Xueru Xie, Yichi Zhang, Jinsong Yang

Professor Schwartzman

MATH 189 Report 2

03 May 2020

How Can a Survey Reveal Useful Information?

In lecture 3, professor Schwartzman had a case study involving a survey conducted in school. This survey was conducted by students that were taking advanced statistics during that time. The motive behind this survey was to give ideas to the design committee to design computer labs for courses. Looking at the survey, we derived that there was an estimate of 38% students who played a video game before the survey, and an exam before the survey would lower our estimation of students playing games in a typical week, and most students enjoyed playing video games overall. Before arriving at our conclusion, we wanted to point out that this survey was considered as good survey research because it contained all the great characteristics. For example, this survey was impartial because it did not favor any kind of opinion in the survey questions. This survey could be replicable because it was a class survey developed by students, which was something that we could do in a normal statistics class. The survey was useful enough to answer each one of our given scenarios with us using analysis with graphs and tables. In our remaining paper, we will discuss each scenario and the answer to it.

Data

Looking at the given datasets, there was one called “videodata.txt” and “videoMultiple.txt”. The first text file contained survey responses that recorded the general

information about the students and their gaming routine. For example, it had recorded students information like their sex, age, grade expectation, and etc. It also had gaming information like the hour of game time before the survey, the scale that measured how much a student liked playing games, the location of playing, and etc. The second dataset "videoMultiple.txt" contained information about whether a student liked to play certain types of games. Students would answer a "1" for affirmative and "0" for negative. Types of games included action, adventure, simulation, sports, and etc.

The first step of our data analysis was data cleaning. We noticed that the data was pre-cleaned already. Some none applicable answers were filled with the number "99". So, we moved on to looking at the scenarios.

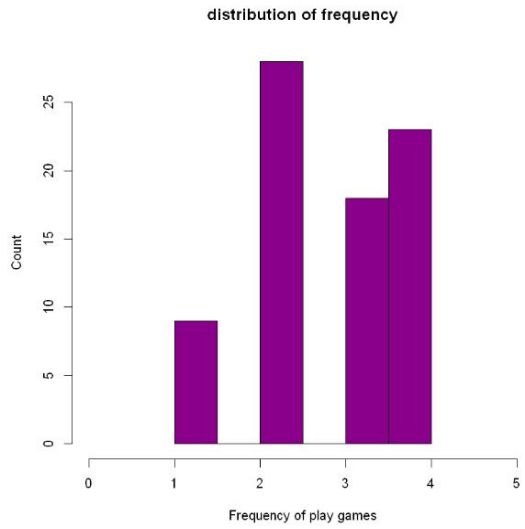
Scenario 1

To provide an estimate for the fraction of students who played a video game in the week prior to the survey, we would use "videodata.txt" to answer this problem. Out of the 91 surveyed students, there were 34 students that played a video game in the week prior to the survey. The proportion of those students would be around 0.38, or 38%. Overall, the average hour of all 91 students playing a video game was about 1.24 hours. Within the 38% of students that played games, the average hour would be about 3.33 hours.

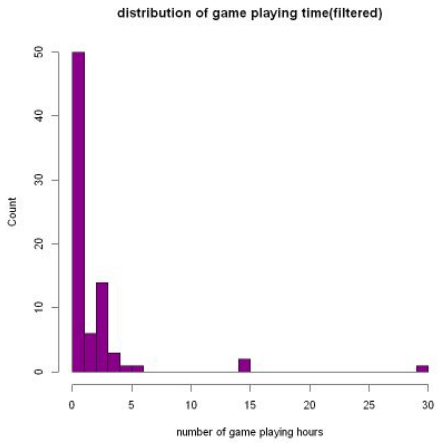
Scenario 2

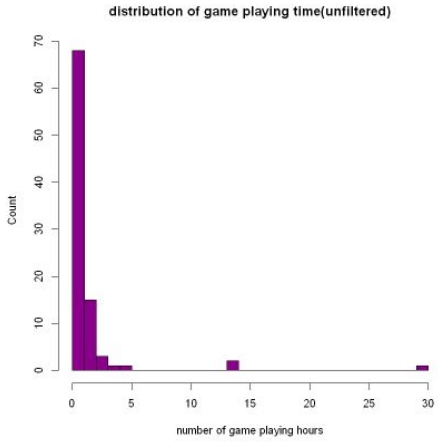
In scenario 2, to know the amount of time spent playing video games in the week prior to the survey compared to the reported frequency, we decided to visualize this using graphs. Using R's unique function, we saw that all students had the frequency of either 1 (daily), 2 (weekly), 3 (monthly), 4 (semesterly), or 99 (unanswered) of playing video games. Calculating the average

(excluding outliers 99 to avoid bias), we noticed that it was about 2.7, which was between 2 (weekly) and 3 (monthly). This meant that most students played games between weekly and monthly. This was graph distribution of the frequencies.



To know whether the exam in the week prior to the survey affected the estimates, we needed to distinguish 3 types of students: students that played games regardless of an exam, students that were gamers but did not play before an upcoming exam, and students that did not play games at all. To show if an exam could impact a student's gaming routine, we had to plot two graphs. One graph excluded students that did not play games at all, the other graph considered all the students. If there was a difference between two graphs, then the difference may be the amount of students that stopped playing games before the exam. This was the result below.





From these two histograms above, the top one represented all the gamers' playing time and the bottom one represented all the students' playing time. This difference of 91 - 78 = 13 students suggested that there were gamers that did not play any games before the exam. This difference would cause an underestimation in our previous scenario about the proportion of students playing video games overall.

Scenario 3

When considering making an interval estimate for the average amount of time spent playing video games in the week prior to the survey, we were going to use the confidence interval formula. Here was the result below.

```
In [12]: sd <- sd(df$time)
         sd
```

```
3.77704007736637
```

```
In [13]: upper_bound <- mean(df$time)+2*sd/sqrt(nrow(df))
         lower_bound <- mean(df$time)-2*sd/sqrt(nrow(df))
```

```
In [14]: c(lower_bound,upper_bound) # construct a 95% interval of estimation
```

```
0.450974374698314  2.03473991101597
```

From this interval, we wanted to see the sample below mean and above mean, the result was this.

```
In [15]: mean(df$time<lower_bound)
```

```
0.637362637362637
```

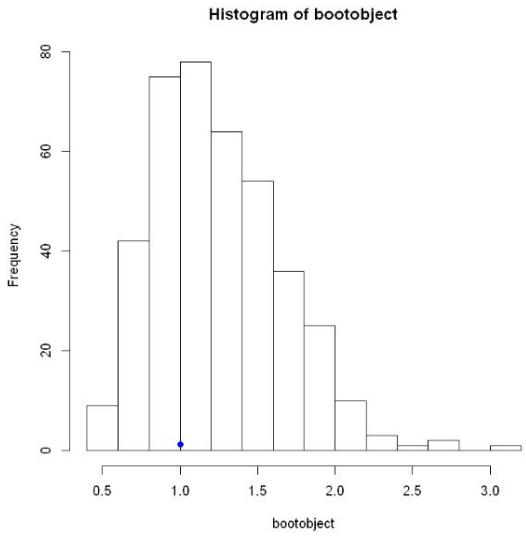
```
In [16]: mean(df$time>upper_bound)
```

```
0.0879120879120879
```

The two outputs meant that there were about 64% of the sample below the mean and about 9% above the mean in this estimation. Such an interval estimation is not so appropriate. This was because there were a lot of responses with 0s, causing a right skewed graph. On the left side(lower bound), about 62% of people did not play the game at all. From the graph above we can observe that the sample is not normally distributed but highly right skewed. In this case, we use bootstrap to help to make an interval estimation of mean.

From the lecture slides, according to the simple random sample probability model, the distribution of the sample should look roughly similar to that of the population. We could create a new population of 314 based on the sample and use this population, which we call the

bootstrap population,to find the probability distribution of the sample average. For every unit in the sample,we make 314/91 = 3.45 units in the bootstrap population with the same time value and round off to the nearest integer. This was the histogram as a result with the new confidence interval. By performing bootstrap with 400 trials, we finalize the 95% interval at (1.209,1.292).




```
c(boot_lower_bound,boot_upper_bound)
```

1.20902285535803 1.29153208969691

Scenario 4

Looking from videodata.txt file, the average number of hours spent on video games is around 1.24 and 0.23 or 23% of the students do not enjoy playing video games so simply from videodata.txt file it is safe to say that most students do enjoy playing video games. Looking from videoMultiple.txt however, videoMultiple.txt indicates that 93.41% of students recorded for a reason that they dislike about video games. Although 93.41% might seem like a high number but focusing on the number of hours they played in the prior week is a more significant factor since

the survey only asks for reasons why students may dislike video games but it cannot be used to conclude that the student dislikes video games.

If we were to create a list of top 3 reasons why students like/dislike video games, the list for why they like video games would be: Relaxation, Feeling of mastery and Bored.

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

The list for why they dislike video games would be : too much time, costs too much, and frustrating.

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

We wanted to further verify that our list of reasons are correct by looking at the correlation table:

	relax	coord	challenge	master	bored
relax	1.000000	0.155230	0.170941	0.071842	0.054554
coord	0.155230	1.000000	0.004423	-0.018122	-0.135496
challenge	0.170941	0.004423	1.000000	0.294738	0.012434
master	0.071842	-0.018122	0.294738	1.000000	-0.050951
bored	0.054554	-0.135496	0.012434	-0.050951	1.000000

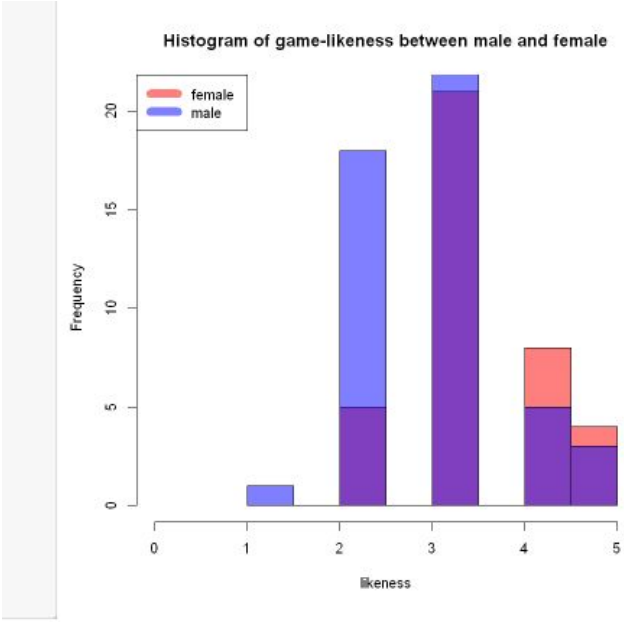
	time	frust	lonely	rules	cost	boring	friends	point
time	1.000000	-0.005396	-0.102255	-0.302060	0.004853	-0.110087	-0.148192	-0.097590
frust	-0.005396	1.000000	0.117298	0.164713	0.092865	-0.120665	-0.091956	-0.258023
lonely	-0.102255	0.117298	1.000000	-0.108185	-0.068182	-0.096138	-0.033674	-0.038808
rules	-0.302060	0.164713	-0.108185	1.000000	-0.167837	0.020856	0.117836	-0.102490
cost	0.004853	0.092865	-0.068182	-0.167837	1.000000	-0.295491	0.030562	-0.182323
boring	-0.110087	-0.120665	-0.096138	0.020856	-0.295491	1.000000	-0.067175	0.486611
friends	-0.148192	-0.091956	-0.033674	0.117836	0.030562	-0.067175	1.000000	0.054233
point	-0.097590	-0.258023	-0.038808	-0.102490	-0.182323	0.486611	0.054233	1.000000

Looking at the correlation table it seems that the top three reasons why students like video games are: ‘relax’, ‘graphic’, and ‘mastery’ and only relaxation and mastery matches our original prediction, concluding that these two reasons are the most important reason why students like to play video games. For reasons why students dislike video games are : ‘boring’, ‘rules’, and ‘frustrating’ so concluding that too much time, boring , and frustrating are still the reasons why students dislike video games.

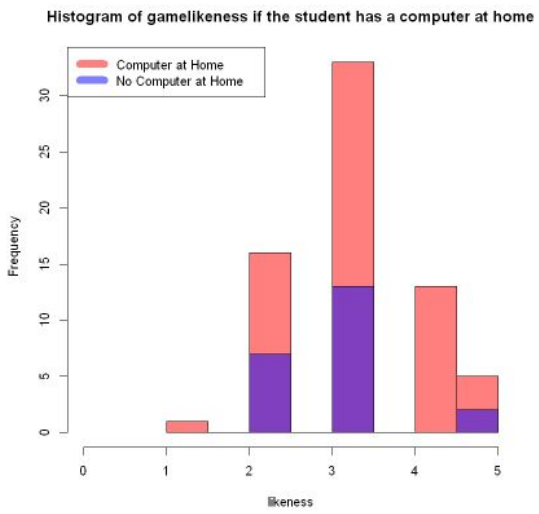
Scenario 5

Looking at the survey data, we wanted to see if there is a significant difference in people in people playing/not playing games. First, we examined if there were more boys playing games than girls in the given sample. To do that, we conducted a two sample z-test. Our null hypothesis

was that there is no difference between the proportion of male gamers vs female gamers, our sample data suggests that 90.5 percent of male played video games and 89.5 percent played video games. Before we actually conduct the z-test, we wanted to note that this z-test is not accurate enough to represent the population because we did not have enough data for the non-gamers, but we still conducted the test anyway to test our hypothesis. After plotting the numbers, we use the significance of $z > 1.96$ or < -1.96 because of the 95% interval. For male vs female, z is 2.24, which is greater than 1.96, which is significant. We rejected our null hypothesis, saying that there is a different proportion between male and female gamers (more male > female).



We used z-test to test similar factors. We tested the factor of working and owning a PC. After running the z-test, there is a difference between games who work and don't. We encounter a different result for PC owners. It turns out that there is a similar proportion for gamers with a PC and without.



We also tested for the factor of computers at home, the result is significant. The work is shown in the coding file.

In summary, male, computers at home, and working gamers had more proportion than their other ones, rejecting the null hypothesis and favoring the alternative. Our tests only show the similarity/dissimilarity between two groups, nothing more can be concluded by our findings.

Appendix

In this report, student Xueru Xie wrote the introduction and scenarios 1 to 3. Student Yichi Zhang had the responsibility to write all the R codes from scenarios 1 to 3. Student Jinsong Yang completed scenarios 4 to 5 and code for those two problems. The coding PDF contained all the work from scenarios 1 to 5.