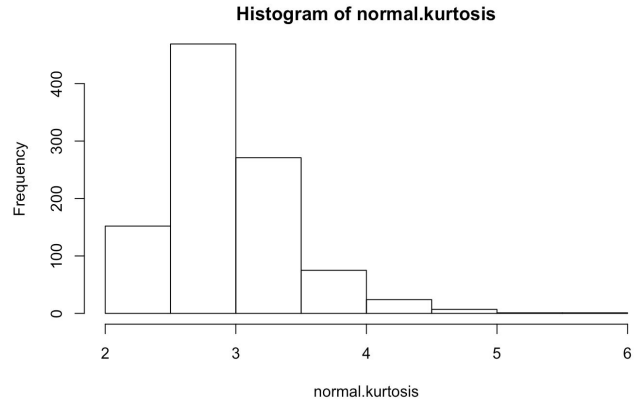


Figure 6: Distribution of 1000 Bootstrapped Sample Kurtosis



Skewness: 0.308
Kurtosis: 2.632

Since the skewness and kurtosis are similar to those of the Monte Carlo simulated Normal distribution, an interval estimate is appropriate. This also suggests that the Central Limit Theorem holds.

We use point estimator to estimate the parameter of the population. Here, the point estimator is the estimate of average amount of time playing video game in the week prior to the survey.

In Finite Sample Corrector, we estimate that the average amount of time playing video game in the prior week is 1.24285714285714 hour, and the probability that the true average

amount of time playing video game last week falls between 0.588861920402482 hour and 1.8968523653118 hour is 0.95.

Using Central Limit Theorem, we get the estimated average amount of time playing video game in the last week is also 1.24285714285714 hour, and there is 95% probability that true average amount of time playing video game in the prior week falls in the interval (0.450974374698314 hour , 2.03473991101597 hour).

Our Bootstrap simulation results that estimated average amount of time playing video game in the last week is 1.14039 hour, and there is 95% probability that true average amount of time playing video game in the prior week falls in the interval (0.6317308hour, 1.8156319hour).

Scenario 4.

"Why do you play the games you checked above? (at most three answers)"

Why?	Percent
Graphics/Realism	26%
Relaxation	66%
Eye/Hand Coordination	5%
Mental Challenge	24%
Feeling of Mastery	28%
Bored	27%

"What don't you like about video game playing? (at most three answers)"

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%

In general, students enjoy playing video games. From the data of additional questions that asked for the reasons why students play the games, 66% of them answered for relaxation, which implies that students enjoy playing game because doing so can relax themselves. On the second table, 48% think that they spend too much time on playing video games, which means that they dislike playing mostly because of the results of playing too much, and it further implies that students actually like playing games that they cannot stop. Therefore, from the data of additional questions, it can be concluded that students like playing video games in general. Moreover, since the main task of going to University is to study, some students may think that playing video games are useless, since it won't boost students' grades.

List of important reasons why people like to play video games:

1. Playing video games help students relax from the busy workload.
2. Playing video games gives students feelings of mastery.
3. Playing video games lets students enjoy visual realization.

List of important reasons why people don't like to play video games:

1. Playing video games take too much time
2. Playing video games cost too much money
3. Playing video games are meaningless

The reasons are listed because from the data of additional questions, the most possible reasons students like/dislike playing video games are the ones that contain the most percentages, so the three most possible reasons students like/dislike playing games are listed above.

Scenario 5.

Look for the differences between those who like to play video games and those who don't. To do this, use the questions in the last part of the survey, and make comparisons between male and female students, those who work for pay and those who don't, those who own a computer and those who don't. Graphical display and cross-tabulations are particularly helpful in making these kinds of comparisons. Also, you may want to collapse the range of responses to a question down to two or three possibilities before making these comparisons.

#1 Like vs Gender

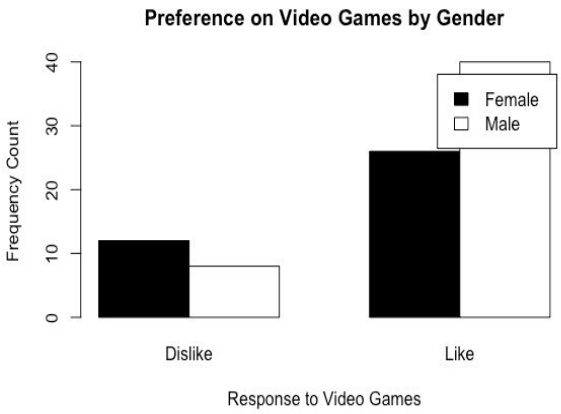
Remark:

Like: 0: people responded 4 and 5 in the survey 1: people responded 2 and 3 in the survey

People responded 1 in the survey indicates they never played video games, which is cleaned out of the calculation.
Gender: 0: Female 1: Male

Total Observations in Table: 86

video.like	video.w$sex: Female	Male	Row Total
Dislike	12	8	20
	1.132	0.896	
	0.600	0.400	0.233
	0.316	0.167	
	0.140	0.093	
Like	26	40	66
	0.343	0.272	
	0.394	0.606	0.767
	0.684	0.833	
	0.302	0.465	
Column Total	38	48	86
	0.442	0.558	



Pearson's Chi-squared test with Yates' continuity correction
data: table(video.clean$like, video.clean$sex)
X-squared = 2.3106, df = 1, p-value = 0.1285
The null hypothesis is gender is independent of responses in the survey, and the alternative hypothesis is otherwise. Since p-value is 0.1285>0.05, under 0.05 significance level we fail to reject null hypothesis (no enough evidence to show like is associated with gender).

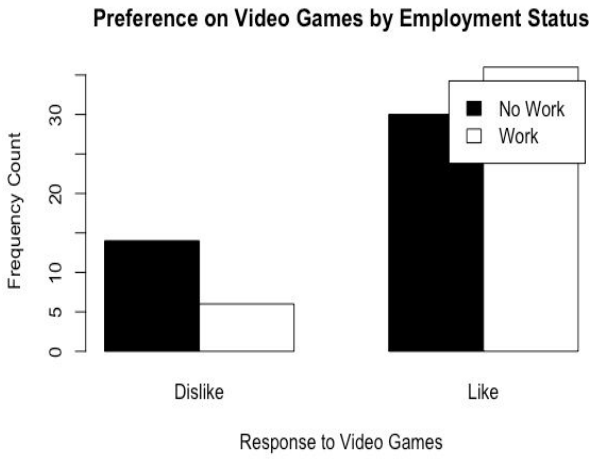
#2 Like vs Work

Remark:

Like: 0: people responded 4 and 5 in the survey 1: people responded 2 and 3 in the survey People responded 1 in the survey indicates they never played video games, which is cleaned out of the calculation.
Work: 0: Doesn't work 1: Work

```
Total Observations in Table:  86

             | video.w$work
video.w$like |    No Work |       Work | Row Total |
-------------|------------|------------|-----------|
     Dislike |         14 |          6 |        20 |
             |      1.387 |      1.453 |           |
             |      0.700 |      0.300 |     0.233 |
             |      0.318 |      0.143 |           |
             |      0.163 |      0.070 |           |
-------------|------------|------------|-----------|
        Like |         30 |         36 |        66 |
             |      0.420 |      0.440 |           |
             |      0.455 |      0.545 |     0.767 |
             |      0.682 |      0.857 |           |
             |      0.349 |      0.419 |           |
-------------|------------|------------|-----------|
Column Total |         44 |         42 |        86 |
             |      0.512 |      0.488 |           |
-------------|------------|------------|-----------|
```



Pearson's Chi-squared test with Yates' continuity correction

data: table(video.w$like, video.w$if_work)
X-squared = 2.7838, df = 1, p-value = 0.09522

The null hypothesis is employment status is independent of responses in the survey, and the alternative hypothesis is otherwise. Since p-value is 0.09522>0.05, under 0.05 significance level we fail to reject null hypothesis (no enough evidence to show like is associated with employment status)

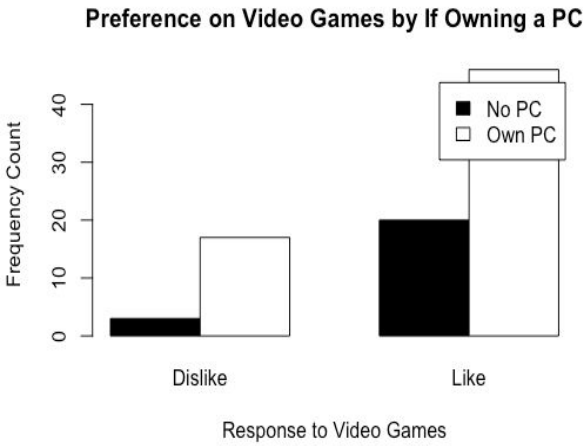
#3 Like vs Own

Remark:

Like: 0: people responded 4 and 5 in the survey 1: people responded 2 and 3 in the survey People responded 1 in the survey indicates they never played video games, which is cleaned out of the calculation.
Own: 0: Doesn't own a computer 1: Owns a computer

```
Total Observations in Table:  86

             | video.w$own
video.w$like |     No PC |    Own PC | Row Total |
-------------|-----------|-----------|-----------|
     Dislike |         3 |        17 |        20 |
             |     1.031 |     0.377 |           |
             |     0.150 |     0.850 |     0.233 |
             |     0.130 |     0.270 |           |
             |     0.035 |     0.198 |           |
-------------|-----------|-----------|-----------|
        Like |        20 |        46 |        66 |
             |     0.313 |     0.114 |           |
             |     0.303 |     0.697 |     0.767 |
             |     0.870 |     0.730 |           |
             |     0.233 |     0.535 |           |
-------------|-----------|-----------|-----------|
Column Total |        23 |        63 |        86 |
             |     0.267 |     0.733 |           |
-------------|-----------|-----------|-----------|
```



Pearson's Chi-squared test with Yates' continuity correction
data: table(video.clean$like, video.clean$own)
X-squared = 1.1738, df = 1, p-value = 0.2786
The null hypothesis is owning a computer or not is independent of responses in the survey, and the alternative hypothesis is otherwise. Since p-value is 0.2786>0.05, under 0.05 significance level we fail to reject null hypothesis (no enough evidence to show like is associated with owning a computer)

All in all, there is no association between gender and the time spent playing video games; there is no association between employment status and time spent playing video games; there is no association between ownership of a computer and time spent playing video games.

Scenario 6.

Just for fun, further investigate the grade that students expect in the course. How will does it match the target distribution used in grade assignment of 20% A's, 30%B's,40% c's and 10%D's or lower? If the nonrespondents were failing students who no longer bothered to come to the discussion section, would this change the picture?

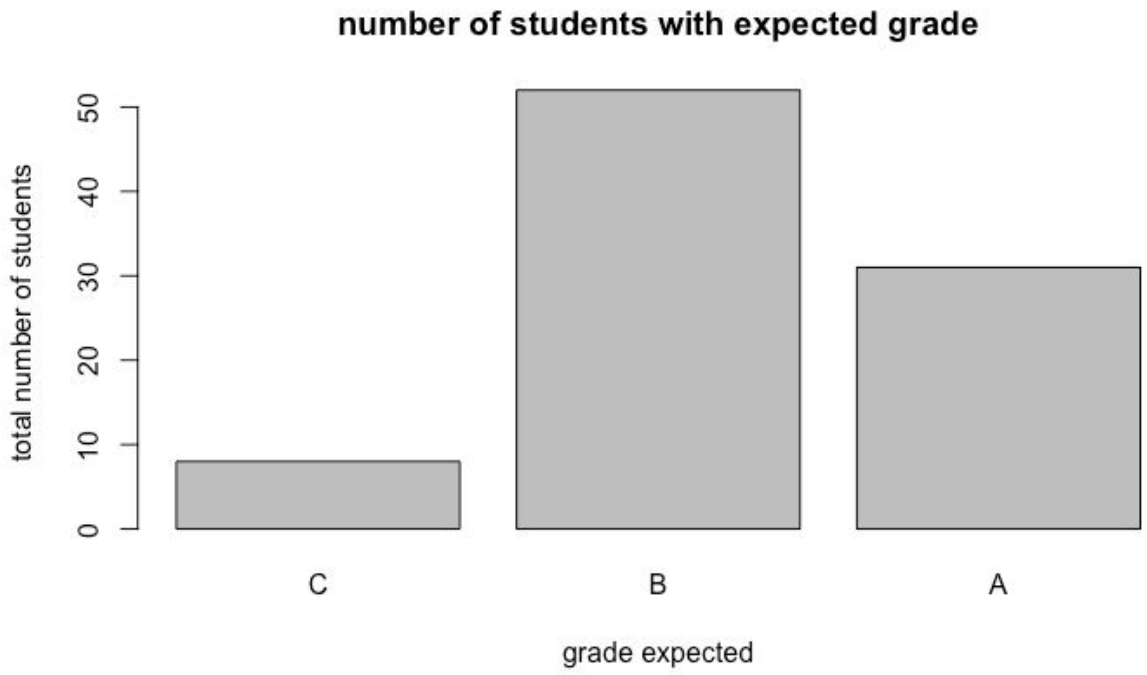


Table 7: Grade distribution of what students' expected and the target

	A	B	C	D
Expected(percentage)	34.07%	57.14%	8.79%	0%
Target(percentage)	20%	30%	40%	10%

From the table 7, the percentages of people who expect to earn A and B are both higher than the target levels, while the percentages of people who expect to earn C and D are both significantly lower than target levels. Thus, generally students' expectations for grades do not match the target distribution used in grade assignment of 20% A's, 30%B's,40% c's and 10%D's or lower.
If the nonrespondents were failing students who no longer bothered to come to the discussion section, the distribution of students' expectation for grades would have been more close to the target grade distribution. Since the expected percentage is calculated by the number of students who expect certain grade divides the total number of students who participate in the survey, if the failing students are added, the deno m of the fraction will be bigger. Thus, if those failing students are added to the sample space, the expected percentage of students earning A or B will

be lower and the expected percentage of students earning C or D will be higher, which changes the picture.

Theory

Goal: By interviewing a subset of the students and using the information collected to provide an approximation to the full group.

The Probability Model: The simple random sample is a probability method for selecting the students. The simple random sample is a very simple probability model for assigning probabilities to all samples of size n from a population of size N. The probability rule that defined the simple random sample is that each one of the $\binom{N}{n}$, samples is equally likely to be selected.

That is, each unique sample of n units has the same chance, $1/\binom{N}{n}$ of being selected. In addition, there is dependence between selections. P(unit # 1 is chosen first and the unit # 2 is chosen second)=(1/N)*(1/(N-1))=1/(N*(N-1)) by conditional probability. The chance is the same for any two units in the population, thus P(# 1 and # 2 are both in the sample)=(n/N)*((n-1)/(N-1))=(n*(n-1))/(N*(N-1)) by independent joint events. In general, for $1 \le j1 \ne ... \ne jn \le N,\ P(I(1) = j1, I(2) = j2, ..., I(n) = jn) = P(I(1) = j1) * P(I(2) = j2) * ... * P(I(n) = jn) = (1/N) * (1/(N-1)) * (1/(N-2)) * ... * (1/(N-n+1)) = 1/(N * (N-1)...(N-n+1))$, since each event I(n) is independent of each other.

Sample Statistics: In our example xi is the time spent playing video games by the student # i: i = 1,...,314. Population average is $\mu = (x1 + x2 + ... + x314)/N = (1/N) \sum_{i=1}^{314} xi$. Let xI(j) represent the value of the characteristic for the j-th unit sampled. In this case, j=1,...,91.E(XI(j))= $\sum_{i=1}^{N} XiP(I(j) = i) = \sum_{i=1}^{N} Xi * (1/N) = \mu$. The sample average is the sample statistic that estimates the population parameter μ, $\bar{x} = (xi(1) + xi(2) + ... + xi(n))/n = (1/n) \sum_{j=1}^{n} xi(j)$. Then, the variance of the sample average $\bar{x}$ is computed as the following:

$$
\begin{aligned}
Var\left(\bar{X}\right) &= \frac{1}{n^2} Var\left(\sum_{j=1}^{n} X_{I(j)}\right) \\
&= \frac{1}{n^2}\left[\sum_{j=1}^{n} Var\left(X_{I(j)}\right) + \sum_{j=1}^{n} Cov\left(X_{I(j)}, X_{I(k)}\right)\right] \\
&= \frac{1}{n^2}\sum_{j=1}^{n} Var\left(X_{I(j)}\right) + \left(\frac{1}{n^2}\right)(2)\sum_{1\le I(j)\le I(k)\le n}\sum Cov\left(X_{I(j)}, X_{I(k)}\right) \\
&= \frac{1}{n^2}\sigma^2 + \left(\frac{2}{n^2}\right)\binom{n}{2} Cov\left(X_{I(1)}, X_{I(2)}\right) \\
&= \frac{1}{n^2}\sigma^2 + \left(\frac{2}{n^2}\right)\left(\frac{n!}{2!(n-2)!}\right) Cov\left(X_{I(1)}, X_{I(2)}\right) \\
&= \frac{1}{n^2}\sigma^2 + \left(\frac{n(n-1)(n-2)!}{n^2(n-2)!}\right) Cov\left(X_{I(1)}, X_{I(2)}\right) \\
&= \frac{1}{n^2}\sigma^2 + \frac{n-1}{n} Cov\left(X_{I(1)}, X_{I(2)}\right)
\end{aligned}
$$

Estimators for Standard Errors:

The common estimator for σ^2 is $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{I(j)} - \bar{x})^2$ and Var($\bar{x}$) can be estimated by $\frac{s^2}{n}\frac{N-n}{N-1}$. The reason to this estimation is that since s^2 is used to estimate σ^2 and since Var($\overline{X}$)=(1/n) $\sigma^2(N-n)/(N-1)$ by plugging in s^2, estimation of Var($\bar{x}$) can be generated. A slightly better estimate for σ^2 is the unbiased estimator $s^2\frac{N-1}{N}$, because when computing the expected value for s^2, it is equal to $\frac{N}{N-1}\sigma^2$. In order to get an unbiased estimator, the equation of expected value for s^2 needs to be multiplied by the reciprocal of $\frac{N}{N-1}$ and that is how the unbiased estimator comes from. Moreover, an unbiased estimator of Var($\overline{X}$) is $\frac{s^2}{n}\frac{N-n}{N}$. When N is very large, the difference between the two estimators will be negligible.

Population Totals and Percentages: When the parameter is a proportion, it is reasonable for the characteristic value x_i to be 0 or 1 to denote the absence or presence of the characteristic respectively. For example, for i=1,...314. x_i is 1 if the ith student in the population owns a PC, otherwise x_i is 0. The summation of all the students who own PCs in the population is denote by τ, where $\tau=\sum x_i$ and π which is the proportion of students in the population who owns PCs is equal to $\pi = \frac{1}{N}\sum_{i=1}^{N} x_i$. In this case $\bar{x}$ is an unbiased estimator of π, the population average and $N\bar{x}$ estimates τ. Therefore an unbiased estimator of Var($\bar{x}$) can be obtained because $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \pi)^2 = \pi(1-\pi)$. Thus the estimator for the standard error is

$$
\widehat{SE(\bar{x})} = \frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{n-1}}\frac{\sqrt{N-n}}{\sqrt{N}}.
$$

Normal Approximation and Confidence Intervals

By the Central Limit Theorem, if sample size n is large and $x_1, ..., x_n$ are independent, identically distributed with mean μ and variance σ^2 , then the probability distribution of sample average nearly follows normal curve, that is, $Z=(\bar{x}-\mu)/(\sigma/n)$ is approximately standard normal. Although $X_{I(j)}$ are not independent in simple random sampling, we can still apply normal approximation here if the sample size n is large enough but not too large comparing to the population size. Moreover, the normal distribution can give us confidence interval for the population parameter. For example, $(\bar{x}-\sigma/\sqrt{n}, \bar{x}+\sigma/\sqrt{n})$ and $(\bar{x}-2\sigma/\sqrt{n}, \bar{x}+2\sigma/\sqrt{n})$ are 68% and 95% confidence intervals respectively, which means, the probability that $\bar{x}$ is with one or two standard error of μ is about 68% or 95%. Thus, we can form a formula that $P(\bar{x}-2\sigma/\sqrt{n} \le \mu \le \bar{x}+2\sigma/\sqrt{n})=P(\mu-2\sigma/\sqrt{n} \le \bar{x} \le \mu+2\sigma/\sqrt{n})=P(2 \le (\bar{x}-\mu)/(\sigma/\sqrt{n}) \le 2)$ which is approximately 0.95. Since σ is unknown in most cases, we replace σ by s in the formula, and the interval using s instead of σ is called approximate confidence interval.

Bootstrap:

Bootstrap applies to finite samples to assess estimators' accuracy via variance estimation and to produce confidence intervals and p-values when dealing with finite populations. Since the distribution of the sample generated by the simple random sample probability model looks roughy similar to that of the population, a new population with the same size based on the sample, defined as the bootstrap population, can be used to find the probability distribution of sampling average. To implement bootstrap, for every unit in the sample, we generate N(population size)/n(sample size) units with the same value in the bootstrap population and round off to the nearest integer. Once the bootstrap population is generated, the average can be calculated from taking a simple random sample of size n from the bootstrap population. This process is repeated for k iterations until k sample averages were generated. Then the histogram of those k bootstrap sample averages is generated to help evaluate the probability distribution of the sample average.

Discussion/Conclusion

From scenario 1, the investigation shows that 37% of sampled students play video games. Thus, the faculty should anticipate around 37% of the total number of enrolled students to attend their labs, and adjust their labs accordingly.

According to Scenario 2, there is a strong association between hours played and whether or not students play when they are busy. In particular, if students are busy and do not play video games when they are busy, then they will play for less hours. It has also been shown that the presence of an exam will negatively affect the number of hours they play, especially for students who play daily or weekly. Thus, the faculty should try to avoid holding labs around exam times.

Results from scenario 3 show that the estimated time of playing video game per week is about 1.24 hour with confidence interval of (0.467,2.019), so we suggest our statistics lab to limit

the lab hour to no more than 2.0 hour which satisfy majority of the population.

According to scenario 4, 66% of sampled students play video game for relaxation and 48% believe they spend too much time on it. Our faculty should expect that the number of students who attend labs in the week prior to the final, the time when they are stressed out, will increase, but faculty should also limit lab hour during this time in order to prevent students from spending too much time on playing video games.

Scenario 5 reveals that there is no association between gender and the time spent on playing video games; there is no association between employment status and time spent on playing video games; there is no association between ownership of a computer and time spent on playing video games.

For scenario 6, students' expectations for grades do not match the target distribution used in grade assignment of 20% A's, 30%B's,40% c's and 10%D's or lower. If those failing students are added to the sample space, the expected percentage of students earning A or B will be lower and the expected percentage of students earning C or D will be higher, which changes the picture.

Works Cited

Dorval, M. "*Effect of Playing a Video Game on a Measure of Spatial Visualization*." *NCBI*, 4 Feb. 1986, doi:10.2466/pms.1986.62.1.159.

Eisenmann, Joey, et al. "*Physical Activity, TV viewing, and Weight in U.S. Youth: 1999 Youth Risk Behavior Survey.*" *Obesity Research*, vol. 10, no. 5, 9 July 2001, onlinelibrary.wiley.com/doi/epdf/10.1038/oby.2002.52.

Gellis, Zvi, et al. "*New York State Case Manager Survey: Urban and Rural Differences in Job Activities, Job Stress, and Job Satisfaction.*" *The Journal of Behavioral Health Services & Research*, Oct. 2004, link.springer.com/content/pdf/10.1007%2FBF02287694.pdf.

Green, Bavelier. "*The Development of Attention Skills in Action Video Game Players*." *ScienceDirect*, vol. 47, nos. 8-9, July 2009, www.sciencedirect.com/science/article/pii/S0028393209000657.

Griffith, JL. "*Differences in Eye-hand Motor Coordination of Video-game Users and Non-users.*" *NCBI*, 21 Aug. 1983, DOI:10.2466/pms.1983.57.1.155.

Nolan, Deborah A, and T P. Speed. *Stat Labs: Mathematical Statistics Through Applications*. New York: Springer, 2000. Internet resource.