Investigation on Surveys for Designing Educational Computer Labs

Chen, Mengting A14782590

Dai, Jixi A91084063

Xiang, Yixiao A92070053

Zhang, Haoran A92036029

Zhang, Su A91115828

Table of Content

0. List of Figures

1. Introduction

Background

Every year, 3, 000 to 4, 000 students enroll in statistics courses at UC Berkeley and half of them take introductory statistics courses to satisfy quantitative reasoning requirements. A committee of faculty and students aim to design a series of computer labs to facilitate instructions and help students study. The labs they intending to design will be an extension of traditional courses and provide an interactive learning environment for learning concepts. Parts of the labs are in the form of video games. The committee wants to better understand the preference of students and effects of video games so a survey of undergraduate students who were enrolled in a lower-division statistics course was conducted. This survey contains three parts and collects information that we will use in this study. First part of survey asks questions regarding frequency of playing video games, where to play, number hours spent on video games and several related aspects. Second part of the survey aims to study whether students like or dislike playing games, what type of video games do they prefer and why. It contains three questions that are different from first part as more than one response may be given. Details of last three questions and responses are listed in the appendix. Last part of the survey collects general information of participants, such as gender and age. The main purpose of this report is to determine the extent to which the students play video games and which aspects of video games they find most and least fun. The information data set utilized in this investigation is "videodata.txt", containing 91 objects, each has 15 features.

Data

Our study uses one data set named "videodata.txt". The set of data is drawn out of 314 students in Statistics 2, Section 1, UC Berkeley, during Fall 1994, while 95 were selected at random to participate in the survey. To make sure students in the survey participated long enough in the course, students on the selection lists are those who had taken the second midterm. The information contained in "videodata.txt" is related to 15-feature vectors. Factors include time, likeness of play, where to play, how often, play if busy, playing educational, sex, age, computer at home, hateness of math, work time, own PC, PS have CD-Rom or not, having email or not, and grade expected. These features are indicated point by point in categorization. For all features in the set of data, the number "99" represents the certain type of question was not answered or improperly answered, so we exclude these data to do the study.

Purpose

The purpose of this study is to investigate the extent to which the students play video games and rank the aspects of video games they find most and least interesting. By investigating the

intention and reasons of playing video games from participants, this study can provide useful information to help to design an interactive learning lab linked to video games.

Limitation

1. There is an artificial correlation involved in our study because most data in dataset are categorical data. Correlation between data variables may not a good measurement and can hardly make some tests for categorical variables.
2. Sample size is small that can hardly represent the universal results.
3. Voluntary questionnaire should be polled with care due to research biases such as non-responses or deceptious responses.

2. Investigation

Scenario 1:

At the beginning of this scenario, we divide the data into two sets, whether played or not played the video game in the week before the survey. There are 34 students who played in the week, the rest of them did not. By a point estimation in the sample, we can get the fraction which is 34/91, equals to 0.3736. Because of this rate, we use it into the success rate of population 'N' 1000. Thus we will have 1000 students in total, and take 300 observations as 'n' from the population 'N'. To get an interval estimate for the fraction, we use the sample mean to subtract the result of critical z score multiple the estimator for the standard error to get the lowest value of this interval, at the same time, we use the sample mean to add the result of critical z score multiple the estimator for the standard error to get the highest value of this interval. By this process, we get the interval which is (0.3144791, 0.4055209).

There is another way to get that interval and fraction, Bootstrap Estimate. To start with the bootstrap, we set the population to N=314, and the sample size will be 91. During the bootstrap estimate, one data can also be sampled from the sample, as a replacement. As the population of bootstrap, we can get the bootstrap samples via taking 1000 times, random samples in size 'n'. After getting these bootstrap samples, we can calculate the sample means, as meanwhile, we can get the distribution via the histogram, figure 1.1. From the histogram, we can find out that the most sample means are mainly distributed within 0.35 to 0.45, which is kinda similar to the interval we get previously.

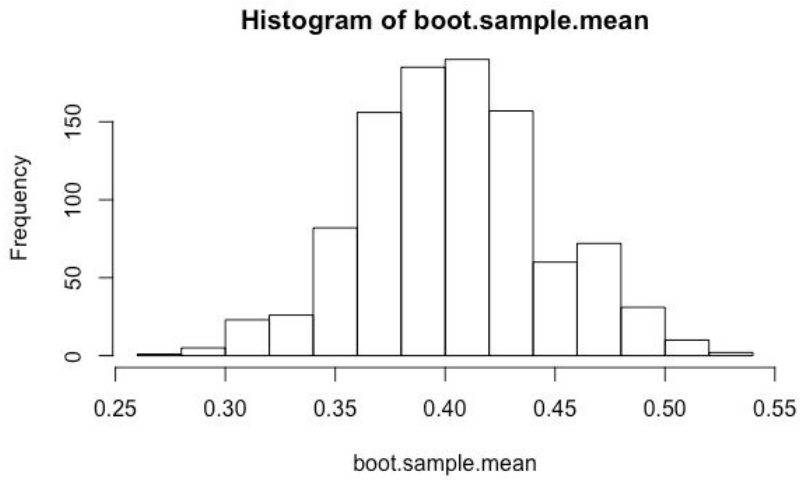

Figure 2.1.1

By the calculation of bootstrap, we will get the point estimate is the mean of bootstrap samples, which will be 0.4038681. At the meantime, we can get a confidence interval by sample mean of bootstrap sample subtract its standard deviation to represent the lowest value of the interval, and the sum of the sample mean from the bootstrap sample and its standard deviation to get the highest value of this interval. (0.3197750 0.4879612) is the result of this method.

There is another way of calculating the interval estimate using the bootstrap, which will be 'simply extract the 0.025-quantile and 0.975 quantiles of the bootstrap sample means and arrive at an interval estimate' from the definition. As the shown below, (0.3186813 0.4835165) will be the result of this method, and we can find out the tiny difference between this one and the previous one.

2.5%	97.5%
0.3186813	0.4835165

In this case, we can easily figure out that sample mean has a certain difference between these methods, which the point estimate can only stand for tiny sample and population, however, by using the bootstrap, we will have the sample mean which can represent the value of the fraction. After calculating these three different methods to get interval estimate, we find out that all of them only have a tiny difference, and set to around 0.3 to 0.4. That is, we can assume that most fraction of students who played the video game is around this percentage plausibly. In other words, we cannot hundred percent trust our database, it may only 30 to 40 percent reliability within this data.

Scenario 2:

Now we are going to compare the time spent playing video games a week before the survey with the reported frequency of play in the survey, in other words, will reported frequency of play affect participants' time spent playing video game a week before the survey.

First, we can see from the boxplot of these two variable that the only notable difference between these distributions is that means of time spent playing video games a week before the survey for participants who have a daily frequency and who have a weekly frequency are slightly above 0 and larger than means of time spent playing video games a week before the survey for participants who have a monthly frequency and who have a semesterly frequency.

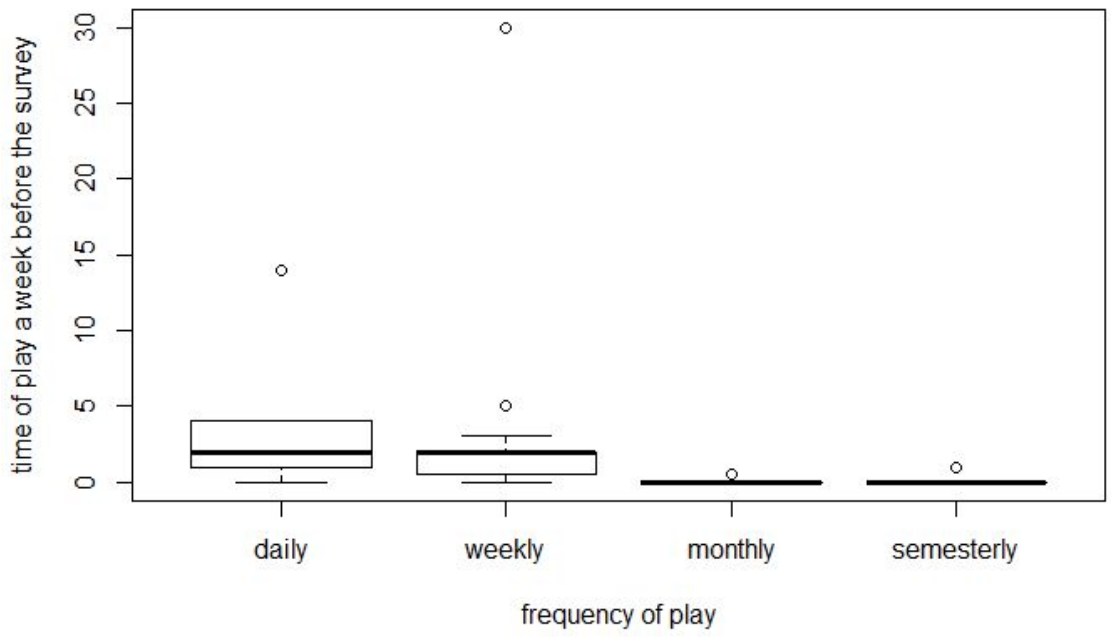

Figure 2.2.1

Then histograms further show that there's no strong relation between frequency of play and time spent playing video games a week before the survey as overall distributions are similar as most density gathers on the left tail in these four histogram.

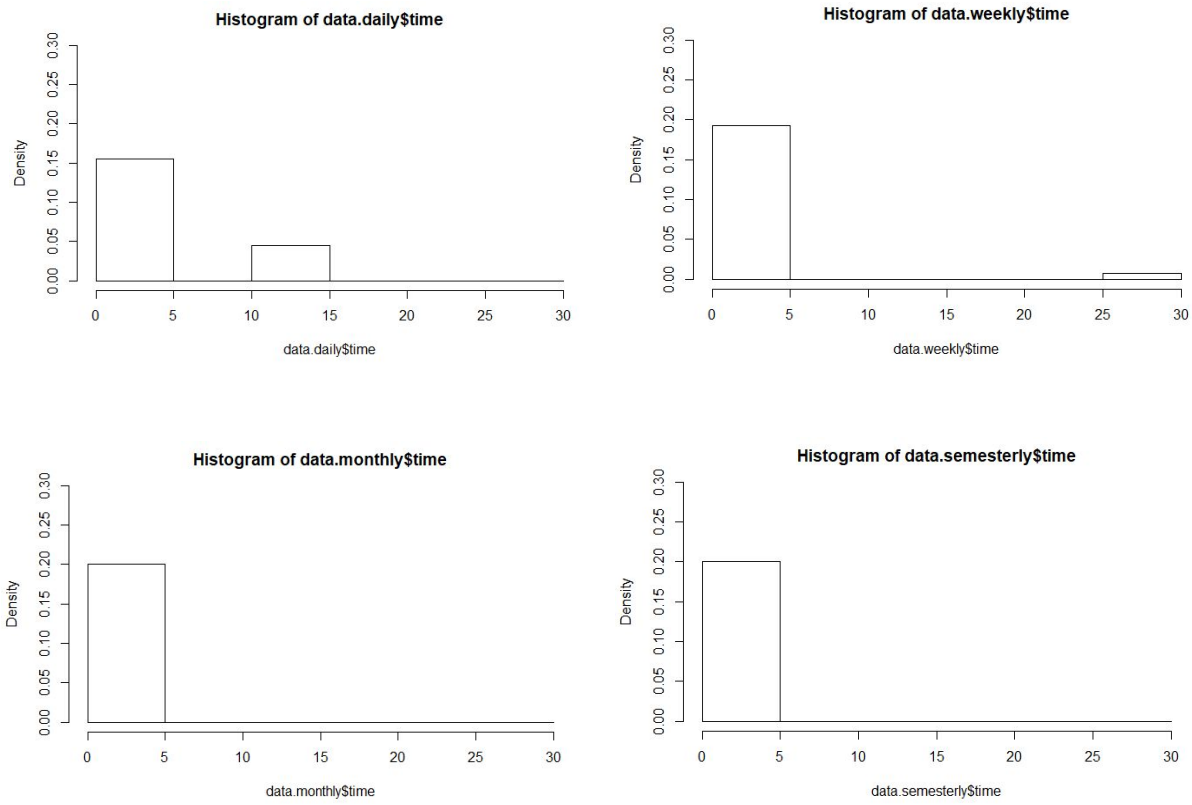

Figure 2.2.2

However, since the frequency of playing video games represents the general pattern of time spent on playing video games as more frequently participants play the game more time they will spend on it. And in this case, this general pattern is violated. If, in fact, there is a exam taking place a week before the survey, then the presence of violation is reasonable. Here, we use bar plot to show the overall pattern of participants' frequency of play and since in the dataset 4 stands for semesterly, which is the least frequent, and 1 stands for daily, which is the most frequent, the scale of horizontal axis should be reverse so that it's ranged from the least frequent to the most frequent on the horizontal axis . We will use bar plot to show the time spent playing video games a week before the survey as well.

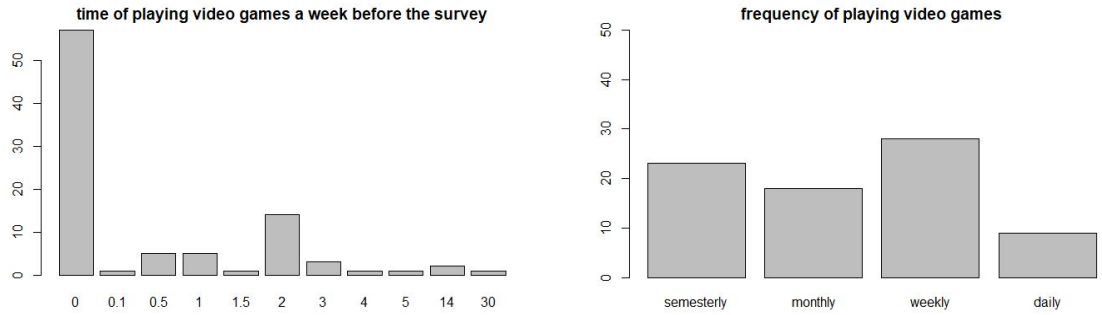

Figure 2.2.3

As we can see in the box plots, most of the participants didn't play video game at all or significantly reduced playtime a week before the survey as the 'evenly distributed' bar plot on the right, which represents the general pattern of playtime, becomes heavily left-tailed on the left. So we can interpret this change in this way: as exam taking place in the week before the survey, participants gave up their playtime to prepare for the exam.

So, the fact that there was a exam a week before the survey made participants' playtime on video games significantly decreased and in turn caused the time spent on playing video games a week before the survey failed to reflect the general pattern of frequency of playing video games. And as the 'evenly distributed' frequency of playing video games became heavily left-tailed, the estimate of the fraction of participants who played video games a week before the survey becomes smaller.

Scenario 3:

Now, we would like to make an interval estimate for the average amount of time spent playing video games in the week from the survey. Therefore, we need to calculate the point estimate of time spent playing video games, check the normality of its distribution and build its confidence interval.

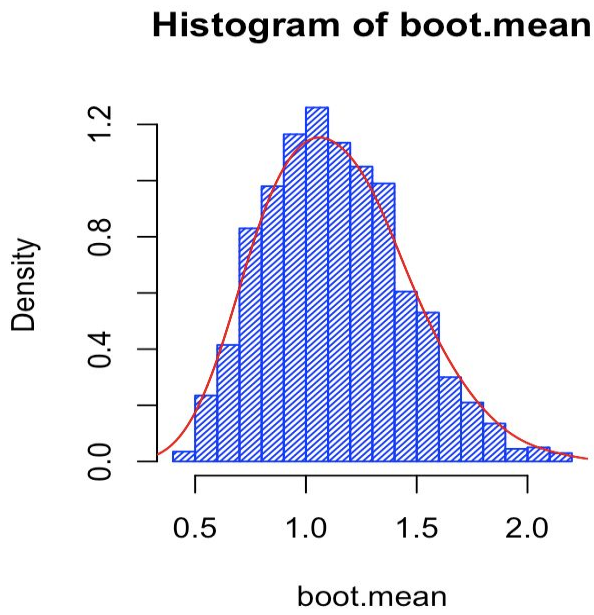

Figure 2.3.1

When considering the set of data from "videodata.txt", we use Bootstrap method to construct a confidence interval of the average amount of time spent playing video games during a week because it is asymptotically more accurate than the standard intervals obtained from sample variance and assumptions of normality. Due to the fact that the sample size n = 91, which is not sufficient enough for straightforward statistical inference, Bootstrap method allows estimation of sampling to control and check the stability of the results. From data set of "videodata.txt", our population size is 314 and sample size is 91. Therefore, each sample occurs about 3 times in the bootstrap population. We can simply duplicate each observation in the sample 3 times, and build the resulting sample as the bootstrap population. With the bootstrap population, we are can generate bootstrap sample means by repeating samples. Now, we will randomly take 500 samples whose size is 91. Then, from Figure 2.3.1, the histogram shows the distribution of Bootstrap sample means of time spent playing video games in a week. From the data set of "videodata.txt", we calculate the sample mean of time played is 1.2428. If we calculate the sample mean for each Bootstrap sample, we can get the bootstrap sample mean. To be more specific, each row of the Bootstrap sample matrix:

```
        One-sample Kolmogorov-Smirnov test

data:  (boot.mean - mean(boot.mean))/sd(boot.mean)
D = 0.045992, p-value = 0.2408
alternative hypothesis: two-sided

Warning message:
In ks.test((boot.mean - mean(boot.mean))/sd(boot.mean), pnorm) :
  ties should not be present for the Kolmogorov-Smirnov test
```

[1] 1.397802 1.227473 1.408791 1.535165 1.115385 1.127473

We now check the Normality by Kolmogorov-Smirnov test and simulation study (Kurtosis coefficient). Since our set of data generated by only one sample, we use one-sample Kolmogorov-Smirnov test. D represents the value of the test-statistic (maximum difference), so the Kolmogorov-Smirnov statistic and the p-value represents the likelihood of observing this particular value of D. From Kolmogorov-Smirnov test, since the p-value is 0.2408, which is bigger than significance level, we conclude that the sample mean is normal distributed. Therefore, we can build the confidence interval properly.

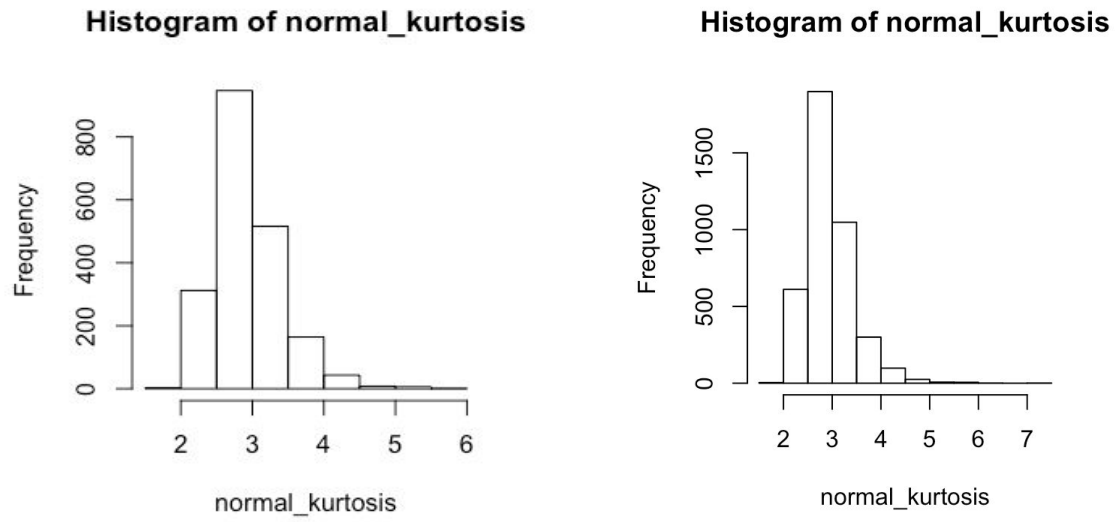

Figure 2.3.2

Next, we do the simulation study relating to Kurtosis coefficient to figure out whether it is appropriate to construct a confidence interval estimate. First, we generate 91 pseudo-random observations from a normal distribution and calculate the kurtosis coefficient, which is 2.92 at n = 2000. From the figure 2.3.2, we see that 2.92 is a typical Kurtosis value of the normal distribution with 91 samples. If we set n = 4000, kurtosis coefficient becomes 2.94. As n goes to infinity, Kurtosis coefficient will be closer to 3. Thus, we conclude that the distribution of sample means of time for playing video games is nearly normal distributed. Finally, we build a

95% confidence interval. There are two ways to construct: the first method is using Bootstrap sample mean to create C.I.; the second method is to extract the 0.025-quantile and 0.975 quantile of the bootstrap sample means.

Method 1	Method 2
95% Confidence Interval	2.5% 97.5%
(0.6289532, 1.8567611)	(0.5966758, 1.8033242)

Scenario 4:

Now we will look into the result of this survey to see whether participants enjoy to play video games or not. First, we should exclude data in which participant never play video games (respond is 1 in like/dislike question) as we are considering the attitude towards playing video games. Also, we rule out the 1 non-respondent to this question. After we come up with the new dataset, it shows 23 participants who think they really like to play video games, 46 participants who somewhat like to play video games, 13 'not really' and 7 'not at all', so in this survey a majority of 51.6% participants somewhat enjoy playing video games, 25.8% participants really enjoy playing video games and the other 22.6% participants dislike playing video games to some extent. In comparison, 77.4% participants like to play video game while 22.6% participants dislike playing video games. We will first draw bar plot of incidence of participants' attitude towards playing video games to better see the overall attitude. Therefore, in general participants like to play video games.

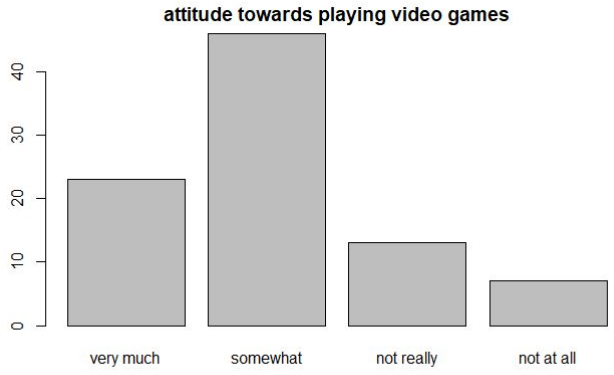

Figure 2.4.1

Then we will investigate what's important reasons that lead participants to like or dislike playing video games. In order to come up with a list of reasons, we will check each possible factor that might influence participants' attitude towards video games.

First factor we check if gender influence attitude towards video games. So we divide dataset into two subset : male participants and female participants. Then we find that 26 of 38 female participants, that is 68.4%, enjoy playing video games somewhat or very much and only 5 of 38 female participants, that is 13.2%, enjoy playing video games very much. In comparison, 43 of 51, that is 84.3% male participants enjoy playing video games somewhat or very much and 35% of them enjoy playing video games very much. So we can see that male participants are more likely to enjoy video games. Therefore, gender is a plausible reason why students like or dislike playing video games. Below is a comparison between bar plots of male and female attitude towards playing video games (left is female and right is male in every adjacent bars).

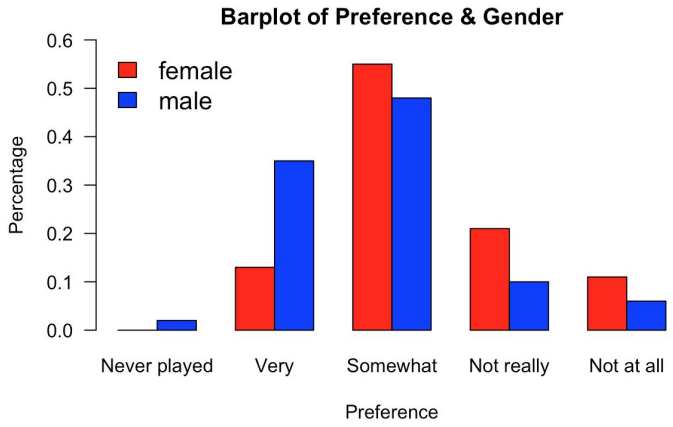

Figure 2.4.2

Next, we will consider the factor that whether student own PCs or not. Using the same method above, we get that 73.8% PC owners like to play video games very much or somewhat while 87.5% participants who don't have a PC like to play video games very much or somewhat. We can see from the bar plots that have the same scale that distributions don't have the same shift from right tail to left tail as we see in the previous comparison., so whether participants own PCs may affect the attitude towards playing video games but it is not as evident as gender factor.

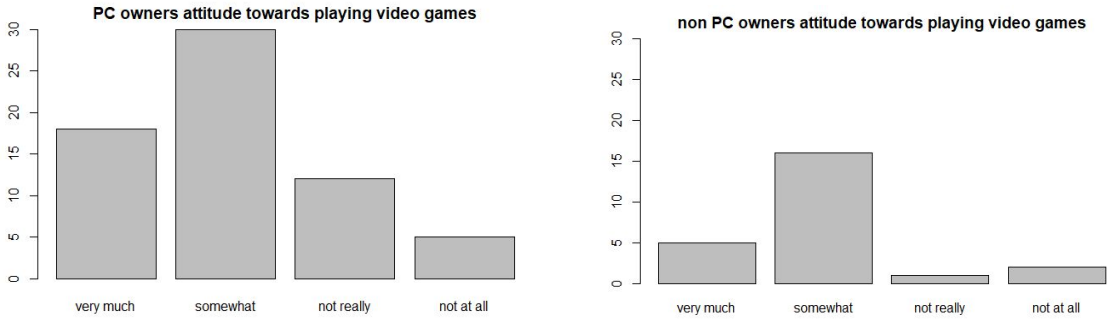

Figure 2.4.3

When evaluating whether considering playing video games as educational, as participants that don't like playing video games at all are non respondents this question, they are all ruled out. So in this comparison, there are only three kinds of attitude towards playing video games: very much, somewhat, not really. Below is the bar plots of this factor. We see from the graph that in fact a larger portion of participants who believe that playing video games is not educational like to play video games.

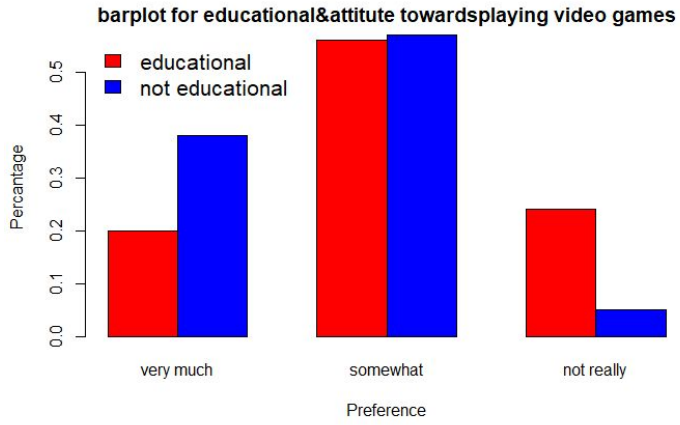

Figure 2.4.4

Also, we will consider whether having PCs at home as a factor because participants may find having PCs at home convenient for them to play video games and hence enjoy playing video games. We can see from the bar plot below that a larger proportion of participants who have PCs at home enjoy playing video games. So it matches our speculation that participants who have PCs at home may be more likely to enjoy playing video games.

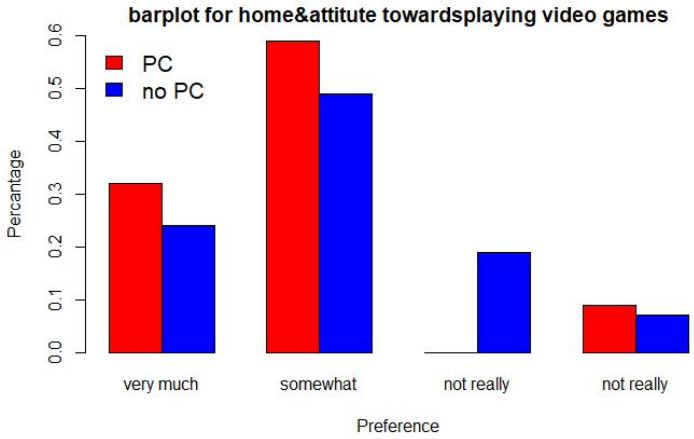

Figure 2.4.5

At last, we consider whether age will influence attitude towards playing video games. To facilitate this comparison, we divide the participants into two groups, participant whose age is greater than 20 and whose age is less than 20. We can see from the bar plot that despite the difference that a larger proportion of participants who have an age less than 20 somewhat like to play video game, the overall preference is similar as the percentage of participants who have an age less than 20 who like to play video game is basically the same as the percentage of participants who have an age greater than 20 who like to play video game.

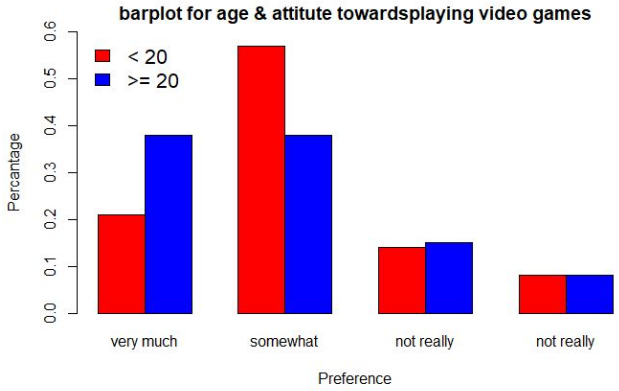

Figure 2.4.6

So in our list of plausible reasons for why students like or dislike playing video games, we have five reasons: whether participants own PCs or not, whether participants' age is less than 20 or not, whether participants have PCs athome or not, whether participants believe that playing video games is educational or not, whether participants are male or female. We will then put these reasons into a decision tree to see their importance. So from top to bottom, we have whether participants believe that playing video games is educational or not, then whether participants have PCs athome or not, then whether participants' age is less than 20 or not, and whether participants are male or female.

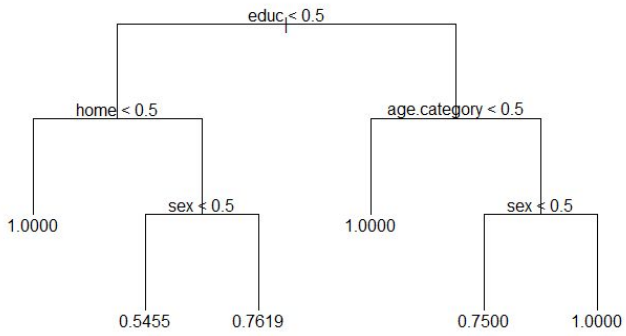

Figure 2.4.7

Scenario 5

This part of the study aims to examine the difference between students who like to play video games and who do not. In the survey, participants were asked to provide their preferences on playing video games by choosing one of preference choices listed in the survey including "Never Played", "Very much", "Somewhat", "Not really" and "Not at all". To see the difference between students who like to play and who don't, we classify students who choose **"Very much" and "Somewhat"** into the category of **"Like playing video games"** and students who choose **"Not really" and "Not at all"** into the category of **"Do not like playing video games"**. One participant who choose **"Never played"** does not fall into two categories, thus we ignore him in this part of the study. We also ignore one participant who did not answer this question. Sorting participants in this way, we get 20 participants who dislike playing video games and 69 participants who like playing video games.

In the following investigation, we want to study how gender, works and owning a computer play roles in the preference of playing video games. Thus, among participants who like playing and who dislike playing, we further separate participants by gender, whether the participant works to pay and owns a computer or not, and study the difference between each group. Participant were asked to choose "0" if they are female and "1" if they are male; choose "0" if they own computer and "1" if they don't. In the section of "Work" , they were asked to fill in number of hours they worked per week. For the purpose of this study, we classify students who work zero hour into a group of no work and students who work for more than 0 hour into a group of have works. We will use two categorical variables in bar plot and use their working hours in histograms. We adopt numerical and graphical analysis in the following study.

Numerical Analysis

We will first look at the numerical summary of students who like playing video games and who don't. Our data on variables are categorical so most numerical statistics such as mean and median do not provide much useful information. Rather, we will look at percentage of participants. Working hours is numerical data so we will look at its average. Among 91 participants in our survey, 69 of them fall into category of like playing video games and 20 of them fall into category of dislike playing video games. Among participants who like playing, 38% of them are female and 62% are male. Among participants who dislike playing video games, 60% of them are female and 40% are male. If we focus on whether participants have works, we find out 45% of participants who like playing do not work and rest 55% of them has a work. For participants who dislike video games, 70% of them do not work and 30% of them has a work. The average working hours are about 8.05 hours per week for those who like playing and 4.55 hours for those who dislike.

Our expectation on average working hours of those who like playing video games is higher than average of those who dislike video games since people who do not work have more time on playing games. However, the result is unexpected and turns out to be the reverse of our expectation. Our hypothesis here is that people who work have more pressure and stress that need to be relieved by playing video games. However, average working hours may be influenced by outliers and may be biased so we will look at the distribution of working hours to check accuracy.

Now we look at the data of owning a computer. We find out that 30% of participants who like playing does not own a computer and 70% of them owns a computer. Among participants who dislike playing, 15% of them does not own a computer and 85% of them owns a computer.

Cross-Tabulation

Cross-tables in the following display numerical summary we mentioned above for two groups of participants in a clear way. All columns in three tables display information for same variables: First column displays percentage of participants who dislike playing video games and second column display percentage of participants who like playing video games. Rows in Table 5.1 displays percentage of female and male in each preference group, respectively. Rows in Table 5.2 displays percentage of no work and has a work in each preference group, respectively. Rows in Table 5.3 displays percentage of not owning a computer and owning a computer in each preference group, respectively.

	Dislike	Like
Female	0.60	0.38
Male	0.40	0.62

Table 2.5.1

	Dislike	like
No Work	0.70	0.45
Has Work	0.30	0.55

Table 2.5.2

	Dislike	Like
No PC	0.15	0.30
Has PC	0.85	0.70

Table 2.5.3

Graphical Analysis

Graphs may give us a more direct and obvious way to make comparison between different groups of participants within each category. First we want to use graphs to study how gender plays a role in video game preferences. We separate participants into two groups of like playing

and dislike playing, same as what we did above, and draw a bar plot to see the preference distribution. We use bar plot here rather than histogram because level of likeness are all categorical data in our dataset. Students were asked to choose "1" if they like to play and "0" if they do not like to play.

Figure 5.1 below displays the gender distribution of two groups of participants. X-axis in the bar plot shows preference responses "Dislike" and "Like", y-axis shows the percentage of participants who fall into one preference. Red bar on the left represents percentage of participants who are female within two categories of preferences. Blue bar on the right represents percentage of participants who are male within each category of preferences. Among participants who dislike playing video games, approximately 60% of them are female and 40% are male. Among participants who like playing, approximately 40% of them are female and 60% are male. Approximations we infer for barplot are close to the result we get from numerical analysis. We can see that majority of participants who dislike video games are female while majority of participants who like video games are male.

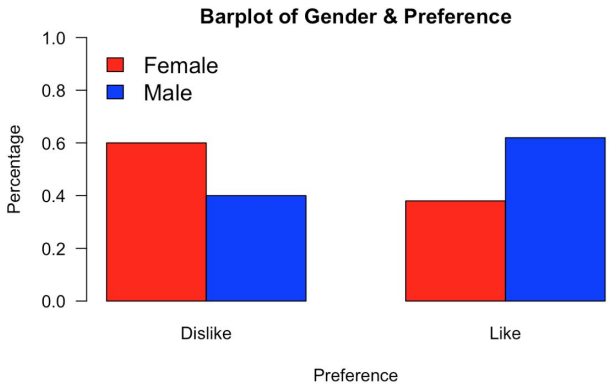

Figure 2.5.1

Next, we start to investigate if there is a difference in employment between participants who like playing and who dislike. Figure 5.2 below displays bar plot of work distribution. X-axis in the bar plot shows preference responses "Dislike" and "Like", y-axis shows the percentage of participants. Students were asked to choose "1" if they work and "0" if they do not work. Red bar on the left represents percentage of participants who work within each category of preferences. Blue bar on the right represents percentage of participants who are do not work within each category of preferences.

From figure 5.2, approximately 45% of participants who like playing do not work and rest 55% of them has a work. For participants who dislike video games, 70% of them do not work and 30% of them has a work. We can see that majority of participants who dislike playing has no

work, while difference between percentages of working and not working is smaller among participants who like playing, which is about 10% difference.

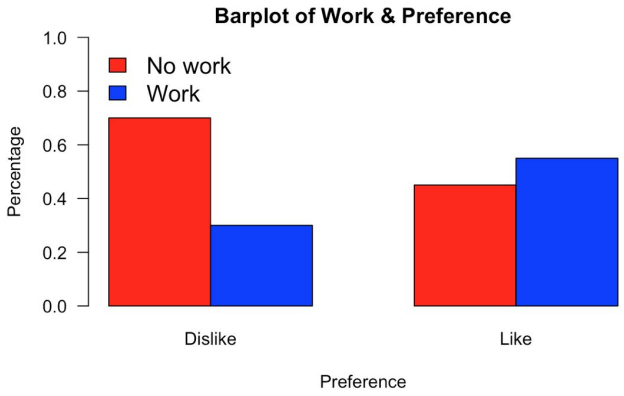

Figure 2.5.2

In the following investigation, we graph histograms to see the distribution of working hours for each group. Here we adopt histograms because working hours is a numerical variable in our dataset. Figure 5.3 displays the distribution of working hours for participants who dislike playing video games and Figure 5.4 displays distribution for who like playing. Working hours of participants who like video games approximately fall into the range of 0 hour to 55 hours while working hours of those who dislike video games approximately fall into the range of 0 hour to 40 hours. We can see that both distributions are skewed to the right, unimodal and have several outliers. Outliers in figure 5.3 for participants who like playing are around 35, 40 and 55 hours. Outlier in figure 5.4 for participants who dislike playing is around 40 hours. We will further check outliers by graphing box plots and check significance of difference in two distributions by qq plot in the following sections.

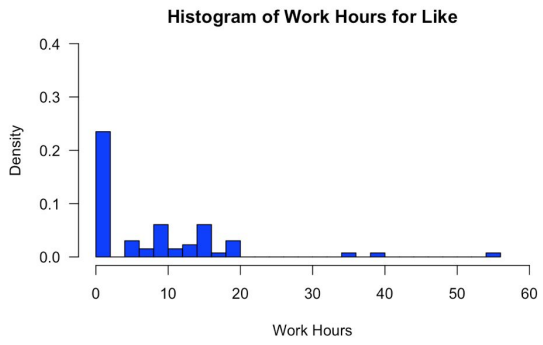

Figure 2.5.3

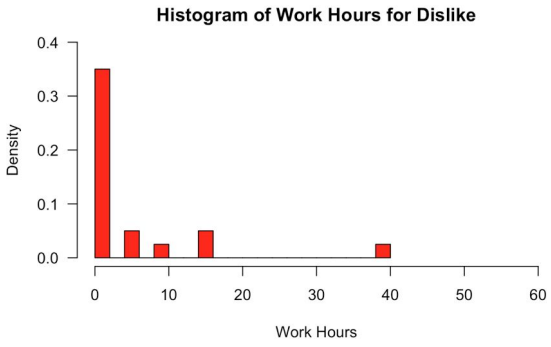

Figure 2.15.4

Next we investigate the difference in owning computers for two groups of participants. Students were asked to choose "0" if they own computers and "1" if they do not. Figure 5.5 below displays the distribution of owning a computer. X-axis in the bar plot shows preference responses "Dislike" and "Like", y-axis shows the percentage of participants. Red bar on the left represents percentage of participants who do not have computers. Blue bar on the right represents percentage of participants who have computers.

From Figure 5.5, approximately less than 20% of participants who dislike playing computers do not have computers and more than 80% of them have computers. Approximately 30% of participants who dislike video games do not have computers while around 70% of them have computers. We see that both majorities of participants who dislike games and like games own computers. The percentage of no computers is higher in group of like games than percentage in group of dislike games.

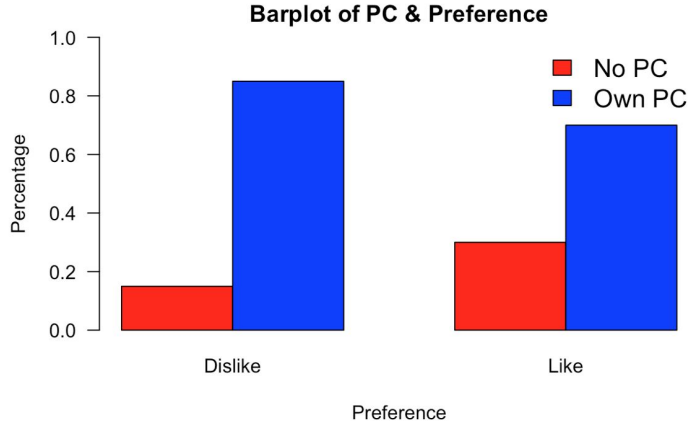

Figure 2.5.5

Boxplot

Figure 2.5.6 below displays boxplots of two groups of participants. Boxplot on the left displays working hour distribution of participants who dislike playing video games and box plot on the right displays that of participants who like games.

From boxplot of "Dislike" group on the left, we can tell that the maximum of working hours is around 10 hours and third quartile is around 5 hours. First quartile, median and minimum working hours overlap at 0 hour. The IQR is around 5 hours. There are two suspected outliers distributed around 15 hours and one outlier distributed around 40 hours. Distribution of working hours ranges from 0 hour to approximately 40 hours.

From boxplot of "Like" group on the right, we can tell that the maximum of working hours is around 35 hours and the median is around 5 hours. The first quartile and minimum working

hours overlap at 0 hours. Third quartile is around 15 hours. The IQR is around 15 hours. There are one suspected outlier distributed around 38 hours and on outlier distributed around 55 houts. Distribution of working hours ranges from 0 hour to approximately 55 hours.

Comparing two distributions shown in figure 5.6, the shape and distribution of "Like" is not very similar to these of "Dislike". Distribution of working hours of "Like" group has larger range, medium and maximum than that these of "Dislike" group. Several participants within two groups do not work at all, indicating by the bottom lines of two boxplots.

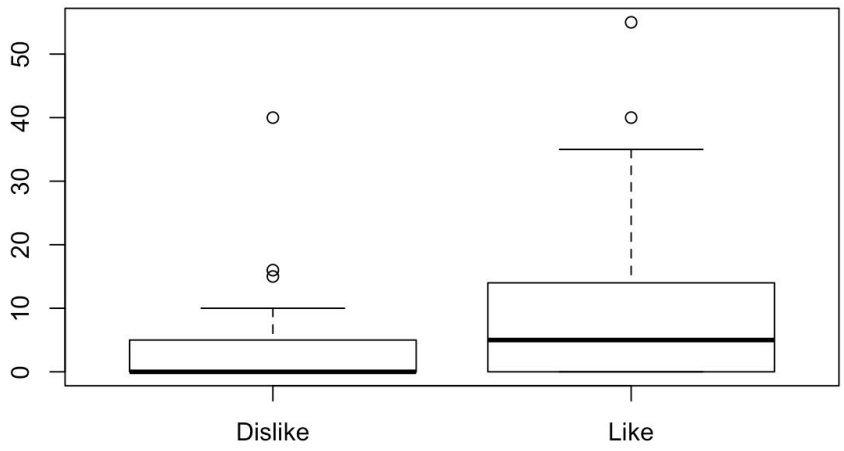

Figure 2.5.6

Check for significance

This part of the study examines the difference in distributions of working hours for participants who like (Figure 5.3) and dislike playing video games (Figure 5.4) that we displayed above. We adopt qq plot to compare the shapes of distributions, providing a graphical observation of how features vary in the two distributions.

Figure 5.7 below is the qq plot between working hours of participants who dislike video games and working hours of participants who like video games, and the reference line is y = x, which has a slope of 1 and intercept of 0. We can see that the plotted points deviate from the reference line and are non-linear. We can infer that two distributions have different means.

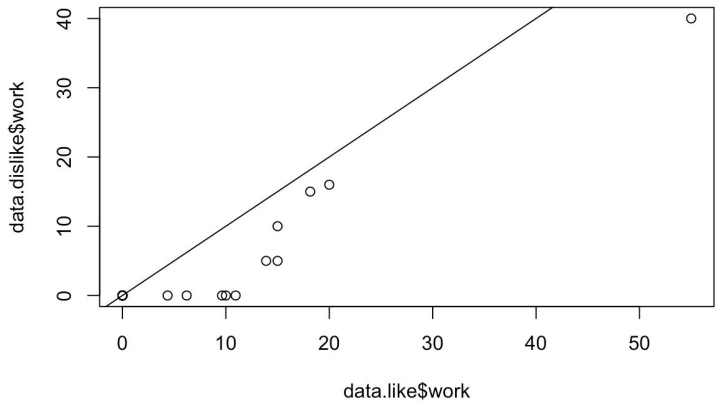

Figure 2.5.7

In this case, where normality of two distributions cannot be assumed, we will use non-parametric test to test if there is a significant difference between working hours of participants who like and dislike playing video games.

We use Mann-Whitney test to test if there is a statistically significant difference between two distributions. Mann-Whitney is a nonparametric test for hypothesis test that the randomly chosen values will be equally likely from one and other. This test does not require the sample variables to be normal distributions. Calculation of the statistic called U is incorporated in this test. Details of test can be found in the Theory section. The null hypothesis we state here is that the mean of working hours who like playing is the same as that of participants who dislike playing, and the alternative hypothesis is the mean of working hours who like playing is different from that of participants who dislike playing. We will reject the null hypothesis if the p-value is less than 0.05 significance level and vice versa.

After we perform the two-sided test, we get a p-value of 0.06395, which is bigger than 0.05 significance level, so we cannot reject the null hypothesis. Therefore, from hypothesis testing, we conclude that a significant difference between two distributions cannot be assumed and average working hours of participants who like playing video games is not significantly different from average of working hours of participants who dislike playing video games.

Scenario 6

In this scenario, we want to further investigate the grade that students expect in the course. We want to test if the grade in videodata.txt matches the target distribution used in grade assignment of 20% A's, 30%B's, 40% C's and 10% D's or lower. Furthermore, if the nonrespondents were failing students who no longer bothered to come to the discussion section, would this change the picture.

Numerical Analysis

We first take a look at the numerical analysis of the difference between expected grade and target grade. We denote 4, 3, 2, 1, 0 as A, B, C, D, F. Among 91 participants in our videodata.txt file, 31, 52, 8, and 0 participants expect to get A's, B's C's D's, so the percentage that participants expect to get A's, B's, C's, D's is 31/91, 52/91, 8/91, and 0. The mean of expected grade distribution is around 3.252747, and the mean of target grade distribution is around 2.6. Compared with 20% A's, 30 % B's, 40% C's, and 10% D's, expected grade distribution is around 34% A's, 57% B's, 9% C's, and 0% D's, so we can see a big difference between target grade and expected grade: expected grade distribution has a higher grade assigned percentage of getting A's and B's, and target grade distribution has a higher grade assigned percentage of getting C's and D's.

When we take the 4 nonrespondents into consideration, we have 4 fails in the new expected grade distribution, since D includes D and lower, so we count our 4 fails as D, then the situation changes. Among 95 participants in our new expected grade distribution, 31, 52, 8, and 4 participants expect to get A's, B's. C's, and D's, so the percentage that participants expect to get A is 31/95, 52/95, 8/91, and 4/95. The mean of new expected grade distribution is around 3.115789, and the mean of target grade distribution is around 2.6. Compared with 20% A's, 30 % B's, 40% C's, and 10% D's, expected grade distribution is around 32.6% A's, 54.7% B's, 8.4% C's, and 5.2% D's, and compared with previous 34% A's, 57% B's, 9 C's, and 0% D's, we have a smaller percentage of A, B, C, but a larger percentage of D. Compared with target grade distribution of 20% A's, 30% B's, 40% C's, and 10% D's, we have a higher percentage of A's, B's, but a lower percentage of C's, D's.

Graphical Analysis

We construct a barplot to show the number of participants with their expected grade. From the bar plot, we can see clearly that majority of the participants choose B and A, only a small amount of people choose C, and nobody chooses D. It has a huge difference compared to the target grade distribution, since target grade has 20% A, 30% B, 40% C, and 10% D.

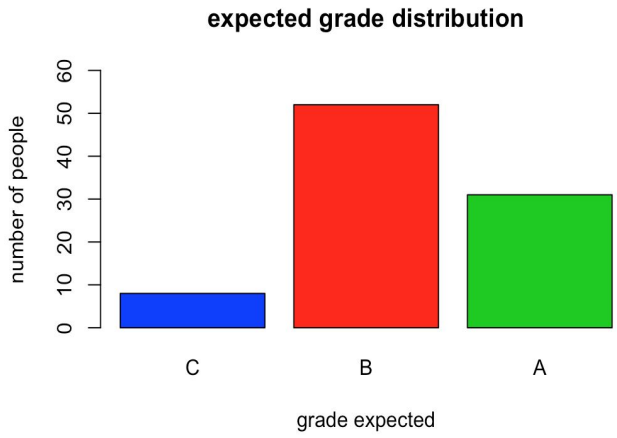

Figure 2.6.1

When we take 4 fails into consideration, we will have a different barplot with new object D, with no change of A, B ,C, since the number of people expect to get A, B, C is the same as previous situation.

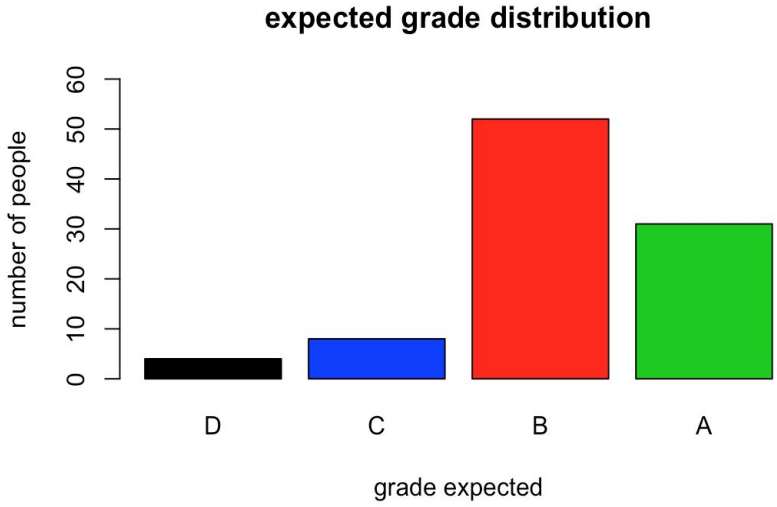

Figure 2.6.2

For this bar plot, we construct a comparison for expected grade and target grade. * represents the target grade. X-axis represents the grade received and Y-axis represents the grade assigned percentage. From this barplot, grade assigned percentage of expected grade for A and B is clearly larger than the grade of assigned percentage of target grade. Grade assigned percentage of expected grade for C and D is definitely smaller than the grade assigned percentage of target grade.

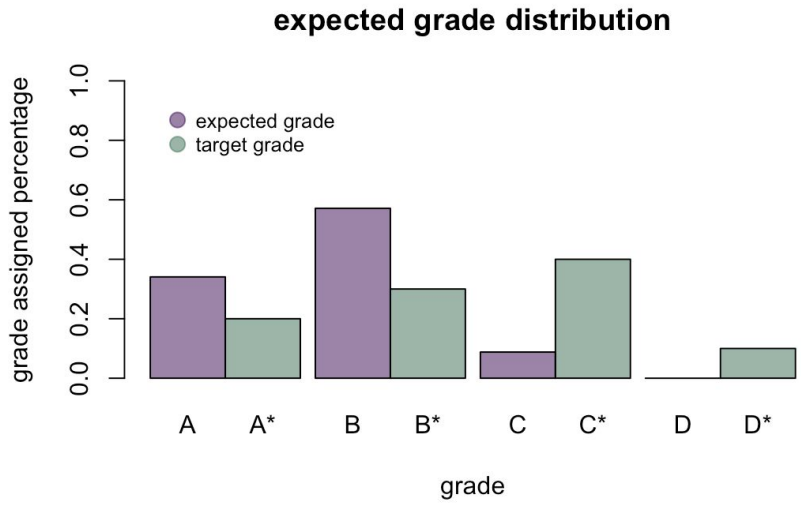

Figure 2.6.3

When we take 4 fails into consideration, we have a different situation with less grade assigned percentage of A, B ,C compared with the old expected grade distribution, but a higher percentage of D compared with the old expected grade distribution. When compared with target grade, we have a higher percentage of A, B, C, but a smaller percentage of getting D or lower.

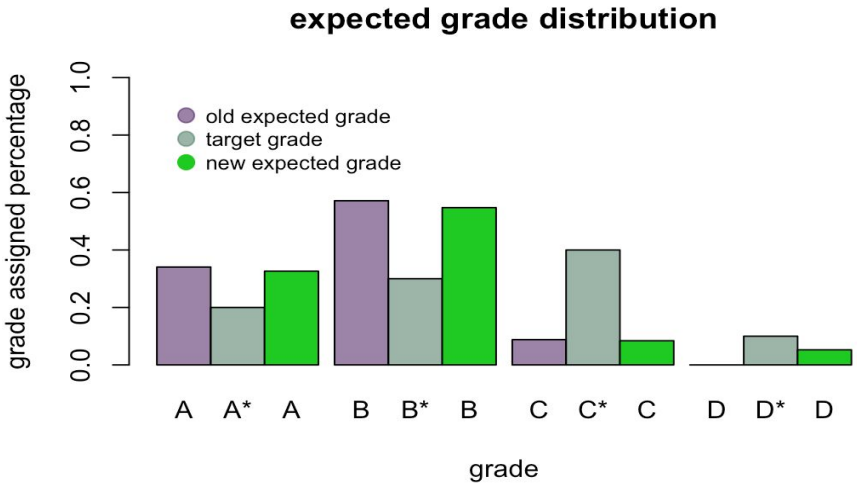

Figure 2.6.4

For this box plot, 1 represents the expected grade distribution and 2 represents the target grade distribution. We can see their distributions have a huge difference: their means are different, and 25th, 75th percentile is totally different.

Two-sample Kolmogorov-Smirnov test

We perform ks test of target grade distribution with old and new expected grade. The null

hypothesis we state here is that the old and new expected grade distribution is the same as the target grade distribution, and the alternative hypothesis is the old and new expected grade distribution is the same as the target grade distribution. We will reject the null hypothesis if the p-value is less than 0.05 significance level and vice versa.

After we perform the two-sided test, we get a p-value of 5.855e-07(old expected grade distribution) and 5.022e-06(new expected grade distribution), which is significantly smaller than 0.05 significance level, so we reject the null hypothesis. Therefore, from hypothesis testing, we conclude that the distribution of old and new expected grade is significantly different from the distribution of target grade.

In conclusion, the old and new expected grade distribution have a huge difference from the target grade distribution. From the numerical and graphical analysis, participants tend to expect receive higher grade, such as A's and B's, since the assigned A and B grade percentage of expected grade is higher than the target grade distribution. The C and D percentage of expected grade is lower than the target grade distribution. When we take 4 fails into consideration, the A's, B's, and C's percentage slightly decrease but still have a higher percentage than the target grade. When 4 fails is taken into consideration, we have more D's but still a lower percentage than the D's percentage of target grade.

3. Discussion

The main purpose of this report is to examine and rank possible factors that may affect preferences of playing video games and study the difference between participants who like playing video games and who dislike. This study aims to provide useful information to help to design an interactive learning lab linked to video games. However, there are certain limitations of this study include small sample size compared to the overall population and there may include an artificial correlation due to categorical data in our dataset. Future studies on factors that may affect preference of playing video games with larger sample size may bring more accurate information to help committee to design an interactive learning lab.

Objectives

As our investigation regarding effect of certain factors on preferences of playing video games has certain limitations, we will perform an additional study to provide more information for the committee in this section of the study. We will examine the association between frequency of playing video games and grades of students, aiming to suggest a suitable frequency of labs that committee hope to design. We want to get a most suitable frequency for lab that can best facilitate study and help students to get good grades.

Reference

The following study refers to the published thesis *The effects of video game play on academic performance* by Jancee Wright at University of the Cumberlands. The study of Jancee Wright aims to examine the association between amount of time spent on playing video games and academic performance of participants indicated by GPA. The hypothesis in this study states that there is no positive association between amount of time spent on playing video and academic achievements (Jancee Wright, 41). Methods involved are series of descriptive statistics analyses, correlation analyses, and one-way ANOVAs (Jancee Wright, 41). Variables studied in the study are amount of time spent on video games, status of players, games types and academic achievements. Conclusion of this study is that "There were no significant correlations concerning the effects of the amount of time spent playing games on GPA, the amount of puzzle or strategy situations faced in the average game on GPA, or gaming mode on GPA. Overall, the only statistically significant correlation was that of player status and GPA" (Jancee Wright, 41).

Hypothesis

Based on the research result of Jancee Wright's study, the following study constructs the following hypothesis: There is a negative association between students' grade and their frequency of playing video games.

Method:

To make the survey more representative, we divide the participants by their frequency of playing video games in this way: participants with daily frequency and weekly frequency make up one group and participants with monthly and semesterly frequency make up another group. In other words, in the dataset, response to the frequency question that is 1 or 2 is changed to 1 and response to the frequency question that is 3 or 4 is changed to 2. Then as grade and frequency are all categorical data, we can only draw bar plot to compare the relative distribution of grade.

Result:

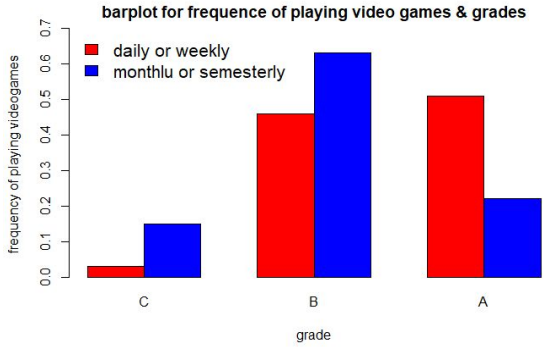

Figure 3.1

As we can see from the bar plot, 51% of the participants who keep daily frequency or weekly frequency are expected to earn A while only 22% of the participants who keep monthly and semesterly frequency are expected to earn A. To a broader sense, 97% of the participants who keep daily frequency or weekly frequency are expected to earn A while only 85% of the participants who keep monthly and semesterly frequency are expected to earn A. Therefore, we can see an improvement in grade when frequency increase from monthly or semesterly to weekly or daily. So the computer lab should be designed to open to students by weekly frequency to daily frequency.

Limitations

This section of study is limited by categorical data. We are not able to perform certain tests and analysis to investigate further on association between frequency of playing video games and grades of students. Only using bar plot would not be persuasive and grounded. We hope to get more information to conduct a new investigation.

4. Theory

1. ***Histogram*:** Histogram gives a clear picture of the data density. Higher bars represent where the data are recorded with higher frequency compared with others. Histogram is very straightforward for describing the shape of the data distribution. In this research, histograms are applied to form a contrast between pregnant mothers who are smokers and who are non-smokers.
2. ***Q-Q Plot*:** A Q-Q plot is used to compare the shapes of distributions, providing a graphical observation of how features such as location, scale, and skewness vary in the two distributions.
3. ***Boxplot*:** A boxplot summarizes a data set using five statistics including minimum, maximum, mean, 25th percentile, and 75th percentile but also plotting unusual observations such as outliers. The median of a boxplot splits the data into the top 50% and the bottom 50%. The total length of the box, shown vertically , is known as the interquartile range (IQR, for short), which, similar to the standard deviation, is a measure of variability in data. If there are more variabilities in a set of data, then the larger the standard deviation and IQR are. The lower bound and upper bound of the box are called the first quartile (25%) and the third quartile (75%), and are often labeled Q1 and Q3, respectively.
4. ***Bar plot***: A bar plot is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be

plotted vertically or horizontally. A bar plot shows comparisons among discrete categories. One axis of the chart shows the specific categories being compared, and the other axis represents a measured value. Some bar graphs present bars clustered in groups of more than one, showing the values of more than one measured variable.

5. ***Mann-Whitney U test:*** is a nonparametric test for hypothesis test that the randomly chosen values will be equally likely from one and other. This test, unlike the t-test, does not require the sample variables to be normal distributions. Calculation of the statistic called U is incorporated in this test. The following is the formula used to calculate Us for two samples.

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \qquad U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

$$U_1 + U_2 = R_1 - \frac{n_1(n_1+1)}{2} + R_2 - \frac{n_2(n_2+1)}{2}.$$

In the above formula, n_1 and n_2 stands for the sample size for each sample, R_1 and R_2 stand for the sum of ranks in each sample.

6. ***Decision tree:*** A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.
7. ***Bootstrap estimate*****:** Bootstrap algorithms are simple and general tools for assessing estimators' accuracy via variance estimation, and producing con dence intervals and p-values. Bootstrap applies to finite samples and provides numerical solutions for non-standard situations so that it is particularly appealing when dealing with finite populations and complex sampling designs.
8. ***Kolmogorov-Smirnov test*****:** can be modified to serve as a goodness of fit test. In the special case of testing for normality of the distribution, samples are standardized and compared with a standard normal distribution. This is equivalent to setting the mean and variance of the reference distribution equal to the sample estimates, and it is known that using these to define the specific reference distribution changes the null distribution of the test statistic. The Kolmogorov–Smirnov statistic for a given cumulative distribution function F(x) is

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[-\infty,x]}(X_i) \qquad D_n = \sup_x |F_n(x) - F(x)|$$

9. ***Kurtosis*:** is a measure of the "tailedness" of the probability distribution of a real-valued random variable. In a similar way to the concept of skewness, *kurtosis* is a descriptor of the shape of a probability distribution and, just as for skewness, there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population. For normal distribution, the kurtosis coefficient is 3.

$$\text{kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s}\right)^4$$

10. ***A point estimate*:** the process of finding an approximate value of some parameter—such as the mean (average)—of a population from random samples of the population. The accuracy of any particular approximation is not known precisely, though probabilistic statements concerning the accuracy of such numbers as found over many experiments can be constructed.
11. ***Interval estimate***: In statistics, the evaluation of a parameter. Intervals are commonly chosen such that the parameter falls within with a 95 or 99 percent probability, called the confidence coefficient. Hence, the intervals are called confidence intervals; the end points of such an interval are called upper and lower confidence limits.

$$(\bar{x} - 1.96s, \bar{x} + 1.96s)$$

12. ***Proportion estimate***:

$$\left(\bar{x} - 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{n-1}\frac{N-n}{N}}, \bar{x} + 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{n-1}\frac{N-n}{N}}\right),$$

13. ***Confidence Interval*:** Sample statistic x is random, so we can think of confidence intervals as random intervals. Different samples lead to different confidence intervals. If we take many simple random samples where for each sample we compute CI, then we expect 95% of CI's to contain μ.

A 95% confidence interval is

$$\left(\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right)$$

5. Conclusion

Scenario 1 shows that the reliability of our database, which can be referred as the confidence interval. From the result we get from all of these methods which are a point estimate, bootstrap sample estimate, the confidence interval is around 3.1 to 4.5, and indicates only one-third of our data are reliable to use in our life. At the meanwhile, especially for the fraction of the student who played the video game prior to the week, the most part we can trust is ranged between 3.1 to 4.5.

Scenario 2 indicates that the exam a week before the survey had significantly decreased participants' playtime on video games and in turn caused the time spent on playing video games a week before the survey failed to reflect the general pattern of frequency of playing video games.

Scenario 3 examines the appropriateness of constructing a confidence interval and builds an internal estimate for the average amount of time spent playing video games in the week relating to the study. After we make sure it follows the normal distribution, we apply two methods to construct the confidence interval: the first method is using Bootstrap sample mean to create C.I.; the second method is to extract the 0.025-quantile and 0.975 quantile of the bootstrap sample means. Therefore, we are 95% confidence that the true mean of time playing video games for population is in (0.6289532, 1.8567611) for method 1 and in (0.5966758, 1.8033242) for method 2.

Scenario 4 investigates possible reasons for why students like or dislike playing video games, and we have five plausible reasons: whether participants own PCs or not, whether participants' age is less than 20 or not, whether participants have PCs athome or not, whether participants believe that playing video games is educational or not, whether participants are male or female. After we rank these reasons, we find that four of them, from top to bottom: whether participants believe that playing video games is educational or not, then whether participants have PCs athome or not, then whether participants' age is less than 20 or not, and whether participants are male or female, are the most important reasons for why students like or dislike playing video games.

Scenario 5 examines the difference between students who like to play video games and who do not. In this part, we classify students into a group of **"Like playing video games"** and a group of **"Do not like playing video games"**. We use numerical, graphical and Mann-Whitney U test to investigate how gender, works and owning a computer play roles on students' preferences toward playing video games. We get several differences in the following: First of all, majorities of participants who dislike video games are female while majorities of participants who like video games are male. Second, majority of participants who dislike playing has no work, while

difference between percentages of working and not working is smaller among participants who like playing. Last, while both majorities of participants who dislike games and like games own computers, the percentage of no computers is higher in group of like games than percentage in group of dislike games.

To conclude, we provide few pieces of advice to help committee to design a interactive learning lab linked to video games. First, it's better not to provide learning lab during the exam week as student may not utilize such opportunity during busy exam week. Second, committee may could provide computers to aid those students who do not have computers and let them have more opportunities to learn. Third, the foremost feature that designers should add into video games is the educational function since educational function has been ranked no.1 among factors affecting participants' preferences of playing video games.

6. Appendix

Question 1 asks that what types of games do you play? In this question, students are asked to check all types of games that apply to them, so the sum of percentage may exceed 100%. Also, students who said thatchy have never played a video game or do not at all like to play video games were instructed to skip this question. Table below displays the types of games and corresponding percent.

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

Question 2 asks that why do you play the games you checked above? Each student may choose at most three reasons for playing video games for this questions. Table below displays reasons and corresponding percent.

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

Question 3 asks that what don't you like about video game playing? All students were asked to answer this question and they may choose at most three answers for this question. Table below displays reasons for not liking video games.

Dislike	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%

7. Reference:

1. Stephanie. “Bootstrap Sample: Definition, Example.” *Statistics How To*, 3 Dec. 2016, www.statisticshowto.com/bootstrap-sample/.
2. “Mann–Whitney U test.” Wikipedia, Wikimedia Foundation, 5 Feb. 2018, en.wikipedia.org/wiki/Mann-Whitney_U_test.
3. Bradic, Jelena. *Math189 Case Study.* math189.edublogs.org/files/2017/01/chp3-2f7bqty-1ac0iv7.pdf.
4. “Kolmogorov-Smirnov test.” *Wikipedia*, Wikimedia Foundation, 22 Dec. 2017, en.wikipedia.org/wiki/Kolmogorov-Smirnov_test.
5. Wright, Jancee. "The Effects of Video Game Play on Academic Performance." *Scholar.utc.edu*. N.p., 1 Nov. 2011. Web.
6. “Bar chart.” *Wikipedia*, Wikimedia Foundation, 5 Feb. 2018, en.wikipedia.org/wiki/Bar_chart.
7. “Https://Www.qualtrics.com/Wp-Content/Uploads/2013/05/Cross-Tabulation-Theory.pdf .” Cross-Tabulation-Theory.

Scenario 1:

```{r}
setwd("/Users/mengtingchen/Desktop/math189/cs2")
data <- read.table("videodata.txt", header = TRUE)
head(data)
```

```{r}
sample(1:10, size = 10, replace = F)
```

```{r}
sample(1:10, size = 10, replace = T)
```

```{r}
played.ind <- which(data['time'] > 0)
nonplayed.ind <- which(data['time'] ==0)
```

```{r}
played <- data[played.ind,]
nonplyed <- data[nonplayed.ind,]
```

```{r}
played.percentage <- 34/91
played.percentage
```

```{r}
boot.population <- rep(data$time, length.out = 314)
length(boot.population)
```

```{r}
sample1 <- sample(boot.population, size = 91, replace = FALSE)
```

```{r}
set.seed(189289)
B = 400
boot.sample <- array(dim = c(B, 91))
for (i in 1:B) {
  boot.sample[i, ] <- sample(boot.population, size = 91, replace = FALSE)
}
```

```{r}
N <- 1000
data.population <- rbinom(n=N, size=1, prob=34/91)
```
```{r}
n <- 300
ind.sample <- sample.int(n=N, size=n)
data.sample <- data.population[ind.sample]
```
```{r}
mean.sample <- mean(data.sample)
mean.sample
```
```{r}
width <- 1.96 * sqrt(mean.sample*(1-mean.sample)*(N-n)/((n-1)*N))
int.sample <- c(mean.sample - width, mean.sample + width)
int.sample
```
```{r}
n<-91
N<-314
ind.boot <- sample.int(n, size=N, replace=TRUE)
data.boot <- data.sample[ind.boot]
```
```{r}
B <- 1000
boot.sample.mean <- rep(NA, B)
for(i in 1:B){
  ind <- sample.int(N, size=n, replace=FALSE)
  boot <- data.boot[ind]
  boot.sample.mean[i] <- mean(boot)
}
```
```{r}
hist(boot.sample.mean)
```
```{r}
mean.boot <- mean(boot.sample.mean)
mean.boot
```

```
```{r}
s <- sd(boot.sample.mean)
int.boot <- c(mean.boot - 1.96*s, mean.boot + 1.96*s)
int.boot
```
```{r}
int.boot <- c(quantile(boot.sample.mean, 0.025), quantile(boot.sample.mean, 0.975))
int.boot
```

Scenario 2:
loading data
```{r}
setwd("/Users/mengtingchen/Desktop/math189/cs2")
data <- read.table("videodata.txt", header=TRUE)
head(data)
```

```{r}
attach(data)
time <- time[time != 99]
freq <- freq[freq != 99]
```

```{r}
counts1 <-table(time)
barplot(counts1,main = "time of playing video games a week before the survey")
counts2 <-table(freq)
barplot(rev(counts2),main = "frequency of playing video games",names.arg=c("semesterly","monthly","weekly","daily"),ylim = c(0,50))

irreg.index <- which(data$freq == 99)
reg.ind <- setdiff(rownames(data), irreg.index)
data <- data[reg.ind,]
boxplot(time~freq,data, xlab = "frequency of play", ylab = "time of play a week before the survey",names=c("daily","weekly","monthly","semesterly"))
```
```

```
daily.ind <- which(data['freq'] == 1)
data.daily <- data[daily.ind,]
hist(data.daily$time,freq = F,breaks = c(0,5,10,15,20,25,30),ylim = c(0,0.3))
weekly.ind <- which(data['freq'] == 2)
data.weekly <- data[weekly.ind,]
hist(data.weekly$time,freq = F,breaks = c(0,5,10,15,20,25,30),ylim = c(0,0.3))
monthly.ind <- which(data['freq'] == 3)
data.monthly <- data[monthly.ind,]
hist(data.monthly$time,freq = F,breaks = c(0,5,10,15,20,25,30),ylim = c(0,0.3))
semesterly.ind <- which(data['freq'] == 4)
data.semesterly <- data[semesterly.ind,]
hist(data.semesterly$time,freq = F, breaks = c(0,5,10,15,20,25,30),ylim = c(0,0.3))
```

Scenario 3:

```{r}
data<-read.table("videodata.txt", header=TRUE)
head(data)
summary(data)
#data[data == 99]<- NA
sum(is.na(data))
summary(data$time)
hist(data$time)
### Bootstrap ####
time.mean <- mean(data$time)
time.mean
boot.population <- rep(data$time,length.out=314)
length(boot.population)
sample1<-sample(boot.population,size = 91, replace = FALSE)
set.seed(189289)
B = 300
boot.sample <- array(dim = c(B, 91))
for (i in 1:B) {
  boot.sample[i, ] <- sample(boot.population, size = 91, replace = FALSE)
}
boot.mean <- apply(X = boot.sample, MARGIN = 1, FUN = mean)
```

```
head(boot.mean)
hist(boot.mean, breaks = 22, probability = TRUE, density = 40, col = 4, border = 4)
lines(density(boot.mean, adjust = 2), col = 2)
### Kolmogorov-Smirnov test ###
par(pty = 's')
qqnorm(boot.mean)
qqline(boot.mean)
ks.test((boot.mean - mean(boot.mean))/sd(boot.mean), pnorm)
### Confidence interval ###
boot.sd<-sd(boot.mean)
time.mean + c(-1,1)*1.96*boot.sd
boot.sd
# method 2
int.boot <- c(quantile(boot.mean, 0.025), quantile(boot.mean, 0.975))
int.boot
### kurtosis ###
install.packages('moments')
library(moments)
### n = 2000 ###
time1 <- data$time
kurtosis(time1)
normal_kurtosis= NULL
for (i in 1:2000)
  normal_kurtosis[i]=kurtosis(rnorm(91))
hist(normal_kurtosis)
mean(normal_kurtosis)
### n = 4000 ###
time1 <- data$time
kurtosis(time1)
normal_kurtosis= NULL
for (i in 1:4000)
  normal_kurtosis[i]=kurtosis(rnorm(91))
hist(normal_kurtosis)
mean(normal_kurtosis)
```

Scenario 4:

```{r}
attach(data)
attitude <- like[like != 1 & like != 99]
count.attitude <- table(attitude)
barplot(count.attitude,main = "attitude towards playing video games",names.arg = c("very much","somewhat","not really","not at all"))

```

```{r}
data <- read.table("videodata.txt", header=TRUE)
reg.ind <- which(data['like'] != 99 & data['like'] != 1)
data <- data[reg.ind,]
male.ind <- which(data['sex'] == 1)
data.male <- data[male.ind,]
female.ind <- which(data['sex']== 0)
data.female <- data[female.ind,]
male.count <- table(data.male$like)
female.count <- table(data.female$like)
barplot(male.count,main = "male attitude towards playing video games",names.arg = c("very much","somewhat","not really","not at all"))
barplot(female.count,main = "female attitude towards playing video games",names.arg = c("very much","somewhat","not really","not at all"),ylim = c(0,25))
```

```{r}
data <- read.table("videodata.txt", header=TRUE)
reg.ind <- which(data['like'] != 99 & data['like'] != 1)
data <- data[reg.ind,]
own.ind <- which(data['own'] == 1)
data.own <- data[own.ind,]
notown.ind <- which(data['own']== 0)
data.notown <- data[notown.ind,]
own.count <- table(data.own$like)
notown.count <- table(data.notown$like)
barplot(own.count,main = "PC owners attitude towards playing video games",names.arg = c("very much","somewhat","not really","not at all"))
barplot(notown.count,main = "non PC owners attitude towards playing video games",names.arg = c("very much","somewhat","not really","not at all"),ylim = c(0,30))
```

```{r}
data <- read.table("videodata.txt", header=TRUE)
reg.ind <- which(data['like'] != 99 & data['like'] != 1 & data['educ'] != 99)
data <- data[reg.ind,]
educ.count <- round(prop.table(table(data$educ,data$like),1),2)
barplot(educ.count,main ="barplot for educational&attitute towardsplaying video games", xlab = "Preference", ylab = "Percantage", col = c("red", "blue"),beside = T, names.arg = c("very much","somewhat","not really"))

```

```{r}
data <- read.table("videodata.txt", header=TRUE)
reg.ind <- which(data['like'] != 99 & data['like'] != 1)
data <- data[reg.ind,]
home.count <- round(prop.table(table(data$home,data$like),1),2)
barplot(home.count,main ="barplot for home&attitute towardsplaying video games", xlab = "Preference", ylab = "Percantage", col = c("red", "blue"),beside = T, names.arg = c("very much","somewhat","not really","not really"),ylim = c(0,0.6))
legend("topleft",c("PC","no PC"),cex=1.3,bty = "n",fill = c("red","blue"))
```

```{r}
data <- read.table("videodata.txt", header=TRUE)
reg.ind <- which(data['like'] != 99 & data['like'] != 1)
data <- data[reg.ind,]
data['age.category'] <- rep(NA, dim(data)[1])
for(i in 1:dim(data)[1]){
  age <- data[i, 'age']
  if(age < 20){
    data[i, 'age.category'] = 0
  }else{
    data[i, 'age.category'] = 1
  }
}
age.count <- round(prop.table(table(data$age.category,data$like),1),2)
barplot(age.count,main ="barplot for age & attitute towardsplaying video games", xlab = "Preference", ylab = "Percantage", col = c("red", "blue"),beside = T, names.arg = c("very much","somewhat","not really","not really"),ylim = c(0,0.6))
```

```
legend("topleft",c("< 20",">= 20"),cex=1.3,bty = "n",fill = c("red","blue"))
```

```{r}
data <- read.table("videodata.txt", header=TRUE)
reg.ind <- which(data['like'] != 99 & data['like'] != 1 & data['educ'] != 99)
data <- data[reg.ind,]
data['age.category'] <- rep(NA, dim(data)[1])
for(i in 1:dim(data)[1]){
  age <- data[i, 'age']
  if(age < 20){
    data[i, 'age.category'] = 0
  }else{
    data[i, 'age.category'] = 1
  }
}
data['dis_like'] <- rep(NA, dim(data)[1])
for(i in 1:dim(data)[1]){
  like <- data[i, 'like']
  if(like==4 || like==5){
    data[i, 'dis_like'] = 0
  }else{
    data[i, 'dis_like'] = 1
  }
}
install.packages("tree")
library(tree)
data.tree <- tree(dis_like~educ+sex+age.category+home+own, data=data)
plot(data.tree, type="uniform")
text(data.tree)
```

```{r}
data <- read.table("videodata.txt", header=TRUE)
reg.ind <- which(data['freq'] != 99)
data <- data[reg.ind,]
data['freq_group'] <- rep(NA, dim(data)[1])
for(i in 1:dim(data)[1]){
  freq <- data[i, 'freq']
  if(freq < 3){
    data[i, 'freq_group'] = 1
```

```
  }else{
    data[i, 'freq_group'] = 2
  }
}
freqgroup.count <- round(prop.table(table(data$freq_group,data$grade),1),2)
barplot(freqgroup.count,main ="barplot for frequence of playing video games & grades", xlab = "grade", ylab = "frequency of playing videogames ",  col = c("red", "blue"),beside = T, names.arg = c("C","B","A"),ylim = c(0,0.7))
legend("topleft",c("daily or weekly","monthlu or semesterly"),cex=1.3,bty = "n",fill = c("red","blue"))
```

Scenrio 5:

```{r}
setwd("/Users/mengtingchen/Desktop/math189/cs2")
data <- read.table("videodata.txt", header=TRUE)
```

```{r}
data[data == 99 |data$like ==1] <- NA
sum(is.na(data))
```

```{r}
female.ind <- which(data['sex'] == 0)
female.ind
```

```{r}
male.ind <- which(data['sex'] == 1)
male.ind
```

```{r}
female <- data[female.ind,]
male <- data[male.ind,]
```

```{r}
prefgender <- round(prop.table(table(data$sex, data$like),2),2)
```

round(prop.table(table(data$sex, data$like), 2), 2)
```
difference in work
```{r}
dislike.ind <- which(data['like'] == 4 | data['like'] == 5)
data.dislike <- data[dislike.ind,]
```

```{r}
like.ind <- which(data['like'] == 2 | data['like'] == 3)
data.like <- data[like.ind,]
```

```{r}
summary(data.like)
summary(data.dislike)
```
```{r}
hist(data.like$work, freq = F, main = "Histogram of Work hours for Like", xlim = c(0,60), ylim = range(0, 0.4), xlab = "Work Hours", las = 1, breaks = 20, col = 4)
```

```{r}
hist(data.dislike$work, freq = F, main = "Histogram of Work hours for Dislike", xlim = c(0,60), ylim = range(0,0.4), xlab = "Work Hours", las = 1, breaks = 15, col = 2)
```

```{r}
attach(data)
data$like[like ==2 | like ==3] <- "Like"
data$like[like ==4 | like ==5] <- "Dislike"
detach(data)
```

```{r}
boxplot(data$work~data$like, data)
```

```{r}

```
genderpref <- round(prop.table(table(data$sex, data$like), 2),2)
round(prop.table(table(data$sex, data$like), 2),2)
```

```{r}
barplot(genderpref, main = "Barplot of Gender & Preference", xlab = "Preference", ylab = "Percentage", ylim = range(0, 1), las = 1, col = c("red", "blue"), beside = T, names.arg = c("Dislike", "Like"))
legend("topleft", c("Female","Male"), cex=1.3, bty="n", fill= c("red","blue"))
```

```{r}
attach(data)
data$work[work > 0] <- "Haswork"
data$work[work = 0] <- "Nowork"
detach(data)
```

```{r}
workpref <- round(prop.table(table(data$work, data$like), 2),2)
round(prop.table(table(data$work, data$like), 2),2)
```

```{r}
barplot(workpref, main = "Barplot of Work & Preference", xlab = "Preference", ylab = "Percentage", ylim = range(0, 1), las = 1, col = c("red", "blue"), beside = T, names.arg = c("Dislike", "Like"))
legend("topleft", c("No work","Work"), cex=1.3, bty="n", fill= c("red","blue"))
```

```{r}
attach(data)
data$own[own = 1] <- "PC"
data$own[own = 0] <- "No PC"
detach(data)
```

```{r}
PCpref <- round(prop.table(table(data$own, data$like), 2),2)
round(prop.table(table(data$own, data$like), 2),2)
```

```{r}
barplot(PCpref, main = "Barplot of PC & Preference", xlab = "Preference", ylab = "Percentage",
ylim = range(0, 1), las = 1, col = c("red", "blue"), beside = T, names.arg = c("Dislike", "Like"))
legend("topright", c("No PC", "Own PC"), cex=1.3, bty="n", fill= c("red","blue"))
```

```{r}
qqplot(data.like$work, data.dislike$work)
abline(c(0,1))

wilcox.test(data.like$work, data.dislike$work)
```

Scenrio 6:

```{r}
data <- read.table("videodata.txt", header = TRUE)
data$grade
data['grade']
attach(data)
grade
i <- 91
while(i < 95){
  grade[i+1] = 0
  i = i +1
  print(grade[i])
  }
grade[92]
length(grade)
count <- table(grade)
barplot(count,main = "expected grade distribution",names.arg=c("D","C","B","A"),ylim = c(0,60), col = c(1,4,2,3), xlab = "grade expected", ylab = "number of people")
mean.grade <- mean(grade)
mean.grade

new_vector <- c(31/91,0.2,31/95,52/91,0.3,52/95,8/91,0.4,8/95,0,0.1,5/95)
```

```
barplot(new_vector,col = c(rgb(0.3,0.1,0.4,0.6) , rgb(0.3,0.5,0.4,0.6), rgb=(3)),xlab="grade",ylab="grade assigned percentage",ylim = c(0,1),main = "expected grade distribution",names.arg=c("A","A*","A", "B","B*","B", "C","C*","C", "D","D*","D"),space=c(0,0,0,0.3,0,0,0.3,0,0,0.3,0,0))
legend("topleft", legend = c("old expected grade","target grade","new expected grade" ) ,
       col = c(rgb(0.3,0.1,0.4,0.6) , rgb(0.3,0.5,0.4,0.6),rgb=(3)) ,
       bty = "n", pch=20 , pt.cex = 2, cex = 0.8, horiz = FALSE, inset = c(0.05, 0.05))

data <- read.table("videodata.txt", header = TRUE)
grade4.ind <- which(data['grade']==4)
data.grade4 <- data[grade4.ind,]
grade3.ind <- which(data['grade']==3)
data.grade3 <- data[grade3.ind,]
grade2.ind <- which(data['grade']==2)
data.grade2 <- data[grade2.ind,]
grade1.ind <- which(data['grade']==1)
data.grade1 <- data[grade1.ind,]
new_vector <- c(31/91,0.2,52/91,0.3,8/91,0.4,0,0.1)
barplot(new_vector,col = c(rgb(0.3,0.1,0.4,0.6) , rgb(0.3,0.5,0.4,0.6)),xlab="grade",ylab="grade assigned percentage",ylim = c(0,1),main = "expected grade distribution",names.arg=c("A","A*","B","B*","C","C*","D","D*"),space=c(0.2,0,0.2,0,0.2,0,0.2,0))
legend("topleft", legend = c("expected grade","target grade" ) ,
       col = c(rgb(0.3,0.1,0.4,0.6) , rgb(0.3,0.5,0.4,0.6)) ,
       bty = "n", pch=20 , pt.cex = 2, cex = 0.8, horiz = FALSE, inset = c(0.05, 0.05))
summary(grade)
grade
target2 <- array(1:95)
i <- 0
while(i < 95*0.2){
  target2[i+1] = 4
  i = i + 1

}
while(i < 95*0.5){
  target2[i+1] =3
  i = i+1
}
while(i < 95*0.9){
```

```
  target2[i+1]=2
  i = i +1

}
while(i < 95){
  target2[i+1] =1
  i = i +1
}
target1 <- array(1:91)
i <- 0
while(i < 91*0.2){
  target1[i+1] = 4
  i = i + 1

}
while(i < 91*0.5){
  target1[i+1] =3
  i = i+1
}
while(i < 91*0.9){
  target1[i+1]=2
  i = i +1

}
while(i < 91){
  target1[i+1] =1
  i = i +1
}
ks.test(grade, target2)
ks.test(data$grade, target1)
mean(target1)
mean(target2)
```