

# Case Study 2: Who Plays Video Games?

*Math 189 Homework 2*

Aditya Anandkumar

Andrea Cabezas

Bethan Lily Wynne-Cattanach

Carlo Mazzafaro

Jake Ehlers

Yi Ma

## Introduction

Well-designed video games can add tremendous value to a statistics course. We demonstrate this dynamic through a case study on 95 UC Berkeley students who were enrolled in Statistics 2 during the Fall of 1994. Understanding that video game interest strongly determines the practicality of an interactive lab, we assessed both the fraction of student gamers and the frequency of gameplay through point and interval estimates. The mentioned statistics were evaluated for both regular and pre-exam instances, and all reported values were verified through bootstrap procedures. We then drew excerpts from other studies that explain what students find most appealing about video games, and also highlight the differences between gaming and non-gaming students. That comparison is followed by a recommendation for game design that would add the most value to a statistics course, and also potential leads for further research.

## Summary of Examined Variables

	Play_Hours	Play_Like	Play_Where	Play_Frequency	Play_If_Busy	\
count	91.000000	90.000000	73.000000	78.000000	80.000000	
mean	1.242857	3.022222	2.972603	2.705128	0.212500	
std	3.777040	0.873811	1.092558	1.020682	0.411658	
min	0.000000	1.000000	1.000000	1.000000	0.000000	
25%	0.000000	2.000000	2.000000	2.000000	0.000000	
50%	0.000000	3.000000	3.000000	3.000000	0.000000	
75%	1.250000	3.000000	4.000000	4.000000	0.000000	
max	30.000000	5.000000	6.000000	4.000000	1.000000	

	Play_Educational	Sex	Age	Home_PC	Hate_Math	\
count	78.000000	91.000000	91.000000	91.000000	90.000000	
mean	0.474359	0.582418	19.516484	0.758242	0.322222	
std	0.502574	0.495893	1.846093	0.430521	0.469946	
min	0.000000	0.000000	18.000000	0.000000	0.000000	
25%	0.000000	0.000000	19.000000	1.000000	0.000000	
50%	0.000000	1.000000	19.000000	1.000000	0.000000	
75%	1.000000	1.000000	20.000000	1.000000	1.000000	
max	1.000000	1.000000	33.000000	1.000000	1.000000	

	Work_Hours_Prior_Week	Owns_PC	PS_CDROM	Have_Email	Expected_Grade
count	88.000000	91.000000	86.000000	91.000000	91.000000
mean	7.352273	0.736264	0.174419	0.791209	3.252747
std	10.313522	0.443099	0.381695	0.408697	0.607242
min	0.000000	0.000000	0.000000	0.000000	2.000000
25%	0.000000	0.000000	0.000000	1.000000	3.000000
50%	1.000000	1.000000	0.000000	1.000000	3.000000
75%	13.250000	1.000000	0.000000	1.000000	4.000000
max	55.000000	1.000000	1.000000	1.000000	4.000000

## Data and Method

Our data consisted of completed survey responses from 91 out of 95 randomly selected students. Questions in the survey assessed a selection of the students' habits, preferences, and traits that we believe pertain to gaming interest [Appendix]. In order to avoid bias from ignorance or negativity, the respondents who had never played video games or possessed a strong distaste were asked to skip a portion of the questions. If a question was not answered or improperly answered, then it was coded as 99. A follow up survey was also conducted, in which students were allowed to pick up to three answers, to gauge the game genres that were most appealing and the reasons behind the liking or disliking of games [Appendix].

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

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

Table 1 summarizes the types of games played.

Why?	Percent
Graphics/Realism	26%
Relaxation	66%
Eye/hand coordination	5%
Mental Challenge	24%
Felling of mastery	28%
Bored	27%

Table 2 summarizes reasons for playing the game.

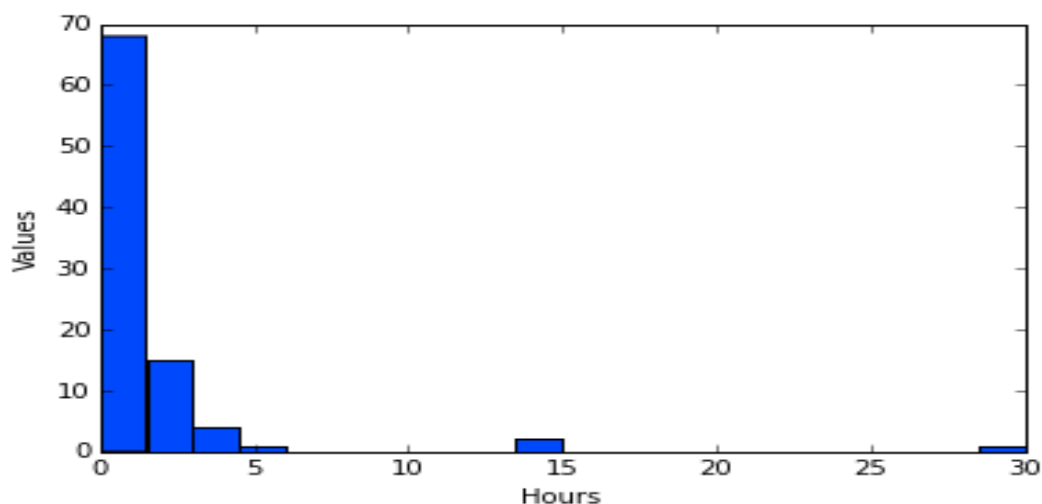
Dislikes	Percent
Too much time	48%
Frustrating	26%
Lonely	6%
Too many rules	19%
Costs too much	40%
Boring	17%
Friend's don't play	17%
It is pointless	33%

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

Table 3 summarizes what students didn't like about the games.

## Fraction of Students Who Play Video Games

From this data, we first separated the respondents who played video games the week prior to the study from those that didn't, to calculate the proportion of students that played the week before the study (0.307). We've included the proof of our estimators being unbiased in appendix 2. We then estimated the variance (0.213), and standard error of the proportion (0.41), and built a confidence interval. Using our results we established that the estimated number of students who played video games a week prior to the survey was in the following interval (0.225694, 0.389690). We also performed a bootstrapping procedure to validate our results. This was done by creating 3.25 bootstrap sample units for every sample unit, and then randomly drawing 91 samples from the 296 bootstrap samples to perform a subsequent series of repeated estimates that yielded a normal distribution of values. Our conclusive summary statistics from the bootstrap were a mean of 0.3104 and a standard deviation of 0.0406, which agree with our previously held results. Here's a histogram of the number of people versus hours played the week before the exam:



## Frequency During Exam Week

Students arguably enjoy video games more if they also play frequently during busy weeks. Therefore, we constructed an algorithm to compare gaming frequencies on random weeks with the frequency prior to an exam, which works as follows. First, the algorithm computes a probability of playing on a random week by assessing each students' frequency response. For example, if a student claimed to play once a semester, then they were given an associated probability of  $1/15$  because there are 15 weeks in a semester. Similarly, someone who plays monthly would have an associated probability of  $7/30$ , whereas someone who plays weekly or daily would sport a probability of 1 because they are expected to play at least once during the week. After gathering all probabilities, we multiplied the probabilities with the number of associated respondents in the sample. That weighted average is the expected number of players on a random week.

Sample size of respondents of frequency variable: 78

Frequency of Play	# of Entries
1	9
2	28
3	18
4	23

Where 1 = daily, 2 = weekly, 3 = monthly, 4 = Semesterly

Expected value of people who played at least once on the week before the survey: 42.206

Hours Played	# of Entries
0.0	57
0.1	1
0.5	5
1.0	5
1.5	1
2.0	14
3.0	3
4.0	1
5.0	1
14.0	2
30.0	1

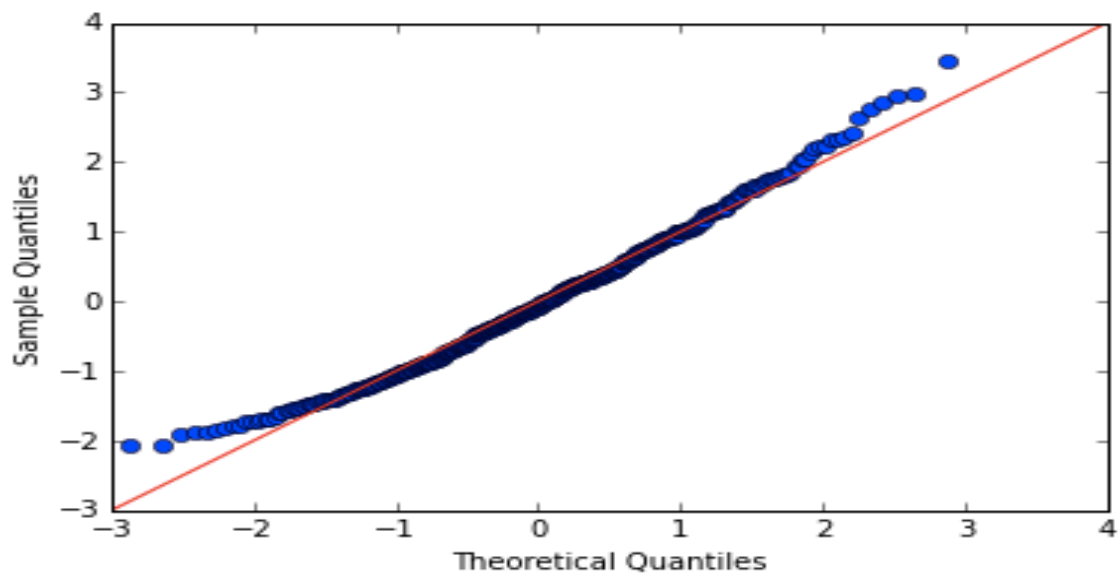
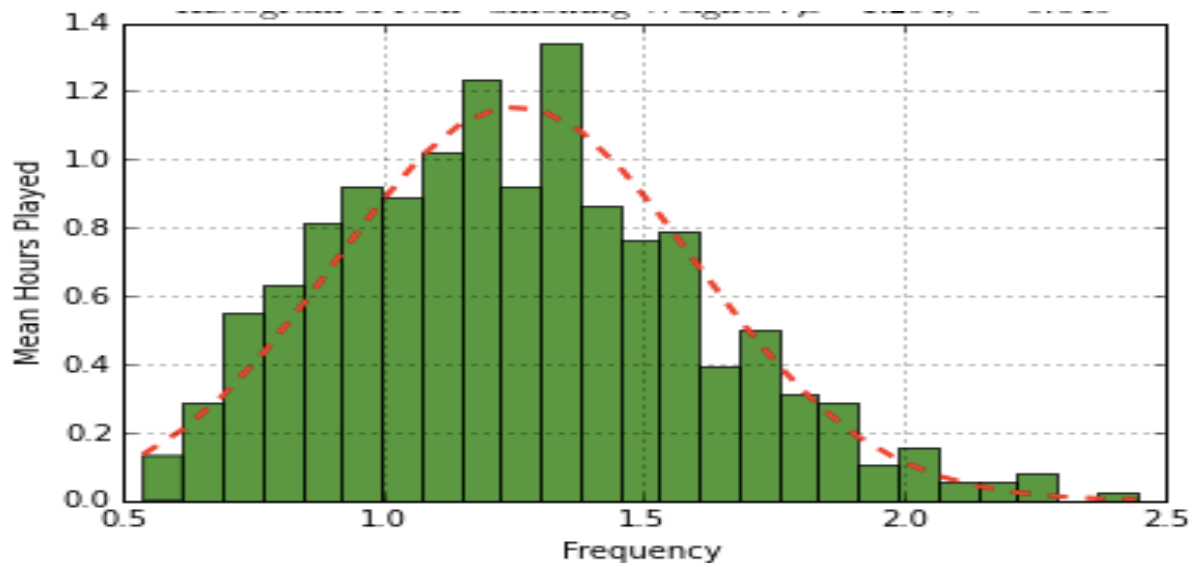
Now for the bootstrapped sample, we have:

Hours Played	# of Entries
0.0	228
0.1	4
0.5	20
1.0	20
1.5	4
2.0	56
3.0	12
4.0	4
5.0	4
14.0	8
30.0	4

Deploying the above algorithm, we arrived at 34.05, as the expected number of people who played video games at least once the week prior to the survey. We used our estimate to compare it to the actual number of students who played video games at least once the week prior to the survey. We assumed that due to the exam, less people played video games than usual. To guarantee a statistical difference, as well as a sound comparison with our previous data, we performed another bootstrapping procedure. This was done in similar fashion to the last, as we multiplied the data size by the same factor of 3.25, and repeatedly drew 77 random samples from the bootstrapped sample to eventually arrive at the mean of people playing video games on a random week as 42.206, and standard deviation of 3.476. These results are clearly larger than the values obtained for the week prior to the survey. We calculated a confidence interval of (26.2, 41.9) for people playing video games the week prior to the exam.

### **Average Hours Played in Week Prior to Exam**

The average number of hours that students spend playing games can also reflect interest. To gauge this statistic, we return yet again to the bootstrap procedure. This time, we granted every sample unit with 4 bootstrap units with identical value. We then drew 91 samples from the bootstrap dataset 500 times, and recorded the mean values in a vector. Assuming that 500 iterations is enough to guarantee convergence to normality, the vector should display a normal distribution. We tested this assumption using a qq plot and by fitting a normal line to a histogram distribution:



Our results from this bootstrap procedure were as follows:

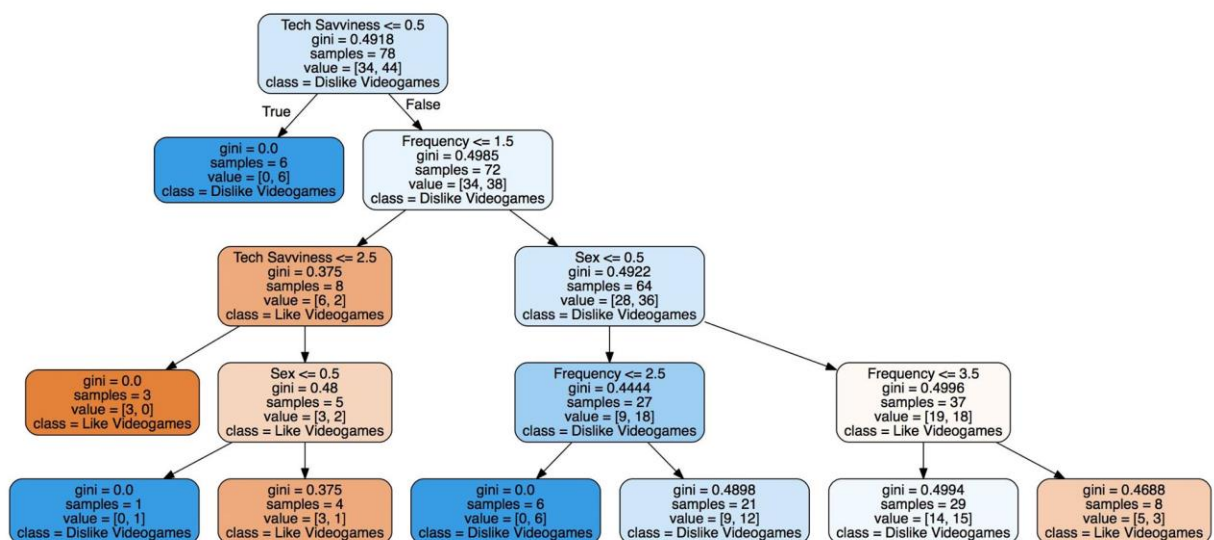
Mean of hours spent playing video games = 1.254070

Standard Deviation of the number of hours spent playing video games = 0.344706

Confidence interval of time spent playing video games in the week prior to survey: (1.5988, 0.9094)

## Comparing those who like and dislike video games

To help us understand the variables involved in making someone like or dislike video games, we built a classification tree. We were specifically interested in understanding the extent to which certain variables are more predicting of liking video games than others. We only selected few variables (frequency of play, sex and tech-savviness) instead of the entire data-set. This last variable, tech-savviness was created on a score system that awards points to students based on whether they have an email, have a PC, if their PC has a CD-drive, or if the response for the variable 'where do you play' is 3 or above. (Meaning the play only on a PC, or PC and arcade, or PC, arcade and a home system)



Since tech savviness was the #1 variable for predicting liking/not liking VGs (from the decision tree), we will compare that and sex to the likelihood of liking VGs through visual methods. We will also compare sex since research indicates that males play more than females.

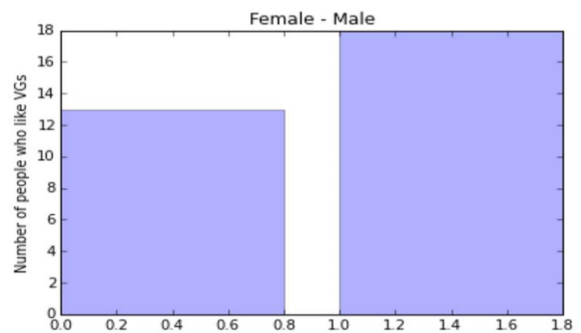
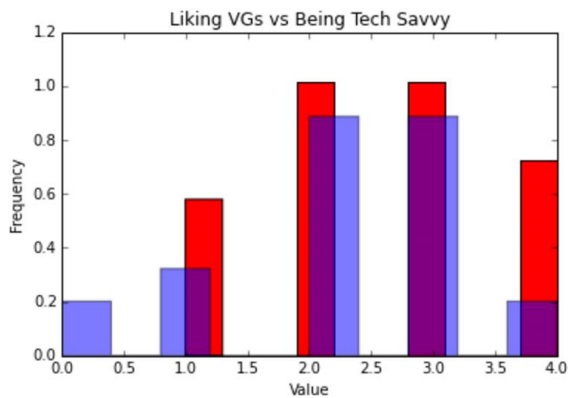


```
x = sum(Dislike_VG[0:23,3])
y = sum(Like_VG[:,3])
lenght = [x,y]

plt.bar(np.arange(2), lenght, alpha=0.3)
plt.ylabel('Number of people who like VGs')
plt.title('Female - Male')
```

```
Like_VG = Like_Savy_Grade_clean[(np.where(Like_Savy_Grade_clean[:,0]==1))]
Dislike_VG = Like_Savy_Grade_clean[(np.where(Like_Savy_Grade_clean[:,0]==0))]

plt.hist(np.ravel(Like_VG[:,1]), normed=True, color='r', label='Like VG and is Tech Savvy')
plt.hist(np.ravel(Dislike_VG[:,1]), histtype='stepfilled', normed=True, color='b', alpha=0.5, label='Like VG and is not Tech Savvy')
plt.title("Liking VGs vs Being Tech Savvy")
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.show()
```



## Analysis and Results

These statistics affirm college students' interest in video games and the practicality of designing an interactive lab. From our classification tree, we extrapolated that tech savviness is the biggest predictor of whether they like or dislike video games. If the parameter is below 0.5, resulting in a score of 0, it implies that they dislike playing video games. Furthermore, looking couple of levels lower in the tree, we were able to infer that women are more likely to dislike video games than men. When sex was below 0.5 (which is women), then it falls into the dislike category as you can see from the picture above. Also, if they're male and play frequently, they're predicted to like video games.

## **Why Do Students Like Video Games?**

When we asked students why they enjoyed playing video games, 66% cited the feeling of relaxation that they derive from gaming as the biggest appeal [Table 2]. That figure was followed by 28% of students claiming to appreciate the feeling of mastery they get from game progression, and 27% of students declaring that they simply used gaming as boredom relief. Our result is supported by the greater body of research. In Steve Jones' 2003 study of 1,162 college students across 27 universities, respondents cited the feeling of relaxation as one of their favorite things about playing video games. Some respondents went as far as to declare video games as an "after-class relaxation ritual." Likewise, Texas A&M professor Christopher Ferguson conducted a similar study on a sample of 103 college students in 2010, and asserted most students enjoyed video games as a method "to relax or to destress." Ferguson also offers a scientific explanation for why video games provide relaxation, suggesting that video games provide "mood management" by mitigating psychological processes that trigger depression and hostility. The logic of college students being drawn to emotional alleviation also makes sense, given that the students have been reporting record levels of stress – UCLA's 2010 national survey reports that the percentage of college students who claim to be "emotionally healthy" now stands at 51.9%, versus 63.6% in 1985. One important caveat for the context of our study is that students derived relaxation from video games particularly because the games allowed them to escape the stresses of school. Therefore, to truly leverage this statistic in the creation of the interactive platform, designers should be careful not to inject elements of learning that induce high stress, such as stringent deadlines or grading policies.

## Differences among Respondents

Using our classification tree we arrived at the conclusion that if sex was female, then the probability of them disliking the game was much higher than if they were male. This gender dichotomy runs parallel to the majority of comparable studies; the body of research citing women as playing fewer hours than men include and Green & McNeese in 2008, Chou & Tsai in 2007, and Buchman & Funk in 1996 (Parsons & David). The last study may offer the most relevant explanation for the pattern in our data given that it was conducted only one year later than ours. Buchman & Funk attribute the dichotomy to the social norms at the time – they discovered that it was more acceptable for males to play video games, and also that a negative correlation existed between females' gaming frequency and their self-esteem. Lucas and Sherry also cited that video games provided men with more of the draws outlined in their framework – men particularly derived much more social interaction benefits than women from video games, a result that runs parallel to Buchman & Funk's. Greenwood also offered the explanation that video games are simply designed for men, with male video game protagonists outnumbering their female counterparts by 3:1.

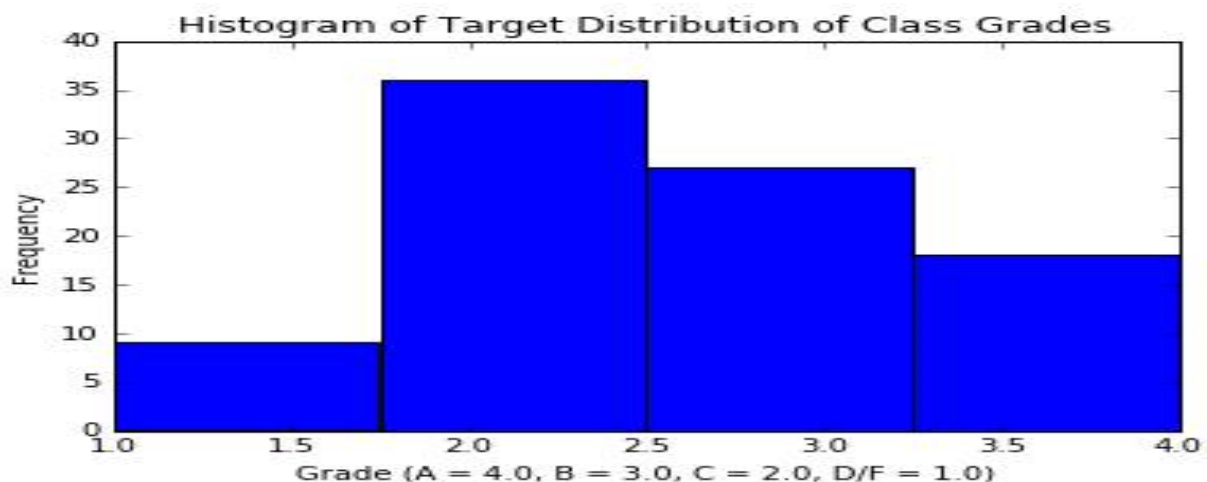
Our cross tabulations show that of the people who like video games only 5/23 were female while 18/23 were male. And of people who liked video games somewhat 21/46 were female, 25/46 were males – our data backs up what we've found and what has been found in the other studies that indicate a higher positive correlation between video game enjoyment and the individuals sex.

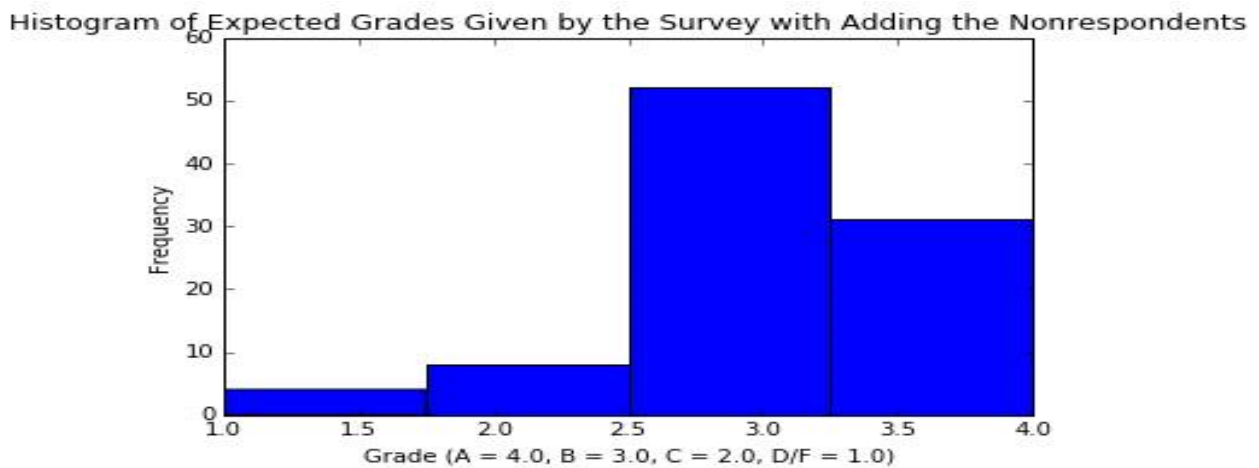
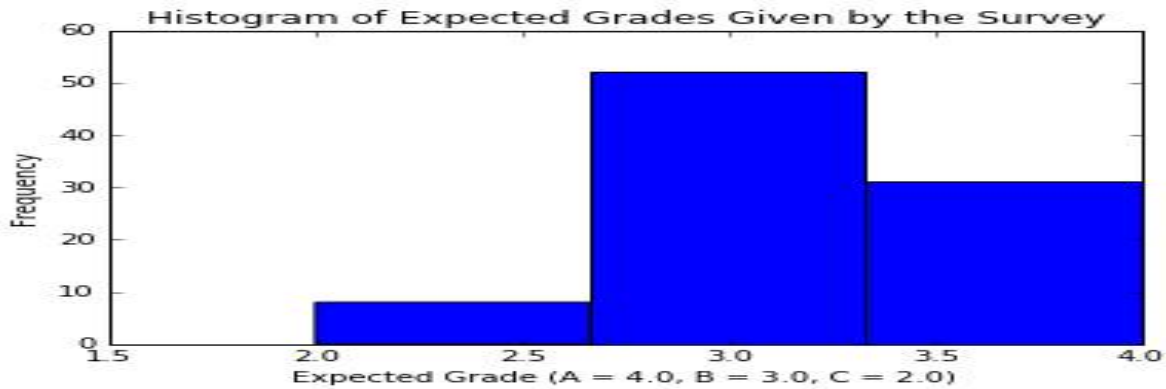
Work for Pay compared to whether they like games: The trend In the tables suggest that when people work less or no hours, a high proportion of them like or somewhat like playing

games more so than the people who work more frequently. Although since such a high percentage do not work at all comparison is not extremely accurate. Comparing people who own and don't own a PC: There is a clear correlation between people who own PCs and how much they enjoy video games with 18/23 people who like videos games, and 30/46 who somewhat like video games. Using our earlier analysis on tech-savviness this would make sense with our data.

## Grade Distributions

Using the survey data we find that the expected proportion of A's, B's, and C's is 34.07%, 57.14%, and 8.79% respectively. Comparing this to the target distribution of 20% A's, 30% B's, 40% C's, and 10% D's or lower, we see that more students expected higher grades than the target distribution allows for as well as far fewer C grades, showing in the histogram as being left-skewed. If we allow the nonrespondent students be added into the data as D or lower grades, our distribution of grades changes to 32.63% A's, 54.74% B's, 8.42% C's, and 4.21% D's or lower, however, this doesn't have a large impact on the proportions when we leave the nonrespondents out of the data. If anything, it reaffirms that more students expected higher grades than the targeted distribution allows.





## Recommendation for Designing the Interactive Platform

Leveraging the fact that the highest preferred gaming genre was “strategy [Table 1],” we recommend blending strategic gameplay with factors that trigger the elements students enjoy most about video games – relaxation and mastery [Table 2]. A particularly strong vehicle for such purposes is the Massive Multiplayer Online (MMO) genre, which teachers have expressed the most interest in for educational gaming (Kirriemuir, 21). The genre allows students to play simultaneously with hundreds of peers from anywhere in the world, making it an excellent platform for cooperative learning. MMO games have also been shown to provide therapeutic effects, with many players citing the virtual worlds within them as “escapes from the stresses of daily life

(Blaha, 6). Furthermore, the genre is characterized by empirical frameworks for progression where players are constantly rewarded for achieving higher levels of skill (Yee, 16), thereby granting students the “feeling of mastery” that our respondents claim to enjoy. The best part is that successful MMOs that reward players for their mastery of statistical concepts already exist; *EVE Online*, a free MMO that allows players to build space colonies using probability theory, recorded 146,000 subscriptions in 2015 (The Nosy Gamer). We recommend drawing inspiration from such commercially successful games while also infusing elements unique to the courses being taught in the design of the interactive platform. One last point worth noting is that our data suggests most college students play games on their home PCs, which is actually the dominant platform for MMOs. This makes it possible students to access the interactive platform outside the lab.

## **Conclusion**

Our summary statistics suggest that it would be lucrative to design an interactive lab. However, the applicability of our results did suffer from some setbacks, one of the largest being the divergence in gender gaming patterns between now and 1995. Whereas our sample displayed a comparatively small portion of female gamers, the present gaming landscape boasts a demographic that is 52% female (Guardian). The new demographics could have spawned changes in prevailing gamer preferences and habits, which in turn can affect gaming frequency and participation among college students. Likewise, the rise of consoles in modern gaming may have strengthened the social benefits students derive from gaming, which could also affect frequency and participation. Future studies should conduct similar tests with updated gender and “play location” statistics to assess the modern efficacy of our recommendation.

## Appendix

1) While estimating the proportion of students that played video games, we checked if there were any non-respondents before categorizing between those who played the week prior to the survey and those that didn't. We estimated the variance and standard errors using the formula indicated in the following code:

```
#Are there NaN values?
print 'Are there non respondents in this entry? %s' % np.isnan(np.min(Hours_Played))

#Separate catagorically between who played (hours played in week prior to survey>0), and who hasn't (hours played in week prior to survey = 0)
Played_Yes = Hours_Played[np.where(Hours_Played > 0)]
Played_No = Hours_Played[np.where(Hours_Played == 0)]

#Proportion
pi_1 = len(Played_Yes)/float(len(Hours_Played))

#For a proportion, the estimator for the variance is pi*(1-pi)
variance_1 = pi_1*(1-pi_1)

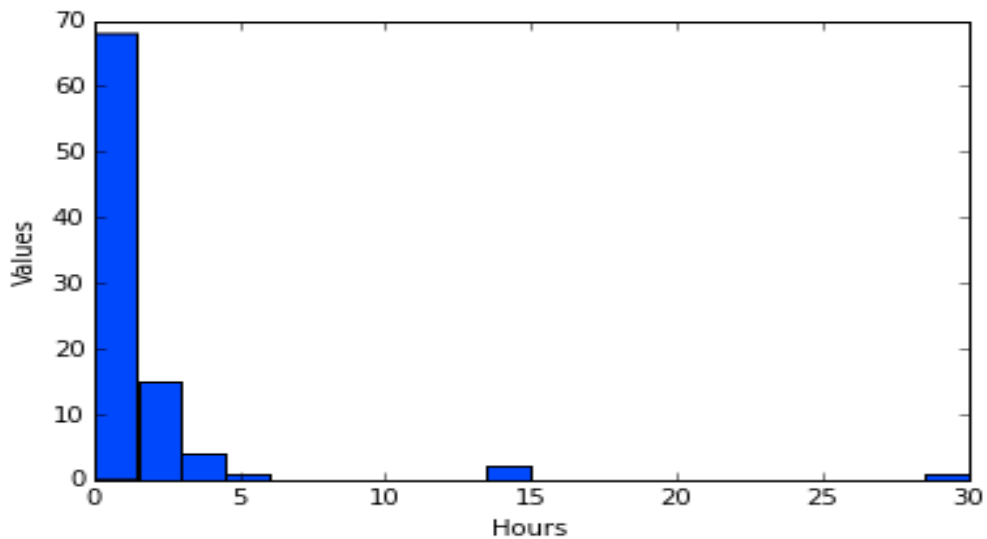
#For a proportion, the estimator for the standard error is: sqrt(variance/n-1) x sqrt((N-n)/N)
#For N = Population Size, n = Population Sample. Clearly, the sqrt((N-n)/N) represent the correction factor for small-sized samples
SE_estimate_1 = (math.sqrt((pi_1*(1-pi_1))/float(n-1)))*(math.sqrt((N-n)/float(N)))

print 'Unbiased estimate of the proportion of population who played videogames: %f' %pi_1
print 'Unbiased estimate of the variance of the porportion of the population who played videogames: %f' %variance_1
print 'Unbiased estimate of the standard error of the porportion of the population who played videogames: %f\n' %SE_estimate_1
```

We plotted a histogram of the number of hours students played video games the week prior to the survey.

```
plt.hist(Hours_Played, bins = 20, label='Hours Played')
plt.xlabel( "Hours" )
plt.ylabel( "Values" )
plt.show( )
```

USING



## 2) Finding the confidence interval.

95% confidence interval:  $\left[ \bar{x} - 1.96\sqrt{Var(\bar{x})}, \bar{x} + 1.96\sqrt{Var(\bar{x})} \right]$

$\bar{x}$  = Sample average, estimator for the population parameter  $\mu$  (the population average)

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad \bar{x} = \frac{1}{N} \sum_{j=1}^n x_{I(j)} \quad \text{where } I(j) = \text{values of the characteristics in our sample}$$

$\therefore$  Need to find  $S.E(\bar{x}) = \sqrt{Var(\bar{x})}$

First need to show that  $\bar{x}$  is an unbiased estimator for the population parameter:



$$\begin{aligned}
 E(\bar{x}) &= E\left(\frac{1}{n} \sum_{j=1}^n x_{I(j)}\right) \\
 &= \frac{1}{n} \sum_{j=1}^n E(x_{I(j)}) \quad \text{by linearity of the expectation} \\
 &= \frac{1}{n} \sum_{j=1}^n \left( \sum_{k=1}^N x_k P(I(j)=k) \right) = \frac{1}{n} \sum_{j=1}^n \left( \sum_{k=1}^N x_k \frac{1}{N} \right) \\
 &= \frac{1}{n} \sum_{j=1}^n \mu = \frac{n\mu}{n} = \mu
 \end{aligned}$$

$\Rightarrow E(\bar{x}) = \mu \Rightarrow \bar{x}$  is an unbiased estimator.

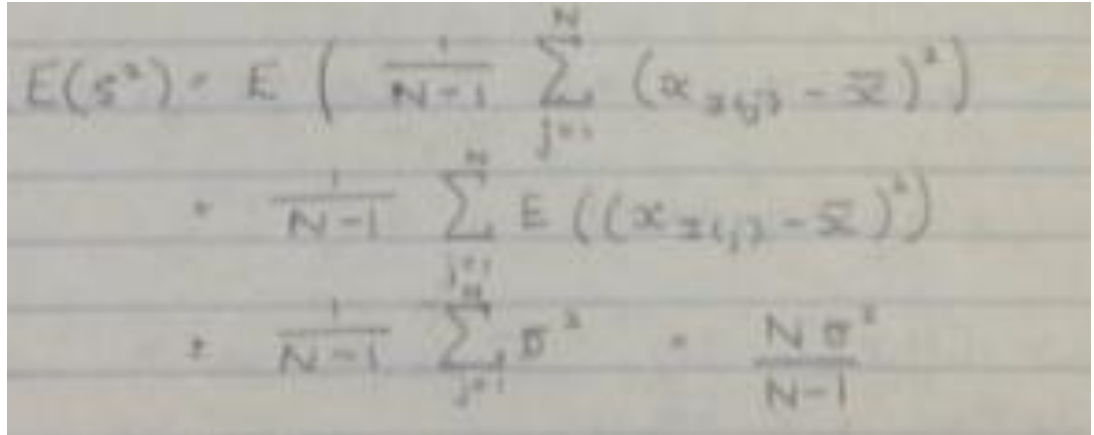
Now to find  $Var(\bar{x})$ :

$$\begin{aligned}
 Cov(x_{I(1)}, x_{I(2)}) &= \frac{1}{N(N-1)} \sum_{k=2}^N (x_k - \mu)(x_k - \mu) \\
 &= \frac{1}{N(N-1)} \left( \sum_k (x_k - \mu) \sum_k (x_k - \mu) - \sum_k (x_k - \mu)^2 \right) \\
 &= \frac{1}{N(N-1)} \left( \left( \sum_k x_k - \sum_k \mu \right) \left( \sum_k x_k - \sum_k \mu \right) - \sum_k (x_k - \mu)^2 \right) \\
 &= \frac{1}{N(N-1)} \left( (N\mu - N\mu)(N\mu - N\mu) - \sum_k (x_k - \mu)^2 \right) \\
 &= \frac{1}{N(N-1)} \left( - \sum_k (x_k - \mu)^2 \right) = - \frac{N\sigma^2}{N(N-1)} = - \frac{\sigma^2}{N-1} \\
 \therefore Var(\bar{x}) &= \frac{1}{n} \sigma^2 + \frac{n-1}{n} Cov(x_{I(1)}, x_{I(2)}) \\
 &= \frac{1}{n} \sigma^2 + \frac{n-1}{n} \left( - \frac{\sigma^2}{N-1} \right) \\
 &= \frac{(N-1)\sigma^2 - (n-1)\sigma^2}{n(N-1)} = \sigma^2 \left( \frac{(N-1) - (n-1)}{n(N-1)} \right) \\
 &= \sigma^2 \left( \frac{N-n}{n(N-1)} \right) \\
 \rightarrow S.D.(\bar{x}) &= \sqrt{Var(\bar{x})} = \sigma \sqrt{\frac{N-n}{n(N-1)}}
 \end{aligned}$$

When  $\sigma^2$  is unknown,  $s^2$  is often used as an estimator

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

$$\Rightarrow \frac{s^2(N-1)}{N}$$



Handwritten derivation showing the expectation of  $s^2$ :

$$\begin{aligned} E(s^2) &= E\left(\frac{1}{N-1} \sum_{j=1}^N (x_{I(j)} - \bar{x})^2\right) \\ &= \frac{1}{N-1} \sum_{j=1}^N E((x_{I(j)} - \bar{x})^2) \\ &= \frac{1}{N-1} \sum_{j=1}^N \sigma^2 = \frac{N\sigma^2}{N-1} \end{aligned}$$

is an unbiased estimator for  $\sigma^2$ .

By the plug in principal:

$$Var(\bar{x}) = \sigma^2 \left( \frac{N-n}{n(N-1)} \right) = \frac{s^2(N-1)}{N} \left( \frac{N-n}{n(N-1)} \right) = \frac{s^2(N-n)}{Nn}$$

$\therefore$  For a sample average  $\bar{x}$  the confidence interval is given by:

$$\left[ \bar{x} - 1.96s \sqrt{\frac{N-n}{Nn}}, \bar{x} + 1.96s \sqrt{\frac{N-n}{Nn}} \right]$$

For students who own PCs

$\tau = \sum x_i$  counts all students who own PCs in the population

$\pi$  = proportion of students who own PCs in the population

$\bar{x}$  remains an unbiased estimator of  $\pi$  (Let  $\bar{x} = \hat{\pi}$ )

$N\bar{x} = N\hat{\pi}$  estimates  $\tau$

A simpler form of the variance can be found in this case:

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

Then, by the plug in principle, an estimator for the standard error is

$$\widehat{SE}(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})(N - n)}{n(N - 1)}}$$

And the confidence interval is given by:

$$\left[ \hat{\pi} - 1.96\sqrt{\hat{\pi}(1 - \hat{\pi})} \sqrt{\frac{N - n}{n(N - 1)}}, \hat{\pi} + 1.96\sqrt{\hat{\pi}(1 - \hat{\pi})} \sqrt{\frac{N - n}{n(N - 1)}} \right]$$

3) We performed a bootstrap to validate results. For every unit in the sample, we made 1/(mean proportion = 0.307692) = 3.25 units in the bootstrap sample with the same time value. This gave us a tool to calculate the distribution of the frequency of play with greater precision. Hence, from a population of 91, we got a bootstrap population of  $91 \times 3.25 \sim 296$  samples from which we randomly chose 91 samples and calculated the proportion of people who actually played. This value is stored in a vector and the process was repeated enough times to guarantee a normal distribution of its values. Using these values we can calculate the mean and standard deviation of the proportion of those that played video games.

```
#Calculate number of people who played: if entry for # of hours played is zero, then he hasn't played, else, he has played
Played_Yes = Hours_Played[(np.where(Hours_Played > 0))]
Played_No = Hours_Played[(np.where(Hours_Played == 0))]
print 'The number of people who has not played is: %i' % int(Played_No.shape[0])
print 'The number of people has not played is: %i\n' % int(Played_Yes.shape[0])

#Multiply these values by 3.25
Non_Play_Bootstrap = int(63*3.25)
Did_Play_Bootstrap = int (28*3.25)

Play_Status_Bootstrap_Sample = np.hstack((np.zeros(Non_Play_Bootstrap), np.ones(Did_Play_Bootstrap)))

#Draw randomly 63 samples, 500 times
proportion_vector = np.zeros(500)

for i in range(0,500):
    sub_sample_play_status = np.asarray(random.sample(Play_Status_Bootstrap_Sample, 91))
    is_zero = np.where(sub_sample_play_status == 0.0)[0]
    is_one = np.where(sub_sample_play_status == 1.0)[0]
    proportion_vector[i] = (float(len(is_one))/91)

print 'Mean of the proportion of people who played video games = %.4f' % np.mean (proportion_vector)
print 'Standard Deviation of the proportion of people who played video games = %.4f\n' % np.std (proportion_vector)

print '\nThis result agrees with the previously held results.'
```

4) To calculate the expected number of people playing in a random week we calculated how likely it is for a random person to play on a weekly average. We devised an algorithm that's a weighted average of the frequency of play. i.e: people who responded that their frequency of play is once a semester will have an associated probability of playing on a random week of  $1/15$  (since the academic semester has 15 weeks). Similarly, someone who plays monthly will have an associated probability of  $7/30$ , someone who plays weekly will have an associated probability of  $1$ , and someone who plays daily will also have an associated probability of playing on a random week of  $1$ . (These are clearly approximations that follow the CLT for large samples). After such probabilities are computed, we get the expectation of the number of players for a random week by multiplying those probabilities by the number of respondents in the survey.

```
#Clean up frequency of play data
Is_NaN_Freq = np.isnan(All_Data_Array[:,3])
NaN_Indices_Freq = np.where(Is_NaN_Freq == True)[0]
Freq_Clean = np.delete(All_Data_Array[:,3], (NaN_Indices_Freq), axis = 0 )

#Get size of cleaned frequency data sample
print 'Sample size of respondents of frequency variable: %i\n' % Freq_Clean.shape[0]

unique_freq, counts_freq = np.unique(Freq_Clean, return_counts=True)

fr = PrettyTable(['Frequency of Play', '# of Entries'])
for i in range(0,len(unique_freq)):
    fr.add_row([int(unique_freq[i]), counts_freq[i]])
print fr

print '\nWhere 1 = daily, 2 = weekly, 3 = monthly, 4 = Semesterly\n'

EXP_Daily = (counts_freq[0] + float(counts_freq[1]) + float(counts_freq[2]/4.28) + float(counts_freq[3]/15))
print 'Expected value of people who played at least once on the week before the survey: %.3f' %EXP_Daily
```

Sample size of respondents of frequency variable: 78

5) Bootstrap to compare the average number of people who played video games the week prior to the exam to that of a random week. We randomly pick 77 samples (size of original data points, excluding non respondents) repeatedly and use this sample to calculate the average number of expected players in a random week.

Sample size of respondents of frequency variable: 78

Frequency of Play	# of Entries
1	9
2	28
3	18
4	23

Where 1 = daily, 2 = weekly, 3 = monthly, 4 = Semesterly

Expected value of people who played at least once on the week before the survey: 42.206

```
x = sp.stats.ttest_ind (np.random.normal(44.8,3.23,20), np.random.normal(34, 7.8,20))
print 'The t-statistic for the test is : %f, with an associated p-value of %.16f' % (x[0], x[1])
```

*#Std and mean for the first test stat was inferred from bootstrap, and for the second one was inferred from the confidence interval calculated previously*

6) Bootstrap to determine the average number of hours spent playing the week prior to the survey.

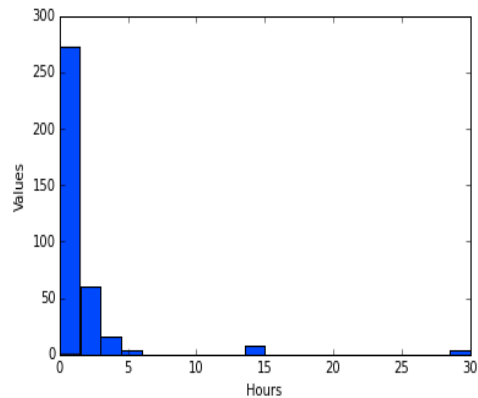
```
import random
from prettytable import PrettyTable
```

Hours Played	# of Entries
0.0	57
0.1	1
0.5	5
1.0	5
1.5	1
2.0	14
3.0	3
4.0	1
5.0	1
14.0	2
30.0	1

Now for the bootstrapped sample, we have:

Hours Played	# of Entries
0.0	228
0.1	4
0.5	20
1.0	20
1.5	4
2.0	56
3.0	12
4.0	4
5.0	4
14.0	8
30.0	4

```
#Redraw histogram of bootstrapped sample
plt.hist(bootstrap_sample, bins = 20, label='Hours Played')
plt.xlabel("Hours")
plt.ylabel("Values")
plt.show()
print 'Mean of bootstrapped sample: %.5f' % np.mean(bootstrap_sample)
```



Mean of bootstrapped sample: 1.24286

We drew 91 sub-samples from the larger bootstrap dataset previously created, 500 times. For each draw, its mean (which is an estimate of the mean hours played weekly) is calculated and stored into a vector. This vector should be normally distributed (assuming that 500 resamples are enough to guarantee convergence from the CLT). This assumption is tested using qq-plots and fitting a normal line to a histogram distribution.

```
#Pick random samples of size 91 from the bootstrap sample and draw a histogram of the means
mean_vector = np.zeros(500)
for i in range(0,500):
    mean_vector[i] = np.mean(random.sample(bootstrap_sample, 91))

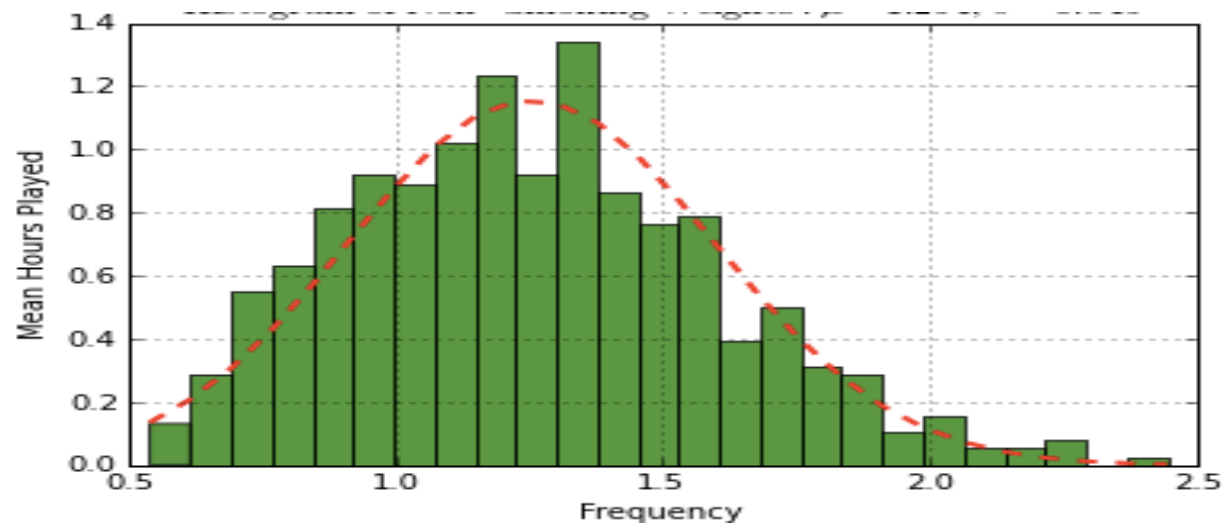
#parameters for best fit normal line
(mu, sigma) = norm.fit(mean_vector)

# the histogram of the data
n, bins, patches = plt.hist(mean_vector, 25, normed=1, facecolor='green', alpha=0.75)

# add a 'best fit' line
y = mlab.normpdf( bins, mu, sigma)
l = plt.plot(bins, y, 'r--', linewidth=2)

#plot
plt.xlabel('Frequency')
plt.ylabel('Mean Hours Played')
plt.title(r'$\mathrm{Histogram\ of\ Mean\ Hours\ Played:}\ \mu=%.3f,\ \sigma=%.3f$' % (mu, sigma))
plt.grid(True)

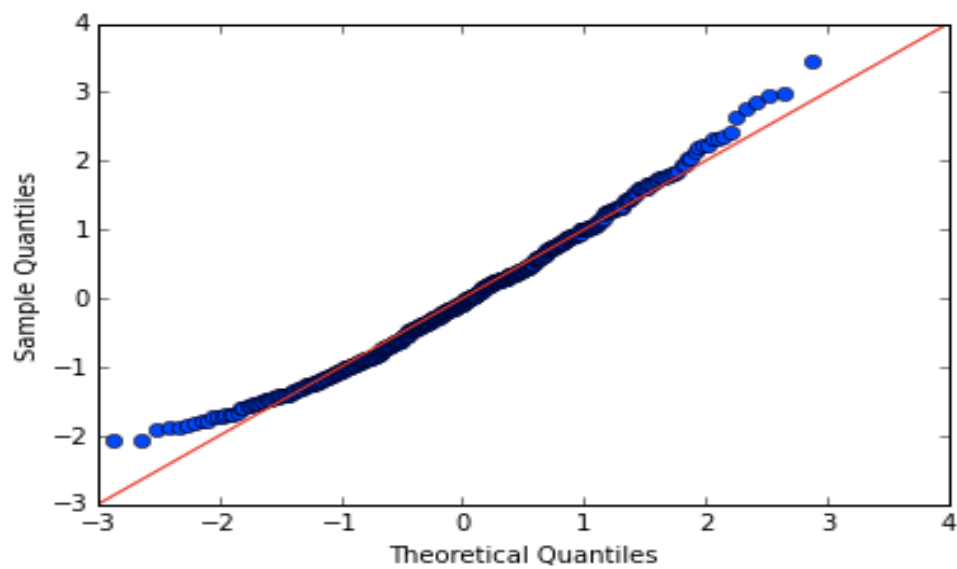
plt.show()
```



```
#qq-plot check for normality
```

```
import statsmodels.api as sm
from matplotlib import pyplot as plt
```

```
fig = sm.qqplot(mean_vector, stats.t, fit = True, line = '45')
```



## 7) Classification tree.

```
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn import tree
#Establish the score for where you play: being tech savy will give you score 1 for this question if response is 3 or greater
Where_Play_Score = np.zeros(91)
for k in range(0,91):
    if All_Data_Array[k,2] > 2:
        Where_Play_Score[k] = 1
    else:
        Where_Play_Score[k] = 0

#Stack up all the columns containing data (binary) that count for tech savviness
Tech_Savvy_Dataset = All_Data_Array[:, (11,12,13)]
Tech_Savy_Column = np.zeros(91)
Dataset_Score = np.nan_to_num(np.column_stack((Tech_Savvy_Dataset, Where_Play_Score, Tech_Savy_Column)))

#Sum all the entries from left to right and assign that number to the score column.
#Since we have solely binary data, a 1 will grant 1 one point, and a zero will grant zero points
#The score column will tell the degree of tech-savviness.

for l in range(0,91):
    Dataset_Score[l, 4] = Dataset_Score[l, 3] + Dataset_Score[l, 2] + Dataset_Score[l, 1] + Dataset_Score[l, 0]

Tech_Savy_Score = Dataset_Score[:,4]

#We assume that not responding on frequency and play if busy are signs of not playing. Hence will be better modeled as zero.
#The function np.nan_to_num will
Tree_Dataset = np.nan_to_num(np.column_stack((All_Data_Array[:,(3,4)], Tech_Savy_Score)))
Like_vector_clean = np.nan_to_num(All_Data_Array[:,1])

Tree_Dataset_Not_Clean = np.column_stack((All_Data_Array[:,1], All_Data_Array[:,(3,4)], Tech_Savy_Score))
x_delete = Tree_Dataset_Not_Clean[~np.isnan(Tree_Dataset_Not_Clean).any(axis=1)]
y_delete = x_delete[:,0]

dt = DecisionTreeClassifier(max_depth = 5, random_state=99)
dt.fit (x_delete,y_delete)

print x_delete
```



## Works Cited

David, and Parsons. "Innovations in Mobile Educational Technologies and Applications." Web.

Dube, Ryan. "Science Proves That Ng Video G Es Yous." *Makeuse*. Web. <<http://www.makeuseof.com/tag/video-game-stress-reduction-need-start-playing-right-now/>>.

Jayanth. "52% of Gamers Are Women - But the Industry Doesn't Know It." *The Guardian*. 2014. Web.

Jones, Steve. "Gaming Comes of Age." *Pew Research Center*. Web. <<http://www.pewinternet.org/2003/07/06/gaming-comes-of-age/>>.

Kirriemuir, John. "Use of Computer and Video Games in the Classroom." Web.

Lucas, and Sherry. "Sex Differences in Video Game Play." (2004). Web.

Yee, Nicholas. "The Psychology of Massively Multi-User Online Role-Playing Games." (2006). Web.

"How Many Subscriptions Does EVE Online Have In 2015?" *The Nosy Gamer*. Web. <<http://nosygamer.blogspot.com/2015/03/how-many-subscriptions-does-eve-online.html>>.

"Why Are College Students Reporting Record High Levels of Stress?" *Time*. Web. <<http://healthland.time.com/2011/01/27/why-are-college-students-reporting-record-high-levels-of-stress/>>.

"Gender and Media: Content, Uses, and Impact." Web.