Case Study 2:

Survey Data Analysis

Caitlyn Bryant - Applied Mathematics, 4th year
Edward Chao - Mathematics: Statistics, 4th year
Kieran Mann - Mathematics/Economics, 4th year
Matthew Kucirek - Applied Mathematics, 4th year

Table of Contents

Introduction

A survey conducted by students enrolled in an advanced statistics course at the University of California, Berkeley investigated who plays video games by selecting students at random to fill out a questionnaire. The survey's goal was to determine how often students play video games and which aspects they find the most and least fun. Our objective is to investigate the responses of the participants in the study with the intention of providing useful information about the students to the designers of the new computer lab. We will do this through six different scenarios.

When investigating the participants' responses, we will: (1) provide an estimate for the number of students who played a video game in the week prior to the survey, (2) check to see how the amount of time spent playing video games in the week prior to the survey compares to the reported frequency of play, (3) make an interval estimate for the average amount of time spent playing video games in the week prior to the survey, (4) examine whether the students enjoy playing video games, (5) look for the differences between those who like to play video games and those who don't, (6) further investigate the grade that students expect in the course.

The primary data that we studied consisted of the survey results from 95 randomly selected participants out of 315 students in a statistics course at the University of California, Berkeley during Fall 1994. 91 out of the 95 randomly selected participants completed the surveys. 3,000-4,000 students enroll in statistics courses at the University of California, Berkeley every year, of which half of these students take introductory statistics courses to satisfy their quantitative reasoning requirement. In order to help aid the instruction of these students, a series of computer labs were designed by both faculty and students. The computer labs are there to provide an alternative method for learning the concepts of statistics and probability. Some people have linked computer labs to video games. In order to help design a new computer lab for these students, our study will use the data collected from the surveys and provide useful information about the students.

The "big questions" that we will answer by our data analyses are: (1) What conclusions can we draw from our case study?; (2) What advice do we offer to the design committee?

In the remainder of our paper, we will summarize our research in all six of our scenarios and use statistical methods to come up with advice to give to the computer lab's design committee.

Research and Analysis

2.1. Proportion Estimation

To begin the analysis of the participants responses, we first took a look at different estimates of students who played video games the week prior to the survey. Our point estimate told us 0.374 of the students played video games that week. To check the accuracy, we created a 95% confidence interval estimate of (0.290, 0.458). From this, we gather that we are 95% confident that the true value of our parameter lies in our interval.

2.2. Frequency vs. Time Spent

Furthering our research, we next observed how the amount of time spent playing video games compared to the frequency of playing.

Frequency Played vs. Time Spent Playing

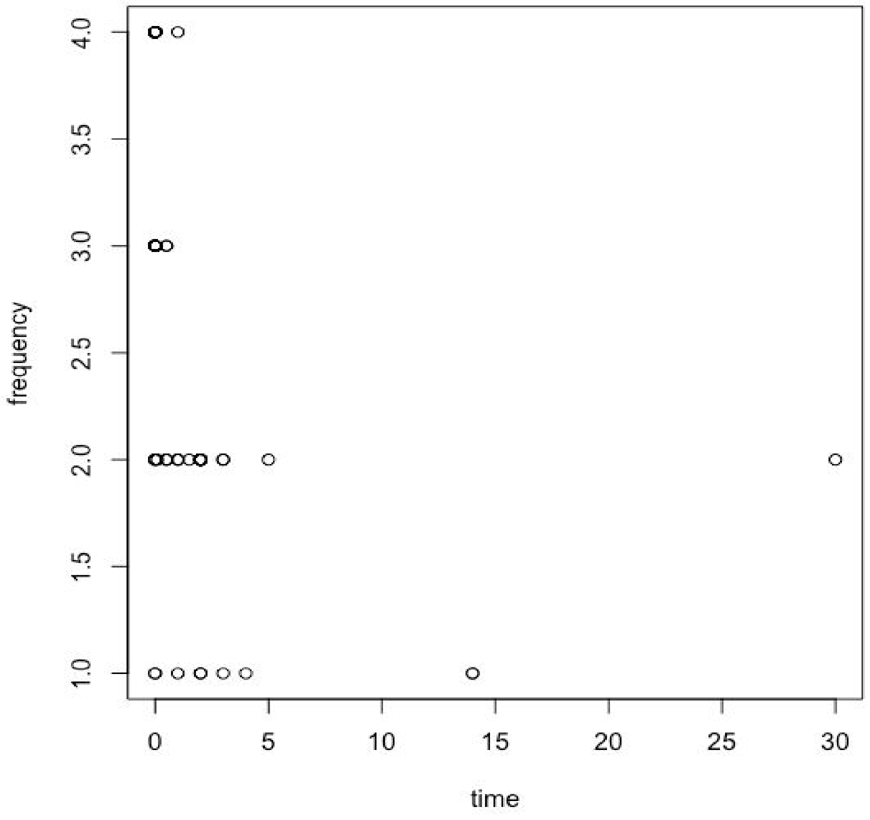


Figure 2.2.1: The frequency of how much the student played compared to the amount of time the student played video games. 1 = daily; 2 = weekly; 3 = monthly; 4 = semesterly

From our graph, students who played daily on average did not spend more than 5 hours a day. And when we take a look at a decreased frequency, the average hours spent weekly was comparable to the group that played daily, giving way to an estimate one hour spent each day on video games. And as the frequency decreased, we noticed a decrease in amount of time played as well. We gather from this data that there is a noticeable correlation between the amount of time played and the frequency that was played. When a student played more frequently, he or she played 1-5 hours daily, with one student even going as far as playing 14 hours a day. And as students reported a lesser frequency, the time spent on video games decreased.

This brought up the question of whether or not an upcoming exam would affect the any previous estimates or our comparison. So to provide more insight on this question, we took a look at whether or not the students were busy, and compared that to the amount of time spent playing.

Busy (or not) vs. Time Spent Playing

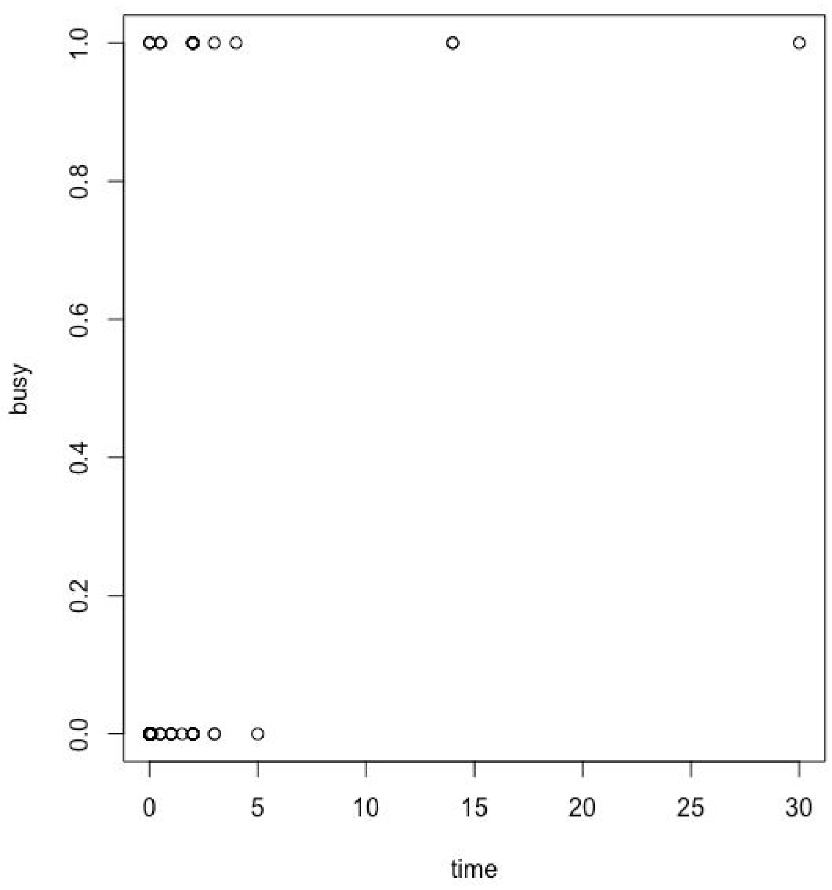
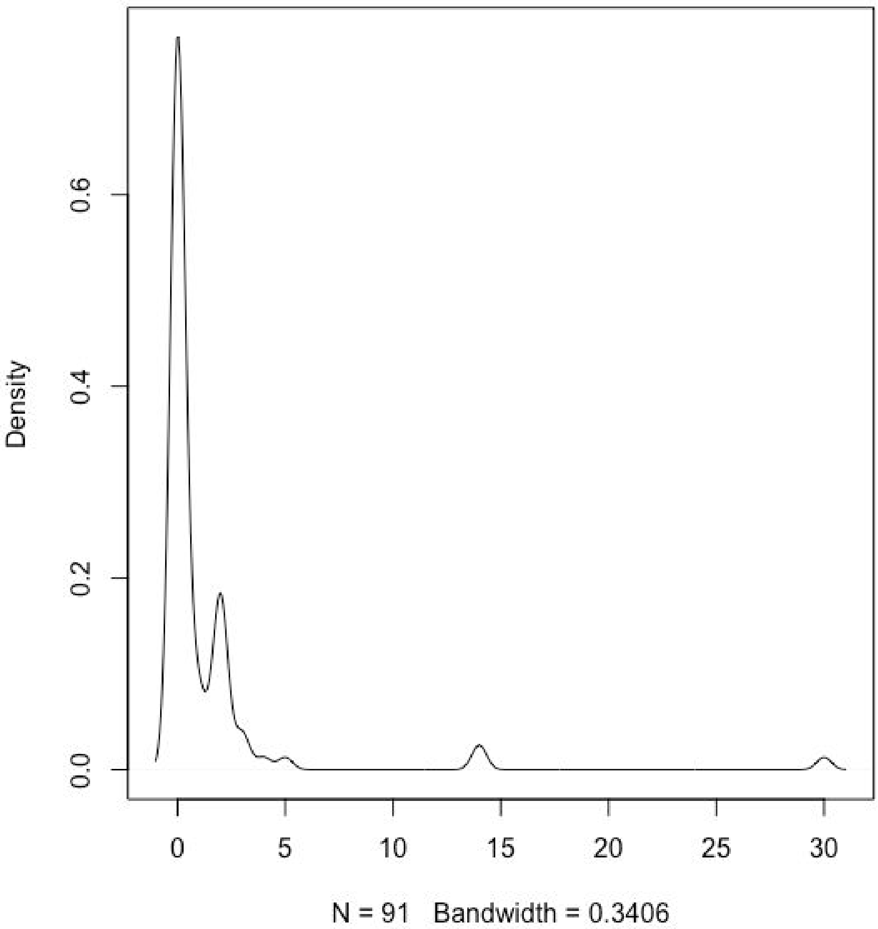


Figure 2.2.2: Whether or not a student was busy and played, and if so, how long the student played. 0 = did not play; 1 = did play

There is a marked difference and decrease in how many students played when busy, but the amount of hours played did vary more for students who answered as busy. More students tended to play video games when they were not busy, and while a handful still played 1-5 hours even when busy, there were more outliers for the busy group, with two students playing for more than 10 hours. So if there was an exam, we would see a decrease in students playing video games, and thus, decrease our point estimate and even our interval estimate.

2.3. Confidence Interval for Sample Mean

Probability Density of Time Spent Playing Week Prior



Looking at our distributions we see that it is unimodal with 4 bumps. The two right-most bumps are from two students who played 14 hours and one student who played 30 hours. These points will pull our mean up, providing an estimate that is too large. However, they are only 3 out of 91 points, and may not affect the mean too much. Nevertheless, we will keep the bias in mind when interpreting our interval.

To calculate the interval, we start by taking the sample mean of time spent playing video games in the prior week. Next, we find the sample standard deviation, applying the bootstrap method; from the standard deviation we obtain the standard error of our data. We then use a student-t distribution with 90 degrees of freedom to obtain the quantiles for a two-sided 95% percent interval. From this information, we assemble our interval: (0.68,1.80). This means that if we were to replicate infinite samples of 91 students from a population of 314, the sample mean would fall in this interval approximately 95% of the time.

2.4. Attitude Toward Video Games

What can we say about students likes and dislikes regarding video games? From the data we have where students play video games, whether or not students feel too busy for video games, and more.

First we looked at where students played video games and found that students who responded "not really" (4) or "not at all" (5) do not play in an arcade while the students who responded "very much" (2) or "somewhat" (3) did play in an arcade. It is possible that playing in an arcade is simply more fun than playing at home or at a friend's house, which would cause students to like video games more. Also, some students may not consider arcade games to be video games and different students may have different conceptions of video games, affecting the results. Nonetheless, we cannot draw any strong conclusions other than students who really liked video games played at arcades.

The following table summarizes the proportion of students who would still play video games even if they were busy:

Do you play when busy?

Busy	No	Yes
Proportion	.727	.273

Like = 2

Busy	No	Yes
Proportion	.82	.18

Like = 3

Busy	No	Yes

Proportion	1	0

Like = 4

Table 2.4.1

Not shocking to say, but people who are busy tend to not play video games. However, the more a student likes video games, the more likely it is that they would play even if they were busy.

The following table summarizes the proportion of students who think playing video games is educational:

Are video games educational?

Educ	No	Yes
Proportion	.36	.64

Like = 2

Educ	No	Yes
Proportion	.52	.48

Like = 3

Educ	No	Yes
Proportion	.833	.167

Like = 4

Table 2.2.2

The more a student likes video games, the more likely it is that they find playing video games to be educational. However, it is not clear whether the students enjoy playing video games because they are educational or if the students find playing video games educational because they already like video games.

Looking at gender, we do not see much variation other than more male students who reported 2. This suggests that boys are more likely to “really like” video games than girls are; however, when it comes to “somewhat” liking video games or “not really” liking video games, gender does not play a prominent role

Similarly, whether a student has a computer at home appears to be independent of how much they reported liking video games. This could be due to larger economic factors at play that regulate whether or not a random home would have a computer. Also, there are many mediums other than computers to play video games on. The same goes for whether or not the students’ computer had a CD-ROM.

Lastly, we look at whether or not the student likes math:

Do you hate math?

Math	No	Yes
Proportion	.78	.22

Like = 2

Math	No	Yes
Proportion	.711	.289

Like = 3

Math	No	Yes
Proportion	.54	.46

Like = 4

Math	No	Yes
Proportion	.57	.43

Like = 5

Table 2.4.3

Looking here, we see that students who like video games more have a higher chance of not hating math. This could be for a variety of reasons.

Overall, it appears students who like math and think video games could be educational tend to like playing the most. Given that most students enjoy playing strategy games, it is reasonable to conclude that students enjoy the mental stimulation and thought-provoking puzzles offered by many video games. In contrast, it appears that students do not like playing video games because they do not have enough time or do not think they will benefit from the game.

2.5. Differences in Preferences

It seems in the case of this dataset that a single model regression is perhaps too simple to explain something as complicated and no doubt influenced by a plethora of interrelated psychological, social and other phenomena as whether or not an individual enjoys video games. For drawing inference out of such multi-faceted and -featured data, then, other forms of interpretation are sometimes useful.

Here, we constructed a regression tree to examine the influence of each of a list of potentially related variables. Looking at the relative impacts of student's sex, how much they work a week, whether or not they own a computer, and whether they have that computer at home. In the end, only working more or less than 5.5 hours per week, having a computer at home and sex ended up having statistically significant influence on the tree constructed for whether or not a given student likes video games.

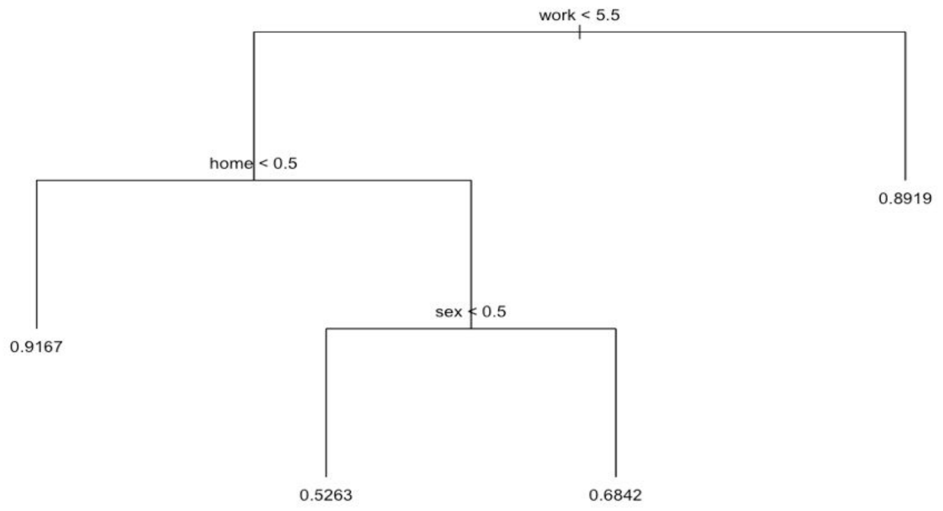


Figure 2.5.1 Regression tree describing influence of hours worked per week, having a computer at home and an individual's sex on liking video games.

The numbers at each "leaf" of the tree represent predictions of the likelihood of liking video games given an observations classification at a given leaf on the tree. For example, a student that works less than 5.5 hours in a week and does not have a computer at home has a 91% chance of liking video games. These people perhaps have more free time and prioritize gaming systems over PC's. Sex at that leaf is not significant. This provides a some insight about the potentially factors that influence whether a given individual likes video games at all, which certainly impacts our core question of inquiry with this dataset.

Among individuals that do own computers at home, males are more likely to like video games than females, a result that some would expect, from anecdotal evidence. Interestingly, working overall does not seem to have an influence over liking video games, perhaps good news for employed gamers looking to eschew stereotypes, or an argument for potential association between playing video games and positive attributes like intelligence and problem solving ability through the proxy of ability to balance employment with scholastic life at Berkeley.

2.6. Grade Investigation

In analyzing the responses received for expected grade for the sample of students surveyed, it is immediately notable that though all students surveyed seem to have answered this question, none of the students surveyed expect to fail the class, that is none answered 0 or 1 (D or F, respectively).

To compare the observed result to a hypothetical normal grade distribution, the golden "bell curve," we created a truncated normal distribution around a mean of 2, then rounded each of the entries to an integer value to resemble the data used in the set. Taking samples of n = 300 from N = 10,000 such identical distributions or something of that nature would offer more of a proof of the central limit theorem over such a small range of possible options and large number of observations than anything else (n > 100). So bootstrapping yields no new information out of this dimension of the data set.

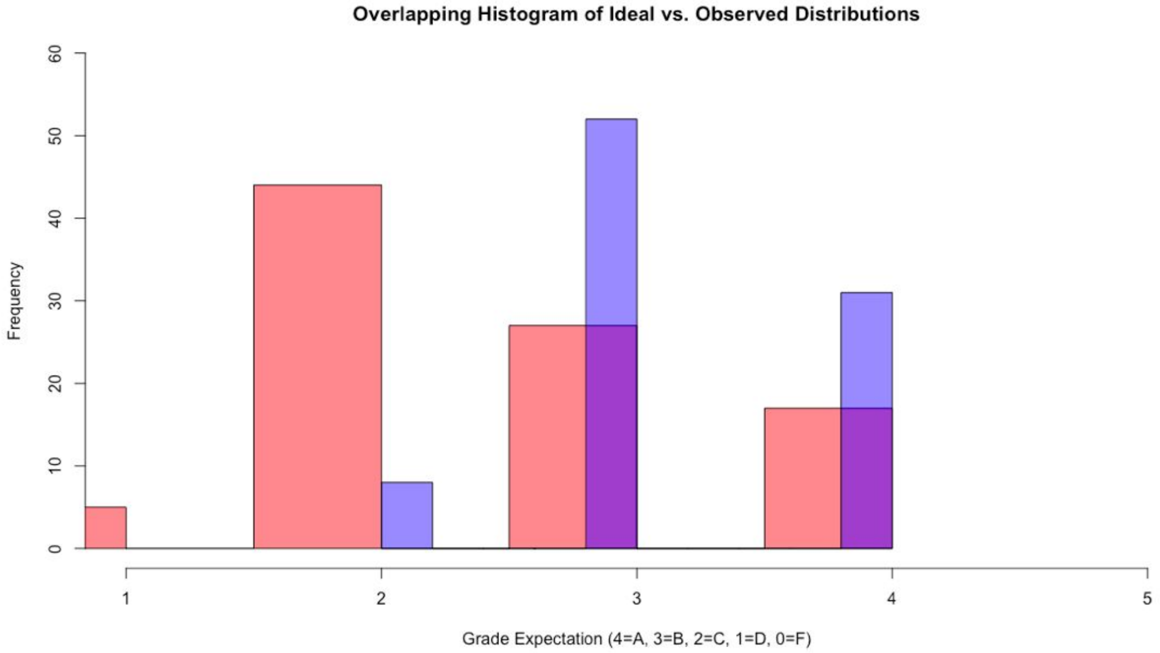


Figure 2.6.1 Overlayed histograms of a theoretical sample of ideal distribution of grades as given by 20% A's, 30% B's, 40% C's and 10% D's (red) and observed student's expectations of final grades after taking the second midterm

Constructing a theoretical perfectly distributed sample by drawing from the sequence 0:4 distributed by ideal grade distribution (here we'll say 20% A's, 30% B's, 40% C's and 10% D's or lower), it is clear that the students expectations have some dissonance with reality. Many students that believe that they are A or B students will likely end up with C's, if the final grade distribution resembles our theoretical one.

Applying a Kolmogorov-Smirnov two-sample test to the survey sample and a sample generated from a theoretical "perfect" distribution yields a p-value of 7.391e-11, so it is virtually impossible that the distributions of student's expectations resemble what grades will actually look like, under ideal.

It is more useful to view the result through the lense of game theory. It is generally practice to compare result like this to theoretical models that seek to explain reality through decision theory. Considering students as decision makers at different time periods throughout the quarter at UC Berkeley, game theorists like to consider the expected value of future quantities to explain decisions, so having a measure of expected grades lends itself to this.

Given that the students are all college students at Berkeley they can (at least somewhat) be assumed to be rational decision makers. As the survey was administered after the second test in the course and the students are studying statistics, it should be that they all have a relatively large amount of information about their true likelihood of passing the class. If one assumes that failing a class rather than withdrawing or dropping from that class yields less utility to a student, it is a Nash equilibria for all students at that time that expect to fail to drop. So, from a game theorist's perspective observing any students that answer that they expect to fail yet stay in the class is an anomaly. It could be that some of the 4 students that were in the class and were not surveyed (i.e. missed several sections of the class at Berkeley) were in fact, the failing students.

Conclusion

The conclusions we draw from our analyses are: (1) the frequency of how much a student played video games decreased as the time the student spent playing video games decreased, (2) more students tended to play video games when they were not busy rather than when they were busy, (3) the more a student likes to play video games, the more they find video games to be educational, (4) males are more likely to "really like" playing video games, (5) whether a student has a computer at home is independent of how much they reported to like video games, (6) students who like video games have a lower chance of hating math, (7) among those students with a computer at home, males are more likely to like video games.

Since students who like math and think video games could be educational tend to like playing video games the most, the advice that we offer to the design committee is to incorporate strategy games in your computer labs. Strategy games provide mental stimulation and require almost autonomous decision-making skills, which keeps the player intrigued. Not only will this be fun for the player, but rewarding as well because they will walk away with some gain of knowledge. Also, incorporating a female main character in the computer lab might make the video gaming experience more enjoyable for the female students playing the game, since the male students have a slightly higher liking for video games than female students. In determining this advice, we used statistical methods to draw conclusions from six different scenarios, each of which looked at different aspects of the data.

Another major reason to incorporate video game-like qualities into a computer lab is because most college student gamers seem to associate positive feelings with gaming, such as pleasant exciting, and challenging. Fewer students reported feeling frustrated, bored, or stressed by gaming[1], and since college courses can be just that sometimes, we believe it will bring a positive aspect to a statistics course at the University of California, Berkeley.

[1] Jones, Steve. Let the Games Begin (2003): Web.
<"http://www.pewinternet.org/files/old-media/Files/Reports/2003/PIP_College_Gaming_Reporta.pdf.pdf">

Appendices

Theory

1. <u>Recursive partitioning regression trees</u>

Sometimes global regressions prove too simple to capture complexities in interrelated datasets, so we use regression trees in an attempt to scare-out insight from potentially influential variables by recursively splitting down a tree and fitting simple regression models.

Partitioning in this way allows quick computation of graphs representing the influence of variables in the model in question individually fit to simple models. The tree can be interpreted by reading from root nodes to terminal nodes at the top as percentage fitting the criteria at each juncture.

Decision trees are useful because they can be easily computed and there are efficient, commonly accepted methods of learning these models to make predictions.

2. <u>Estimations</u>

Point estimations are used to give an estimate value of an unknown parameter. It takes the sum of all X's, and divides by the population sample, n:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The interval estimate was used to check the confidence interval of the value obtained from the point estimation. With population proportion *p*, sample size *n*, and $z_{a/2}$ as the *100(1-a/2)* percentile, then we can gather the endpoints as:

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Code

(See attached r markdown file for replicable code.)