# Suggestions for Computer Lab Design Given Survey on Video Games

Mitesh Gadgil, Saurabh Kulkarni, Kyle Kole

February 8, 2016

# 1 Introduction

Every year, 3000 to 4000 students enrol in statistics courses at UC Berkeley. Half of these students take introductory statistics courses to satisfy quantitative reasoning requirement. To aid the instruction of these students, a committee of faculty and students have designed a series of computer labs. The labs are meant to extend the traditional syllabus for a course by providing an interactive learning environment that offers students an alternative method for learning the concepts of statistics and probability. Some have linked labs to video games. To help committee design the labs a survey of undergraduate students who were enrolled in a lower-division statistics course was conducted. The survey's aim was to determine the extent to which the students play video games and which aspects of video games they find most and least fun.

## 1.1 Data Collection

Students who were enrolled in advanced statistics course conducted the study. They developed the questionnaire, selected the students to be sampled and collected the data. In this study you will have the opportunity to analyze the results from the sample survey to offer advice to the design committee.

# 2 Analysis

A series of 6 analysis will be performed on this data. We begin with estimation of the summary statistics.

## 2.1 Scenario 1

### 2.1.1 Estimation

The aim of statistical estimation is to arrive at the value of population parameters using a sample of the population. This is done because at most times gathering data about all units in the population is intractable and hence analysis is done on a sample collected from the population. Following are a few terminologies which are required to build this theory of estimation:

Population: The collection of subjects with a particular attribute about which we want to conduct analysis and obtain inferences or recommendations. In this particular case, the population is the set of students who would be taking the introductory statistics course at UC Berkeley.

Parameters: There are certain things that we would like to know about the population which would aid us in our decision making. These measures are computed over the entire population and are called parameters. Since, populations are usually very large gathering data for the entire population and computing values of parameters from this data is not viable. In the scenario at hand, the parameter is the proportion of students who played video games in the week prior to the survey.

Sample: In order to estimate the parameter of interest of a population we collect a sample i.e. gather data from a small subset of the population which is diverse enough to be representative of the actual population. One method of sample collection is called survey.

Statistic: The value of the parameter of interest computed over all data present in the sample is called a statistic. Hence, note that the statistic is a random variable and depends on the sample collected, while the parameter is a constant value. In this scenario, the proportion of students in the sample who played video games in the week prior to the survey is the statistic.

Sample statistic = population parameter + bias + chance variation

In the above relationship, we would like to estimate the parameter value from the statistic value. We can eliminate the bias term in this relation by choosing a representative sample using various sampling techniques so as to not exclude any particular class of the population from the sample. We cannot eliminate the chance variation, however we can get quantify this chance variation in the form of a confidence interval.

### 2.1.2 Point Estimation of a Parameter

A point estimate of a parameter is a single value of a statistic. For example, the sample mean x is a point estimate of the population mean $\mu$. Similarly, the sample proportion p is a point estimate of the population proportion P.

In the particular dataset we have 91 sample points and 34 of those have played video games in the week prior to the survey i.e. the time attribute of 34 samplep points is greater than zero. Thus the value of the sample proportion $p = 0.4359$ is the point estimate for proportion of students who played video games last week.

### 2.1.3 Interval Estimation of a Parameter

An interval estimate is defined by two numbers, between which a population parameter is said to lie. For example, a < x < b is an interval estimate of the population mean $\mu$. It indicates that the population mean is greater than a but less than b.

Confidence Interval is a terminology used frequently to quantify how confident we are that the population parameter lies in the interval that we estimate. A confidence interval is typically stated as an interval of sample statistic (point estimate) ± margin of error with x% confidence. It can be interpreted as, if we were to take all possible random samples of size 91 from the population of students in the statistics class and then calculate an interval estimate, then x% of those calculated intervals would contain in them the true population parameter. Said differently, we can be at least x% sure that the interval we have computed contains the true parameter.

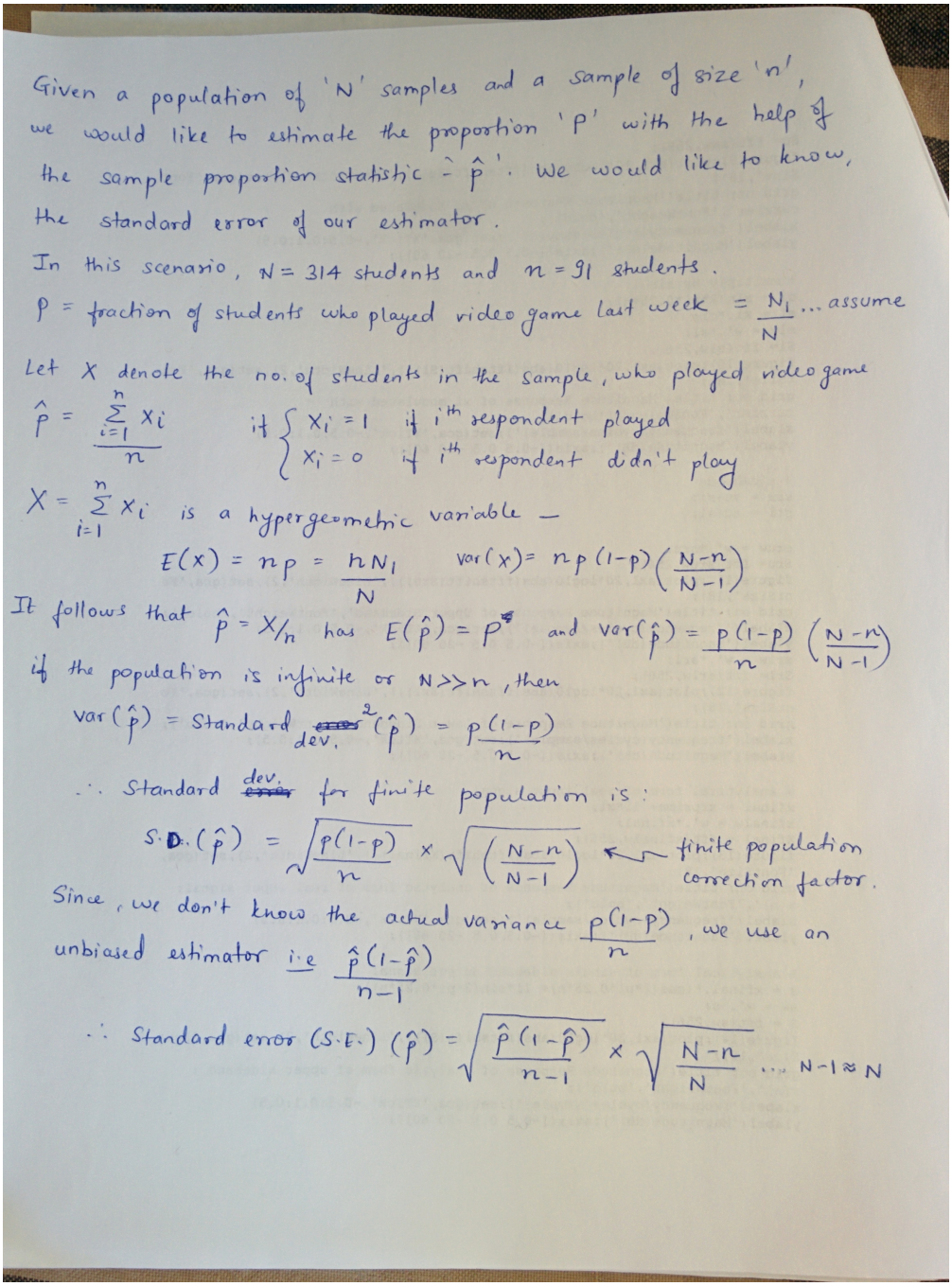
Calculation of the margin of error involves the standard deviation of sample statistic and a multiplier factor which is determined by the confidence level we wish to have. Let us look at this process and the concepts involved:

Sampling distribution: As we have seen before, the statistic is a property of the specific sample and hence will vary from sample to sample. The probability distribution of a statistic like sample mean helps us determine the mean and variance of our statistic and how good of an estimate it is for the population parameter. This probability distribution of the sample statistic is called sampling distribution. The central limit theorem helps us comment on the sampling distribution under some criterion.

Central Limit Theorem: The central limit theorem states that the sampling distribution of the mean of any independent, random variable will be normal or nearly normal, if the sample size is large enough. Usually, a sample size of 30 is large enough if the underlying population has a normal distribution or 40 if the underlying population has a different distribution.

Standard deviation and standard error: The central limit theorem helps us quantify the interval in which the sample statistic (sample mean or sample proportion) will lie with a certain degree of confidence. If the sampling distribution is normal, then we can use the standard deviation and z-score to compute the confidence interval. If the standard deviation of population is unknown, then standard error is used to approximate this quantity and used to compute the t-score using a t-distribution. The standard error computation depends on the relative sizes of sample and population, called the sampling fraction n/N.

Finite population correction factor: If the population is assumed to be infinite then the samples drawn after the first few samples have been drawn are not correlated to the ones drawn already. However if the population is finite then we can expect a correlation between what is to come and what has gone before. This bias is accounted for by the finite population correction factor. This correction comes up especially in the standard error of sample statistics like sample mean, proportion. Standard error of sample proportion: 0.0433




Calculation of interval: Let us first choose a confidence level with which we would like to predict the interval, say 95%. Now we will look at the critical value i.e. what is the value in terms of a normal curve with mean 0 and variance 1, that will cover 95% of the area under the curve i.e. value with cumulative probability of 0.025 on each side. This value turns out to be 1.96, which can be interpreted as follows: Give a standard normal curve, we can say that 95% of all samples generated will lie in the interval [-1.96 1.96]. Similarly, for our normal distribution with mean and standard error given by [0.3736 0.0795], a 95% confidence interval would be [0.3736 + 1.96*0.0795] = [0.3736 + 0.1558].

#### 2.1.4 Conclusion

Point estimate of fraction of students who played video game in the week prior to the survey is 0.3736264. Interval estimate of fraction of students who played video game in the week prior to the survey is $0.3736264 \pm 0.1558651$ with 95% confidence level.

## 2.2 Scenario 2

As we can see from the summaries of the data samples for each of the reported playing frequencies namely- daily, weekly, monthly, semesterly, there seems to be a good correlation between the reported frequency and the actual time spent playing video games. The mean for each of these categories is monotonically decreasing as we would expect to see. The mean for daily category is around 4.5 hrs, which is close to what we would expect and so also is the mean of 2.5 hrs for the weekly category. The monthly and semesterly categories show a sub-zero average, which is understandable since the probability that students of these categories played last week is bound to be low.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   2.000   4.444   4.000  14.000
```

Figure 1: Summary Statistics for Daily Players

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.500   2.000   2.539   2.000  30.000
```

Figure 2: Summary Statistics for Weekly Players

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.00000 0.05556 0.00000 0.50000
```

Figure 3: Summary Statistics for monthly Players

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.00000 0.04348 0.00000 1.00000
```

Figure 4: Summary Statistics for semesterly Players

The purpose of this comparison would be to cross verify if there is correlation between the playing behaviour of the sample students and the amount of time that they played video game last week. If for example, we observed a lot of students who usually play video games daily or weekly who didn't play the last week owing to exams, then this would mean our survey hasn't captured the general playing habits of the students. Instead what we have is a sample which is affected by its time of collection i.e. proximity to the examination. Any analysis made with such a dataset and the recommendation or inferences drawn wouldn't necessarily generalise to the population on a normal week and hence fail its purpose.

Since, there seems to be no effect of the examination on the playing habits of the students in the past week in spite of exams, we can be confident that the dataset can be used to make useful recommendations and analysis. Our estimate of the fraction of students playing video game last week won't be affected since the sample itself is unaffected by the fact that it was taken at a time which was unique in the sense that there was an examination soon.

### 2.2.1 Conclusion

The data stating frequency of playing and the average time spent playing video game last week is consistent although it was the week prior to examinations, and hence all analysis, estimates and recommendations drawn using this data are valid and generalizable to all times. This check of validity was essential since the data was collected at a time which could have possibly affected the sample data that we observe. Hence, checking if the data obtained is independent and general and not a result of particular external factors like timing, sampling bias, etc. is essential.

## 2.3 Scenario 3

In statistics, a confidence interval (CI) is a type of interval estimate of a population parameter. It is an observed interval (i.e., it is calculated from the observations), in principle different from sample to sample, that frequently includes the value of an unobservable parameter of interest if the experiment is repeated. Confidence intervals consist of a range of values (interval) that act as good estimates of the unknown population parameter; however, the interval computed from a particular sample does not necessarily include the true value of the parameter. When we say, "We are 99% confident that the true value of the parameter is in our confidence interval", we express that 99% of the hypothetically observed confidence intervals will hold the true value of the parameter. After any particular sample is taken, the population parameter is either in the interval realized or not; it is not a matter of chance.

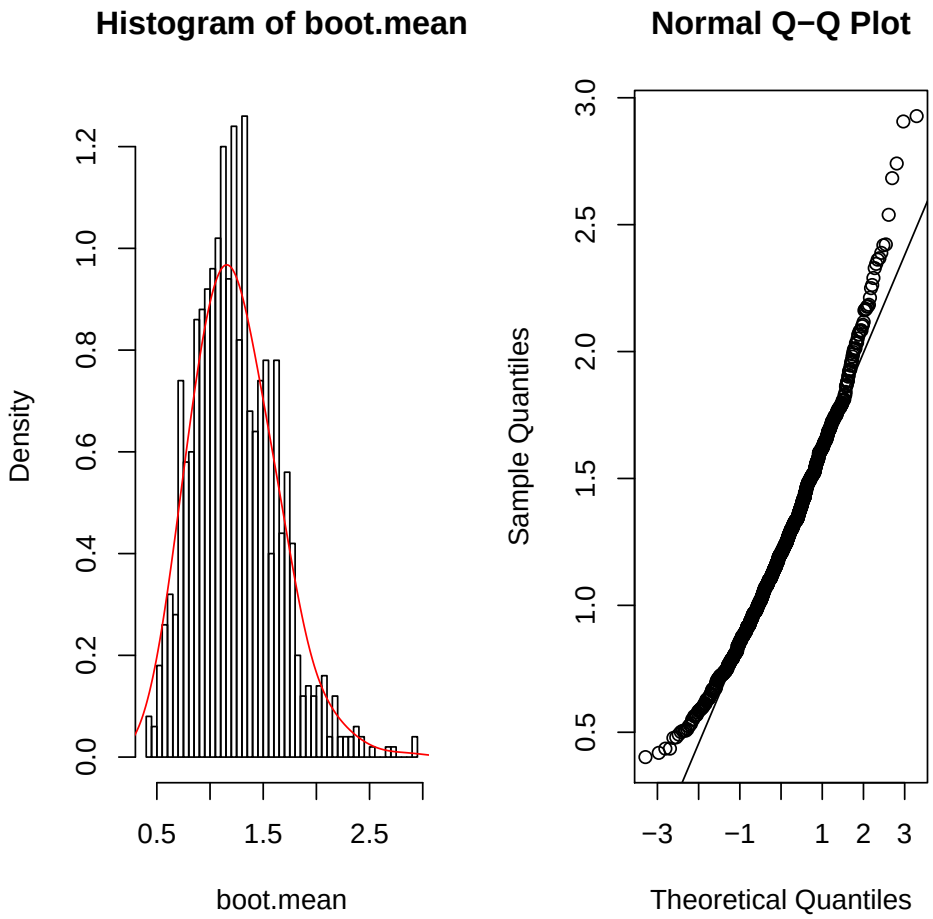
In order to find the average amount of time spent playing video games in the entire dataset in week prior to the survey we use three approaches.

1. Using Bootstrap to increase the number of random samples (without using fpcf)
2. Using the finite population correction factor
3. Using T estimation

### 2.3.1 Using Bootstrap and Normal Approximation

In statistics, bootstrapping can refer to any test or metric that relies on random sampling with replacement. Bootstrapping allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates. This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods. Generally, it falls in the broader class of resampling methods.

Using bootstrap the confidence interval is [0.4680685, 2.0452648]. The standard error is 0.3975.




```
[1] 0.4966399 2.0166934
```

### 2.3.2 Finite Population Correction Factor

The normal distribution can be used to provide confidence intervals for the population parameter. If the sample size is large, then the probability distribution of the sample average is often well approximated by the normal curve, by the central limit theorem. However, the normal approximation can still hold if, in addition to the sample size being large, it is not too large relative to the population size. In this example n is large but n=N = 91=314 is not small. We will use bootstrap to check if we can use normal approximation. Using normal approximation we get that a 95% confidence interval will have z value 1.96

The central limit theorem and the standard errors of the mean and of the proportion are based on the premise that the samples selected are chosen with replacement. However, in virtually all survey research, sampling is conducted without replacement from populations that are of a finite size N. In these cases, particularly when the sample size n is not small in comparison with the population size N (i.e., more than 5% of the population is sampled) so that n/N $>$ 0.05, a finite population correction factor (fpc) is used to define both the standard error of the mean and the standard error of the proportion. Finite Proportion Character: $f_{pc} = \sqrt{\frac{N-n}{N-1}}$

The confidence interval using the normal approximation with finite population correction factor is [0.5942862, 1.9190472], the standard error is 0.3379.

### 2.3.3 T-test

The t.test() command of R is a useful one for finding confidence intervals for the mean when the data are normally distributed with unknown variance. The degrees of freedom, confidence interval are all provided in the summary after implementing the function.

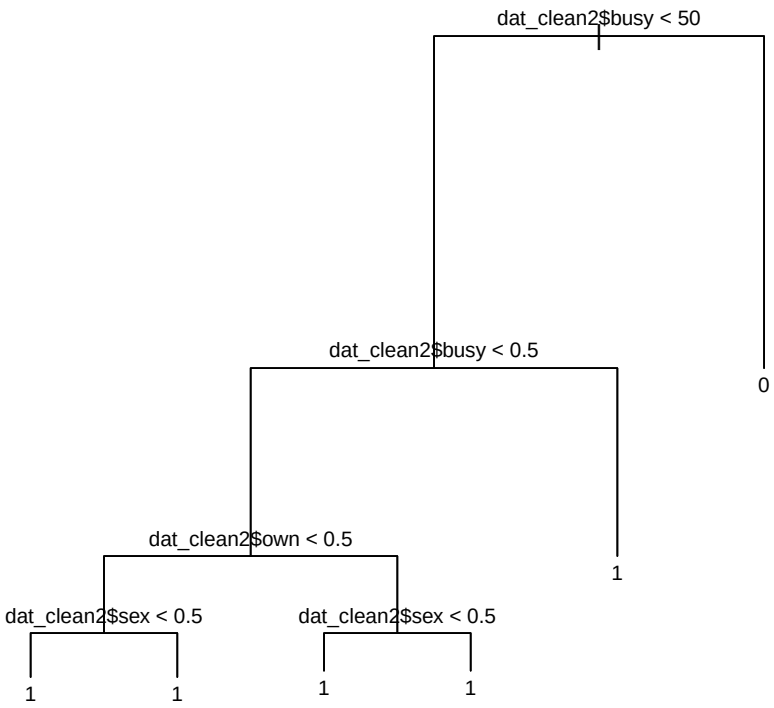
```
        One Sample t-test

data:  dat_clean$time
t = 3.1407, df = 89, p-value = 0.002288
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.4616331 2.0517003
sample estimates:
mean of x
 1.256667
```

Using Bootstrap and T-test, we see that the confidence intervals are close enough. Using Normal approximation with finite population correction factor we get a slightly different answer. This is because the data is not ideally Gaussian as the number of samples is very less and the population is finite and small as well. Bootstrapping leads to a better result.

## 2.4 Scenario 4

Next consider the "attitude" questions. In general, do you think the students enjoy playing video games? If you had to make a short list of the most important reasons why students like/dislike video games, what would you put on the list? Don't forget that those students who say that they have never played video games or do not at all like video games are asked to skip over some of these questions. So, there may be many non-respondents to the questions as to whether they think video games are educational, where they play video games, etc.

Classification Tree: Classification and regression trees are machine-learning methods for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree. Classification trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost.




```
Classification tree:
tree(formula = (as.factor(dat_clean2$like)) ~ (dat_clean2$own) +
    (dat_clean2$busy) + (dat_clean2$sex), data = dat_clean2,
    mindev = 0.001)
Number of terminal nodes:  6
Residual mean deviance:  0.8401 = 70.57 / 84
Misclassification error rate: 0.1778 = 16 / 90
```

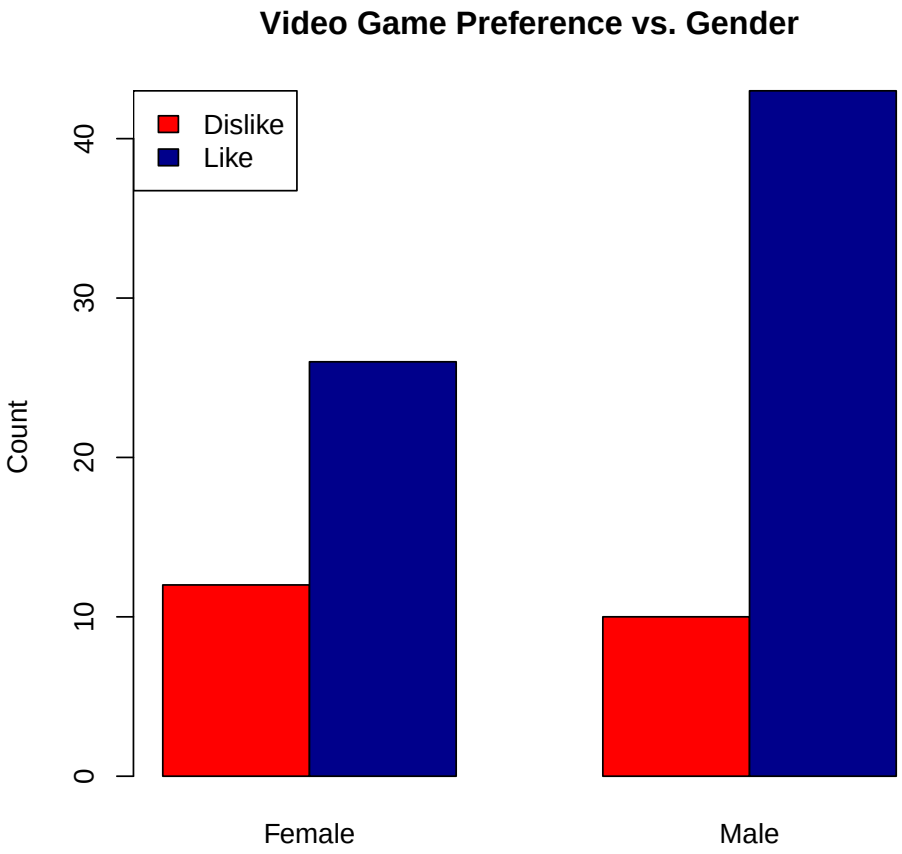
Here we have built a tree to observe that which factors lead to a student enjoying video games. We use the key factors on how busy the person is, does he own a PC or not and how educated the person it etc. For the first question we see that people who dislike video games (labeled = 0) have not answered question to whether they were busy or not and hence are bifurcated in the first split. (Note that they were NOT removed while cleaning data as they were key subset consisting of point who specifically dislike video games) We see that majority of the classes of the leaves imply that majority of kind of students like video games whether they are busy or not, or whether they own PC or not. The misclassification rate is relatively low (17.78

## 2.5 Scenario 5

Consider the differences in number of respondants for those who like and dislike video games versus several attributes. We compare the population based on preferences for video games versus gender, working for pay, and owning a computer. In particular, we are searching for glaring differences in our sample and trying to determine strong attributes contributing to an individual's preferences to video games.

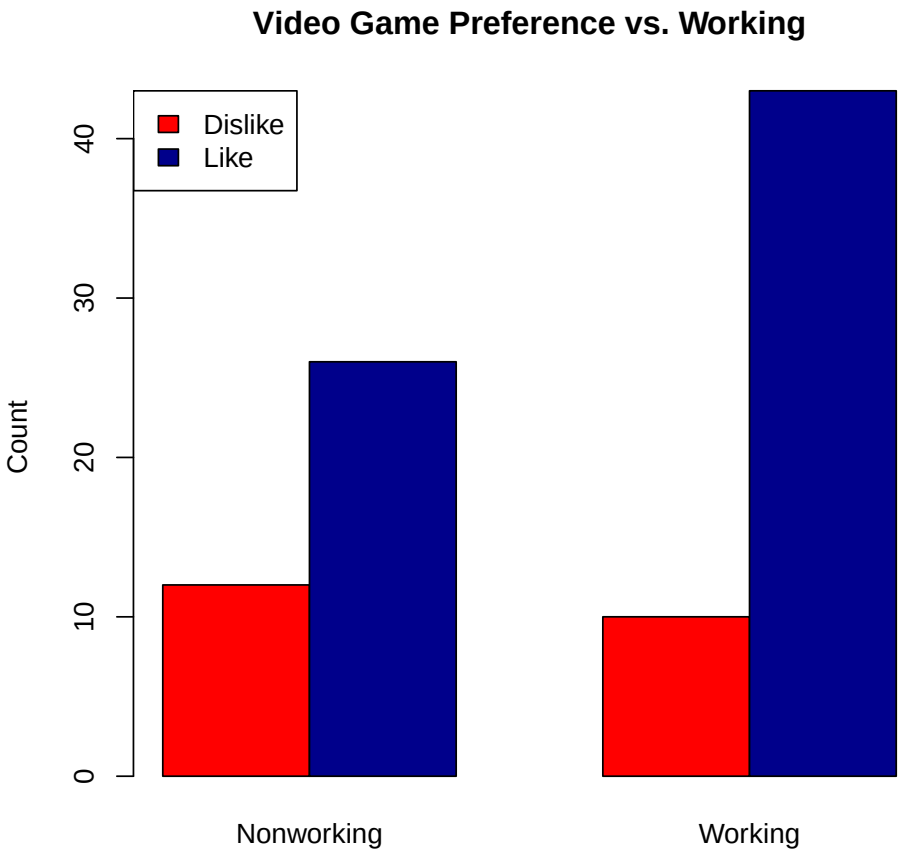
```
         Female  Male
Dislike  13.19 10.99
Like     28.57 47.25
```

Figure 5: Percentages for Preferences versus Gender



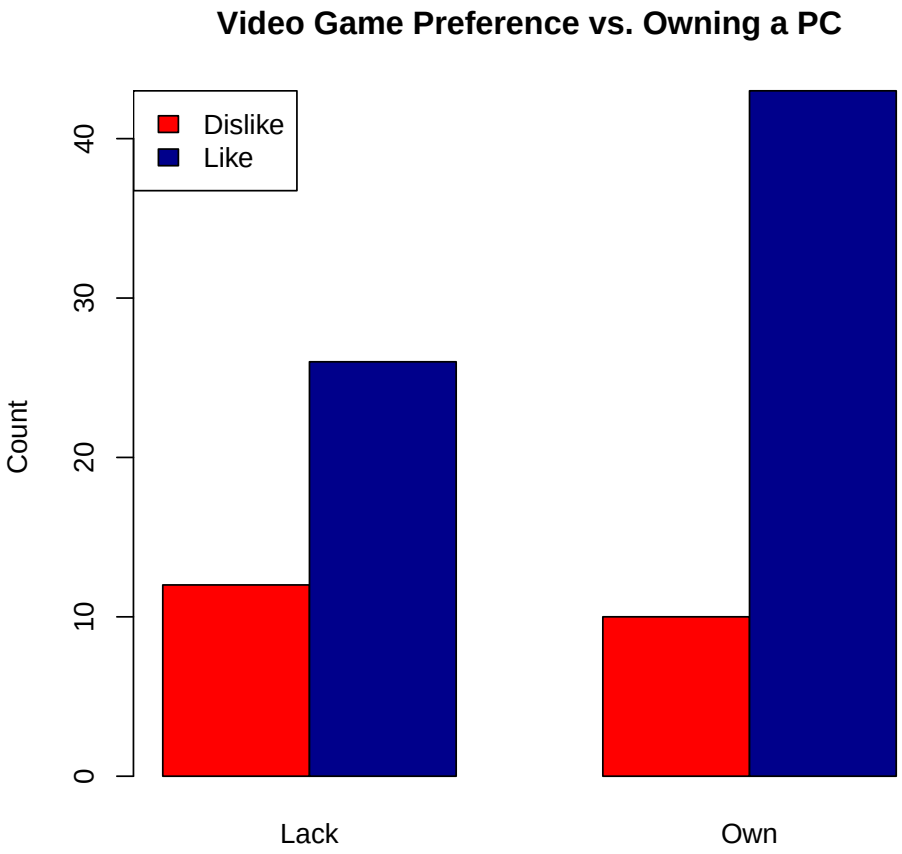
```
         Nonworking  Work
Dislike       15.38  8.79
Like          36.26 39.56
```

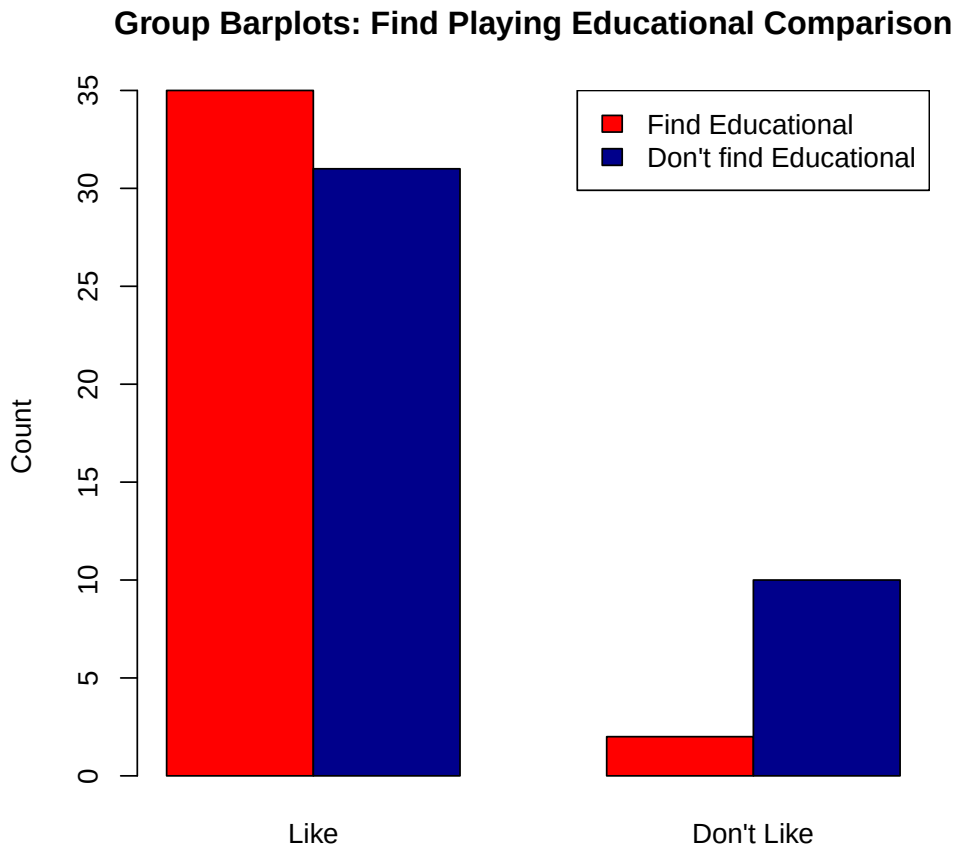
Figure 6: Percentages for Preferences versus Working



```
          Lack   Own
Dislike   3.30 20.88
Like     23.08 52.75
```

Figure 7: Percentages for Preferences versus Owning a PC






As we can see, the differences between liking and disliking depending on a particular attribute looks significant. Hence, these attributes most likely contribute to liking to play video games. Unfortunately, independent of whether or not the individual likes to play video games, the survey data suggests there is an even split between finding video games educational or not. In light of this, our analysis focuses on whether designing a computer lab based on attributes of video games will be conducive to a more enjoyable lab session. The educational factor of video games will be satisfied through the fact that the lab session is for the purpose of education.

### 2.6 Scenario 6

```
     A      B      C   D/F
 34.07  57.14   8.79  0.00
```

Exploring the question of expected grades, we have summarized the number of students who expect a certain letter grade. Notice how none of the respondants expect to fail the class (get a D or lower). Hence, it is clear that the distribution of expected grades is not at all close to the common grade assignment distribution of 20% A's, 30% B's, 40% C's, and 10% D or below. This is easily explained by the fact that none of the students take a class expecting to fail. The fact that there are no students with expected grades of D or lesser points to the fact that there is a response bias in the sample. The fact that failing students might not have attended the class for collecting their exam papers or the class following that session might be responsible for this bias. Hence, the data collected and the inferences drawn from the same apply to students who are doing decently well in the course and exclude the students who might not be doing too well. This non-representation of students is not desirable and should be rectified to make better and accurate recommendations for designing the lab for all students- those doing well and those doing not so well.

## 2.7 Conclusions

In scenario 2, we do a sanity check of the data to see if the fact that the survey was collected during the week the examination was held has affected the data. While analysing data and drawing actionable inference from it, we would like the data to be generalisable and not a rare sample. Hence, we verified whether there was a correlation between the playing frequency reported by students and the time they spent playing video game the week prior to the survey. This helped us establish that the data obtained can be used reliably.

In scenario 3, using Bootstrap and T-test, we see that the confidence intervals are close enough. Using Normal approximation with finite population correction factor we get a slightly different answer as the data is not ideally Gaussian as the number of samples is very less and the population is finite. We further see from scenario 4, the classification tree, from the bar plots that people are inclined to like video games, but the follow up survey insists that it is not because of educational feature of the game, or the sex of the students or whether he has a PC at home or not. People who like games, want them to challenge them mentally as well as help them relax. Lastly we see that the major drawbacks of playing games which are mentioned in the follow-up survey are essentially some which we can overcome during our design of the lab.

In conclusion for scenario 4 and 5, we see from the classification tree, from the bar plots that people are inclined to like video games, but the follow up survey insists that it is not because of educational feature of the game, or the sex of the students or whether he has a PC at home or not. People who like games, want them to challenge them mentally as well as help them relax. Majority enjoy playing games involving strategy. Lastly we see that the major drawbacks of playing games which are mentioned in the follow-up survey are essentially some which we can overcome during our design of the lab.

In Scenario 6, we see that all students expect higher grades than the expected distribution. This can mean two things, either the students have a biased view of their performance in the exam in the prior week. Or else, when the survey was conducted the student's who did not perform well were not present for the survey.

## 2.8 Suggestions for the Computer Lab

Given the analysis on the survey data, we have come to a few recommendations on how to design the computer lab. Our recommendations include:

1. The lab should be designed for computer games which involve strategy relevant to the subject.
2. Strategy games can be designed based on experimentation and optimization, which can pose as mental challenges for students.
3. The games should not be purely educational (like a quiz), they should involve some aesthetic components like graphics to attract students.
4. Since the games would have a constructive purpose, many students would feel that they are utilizing their time well and would also help them learn.