Alex Hsieh - A11651288 - 4th Year Math and Economics Major
Atharva Fulay - A13140630 - 4th year Applied Math Major
Peter Yao - A12286133 - 3rd year Math-Computer Science Major
Justin Glommen - A13045994 - 2nd year Math-Computer Science Major
Victor Angulo - A99088060 - 4th year Math-Statistics Major

Case Study 2

Introduction:

At UC Berkeley, a committee of faculty and students designed a series of labs to assist with class instruction and to provide useful insights to the designers of the new computer lab. These labs provide an on-hands approach as an alternative method for students to learn statistics and probability. A survey was then conducted , using an independently and identically distributed sample of students from “Introductory Probability and Statistics, Section 1” in Fall of 1994, a prerequisite for all students wanting to major in business. These students had all completed the second exam of the semester, and all 314 students enrolled were assigned a number from 1 to 314, and then selected via a pseudo-random number generator. Those selected were all informed of what the study would be about before participating in it, and 91 participated of the 95 selected.

Metrics that were analyzed include: how much time was spent playing & how many students played video games prior to the survey, whether or not they like to play video games, how often they play video games, gender, whether or not they work for pay or not, and whether or not they have a computer. Respondents that answered that they did not like video games were asked not to respond to some of the other survey questions.

We acknowledge the existence of other factors, not integrated in the data, that could also be responsible for a change in video game habits. Students that are enrolled in the class are mostly business majors, but students enrolled in other majors may have statistically different responses. The type of computer the students owned could skew whether or not the students liked video games as well, since Macintosh owners most likely would not purchase their computer with the intent of playing video games. Accordingly, the time period when this study was taken was before video game culture was prevalent, which might skew results slightly.

Data:

Cleaned up

```
      time              like            where            freq            busy
 Min.   : 0.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :0.0000
 1st Qu.: 0.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:0.0000
 Median : 0.000   Median :3.000   Median :3.000   Median :3.000   Median :0.0000
 Mean   : 1.243   Mean   :3.022   Mean   :2.973   Mean   :2.705   Mean   :0.2125
 3rd Qu.: 1.250   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:0.0000
 Max.   :30.000   Max.   :5.000   Max.   :6.000   Max.   :4.000   Max.   :1.0000
                  NA's   :1       NA's   :18      NA's   :13      NA's   :11
      educ              sex              age             home             math
 Min.   :0.0000   Min.   :0.0000   Min.   :18.00   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:1.0000   1st Qu.:0.0000
 Median :0.0000   Median :1.0000   Median :19.00   Median :1.0000   Median :0.0000
 Mean   :0.4744   Mean   :0.5824   Mean   :19.52   Mean   :0.7582   Mean   :0.3222
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:20.00   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :33.00   Max.   :1.0000   Max.   :1.0000
 NA's   :13                                                         NA's   :1
      work              own             cdrom             email            grade
 Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :2.000
 1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:3.000
 Median : 1.000   Median :1.0000   Median :0.0000   Median :1.0000   Median :3.000
 Mean   : 7.352   Mean   :0.7363   Mean   :0.1744   Mean   :0.7912   Mean   :3.253
 3rd Qu.:13.250   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:4.000
 Max.   :55.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :4.000
 NA's   :3                         NA's   :5
```

Analysis and Methods:

Scenario 1:

Using the provided time factor of the statistics class sample dataset, we are able to estimate the proportion of students who played games within the last week prior to the survey. This is done by comparing how many students spent more than zero hours playing video games, to the total amount of students in the class. That ratio of players to total students is our point estimate for the proportion, which we can use to generate a confidence interval. That said, our point estimate for the proportion of students who have played video games in the last week is approximately **37.36%**. Therefore, with 95% confidence we report that the true proportion lies within (28.94%, 45.79%). This yields a finite population corrected standard error of roughly 4.30%, and a margin of error of about +/-8.42%.

```
[1] "Corrected Standard Error: 0.0429736108569751"
[1] "Margin of Error: 0.0842282772796712"
[1] "Player Proportion 95% CI: (0.289398096346702, 0.457854650906045)"
```

Scenario 2:

Students were assessed on how much video games they play in the week prior to the study by hour and frequency of play. How often they played was measured either daily, weekly, monthly, or semesterly corresponding respectively to 1, 2, 3, and 4. We took the liberty of categorizing students by how many hours they played in the week prior. 57 students (62.64%) responded that they played zero hours in the week prior, 31 (34.07%) students played a few hours (less than five hours), and 3 students (3.30%) played many hours (more than five hours). Of the students who played zero hours, the mean of how frequently they played is 3, which corresponds to monthly. Of the students who played a few hours, the mean of how frequently they played is 2.207, which is in between weekly and monthly but leaning towards weekly. Lastly, the three students who played many hours, had a mean of 1.333 which means two of them are playing daily while the third is playing weekly.

It can clearly be seen that students who did not play the week prior to the survey played less overall than those who played during that week. Moreover, students who inputted daily on how frequently they played averaged 4.44 hours of play in the week. Students who inputted weekly averaged 2.53 hours, monthly averaged 0.556 hours, and semesterly averaged 0.435 hours in the week prior. This shows a very strong correlation between time spent playing in the last week and time spent playing in general. Similarly, students who play as frequently as semesterly played an average of less than a half hour in the week prior. It is clear that students who play more, on average during the semester, played more in the week prior to the survey and students who played more in the week prior usually play more frequently than students who didn't play that much.

The fact that there was an exam in the week prior could have affected the number of hours students played video games. 17 of 91 students play video games even if they are busy, while the rest do not. The average amount of hours played in the week prior by the students who play even if busy is 4.71 hours, while the students who don't play when busy averaged 0.510 hours of play. Since there was a test the week prior, assuming students are busy studying, a majority of students would be more likely to not play, decreasing the amount of hours played in the week prior. This would also impact this comparison as students who 'somewhat' like to play or students who play weekly are less likely to play due to being busy studying for the exam. These students are more likely to play more during weeks where there are no exams. Therefore, we would expect the number of hours that students played video games to rise if the exam is not held in the week prior.

Scenario 3:

First, we analyze the kurtosis of the distribution. The value of the sample's kurtosis is 38.8. After simulating random normal samples of the same size we obtained a mean value of -0.133 as seen in **Figure 6**. Calculating the third moment, we obtain a skewness value of 5.69, which is far off from 0. Due to the skewness of the sample in terms of the time spent playing video games, we used a bootstrap simulation to create random samples that assimilate the nature of the observed sample (See **Figure 1**). We created 500 bootstrap samples that contained the same characteristics as the original sample. Through this simulation, we obtain a mean value of 1.13 hours spent playing video games. Using the 0.025 and 0.975 quantiles, we obtain the following interval (0.544, 1.81) (see **Figure 5**). The unimodal distribution is centered at 1.1. Thus, we believe the true mean time spent playing video games lies within the interval.

Scenario 4:

In general, it would seem that the majority of students enjoy playing video games, since the percentage of those who chose "very much" or "somewhat" from Scenario 2 is **75.82%.** We believe in order to get a more precise estimate of why students like/dislike students, there should be specific questions about the amount of time spent playing, since the amount of hours playing video games has a high correlation with how much the student likes video games. If the student answered that they did not like video games, it would be appropriate to have a question following which asked the reason why not.

A follow up survey was conducted to the students who participated in the study, where video game players were asked about what video games they liked and why they play. According to the results of this study in the lecture slides, the majority of video game players preferred strategy and action games. We can infer that video game players from the study enjoy fast-paced games that require large amounts of critical thinking, as opposed to games that require some time to fully immerse players into an activity, like many adventure and simulation games do.

The most significant reasons for why students played video games are to relax and to achieve mastery of their games. Therefore, we cannot make a lab environment where students are extremely stressed about their grades or put them in an uncomfortable position, such as having quizzes.

Every student who participated in the study was asked what they did not like about playing video games, and the most common responses were that they took up too much time or were pointless. Thus, in order to generalize gaming to a statistics lab where a plurality of students feel this way, the lab will have to find ways to combat these issues.

Scenario 5:

To distinguish between students who like or don't like playing video games, we put every respondent into two groups. Those who answered 2 or 3 for whether they like to play video games are said to like video games, while those who answered 1, 4 or 5 are said to not like video games.

From Figure 2, 48 people who own a PC like playing games, while 18 of those who own a PC don't like playing games. 21 people who don't own a PC like playing games, while 3 people who don't own a PC don't like playing games. Most students in the section own a PC, but there doesn't seem to be a correlation between owning one and not liking video games.

By Figure 3, 36 people who like games worked the week prior to the survey, while 7 people who didn't like games worked. 30 people who like games didn't work, while 14 people who didn't like games didn't work. A high number of people who liked games worked, while many people who didn't like games didn't work. This is in correlation to the high number of respondents who said they liked playing games to relax and those who said games costed too much. It is possible that many students use the money they earn from their jobs to buy their games and play games to relax after working, while those who didn't want to play any games didn't need to work to buy any.

Looking at Figure 4, 43 males like playing games, while 9 males didn't like playing games. 26 females like playing games, while 12 females don't like playing games. Because there is a bigger proportion of males who like playing games compared to those who don't as opposed to that of females, males have an easier time finding other males to play games with, which may mean that they generally won't find games lonely or boring.

Scenario 6:

The grade distribution that is usually assigned during grading is 20% A's, 30% B's, 40% C's and 10% D's and lower. To understand if the grade that the students expect is different to the distribution that was described before, we do a chi-square test for contingency tables (See Table 1).

Grade	Observed	Expected
A	31	18
B	52	30
C	8	36
D or lower	0	9

Table 1. Shows the observed grade that students expect within the sample of 91 compared to the expected number of students based on the grading scale.

Using a chi-square test for contingency tables, we obtain a p-value = 1.617e-12. Based on this p-value, at an alpha level of 0.05, we are confident that the observed grade distribution does not match that of the expected grading distribution. That is, there are too many students expecting a grade that they will likely not receive. If the nonrespondents were students who were failing the class, our conclusion would be the same. There would not be enough evidence to expect both distributions to be the same, or similar, due to the large number of students expecting the grade of B.

Conclusion:

From the responses of the students in the discussion session in the survey, we can infer that most of the students enjoy playing video games, but did not play very much in the week prior to when the survey was conducted. Students that answered whether they played games daily, weekly, monthly or semesterly played around the proportionate amount of hours compared to how many hours they played in the week prior to the survey. From this survey, it seems that there is a strong correlation between the amount of video games played prior to the survey and the amount an individual plays in general.

Ultimately, the most important things to consider when making the new computer lab is to promote a high paced environment that keeps students interested and occupied, yet a relaxed one where not much is at stake. It is essential to design a computer lab where students can immediately interact with the material hands-on instead of simply giving them a list of tasks to complete, like in some adventure games. It is also vital that students do not get bored in the lab and do not feel lonely or frustrated with the content, which we can prevent by encouraging more group-based activities. Because students in the discussion section expected higher grades than what the target grade distribution is, we also need to incentivize students who are not doing as well in future classes to attend the labs.

Theory:

Computing the expected value of $X_{I(J)}$:

E[$X_{I(j)}$] = 0 * **P**(otherwise) + X_i * **P**(I(j) = i) for every i, j because of how $X_{I(j)}$ is distributed. This simplifies to Summation from 1 to *N* (X_i * **P**(I(j) = i)), where j is some value between 1 and *n* (which is 91 in this case) and i is between 1 and *N* (which is 314). Due to every **P**(I(j) = i) = 1/*N*, we can replace **P**(I(j) = i) with 1/*N*. This simplifies to Summation from 1 to *N* (X_i * (1/*N*)). This is equal to *mu* because it is known that the population average is exactly that equation. We can factor out the 1/*N* statement out front to get the final form: *mu* = (1/*N*) * Summation from 1 to *N* (X_i).

Computing the Variance of Xbar:

The common estimator for the variance of the population is the sample variance, where s^2 = (1/n-1)sum(x(I(J)) - xbar). However, **E**[s^2] = (N/N-1)(sigma)^2, thus it is biased. In order to achieve an unbiased estimator of sigma, we need to adjust s^2. Then, **E**[(N-1/N)s^2] = (sigma)^2. Hence, an unbiased estimator for Var(xbar) is (N-n/N)(s^2/n). When xbar is a percentage or proportion, when can treat it as a Bernoulli random variable and apply the correction factor. That is, Var(xbar) = xbar(1-xbar)(N-n)/(N(n-1)).

Creating a confidence interval for Xbar:

In this homework, we used confidence intervals to give bounds for xbar. The standard format of a confidence interval for a desired estimator O is (O - a*SE(O), O + a*SE(O)), that is, our estimator adjusted by the product of a quantile value and the standard error of our estimator. For the scenarios explored, we created confidence intervals for xbar (an average or percentage value). When xbar is an average, we can adjust xbar by its standard deviation multiplied by two.

This is a reasonable method since we are estimating all possible values that are 2-standard deviations away from our estimator. When xbar represents a percentage, we can use Bootstrap simulation to obtain reasonable 25 and 95 quantile values since the distribution of the data can be skewed.

Appendix:

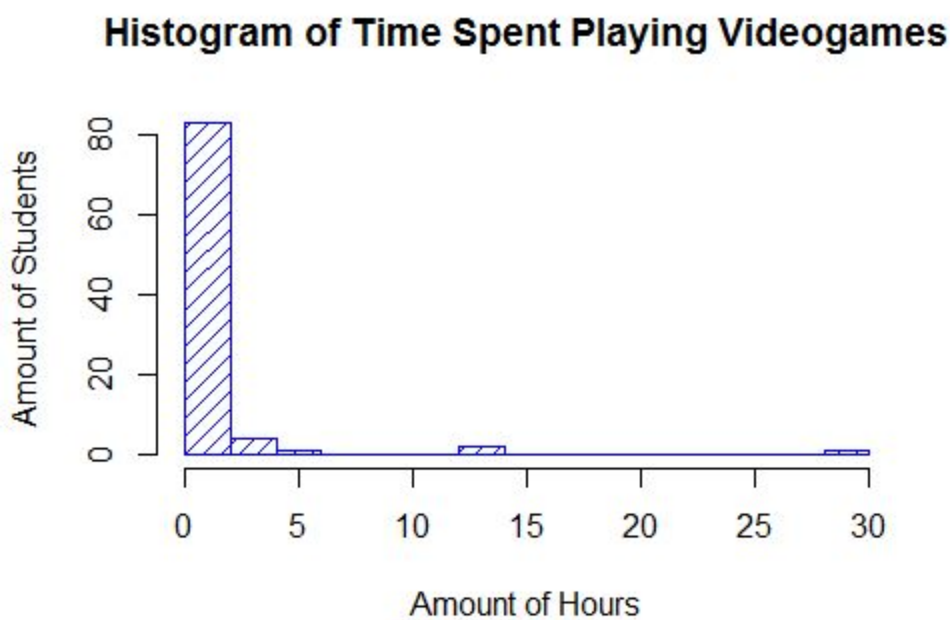


Figure 1. Illustrates the distribution of the number of hours students spent playing video games. It is easy to observe the skewness of the distribution.

```
Total Observations in Table:  90

             | editData$own
editData$like |         0 |         1 | Row Total |
--------------|-----------|-----------|-----------|
            1 |         0 |         1 |         1 |
              |     0.267 |     0.097 |           |
              |     0.000 |     1.000 |     0.011 |
              |     0.000 |     0.015 |           |
              |     0.000 |     0.011 |           |
--------------|-----------|-----------|-----------|
            2 |         5 |        18 |        23 |
              |     0.209 |     0.076 |           |
              |     0.217 |     0.783 |     0.256 |
              |     0.208 |     0.273 |           |
              |     0.056 |     0.200 |           |
--------------|-----------|-----------|-----------|
            3 |        16 |        30 |        46 |
              |     1.136 |     0.413 |           |
              |     0.348 |     0.652 |     0.511 |
              |     0.667 |     0.455 |           |
              |     0.178 |     0.333 |           |
--------------|-----------|-----------|-----------|
            4 |         1 |        12 |        13 |
              |     1.755 |     0.638 |           |
              |     0.077 |     0.923 |     0.144 |
              |     0.042 |     0.182 |           |
              |     0.011 |     0.133 |           |
--------------|-----------|-----------|-----------|
            5 |         2 |         5 |         7 |
              |     0.010 |     0.003 |           |
              |     0.286 |     0.714 |     0.078 |
              |     0.083 |     0.076 |           |
              |     0.022 |     0.056 |           |
--------------|-----------|-----------|-----------|
 Column Total |        24 |        66 |        90 |
              |     0.267 |     0.733 |           |
--------------|-----------|-----------|-----------|
```

Figure 2. The cross-tabulation between the number of people who like video games and those who own a PC.

```
Total Observations in Table:  87

             | editData$work == 0
editData$like |     FALSE |      TRUE | Row Total |
--------------|-----------|-----------|-----------|
            1 |         1 |         0 |         1 |
              |     0.518 |     0.506 |           |
              |     1.000 |     0.000 |     0.011 |
              |     0.023 |     0.000 |           |
              |     0.011 |     0.000 |           |
--------------|-----------|-----------|-----------|
            2 |        14 |         9 |        23 |
              |     0.609 |     0.596 |           |
              |     0.609 |     0.391 |     0.264 |
              |     0.326 |     0.205 |           |
              |     0.161 |     0.103 |           |
--------------|-----------|-----------|-----------|
            3 |        22 |        21 |        43 |
              |     0.026 |     0.026 |           |
              |     0.512 |     0.488 |     0.494 |
              |     0.512 |     0.477 |           |
              |     0.253 |     0.241 |           |
--------------|-----------|-----------|-----------|
            4 |         3 |        10 |        13 |
              |     1.826 |     1.785 |           |
              |     0.231 |     0.769 |     0.149 |
              |     0.070 |     0.227 |           |
              |     0.034 |     0.115 |           |
--------------|-----------|-----------|-----------|
            5 |         3 |         4 |         7 |
              |     0.061 |     0.060 |           |
              |     0.429 |     0.571 |     0.080 |
              |     0.070 |     0.091 |           |
              |     0.034 |     0.046 |           |
--------------|-----------|-----------|-----------|
 Column Total |        43 |        44 |        87 |
              |     0.494 |     0.506 |           |
--------------|-----------|-----------|-----------|
```

Figure 3. The cross-tabulation between the number of people who like video games and those who worked some number of hours in the week prior to the survey.

Total Observations in Table: 90

editData$like	editData$sex 0	1	Row Total
1	0	1	1
	0.422	0.309	
	0.000	1.000	0.011
	0.000	0.019	
	0.000	0.011	
2	5	18	23
	2.285	1.670	
	0.217	0.783	0.256
	0.132	0.346	
	0.056	0.200	
3	21	25	46
	0.128	0.094	
	0.457	0.543	0.511
	0.553	0.481	
	0.233	0.278	
4	8	5	13
	1.149	0.840	
	0.615	0.385	0.144
	0.211	0.096	
	0.089	0.056	
5	4	3	7
	0.369	0.270	
	0.571	0.429	0.078
	0.105	0.058	
	0.044	0.033	
Column Total	38	52	90
	0.422	0.578	

Figure 4. The cross-tabulation between the number of people who like video games and their sex.

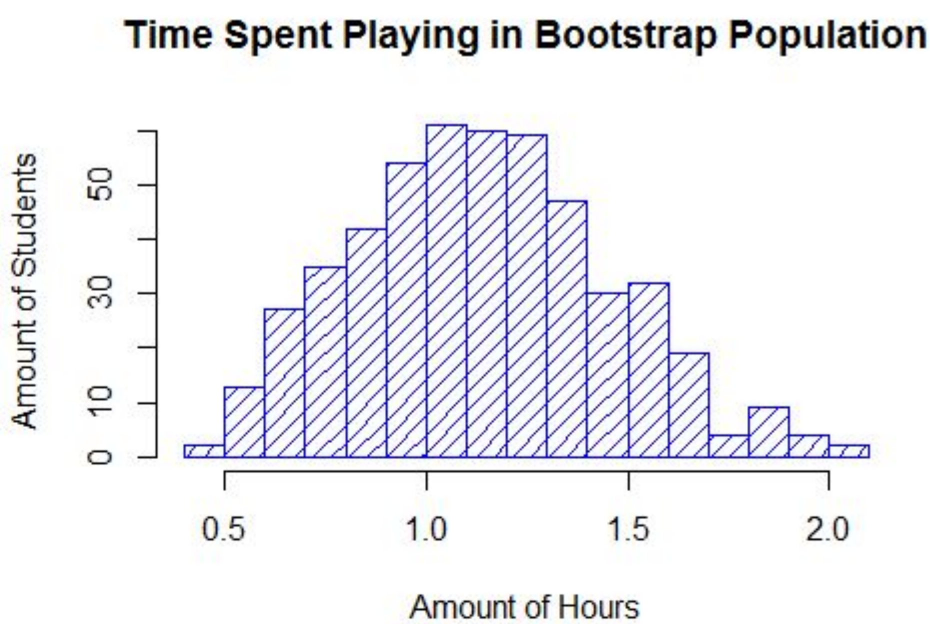


Figure 5. Depicts the average time spent playing video games in the Bootstrap population. The distribution appears to be have a bell shaped curve but it still shows some skewness.

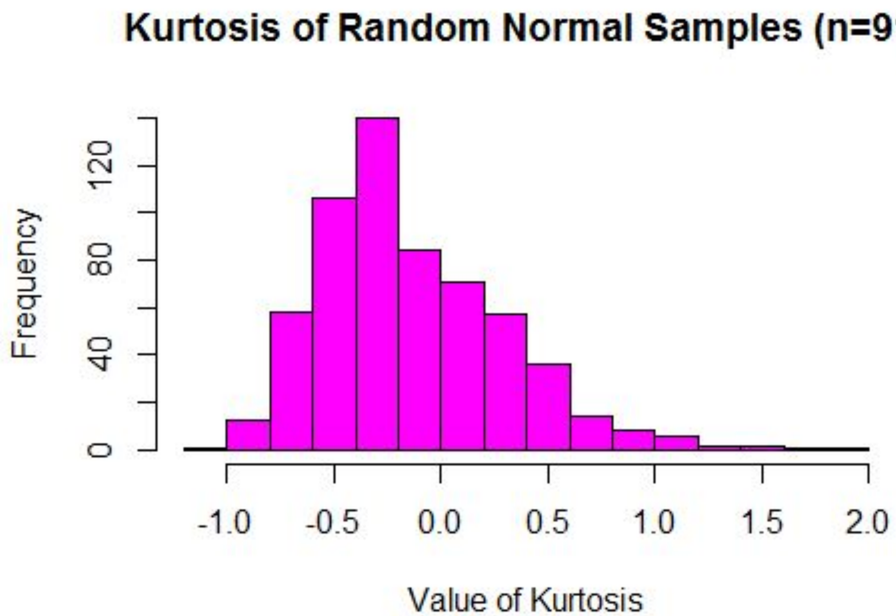


Figure 6. Shows the distribution curve for the kurtosis value of 600 random Normal distributions of sample size 91. We can observe the peak to be approximately -0.3. This supports, more than just visually, that the distribution is not symmetric.

StatsLab RCode Report

Justin Glommen (Scenarios 1,4)
Victor Angulo (Scenarios 3,6)
Atharva Fulay (Scenario 2)
Peter Yao (Scenario 5)

2/5/2017

Data Management

Loading data from current directory

```
data <- read.table("videodata.txt", header=TRUE)
data.population <- 314      # True population
data.samples <- 91          # Number of samples
head(data)
```

```
##   time like where freq busy educ sex age home math work own cdrom email
## 1  2.0    3     3    2    0    1   0  19    1    0   10   1     0     1
## 2  0.0    3     3    3    0    0   0  18    1    1    0   1     1     1
## 3  0.0    3     1    3    0    0   1  19    1    0    0   1     0     1
## 4  0.5    3     3    3    0    1   0  19    1    0    0   1     0     1
## 5  0.0    3     3    4    0    1   0  19    1    1    0   0     0     1
## 6  0.0    3     2    4    0    0   1  19    0    0   12   0     0     0
##   grade
## 1     4
## 2     2
## 3     3
## 4     3
## 5     3
## 6     3
```

```
summary(data)
```

```
##       time             like            where            freq
##  Min.   : 0.000   Min.   : 1.000   Min.   : 1.00   Min.   : 1.00
##  1st Qu.: 0.000   1st Qu.: 2.000   1st Qu.: 3.00   1st Qu.: 2.00
##  Median : 0.000   Median : 3.000   Median : 3.00   Median : 3.00
##  Mean   : 1.243   Mean   : 4.077   Mean   :21.97   Mean   :16.46
##  3rd Qu.: 1.250   3rd Qu.: 3.000   3rd Qu.: 5.00   3rd Qu.: 4.00
##  Max.   :30.000   Max.   :99.000   Max.   :99.00   Max.   :99.00
##       busy            educ            sex              age
##  Min.   : 0.00   Min.   : 0.00   Min.   :0.0000   Min.   :18.00
##  1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.:0.0000   1st Qu.:19.00
##  Median : 0.00   Median : 1.00   Median :1.0000   Median :19.00
##  Mean   :12.15   Mean   :14.55   Mean   :0.5824   Mean   :19.52
##  3rd Qu.: 1.00   3rd Qu.: 1.00   3rd Qu.:1.0000   3rd Qu.:20.00
##  Max.   :99.00   Max.   :99.00   Max.   :1.0000   Max.   :33.00
##       home             math             work            own
##  Min.   :0.0000   Min.   : 0.000   Min.   : 0.00   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.: 0.000   1st Qu.: 0.00   1st Qu.:0.0000
##  Median :1.0000   Median : 0.000   Median : 5.00   Median :1.0000
```

```
##  Mean   :0.7582   Mean   : 1.407   Mean   :10.37   Mean   :0.7363
##  3rd Qu.:1.0000   3rd Qu.: 1.000   3rd Qu.:14.50   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :99.000   Max.   :99.00   Max.   :1.0000
##      cdrom            email            grade
##  Min.   : 0.000   Min.   :0.0000   Min.   :2.000
##  1st Qu.: 0.000   1st Qu.:1.0000   1st Qu.:3.000
##  Median : 0.000   Median :1.0000   Median :3.000
##  Mean   : 5.604   Mean   :0.7912   Mean   :3.253
##  3rd Qu.: 0.000   3rd Qu.:1.0000   3rd Qu.:4.000
##  Max.   :99.000   Max.   :1.0000   Max.   :4.000
```

Cleaning Data

Replacing 99 values (the unanswered/improper results) with NAs

```
data[data == 99] <- NA
numSamples <- NROW(data)
head(data)
```

```
##   time like where freq busy educ sex age home math work own cdrom email
## 1  2.0    3     3    2    0    1   0  19    1    0   10   1     0     1
## 2  0.0    3     3    3    0    0   0  18    1    1    0   1     1     1
## 3  0.0    3     1    3    0    0   1  19    1    0    0   1     0     1
## 4  0.5    3     3    3    0    1   0  19    1    0    0   1     0     1
## 5  0.0    3     3    4    0    1   0  19    1    1    0   0     0     1
## 6  0.0    3     2    4    0    0   1  19    0    0   12   0     0     0
##   grade
## 1     4
## 2     2
## 3     3
## 4     3
## 5     3
## 6     3
```

```
summary(data)
```

```
##       time             like           where            freq
##  Min.   : 0.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.: 0.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median : 0.000   Median :3.000   Median :3.000   Median :3.000
##  Mean   : 1.243   Mean   :3.022   Mean   :2.973   Mean   :2.705
##  3rd Qu.: 1.250   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :30.000   Max.   :5.000   Max.   :6.000   Max.   :4.000
##                   NA's   :1       NA's   :18      NA's   :13
##       busy             educ             sex              age
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :18.00
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:19.00
##  Median :0.0000   Median :0.0000   Median :1.0000   Median :19.00
##  Mean   :0.2125   Mean   :0.4744   Mean   :0.5824   Mean   :19.52
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:20.00
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :33.00
##  NA's   :11       NA's   :13
##       home             math             work             own
##  Min.   :0.0000   Min.   :0.0000   Min.   : 0.000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:0.0000
```

```
##  Median :1.0000   Median :0.0000   Median : 1.000   Median :1.0000
##  Mean   :0.7582   Mean   :0.3222   Mean   : 7.352   Mean   :0.7363
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:13.250   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :55.000   Max.   :1.0000
##                   NA's   :1        NA's   :3
##      cdrom            email             grade
##  Min.   :0.0000   Min.   :0.0000   Min.   :2.000
##  1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:3.000
##  Median :0.0000   Median :1.0000   Median :3.000
##  Mean   :0.1744   Mean   :0.7912   Mean   :3.253
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :4.000
##  NA's   :5
```

Scenario 1

Sample Proportion of Students Who Played a Video Game in the Last Week

The individual variables measured here are Bernoulli since time is being converted to a binary 'did' or 'did not' play.

```r
# Create 'numPlayers' variable to count number of players in the last week.
# This is done by counting the number of people with time spent over 0, which represents the
# people who played something in the last week since they spent time on it. 0 indicates no time
# spent.
numPlayers <- NROW(which(data$time > 0))
paste("Number of players:", numPlayers, sep=" ")
```

```
## [1] "Number of players: 34"
```

```r
# Sample proportion is the ratio of numPlayers to total students (rows in data)
data.playersSampleProportion <- (numPlayers/numSamples)
paste("Sample proportion:", data.playersSampleProportion, sep=" ")
```

```
## [1] "Sample proportion: 0.373626373626374"
```

Players Sample Proportion Confidence Interval

Since the sample Bernoulli variables are NOT identically independentally distributed, the confidence interval itself will be computed utilizing the finite population correction factor.

```r
# Sample proportion is nearly Binomial, except not iid.
playersCorrectionFactor <- sqrt((data.population - numSamples)/data.population)
# Binomial standard error formula without correction
playersIndepStandardError <-  (sqrt(data.playersSampleProportion*(1-data.playersSampleProportion))/sqrt
# Standard error with finite population correction
data.playersStandardErrorEstimate <- playersIndepStandardError*playersCorrectionFactor
paste("Corrected Standard Error:", data.playersStandardErrorEstimate, sep=" ")
```

```
## [1] "Corrected Standard Error: 0.0429736108569751"
```

```r
# Since the sample proportion follows a normal distribution by the Central Limit Theorem,
# we need to multiply the corrected standard error by 1.96 to generate the interval.
```

```r
data.playersMarginOfError <- 1.96*data.playersStandardErrorEstimate
paste("Margin of Error: ", data.playersMarginOfError, sep="")
```

```
## [1] "Margin of Error: 0.0842282772796712"
```

```r
# Therefore, the confidence interval:
playersLowerBound <- data.playersSampleProportion - data.playersMarginOfError
playersUpperBound <- data.playersSampleProportion + data.playersMarginOfError
data.playersSampleProportionConf95 <- c(playersLowerBound, playersUpperBound)
paste("Player Proportion 95% CI: ", "(",playersLowerBound, ", ", playersUpperBound,")", sep="")
```

```
## [1] "Player Proportion 95% CI: (0.289398096346702, 0.457854650906045)"
```

Scenario 2

```r
smalltime.ind <- which(data$time < 6)
data.smalltime <- data[smalltime.ind,]

zerohours.ind <- which(data.smalltime$time ==0)
data.zerohours <- data[zerohours.ind, ]
mean(data.zerohours$freq, na.rm=TRUE)
```

```
## [1] 3
```

```r
fewhours.ind <- which(data.smalltime$time > 0 & data.smalltime$time <=5 )
data.fewhours <- data[fewhours.ind, ]
mean(data.fewhours$freq, na.rm=TRUE)
```

```
## [1] 2.206897
```

```r
manyhours.ind <- which(data$time > 6)
data.manyhours <- data[manyhours.ind, ]
summary(data.manyhours$freq, na.rm=TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.333   1.500   2.000
```

```r
daily.ind <- which(data$freq == 1)
weekly.ind <- which(data$freq == 2)
monthly.ind <- which(data$freq == 3)
semester.ind <- which(data$freq == 4)

data.daily <- data[daily.ind, ]
data.weekly <- data[weekly.ind, ]
data.monthly <- data[monthly.ind, ]
data.semester <- data[semester.ind, ]

mean(data.daily$time)
```

```
## [1] 4.444444
```

```r
mean(data.weekly$time)
```

```
## [1] 2.539286
```

```r
mean(data.monthly$time)
```

```
## [1] 0.05555556
```

```r
mean(data.semester$time)
```

```
## [1] 0.04347826
```

```r
busy.ind <- which(data$busy == 1)
data.busy <- data[busy.ind, ]

notbusy.ind <- which(data$busy == 0)
data.notbusy <- data[notbusy.ind, ]

mean(data.busy$time)
```

```
## [1] 4.705882
```

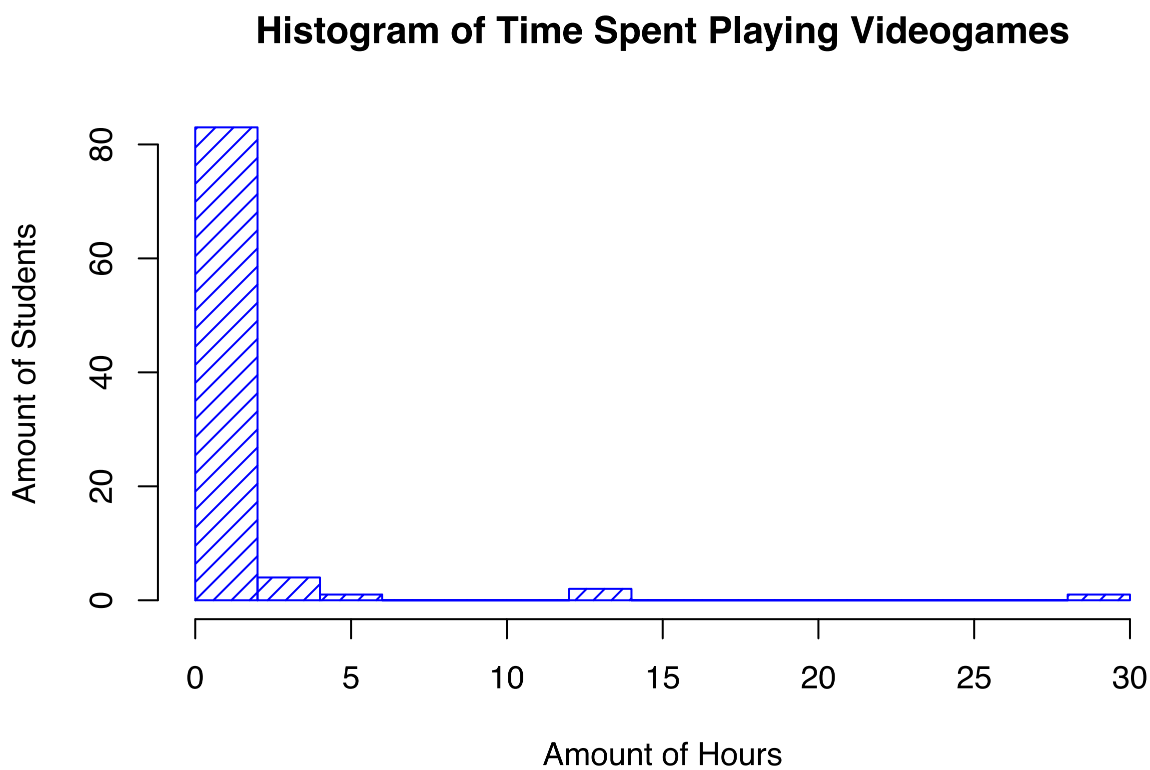
```r
mean(data.notbusy$time)
```

```
## [1] 0.5095238
```

Scenario 3

```r
#First we calculate the estimate for the # of students that played a video game:
nogame.ind <- which(data['time'] == 0.0) #Identify those who did not play video games the week prior
data.nogame <- data[nogame.ind,] #Create a data frame with no gamers
n1 <- length(data.nogame$time) #Calculates the # of students that played video games
#Calculates the proportion (# that don't play/sample size)
prop.nogame <- (n1)/91
#Calculates the sd of those that don't game
sd.prop.nogame <- sqrt( (.6263736)*(1-.6263736)/90 )*sqrt((314-91)/314 )
prop.nogame.ci <- prop.nogame + c(-1, 1)*2*sd.prop.nogame #Creates the CI

#Histogram of sample time spent playing
hist(data$time, main = "Histogram of Time Spent Playing Videogames", xlab = "Amount of Hours", ylab = "
     col = 4, density = 15, breaks = 15)
```




```
#Here we do Bootstrap
boot.population <- rep(data$time, length.out = 314) #Creates the population
sample1 <- sample(boot.population, size = 91, replace = FALSE) #creates the sample populations
B = 500 # the number of bootstrap samples we want
boot.sample <- array(dim = c(B, 91))
for (i in 1:B)
  {
  boot.sample[i, ] <- sample(boot.population, size = 91, replace = FALSE)
}
#Here we take the sample mean of each sample
boot.mean <- apply(X = boot.sample, MARGIN = 1, FUN = mean)
ci.boot <- c(quantile(boot.mean,0.025), quantile(boot.mean, 0.975))
```

Scenario 4

Getting proportion who likes games.

```
# Initializing variables corresponding to responses from students on the survey
likeVeryMuch <- 2
likeSomewhat <- 3
# Fetching all students who responded with positive game likeness
data.likeColumns <- which(data$like == likeVeryMuch)
data.likeColumns <- c(data.likeColumns, which(data$like == likeSomewhat))
# Calculating percentage
numOfLikes <- NROW(data.likeColumns)
proportionLike <- numOfLikes/data.samples
paste("Proportion of Like: ", proportionLike, sep="")
```

```
## [1] "Proportion of Like: 0.758241758241758"
```

Scenario 5

```r
# Using gmodels library
library(gmodels)

#Cross-Tabulation for owning a computer/like playing games
CrossTable(data$like, data$own)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  90
##
##
##              | data$own
##    data$like |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            1 |         0 |         1 |         1 |
##              |     0.267 |     0.097 |           |
##              |     0.000 |     1.000 |     0.011 |
##              |     0.000 |     0.015 |           |
##              |     0.000 |     0.011 |           |
## -------------|-----------|-----------|-----------|
##            2 |         5 |        18 |        23 |
##              |     0.209 |     0.076 |           |
##              |     0.217 |     0.783 |     0.256 |
##              |     0.208 |     0.273 |           |
##              |     0.056 |     0.200 |           |
## -------------|-----------|-----------|-----------|
##            3 |        16 |        30 |        46 |
##              |     1.136 |     0.413 |           |
##              |     0.348 |     0.652 |     0.511 |
##              |     0.667 |     0.455 |           |
##              |     0.178 |     0.333 |           |
## -------------|-----------|-----------|-----------|
##            4 |         1 |        12 |        13 |
##              |     1.755 |     0.638 |           |
##              |     0.077 |     0.923 |     0.144 |
##              |     0.042 |     0.182 |           |
##              |     0.011 |     0.133 |           |
## -------------|-----------|-----------|-----------|
```

```
##            5 |         2 |         5 |         7 |
##              |     0.010 |     0.003 |           |
##              |     0.286 |     0.714 |     0.078 |
##              |     0.083 |     0.076 |           |
##              |     0.022 |     0.056 |           |
## -------------|-----------|-----------|-----------|
## Column Total |        24 |        66 |        90 |
##              |     0.267 |     0.733 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```
#Cross-Tabulation for working/like playing games
CrossTable(data$like, data$work==0)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  87
##
##
##              | data$work == 0
##    data$like |     FALSE |      TRUE | Row Total |
## -------------|-----------|-----------|-----------|
##            1 |         1 |         0 |         1 |
##              |     0.518 |     0.506 |           |
##              |     1.000 |     0.000 |     0.011 |
##              |     0.023 |     0.000 |           |
##              |     0.011 |     0.000 |           |
## -------------|-----------|-----------|-----------|
##            2 |        14 |         9 |        23 |
##              |     0.609 |     0.596 |           |
##              |     0.609 |     0.391 |     0.264 |
##              |     0.326 |     0.205 |           |
##              |     0.161 |     0.103 |           |
## -------------|-----------|-----------|-----------|
##            3 |        22 |        21 |        43 |
##              |     0.026 |     0.026 |           |
##              |     0.512 |     0.488 |     0.494 |
##              |     0.512 |     0.477 |           |
##              |     0.253 |     0.241 |           |
## -------------|-----------|-----------|-----------|
##            4 |         3 |        10 |        13 |
##              |     1.826 |     1.785 |           |
##              |     0.231 |     0.769 |     0.149 |
##              |     0.070 |     0.227 |           |
```

```
##               |     0.034 |     0.115 |           |
## -------------|-----------|-----------|-----------|
##            5 |         3 |         4 |         7 |
##              |     0.061 |     0.060 |           |
##              |     0.429 |     0.571 |     0.080 |
##              |     0.070 |     0.091 |           |
##              |     0.034 |     0.046 |           |
## -------------|-----------|-----------|-----------|
## Column Total |        43 |        44 |        87 |
##              |     0.494 |     0.506 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```r
#Cross-Tabultion for sex/like playing games
CrossTable(data$like, data$sex)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  90
##
##
##              | data$sex
##    data$like |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            1 |         0 |         1 |         1 |
##              |     0.422 |     0.309 |           |
##              |     0.000 |     1.000 |     0.011 |
##              |     0.000 |     0.019 |           |
##              |     0.000 |     0.011 |           |
## -------------|-----------|-----------|-----------|
##            2 |         5 |        18 |        23 |
##              |     2.285 |     1.670 |           |
##              |     0.217 |     0.783 |     0.256 |
##              |     0.132 |     0.346 |           |
##              |     0.056 |     0.200 |           |
## -------------|-----------|-----------|-----------|
##            3 |        21 |        25 |        46 |
##              |     0.128 |     0.094 |           |
##              |     0.457 |     0.543 |     0.511 |
##              |     0.553 |     0.481 |           |
##              |     0.233 |     0.278 |           |
## -------------|-----------|-----------|-----------|
##            4 |         8 |         5 |        13 |
##              |     1.149 |     0.840 |           |
```

```
##              |     0.615 |     0.385 |     0.144 |
##              |     0.211 |     0.096 |           |
##              |     0.089 |     0.056 |           |
## -------------|-----------|-----------|-----------|
##            5 |         4 |         3 |         7 |
##              |     0.369 |     0.270 |           |
##              |     0.571 |     0.429 |     0.078 |
##              |     0.105 |     0.058 |           |
##              |     0.044 |     0.033 |           |
## -------------|-----------|-----------|-----------|
## Column Total |        38 |        52 |        90 |
##              |     0.422 |     0.578 |           |
## -------------|-----------|-----------|-----------|
##
##
```

Scenario 6

```r
#Chi-square test
observed <- c(31, 52, 8, 0)
expected <- c(.2, .33, .4, .1)
chisq.test(observed, p = expected, rescale.p = TRUE)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 57.942, df = 3, p-value = 1.617e-12
```