Who plays video games

Lu Xu and Zehui(Barry) Zhang

Author contributions

Both Lu Xu and Zehui (Barry) Zhang contributed the doc wriation. We talked about ideas about every single question in this analysis. Every time we have some missing points, the other one will make the complements. Then attributing to a comprehensive idea, methods. In the code part, Barry wrote the majority and Lu did the complement work. In advanced analysis, Lu brought up the idea, and Barry processed it.

Index

1.1 Introduction

One of the designers of the new computer lab wants to get useful information or responses from the students about their intention of playing video games. During Fall 1994, there was an investigation happening at statistics courses at University of California, Berkeley. This investigation targets a population of 3000 to 4000 students that are taking statistics courses at UC Berkeley. There are 314 students in Introductory Probability and Statistics, Section 1 during that year and Fall semester. Out of the 314 students, they randomly selected 95 students to participate in the survey and only 91 students actually completed the survey. This survey is about how often they play video games and what they like and dislike about the games? Which determine the extent of the students playing video games and the most fun and least fun aspects of video games.

The main target of this analysis is to investigate the motivation of playing video games and compare the differences between people who like to play games and those who don't like to play video games. In this analysis, we use histogram, bootstrap, qq plot,hypothesis test in Advanced Analysis, and etc... In R we find their point estimate and the confidence interval. For the reasons why people play or not play video games , it is categorical variables. Therefore we use the pie chart, bar plot and segmented bar to compare to make a clear comparison.

Data

The data set is based on those students who responded to the questionnaire so there is some missing data. Such as, those participants who did not like playing video games and who had never played video games were skipped many questions. Those missing values (no answered or improperly answered) are shown as “99” in the data set. There was also a follow up survey which gave more response and to give a reason why they like or dislike playing games. Those selected students were the ones who had taken the second exam of that semester and the exam was taken a week prior to the survey. The pseudo random generator was used to select 95 students from 314 students. All the answer was anonymity. All of the students who complete the survey are all briefly informed of their reason or goal to take the survey and know it guarantees anonymity.

2. Basic Analysis

2.0 Data Processing

Method

Loaded the data to R. Our basic analysis focuses on unit characteristics of the students who played video games in the week prior to the survey. The data set is randomly sampled. The missing values have to change to NA (not available). Since as well to reduce the errors, we cannot simply remove all missing datas here. Because not all variables are missed in single data observation. Therefore, the value “99” in the different columns has to be replaced by NA.

Analysis

The original sample set has 91 observations. After being replaced by NA, the number of observations does not change. From the new dataset, there are seldom students who never played and do not like it at all which are represented by “1” and “5” respectively(totally 8.8%). 23 students (25.3 %) very like playing video games, 46 students (50.5%) somewhat like playing video games, and 13 students (14.3%) do not really like playing video games. Most of the students somewhat somewhat like playing video games.

Conclusion

Replacing the values “99” from the data set with NA does not induce the reduction of the number of observations of the data sets. There are still 91 observations. Compared with the population size 314, this is an ideal sample size, large enough here(almost a third).Also, those “99” values could make distrubed to the right information the data set initially conveyed. Moreover, after simply analyzing the degree of likeness for students with numbers. Most of the students somewhat like playing video games.

2.1 Fraction Analysis

Methods

Separate the data into 2 parts, people played video games and not played video games prior to the survey. Then find the fraction X bar, and this fraction to estimate the overall portion (pi) of students who played video and not. Then utilize the X bar to find the estimator of Standard Error. Therefore construct a confidence interval with default significance level of 5%.

Analysis

Here in this case, the people who played video games are considered as those whose playing times are higher than 0. And the people who didn't play video games at all are those whose playing times are equal to 0. The number of people who played video games in the data set is 34. The number of people who did not play video games is 57. The X bar is 0.374(37.4%) representing the fraction of students who played video games in the week prior to the survey. Also, the standard error of playing video games is about 0.051. Since we have a really small number of standard error which is pretty close to 0, there is high credibility and reliability of this estimate. Because the sampling distribution of the proportion indeed follows a normal distribution by Central Limit Theorem. It is able to represent around 37.4% students who have indeed played video games. The majority of students did not play video games prior to the survey even though they have a high likeness of playing video games, which is 3.289. Here, we constructed a 95% confidence interval for the real proportion which is 0.272 to 0.475. It means that about 95% of the real value of proportion will lay under this interval.

Conclusion

By calculating the X bar, the point estimate, and Standard Error from the sample set, it shows that the students are not that much playing video games prior to the survey. This could be due to the fact that students were affected by the exam during the week. Moreover, the tiny standard error, 0.051, here makes our point estimate and interval estimate, confidence interval, super powerful and persuasive. This estimate could represent the overall behavior of the total population, that minor students played video games in the week prior to the survey -- during the exam week.

2.2 Numerical Analysis for Playing Time vs. Frequency of Play

Methods

Separating the original data set into 4 subset, which is grouped by the frequency of the play video games. Then calculate the mean of playing time of each group. Then go check the difference between these 4 differences.

Analysis

Subset	Groups	#of Total Student	#of People Played Games and Percentage	Percentage of People	Mean	Standard Error

1	Students who play video games every single day	9	7 (77.7%)	9.9%	4.45	1.86
2	Students who play video weekly	28	24 (85.7%)	30.8%	2.54	1.04
3	Students who play video games monthly	18	2 (11.1%)	9.8%	0.06	0.038
4	Students who played semesterly	23	1 (4.34%)	25.3%	0.04	0.043

From the above table, the students who play video games weekly have a higher percentage of people playing video games. The highest amount of time spent playing games are those students who play video games every single day which have a mean 4.45 hours. Compared to those who play video games weekly, they play 1.91 hours more. And the students playing monthly and semesterly is 4.4 lower than the people who play video games every single day. This is really consistent with the frequency. Because people who played video games more frequently means people are more likely to play video games.

There is a reduction in the number of people playing video games in each subset prior to the survey. Students who play video games every day are reduced to 7 from 9 and students who play weekly are 28 before and now 24. For students who played weekly and everyday, it still has 75% above of students still playing video games prior to the survey. However, the students played monthly have changed from 18 to 2 and this number changed from 23 to 1 for students played semesterly . Those who played monthly and semesterly are only left with less than 12 % of students playing video games prior to the survey. Especially the students who play semesterly, there are less than 5 % of them still playing video games. Also, from the standard error we also know that the monthly and semesterly SE are more concentrated than play everyday and weekly. Therefore those who play video games more frequently are those people who spend more time on playing video games. It has less effect on them whether there is an exam or not. However, it is totally different to those who play video games less frequently. Students who play video games not that often are having a huge effect on students and the time playing video games. Overally, it is a similar fact that there was an exam in the week prior to the survey. It will reduce the number of people playing video games but there are more reductions on the people who play video games less frequently than those who play video games at a high frequency.

Conclusion

By different variables, we have calculated out and input in the data, it shows there is a reduction of the average time spent playing video games prior to the survey. The people who have less frequency to play video games are less likely to play video games prior to the survey. Due to the existence of an exam in the week prior to the survey, we could conclude that those who do not like playing video games are less likely to play video games. The interception could be they have some other task to complete such as an exam. On the other hand, the people who do really like to play video games(ex. those people who play video games everyday) are harder to concede. They will still go to play video games really frequently.

2.3 Average Amount of Time Spent Playing Video Games

Methods

We are going to find out sample statistics (average value of the characteristic): average amount of time students play video games the week prior to the survey. With the help of confidence intervals, we could also have an interval estimate. Finally the bootstrap will be used as a simulation study to determine the appropriateness of an interval estimate.
Mean(unbiased estimator) is defined as:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} Xi$$

Analysis

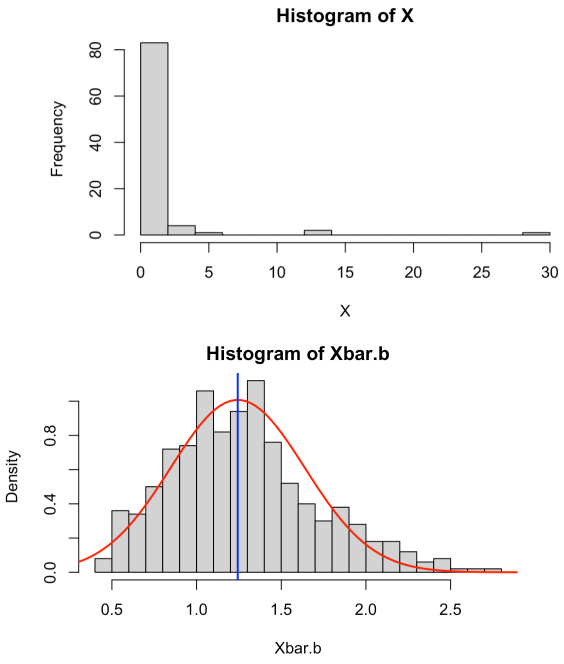


Figure 1: Histogram of X by playing video games status, X represents times (up left). Histogram of Xbar.b is about bootstrap by repeating 500 times (bottom left)and also we have a Normal Q-Q plot for this simulation study.(up right)

The histogram of playing video games times are extremely skewed to the right. By finding the skewness and kurtosis coefficients, which are 5.7, and 39.8 respectively, they are corresponding with what showed in the histogram. The higher these coefficients are, the more extreme skewed they are. Also, That is pretty obvious that there is a high frequency of people who are playing 0-2 hours of video games prior to the survey. Moreover, the real average time of this sample is 1.243, the point estimator, which indeed shows that this 0 to 2 interval covers the majority of students. Similarly the mode here is 0 which is also in the interval 0 to 2. The outlier does have a huge impact on the standard division which is 3.78, but they do not vary the mean and the point estimator too much in this data. Because the number or the frequency is too rare that can be ignored. Thus it is full of credibility that this estimator is a good point estimator.

By setting down the significance level as 5% here, we could construct a confidence interval (0.451, 2.034). It means that about 95% of them would cover the true value of the mean of the overall population. So as to check the validity of this confidence interval, we do a bootstrap.

From the bootstrap, we sample repeatedly 500 times and get the mean of the bootstrap distribution is 1.274, the standard Error is 0.019 and 95% confidence interval is 0.576 to 2.258 quantile. Compared to the previous mean and variance we got, it is pretty much the same. Because the bootstrap distribution should approximate the sampling distribution of the sample average and variance. The standard deviation is 0.388 which is quite close to the sample average of standard error, 0.395. Both of these represent that we got good estimators and accurate bootstrap. Bootstrap are suing for the provides number solution for non-standard situation so this was been used to solve for this complex sampling design. From the graph of histogram of Xbar. b, it seems to follow the shape of the right skewed distribution and the red line is used to represent the standard normal distribution with the same mean and the standard deviation.The Q-Q plots also verify this information because in the right line we could see that it has not really been touched after the theoretical quantiles go up.

Conclusion

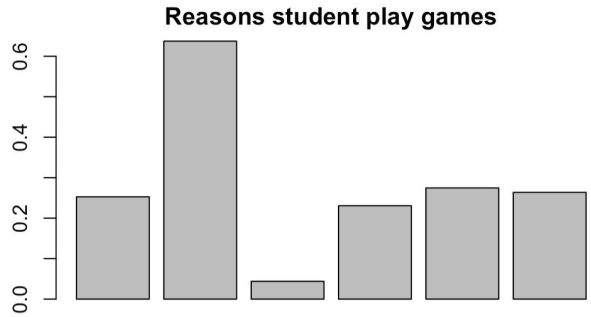
From the histogram and Q-Q plot, they have slightly skewed to the right distribution of the average amount of time spent playing video prior to the survey. The mean and the confidence interval for it is 1.24 and (0.451, 2.034). Then it is a non-standard numerical solution then bootstrap has checked that it is actually skewed to the right by the standard deviation 0.388 is similar to the standard error 0.395 and other numbers are also the same. It could mean that we did have actual estimates of the interval. The shape of the graph also can be seen from the skewness and kurtosis coefficients. They are really larger than 0 and 3 so it must not be the standard normal distribution.

2.4 “Attitude” for Student Playing Video Games

Methods

Find the mean and mode of likeness of the sample sets. And this mean and mode ’s meaning by comparing it with different meanings of likeness. So as to find the reason why students like/ dislike video games, we could go and find the proportion of different reasons for both like and dislike. Put several most common and popular reasons into my shortlist, and leave the other reason option. For those who skipped some questions, treat them as NA.

Analysis



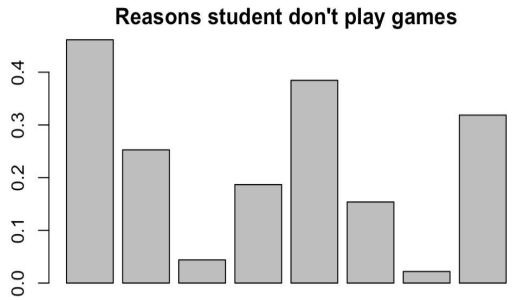


Figure 1: From left to right, graphics/realism is 26%, Relaxation is 66%, Eye/hand coordination is 5%, Mental challenge is 24%, feeling of mastery is 28% and bored is 27%.(Left)
Figure2:From left to right, too much time is 48%, frustration is 26%, lonely is 6%, too many rules is 19%, costs too much is 40%, boring is 17%, friend’s don’t play is 17%, and it is pointless is 33%.(Right)

Based on the bar plot, most of the students played video games for relaxation, feeling of mastery and boredom. But the percentage of playing games for relaxation is much higher than other attitudes. On the other hand, the other majority of students do not play video games because they think playing video games is spending too much time, and costing too much, and are really pointless, and frustration while playing video games. The mean of the likeness is 3.02 and the mode of the likeness is 3. This means that students are somewhat liked to play video games. Most students enjoy playing video games, but they probably do not have much time or do not have too much to spend on the games. By making a comparison that students are relaxed to playing video games is 66% but the students who are frustrated to play video games is 26%. There are 40% higher on the students feeling video games are more relaxing. Somewhat like playing video games means more students prefer to play video games so that the attitude percentage actually shows that most students feel playing video games are helping them relax then those who think playing video games are spending too much time. Some people do not like playing video games because they feel it is pointless but some who really like playing video games feel that video games could bring the feeling of mastery. There are a similar percentage who choose those two with only a 5 % difference. It also shows that students somewhat liked to

play video games. Because they achieve more achievement in the game world than the real world but the people who do not play video games might not know. Some of the students were asked to skip those questions because they do not play video games or never play video games before. Video games are not really educational to people because 66% of people like to play video games for relaxation are 66%. If the video games make them so frustrated (26%) then they might not play it at all. Playing video games is educational, and may already include the percentage of people playing video games for mental challenges. My short list is that students played video games for relaxation, feeling of mastery and boredom and disliked playing video games caused by too much time, pointlessness, cost too much and frustration. Those are the most common reasons and it did show most people select those reasons in the graph.

Conclusion

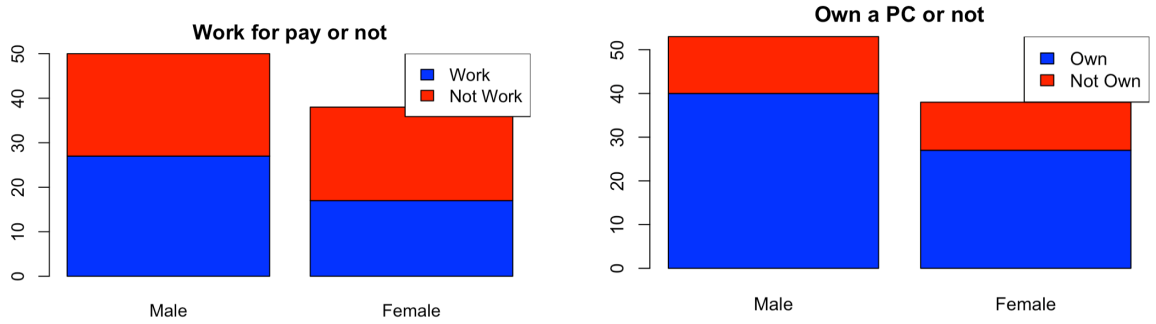
The mean of the likeness is around 3 which means people are somewhat like to play video games. Based on the percentage of playing video games it also shows the people who selected to play video games is because of relaxation. And others think playing video games is too frustrating, which is totally opposite. By this too opposite choice option, we have more people selected playing video games could make people relax. Therefore people are somewhat likely to play verified video games. Students are enjoying playing video games and the reasons for them playing video games are relaxation, feeling of mastery and boredom because those nonrespondents think video games are education. If video games become educational then there are no people who feel relaxed while playing video games. The reasons they dislike playing video games are spending too much time, costing too much, and frustrating. Frustrating also points out here because when the video games are educational then at least 28 % of people feel it is frustrating.

2.5 Differences for Students Who Like to Play Video Games vs. Who do not Like

Methods

First, in this case we define that those who work for pay are represented as work is greater than 0 and not NA shown in the data set; those who work did not pay is shown as 0.By using the bar plot we could construct and show the difference apparently and straightforwardly.

Analysis



By converting the data into two bar plots shown above, we could easily see and compare the difference between those who like to play video games and those who don't. As defined in the methods part, we consider people whose work time more than 0 is work for pay, and rest of them are not. Roughly, the proportion of students who work for pay and not is similar in both male and female. Even though the total number of male in this sample is larger than females, female students who do not work for pay are slightly weight more in the female group. Instead this proportion is altered in the male group. By looking at the area of red blocks of both male and female in the first bar plot, we can find out that the area is almost the same, but the area of blue blocks varied a lot. In other words, the number of male students who work for pay is higher than female students.

For the second bar plot, it represents the proportion of male and female students who own or not own a personal computer(PC). Still, due to the unchanged data set, the total number of male students is higher than female students. Apparently, due to the different base, the number of the students in male and female groups, the proportion of students who do not use a computer is higher in the female group. Same as above, the red blocks in both males and females are about the same, which means the number of people who don't have computers is almost the same in male and female students.

Conclusion

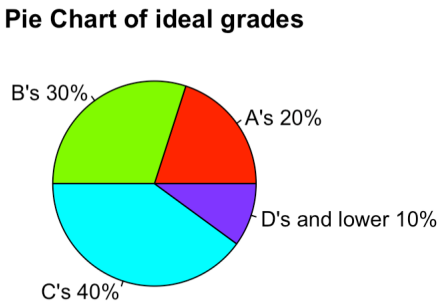
By the bar plots, we clearly see the difference between male and female students and how variables behave differently within the male and female in the same graph. Since the area of the blocks in the bar plot means the numbers. We conclude that the numbers of students who do not work for pay and who do not own a computer are similar in both males and females. But due to the total number of students in male and female groups is different. So the proportions will be different as well.

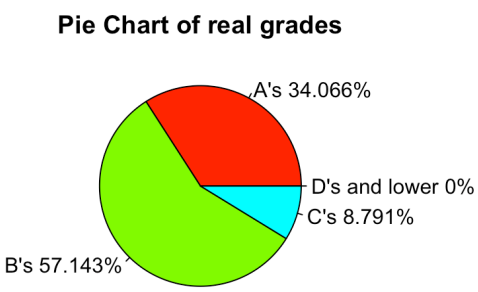
2.6 Grade Analysis

Method

Constructing two pie charts graphically shows the proportion of target grading distribution and the real condition.

Analysis





These two pie charts show the difference in the distribution pretty clearly. It is obvious that this is not a good match with the target distribution. In this sample set, there are no D's and lower. They occupied 0% of overall. Compared with the targets, it varies a lot. What's more, in this sample, the majority of students gained a B which is 57% of the overall. In target grade distribution, the majority should be the C's which should be 40%. While, in the real condition, C's are less than 9%. And the same thing happened on A's. The target grade distribution contains 20% of A's, but in reality, this number goes up to 34%. If we consider A's and B's as good grades, then, overall the good grades dominate this sample. More than 90% of students gained at least B's in this class. Oppositely, in the ideal or the target grade distribution, this number should only be 50%, which is quite fair and reasonable. Because, separating students in a course to half good and half not pretty good is pretty reasonable and fair to every single student in this course. Moreover, 90% of students who gained a great grade finally in this course means nothing. It reflects nothing about who spends more time and pays more attention on this course. Almost everyone is good. This fact makes students common, not pretty distinctive.

If we count all nonrespondents as failing in this course, definitely, the number of students in the D's and lower will be increased. But in this sample, there are no unavailable data points in grades.

Conclusion

The grade distribution is pretty different from the targets. In the target grading distribution, students who gained good grades and not very good grades are equally distributed. But in the sample, the real condition, students who have good grades almost dominate the courses. More than 90% of the students who got at least B's in this course.

3.0 Advanced Analysis

In the prior analysis, we have estimated the average playing time. So here in this part we will go beyond that. What is the average playing time of male students and female students? Are they the same, or different. In this part we will find out this problem by hypothesis test with the assistant by R.
Null hypothesis: the average playing time is the same in male and female students ($\mu m = \mu f$)
Alternative hypothesis: the average playing time is the different ($\mu m \neq \mu f$)

Methods
So as to test the mean hypothesis test of two independent populations, we have no idea about the variance of these two groups. Since we only have a sample rather than the whole population. Thus we have to do a variance hypothesis test first. Based on the result of the variance test, we can make the decision to choose which strategy of our mean hypothesis test will use here(determine the degree of freedom)

Analysis
Variance HT:
Null hypothesis: variances of male and female are same($\sigma m^{(2)} = \sigma f^{(2)}$)
Alternative hypothesis: variances are not same($\sigma m^{(2)} \neq \sigma f^{(2)}$)

```
        F test to compare two variances

data:  Male$time and Female$time
F = 3.4393, num df = 52, denom df = 37, p-value = 0.0001538
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.850596 6.196606
sample estimates:
ratio of variances
          3.439343
```

By doing the F test(or variance test) in R we can find out that the p-value here is too small so that we can reject the null hypothesis in preferring the alternative hypothesis that the variance is not the same. Thus we can do the means hypothesis test with unequal variances.

```
        Welch Two Sample t-test

data:  Male$time and Female$time
t = 1.156, df = 83.441, p-value = 0.251
alternative hypothesis: true difference in means is not equal to 0
98 percent confidence interval:
 -0.8900313  2.5824841
sample estimates:
mean of x mean of y
 1.596226  0.750000
```

So, in R, we find the p-value here is 0.251 which is much bigger than the significance level, 0.01,we chose. Thus we keep the null hypothesis, which is that the means are the same in male and female students.

Conclusion

By doing these two hypothesis tests from this data, we can conclude that the time of playing video games for male and female students are almost the same. Since the p-value is quite larger than the significance level we choose, the conclusion is quite reliable.

4.0 Discussion and Conclusion

For this analysis, it randomly selected 95 students from the 314 students. Only 91 of them participated in the survey, so we comprehensively analyze the data set with this 91 participated in the survey.We find out that most people are somewhat playing video games. The top three reasons they like to play video games are relaxation, feeling of mastery and boredom. People who do not like playing games think it takes too much time, costs too much and is pointless. From those top choices of the reason why they play or not playing video games, it seems they are most enjoying playing video games but there are some obstacles for them to playing video games. Such as, based on the point estimate and interval estimate from the proportion, it shows the most students are not playing video games prior to the survey. This is because of the exam during that week. Even by comparing the frequency, it also shows that there is a huge impact on the students who are playing video games. Especially the students who only play video games semesterly less than 5% of them are playing video games prior to the survey. Also, video games for education would not have a big impact on the students because most of the reason they play video games is for relaxation. If they play video games that are educational then it will lose many players. Another factor is that male students work for pay slightly more than the female students and this is the same result to those who own a computer.

Based on the grades, there is no really huge impact on their expectation of their grades. Some of them are also getting higher grades than what they expected but those who are thinking about getting A are mostly getting the grade they want. Also female and male students are having exactly the same time playing video games. The fact of gender did not really impact the likeness of playing video games.

Based on our current result, we are only looking for the reason why they are playing video games and why they do not play video games. There are some people who have never played video games so his group of students have been put in NA in our data. My suggestion is that we could let those students who have never played video games teach them for a week and test whether they like to play video games or not and what they like or dislike about the video games. Those people could also be included in our data set.