Project 2 Report

Alison Camille Dunning [1], Bryan Talavera [2], Jared Dishman [3], Abdiaziz Weheliye [4], Yaoxin Li [5], Ruotian Gao [6], Ye Yint Win [7]

[1] Second Year Undergraduate: Data Science, A16166698
[2] Second Year Undergraduate: Data Science, A13278593
[3] Transfer Undergraduate: Data Science, A16688060
[4] Transfer Undergraduate: Cognitive Science and Machine Learning, A15690733
[5] Transfer Undergraduate: Data Science, A16542726
[6] Fourth Year Undergraduate: Data Science, A15230548
[7] Transfer Undergraduate: Data Science, A16688105

Introduction

The main question that we are engaging with is “Is there too much technology in the classroom?”. This is a multifaceted question, which makes it difficult to answer. This requires us to break down the question into smaller pieces that we can feasibly answer, such as “is technology beneficial in the classroom for students?”. In this instance, our target audience consists of students. However, the difficulty with this stems from the fact that we cannot accurately assess how technology affects students, for moral reasons. For this reason we have to again reframe our question.

For many students, video games are an established part of their lives. Some use video games as outlets in order to escape reality, while others simply enjoy the thrill that comes with gaming. There are many opportunities for instructors to incorporate the positive aspects of video games in order to make lab sections more enjoyable and make the learning experience more engaging for students. This is the idea that a statistics course at UC Berkeley had, in an attempt to increase student engagement in discussion labs. In order to make this a reality, the students taking the course were given a survey which asked them about their habits when it came to video games, in addition to general information about the student.

The survey was given to students at the University of California, Berkeley during the Fall Semester of 1994. The students who were initially chosen were students who had completed the second midterm of the course, this was done in order to limit the amount of non-responding students. From the 341 students in the course, 95 were chosen from a pseudo-random number generator, 91 of which successfully completed the survey. The first section of our survey highlights the hours spent playing video games during the week of exams, as well as other general questions about the student. The second section of the survey focuses on what students liked or disliked about video games, and why. The purpose of this survey was to use video games as a proxy for certain discussion topics, in order to increase engagement in future discussion labs. In this report, we will be using the dataset that was a product of this survey, in order to provide better insight to the Instructors who are creating the new labs.

Background

In this day and age, technology has become an integral part of everyday life and especially when it comes to education. As effective as it is, it raises questions and concerns such as "Are we relying on it too much? Should it be limited? And is this going to be the future of education? Does it truly benefit the students?" Modern day technology transforms students' daily routines and their developments. Students used to take notes on paper with pens and pencils but nowadays, the majority of them are using laptops and tablets for the same purpose. Recent events, such as global pandemic, have shifted in-class lectures to remote learning where the students are learning the same material in the comfort of their own rooms or even across the world. Even though students are benefiting from technology and basking in the glory of it, it also comes with a price.

In the article *Technology and Its Impact in the Classroom* written by Rozalind G. Muir-Herzig and published in Computers & Education journal in 2004, Muir-Herzig conducted the four-part survey that measures technology proficiency and frequency of technology use of the teacher and the students in the classroom. All four research questions in the survey were in focus to how the use of technology by both teachers and students in the classroom have effect on at-risk student's classroom grade where the term "at-risk" students is used to identify students with failing grades, low GPA, and/or with high absence. The data is collected from 66 at-risk students of the 2000-2001 school year from a Northwest Ohio high school. According to the surveys he conducted, "the overall grade mean for the high users of technology dropped from 1.70 to 1.24 from first quarter to second quarter meanwhile, the low users' dropped from 1.52 to 1.31"(Pg.126)[4] which is our main focus and with the possible explanation being the decline in grades is due to the greater usage of technology in the first quarter.

In the article *Students' Attitudes Toward Playing Games and Using Games in Education: Comparing Scotland and the Netherlands* by Thomas Hainey and several other authors published in Computers & Education journal in 2013, a survey was conducted with 887 students from 13 different higher education institutions in Scotland and Netherlands that examines students' characteristics related to their gaming preferences, game playing habits, and their perceptions and thoughts on the use of games in education. According to their findings, a large number of of participants (459 out of 632) overwhelmingly agreed that "computer games can be used as educational mechanism in higher education"(Pg. 483)[5] The majority of the participants believed that computer "can encourage the importance of cooperation and teamwork between students" with only a small percentage of them had the opinion of it being "too distracting"(481)[5]

With the insights from both perspectives of how technology can potentially have both positive and negative impacts in education, we will be looking at the data set to answer our big question of "Is there too much technology in the classroom?"

Data

The data in our dataset contains $n = 91$ total observations, with an additional 4 nonrespondents. Our data was collected from the University of California, Berkeley from an Introductory Probability and Statistics course in the Fall of 1994. The class itself is a lower-division

statistics course meant for students who intended to major in business. Our 95 observations were randomly selected from a population of 314 students, all of whom had taken the second midterm of the course. The students were selected via a pseudorandom number generator, where each student was assigned a number from 1 to 314. This extra step was taken in order to limit the number of nonrespondents to our survey.

It is important to note that this sample is relatively small, with $n < 100$. A small sample size might be detrimental to statistical analysis because it runs the risk of finding unusual trends just by chance alone. Small sample sizes also pose other disadvantages such as increased bias arising from non-responses, and lack of representativity of the whole population (larger margin of error) (Simmons, 2019).

It is also important to note that our samples were taken without replacement, which means that our sample values are *not independent* of each other. This means that, for instance, the number 'a' that was sampled first will have an impact on the probability of the number 'b' being sampled next. Additionally, this will yield a covariance of 0 between the two numbers.

Survey Variable Descriptions

The variables collected from the survey are displayed in the table below, along with their types and descriptions. The answers to each of the questions were encoded numerically.

Variable Name	Type	Description
Time	Numerical Discrete	Number of hours played in the week prior to survey
Like to play	Nominal Categorical	Rating of 1-5 on whether or not a student likes playing video games ● 1 = Never played video games ● 2 = Very much ● 3 = Somewhat ● 4 = Not really ● 5 = Not at all
Where played	Nominal Categorical	Where does the student play video games, when they do ● 1 = At the arcade ● 2 = Home system ● 3 = Home computer ● 4 = Arcade and either home system or home computer ● 5 = Home computer and system ● 6 = All three

Often	Ordinal Categorical	How often does the student play video games, on a scale of 1-4
Busy	Nominal Categorical	Does the student play when they're busy with other things?
Educational	Nominal Categorical	Does the student play educational video games
Sex	Nominal Categorical	0 if female, 1 if male
Age	Numerical Discrete	Student age in years
Computer at home	Nominal Categorical	0 if No, 1 if Yes
Hate math	Nominal Categorical	Does the student hate math?
# hours of work week prior to survey	Numerical Discrete	How many hours did the student work prior to the survey?
Own PC	Nominal Categorical	Does the student own a PC?
CD-Rom in PC	Nominal Categorical	(Subset of students who own a PC) Does the student's PC own a CD-ROM?
Have email	Nominal Categorical	Does the student have an email address?
Grade	Ordinal Categorical	What grade does the student expect to earn in this course?

Table 1: Survey variable types and descriptions. The majority are categorical (nominal). A numerical answer of '99' denotes a non-existent answer by a respondent to a particular question.

This survey aimed to use the data to help design new discussions embracing developments in technology and gaming. Many of these variables can add some insight as to how to reform the structure of discussion sections. We can derive several examples from the natures of the variables in this dataset. First of all, gathering variables such as the students' work schedules and whether they play video games while occupied with other tasks will provide a sense of how committed students are to attending extra class sessions such as discussions. It may be possible that if students are too busy, there may be no reason to collect such data and invest so many resources into improving discussion sections. Another example would be whether or not students hate math: can the visual and interactive nature of video games reduce mathematics difficulty?

We could draw another potentially exciting element from where students typically play video games. In the proceeding tables, we see that some genres of video games, such as Sports and

Adventure, may have outdoor settings. Could discussions take place outdoors to engage students further? Collecting data on whether or not students have e-mails may yield information about communicating online as a viable option, although this doesn't directly pertain to video games. As a final example, knowing how often students play video games can tell us how long or how often students are engaged in a particular task. With a comprehensive survey such as this, data collectors are now in the face of many options for how to reform discussion sections.

A follow-up survey that was given related closely to video games themselves. They also tapped into the sentimental and practical reasons why students like or dislike video games. Several tables below display the frequency of different answers to these additional questions. First, we observe the most common genres of video games among the respondents:

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

Table 2: What types of games do you play? Pick at most three answers. Note that the originally sampled students who said they didn't or have never played video games were instructed to skip this question.

Why play video games?	Percent
Graphics/realism	26%
Relaxation	66%
Eye/hand coordination	5%
Mental challenge	24%
Feeling of mastery	28%
Bored	27%

Table 3: What best describes why you play video games? Pick at most three answers.

Dislikes	Percent
Too much time	48%
Frustrating	26%

Lonely	6%
Too many rules	19%
Costs too much	40%
Boring	17%
Friends don't play	17%
Pointless	33%

Table 4: What best describes what you dislike about playing video games? Pick at most three answers. All students surveyed originally were asked to answer this question.

The answers of the last two questions in particular are reminiscent of why students may like or dislike discussion sections. Should discussions be more relaxing, or should they involve mental challenges that resemble those seen in games? Should discussions be shorter, since the most common response in Table 4 was that video game tasks take too much time?

Investigations

We present a number of scenarios that will bear statistical information exploring the questions at hand.

Scenario 1

We seek to determine a point estimate of the proportion of students who played video games for the week prior to the survey. We acquire this proportion by dividing the number of students who played video games for more than zero hours in the prior week by the total number of students who responded to the survey. 34 students out of 91 played video games for more than zero hours, and we obtain the following:

$$\frac{\#\ students\ played\ video\ games}{\#\ students\ responded} = \frac{34}{91} = 37.362\%$$

Next, we want to verify that this proportion is close to the true population metric. We apply Confidence Intervals because they are capable of showing *likely* values of the metric. Conforming to the Central Limit Theorem, our sample size is greater than 30 [2] so our approximate interval estimates will be good. Since our sample size is large, we can assume that when we sample the proportions our observations $y_1, y_2, \ldots, y_n$, those proportions will form a normal distribution. We calculate Confidence Intervals using the following formula:

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}},$$

where $\bar{x}$ = our estimate of the sample statistic, z = our confidence level value, and s = our sample standard deviation. The second term to be added to our estimate, $z\frac{s}{\sqrt{n}}$, is called the **margin of error** and. In the margin of error, the z-score is multiplied by the **standard error**. We acquire the following bounds of the proportion confidence interval: (0.274, 0.437). Thus, we can surmise that the true population parameter has a 95% probability of falling within this interval.

Scenario 2

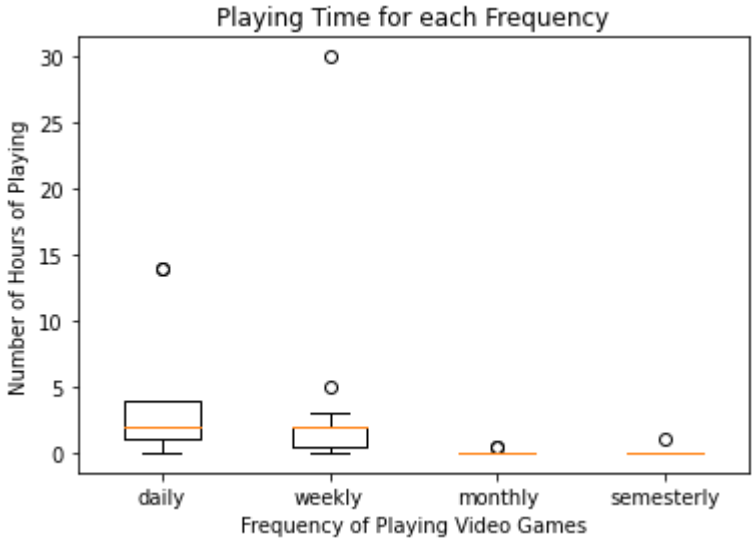

Figure 1. Box plot of time spent playing video games, split by frequency of which games are played.

The four box plots above demonstrate that the time of playing video games for various frequencies including daily (1), weekly (2), monthly (3), and semesterly (4). By the way, there are 13 observations containing frequency of 99, and these observations are excluded in this scenario. To be specific, the medians of daily and weekly groups are relatively higher. Therefore, students who play video games frequently tend to play video games longer than students who play less often. Furthermore, by comparing daily and weekly groups, the third quartile of daily group is higher than the weekly group, and, in other words, most students who play video games daily spent more time than students who play video games weekly. It is coincident with the conclusion by comparing four groups. Therefore, we can conclude that there is a positive association between frequency of playing video games and time spent playing video games. In other words, students who play video more frequently tend to spend more time on playing video games.

Frequency of Playing Video Games	Average Time of Playing in Hours
Daily	4.444
Weekly	2.539

Monthly	0.056
Semesterly	0.043

Table 5. How frequently students play video games vs. the average time spent in hours playing video games.

This table shows the average time spent playing video games for each frequency. As we can see, the average time spent playing video games increases with increasing frequency. Thus, there is a positive association between frequency of playing video games and time spent playing video games

Frequency of Playing Video Game	Mean Time of Students who Play if Busy	Number of Students who Play if Busy	Mean Time of Students who Do Not Play when Busy	Number of Students who Do Not Play if Busy	Proportion of Students who Play if Busy
Daily	7.20	5	1.00	4	0.56
Weekly	4.00	11	1.59	17	0.39
Monthly	0	1	0.06	17	0.06
Semesterly	N/A	0	0.00	22	0.00

Table 6. Numerical comparison between students who play video games if they are busy and not busy, according to how frequently they play video games.

To determine how might the fact that there was an exam in the week prior to the survey affect our results, we got this table which includes the frequencies, whether students play games when they are busy, according average time spent playing games, and proportions of students who play when they are busy. As the table shows, the more frequent playing students have a higher proportion of students who even play when busy. Moreover, we plotted a bar plot for comparing mean times graphically below.

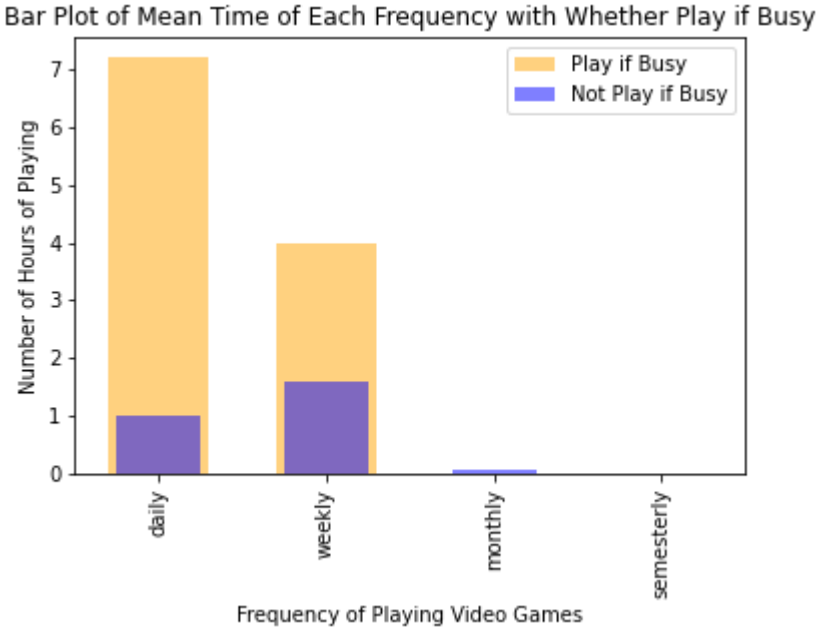

Figure 2. Bar plot showing the mean time spent on video games by students while busy or not busy, divided by how frequently they play video games.

This bar plot shows the difference of mean time spent playing between students who play if busy and students who do not play if busy for each frequency respectively. Accordingly, for daily and weekly groups, they are affected by the exam mostly, because the mean time of students who do not play if busy is obviously lower than the mean time of students who play if busy. In other words, for students who play games daily and weekly, the results for these students are affected by the presence of the exam, and these two groups of students should have higher mean time spent playing if there were no exam prior to the survey.

Scenario 3

This histogram illustrates the length of playing video games prior to the survey. The x-axis reflects the amount of time of playing video games, while the y-axis represents the according frequency. Furthermore, the red line indicates the mean of the time of playing video games which is approximately 1.24 hours. In other words, most students spent 0 to 2.5 hours playing video games, and the mean time lies in this interval. Meanwhile, the distribution of data is not normally and right skewed.

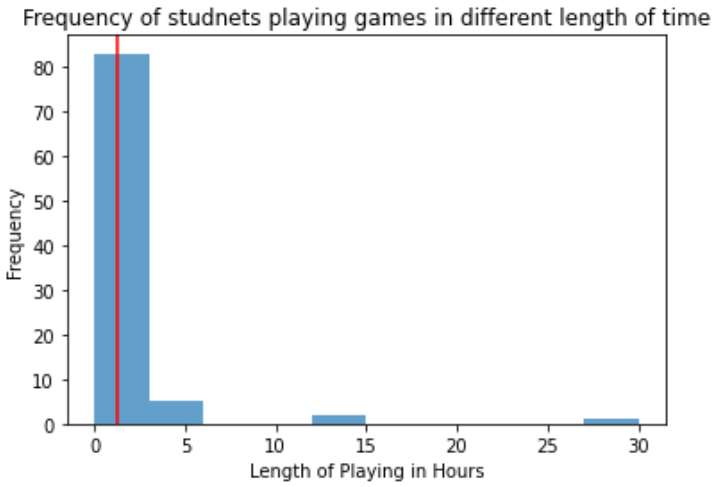

Figure 3. Histogram of time spent playing video games in hours. This is clearly a skewed variable.

We calculate the confidence interval by three methods including simple confidence interval by central limit theorem, confidence interval with correction factor, and confidence interval by bootstrap. More specifically, we use a two-sided 95% confidence interval as the results.

- Simple confidence interval by central limit theorem
  - Point estimator: 1.243
  - Confidence interval: (0.471, 2.015)
- Confidence interval with finite population correction factor
  - Point estimator: 1.243
  - Confidence interval: (0.592, 1.893)
- Bootstrap Confidence interval
  - Point estimator: 1.237
  - Confidence interval: (0.660, 1.876)

Missing Jackknife C.I. -2.

For the simple confidence interval by central limit theorem, we need our data points to be independent to each other. Nevertheless , the data is collected by a survey study, the data is not independent to each other. Therefore, this confidence interval is not appropriate for this data set.

Then, for the confidence interval with correction factor, we use the finite sample population correction factor to help the confidence interval.

Eventually, We use bootstrap to calculate the confidence interval. To be specific, our bootstrap population is the population size divided by the sample size and round it to the nearest integral, and the size of the bootstrap sample is the same as the sample size which is 91. We run 1,000 times bootstrap to get the point estimator and confidence interval for the time of playing video games. Finally, we get the below bootstrap histogram.

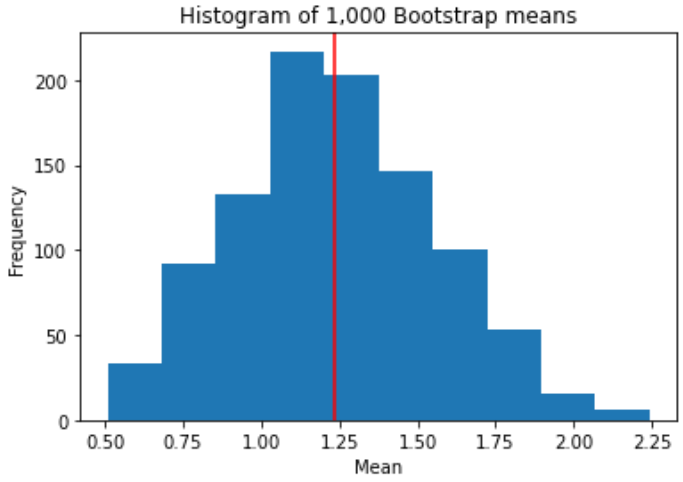

Figure 4. Histogram of bootstrap means

This histogram represents the distribution of 1,000 bootstrap means. Specifically, the x-axis represents the value of bootstrap means, and the red line indicates where the pointer estimator is.

For confidence intervals, we need our means to have normal distribution. Thus, we need to check the normality of bootstrap means.

Firstly, we use the Q-Q plot to check the normality as an empirical test, and I got the plot below. The theoretical quantiles are the quantiles of normal distribution. Thus, if we get the straight line of, with slope 1 and intercept 0, on the q-q plot, the distribution of bootstrap means is normal. As a result, the Q-Q plot demonstrates that the bootstrap means is normally distributed.

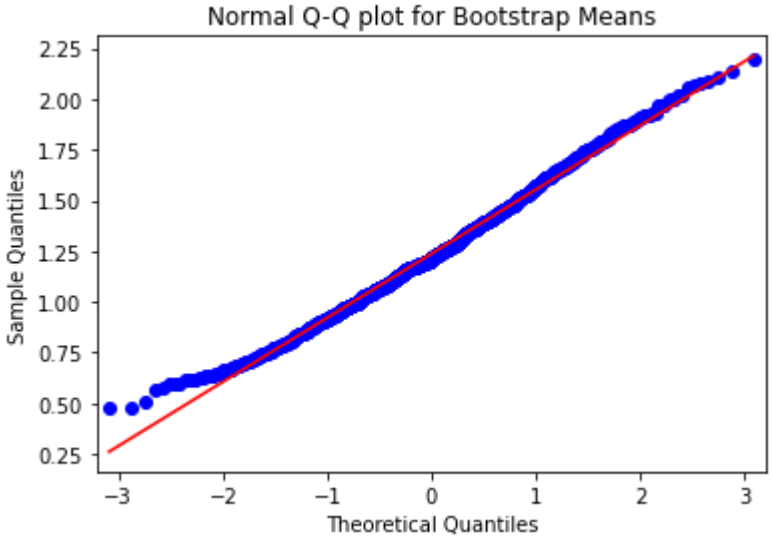

Figure 5. Normal Q-Q plot of bootstrap means.

We use a simulation test to test the normality. Firstly, we calculate the Kurtosis and Skewness coefficients of bootstrap means.

- Kurtosis is: -0.3162942029344378
- Skewness is: 0.2102997653604791

Then, we generated the histograms for pseudo normal samples with 91 sample size, and we ran the simulation 1,000 times. To be specific, the red line is where the kurtosis and skewness of the bootstrap means are. As we can see, the histogram shows that the kurtosis and skewness coefficients of bootstrap means are not rare to see, and, thus, the distribution of bootstrap means is almost normal.

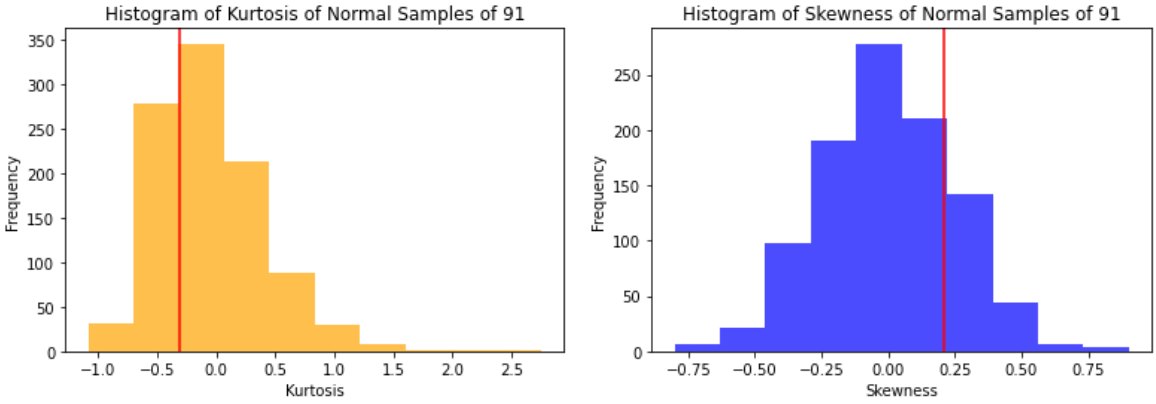

Figure 6. Kurtosis and Skewness plots of bootstrap means

Using the Q-Q plot and the simulation test, we can say that the bootstrap means are normally distributed. Therefore, since we can apply the central limit theorem to it, the confidence interval calculated by the central limit theorem is also valid. Eventually, since all of the three confidence intervals are valid, the most narrow confidence interval is our interval estimate for the average time spent playing. Thus, (0.660, 1.876) is the interval estimate, and it means there is 95% probability that the true average of time spent playing falls in the interval (0.660, 1.876).

Scenario 4

For this scenario, we used the additional dataset videoMulti.txt to explore the reasons of students' likes or dislikes about video games.

First, we know that students enjoy playing video games in general. According to the pie chart, we can see that the number of students who do not like video games or never played is less than 25%, so we can say that most of the students like video games to varied extents.

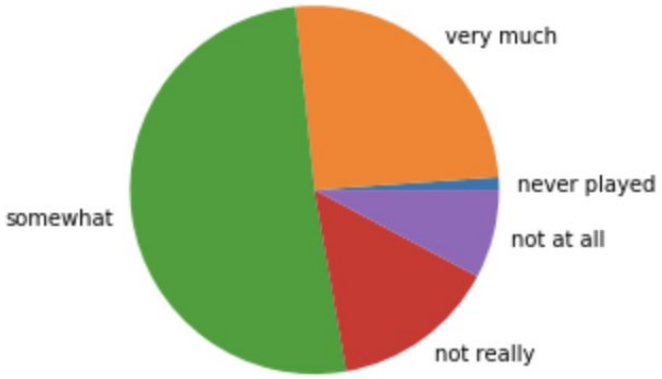

Figure 7. Percentage of students' attitude towards video games

For the most important reasons why students like video games, the top 3 on the lists are: relaxing, feeling of mastery and bored.

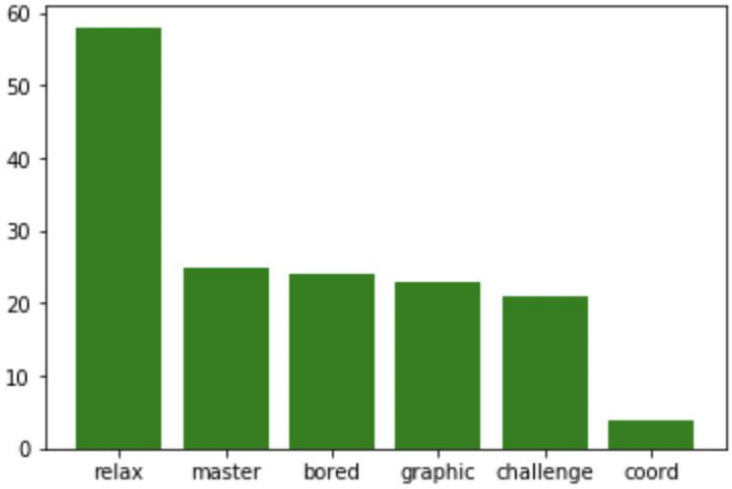

Figure 8. Reasons why students like to play video games

For the most important reasons why students do not like video games, the top 3 are: too much time, cost too much and frustrating.

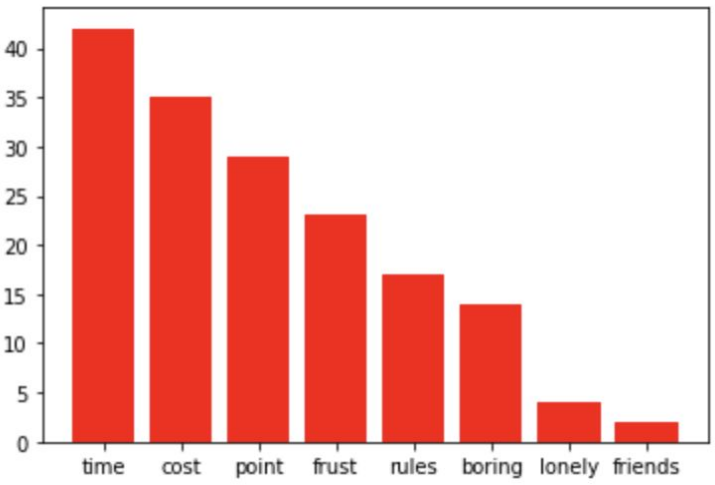

Figure 9. Reasons that students do not like video games

t 2 correct use of tests.

Scenario 5

In this scenario, we explore the underlying differences between those who like video games and those who don't. In doing so, we run statistical tests to determine the degree of association between various factors and whether or not they like games. We also employ additional results from the follow-up survey (Tables 2-4). Note that for seamless analysis, we group individuals who like video games very much (coded as 2) and somewhat like video games (coded as 3) into one category. Similarly, we group individuals who never played video games (coded as 1), do not really like them (coded as 4), and do not like them at all (coded as 5) into another category.

Sex

The first of the factors we explore is sex of the respondents. Displayed below is a heatmap of the two categories based on sex.

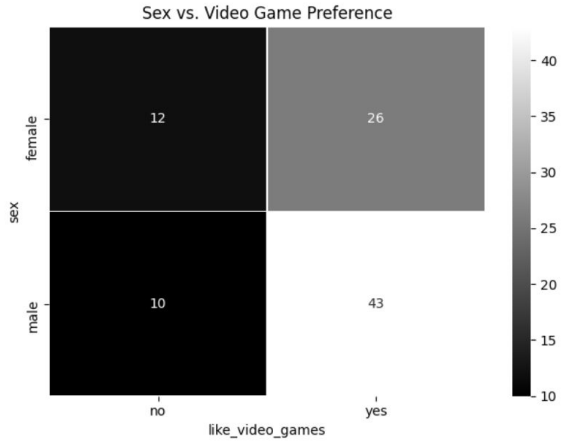

Figure 10. How many females and males like or dislike video games? A greater proportion of both sexes like video games, although the sample size for males is greater, with them being represented in approximately 52.8% of the sample.

We apply a Chi-square contingency test of independence/association to decide whether there is a strong relationship between sex and whether or not an individual likes video games. We hypothesize that there will be no strong correlation between these two variables. Our null hypothesis, H_0, is that there sex does not influence whether or not one likes video games. The alternate hypothesis, H_1, is that these two variables are associated. The gathered statistics are for the sex vs. video game preference are shown below:

χ^2 value: 1.319, p-value: 0.25, df: 1 (number of categories - 1)

We observe that the p-value is 0.25, which is more than our significance level of 0.05. Thus, we will fail to reject the null hypothesis at the 5% level of significance, and conclude that there is not sufficient evidence that whether or not one likes to play video games is dependent on their sex.

Hours Working

We first provide a histogram of the number of hours the students worked the week prior to the survey, in order to understand how many students don't work.

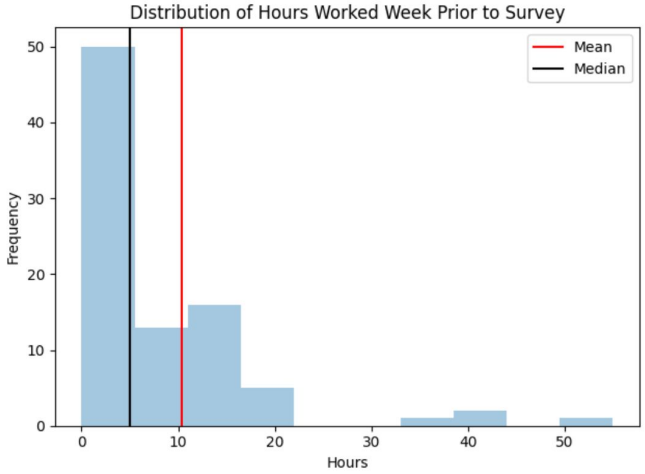

Figure 11. Distribution of hours students worked the week before the survey. There are clearly outliers, with students working up to 55 hours a week.

A job is considered part-time if its employee works fewer than 30 hours per week on average [3]. We partition the 'work' variable in the dataset into 'unemployed', 'part-time', and 'full-time' categories. The relationship between the students' work schedule and whether or not they like video games is shown below.

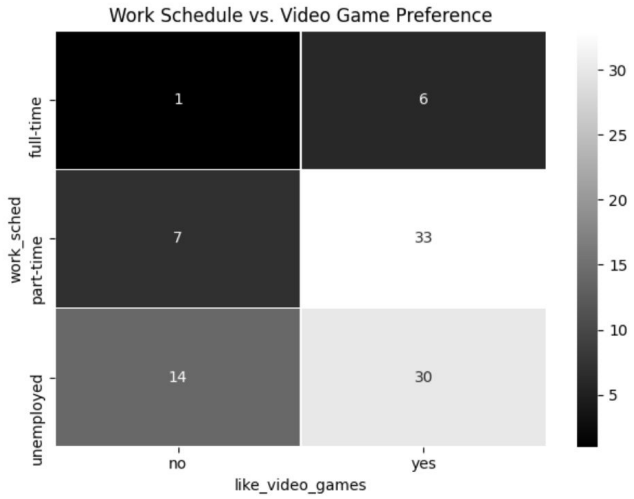

Figure 12. How many of those who don't work, work part-time, or full-time like or dislike video games? Again, a greater proportion in each category like video games.

We test the possibility of an association with another Chi-square contingency test. Our null hypothesis is $H_0 =$ there is no relationship between the students' work schedule and whether or not they like video games. The alternative hypothesis is $H_1 =$ there is a relationship between the aforementioned variables. The Chi-square contingency test results are as follows:

χ^2 value: 2.748, p-value: 0.253, df: 2

We observe that the p-value is 0.253, which is more than our significance level of 0.05. Thus, we will fail to reject the null hypothesis at the 5% level of significance, and conclude that there is not sufficient evidence that whether or not one likes to play video games is dependent on their work schedule.

Computer Ownership

Next, we compare computer ownership with liking of video games. Below shows the numbers of students who own computers and don't own computers.

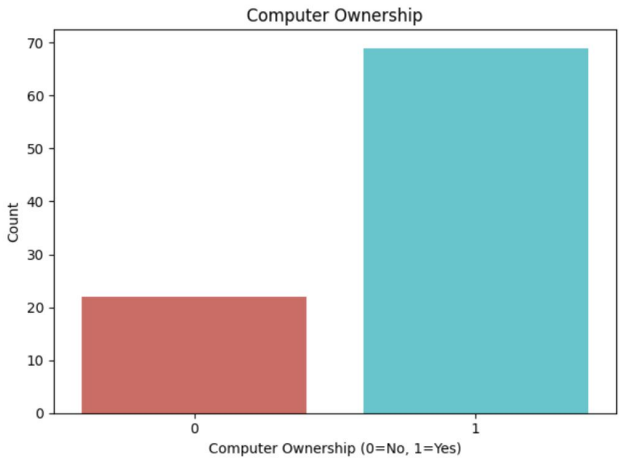

Figure 13. Most of the students own computers at home.

Once again, we apply a Chi-square contingency test, with H_0 = there is no association between computer ownership and liking of video games, and H_1 = there is an association between the two. We hypothesize that there may be an association because those who own computers at home may have access to a variety of games. Interestingly, however, by the given heatmap, there is a greater proportion of students who don't like video games in the group of students who do own computers than in those who don't. This could be attributed to a lesser need to go to the arcade. We discuss this in the Additional Question II section of this report.

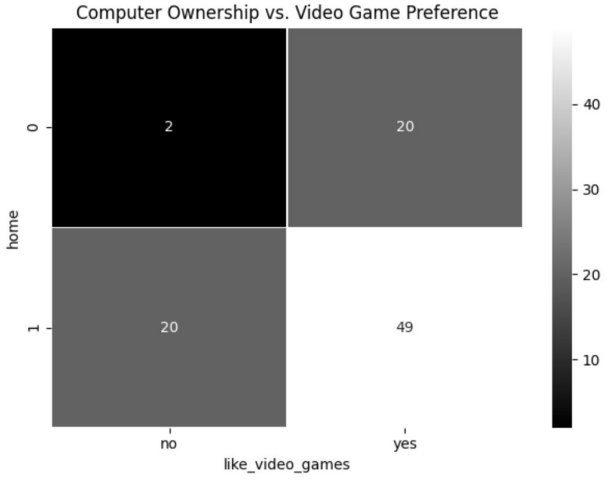

Figure 14. How many students who own a computer and don't own a computer like and dislike video games? A greater percentage of those who own a computer don't like video games than those who don't own a computer.

The results for this Chi-square contingency test are given below:

χ^2 value: 2.598, p-value: 0.107, df: 1

Since $p = 0.107 > 0.05$, we can again fail to reject the null hypothesis, and surmise that there is no sufficient evidence to demonstrate that there is a relationship between computer ownership and whether or not the student likes playing video games.

We observe that there is no evidence to suggest that any of the above factors contributes to whether or not one likes video games, based on the heatmaps and the tests of independence.

Scenario 6

+1 +2 correct use of tests.

For this scenario, we also decide to adopt a chi-square test to find out the grade distribution of the students. Our null hypothesis is that "the observed grade data is consistent with the 20% A, 30% B, 40% C and 10% D" and we set the significance level to be 0.05. After the test, we find out that the p-value based on the 91 students' test is:

```
p-value of chi-square test using 91 students is: 1.6287921892829858e-13
```

This is far less than the significance level of 0.05, so we can reject the null hypothesis and conclude that the current student grade is not consistent with the expected distribution.

Even after we add the 4 non-responders students, the p-value we have is :

```
p-value of chi-square test using 95 students is: 1.2230108335137166e-11
```

Adding the 4 non-responders student do not influence the result we got from the previous test which is that the students' grade distribution does not match target distribution of 20% A, 30% B, 40% C and 10% D

+2 correct use of tests.

Additional Question I - Educational Video Games

This additional question is how students think of their playing, whether they play educationally, affects the time spent playing. As the big question for this project is to investigate how the new discussion labs should look, whether students tend to play games when they play educationally is important for the big question. In other words, we need to determine whether the "play education" categorical variable affects the time spent playing. To be specific, we need to explore whether the time spent playing of students who play educationally and the time spent playing of students who do not play educationally come from the same distribution.

Therefore, let us divide the time spent playing into two categories including "play educational" and "not play educational". Then, compare the mean times spent playing of two groups.

- Playing Educational Mean Time: 1.31
- Not Playing Educational Mean Time: 1.52

The "not playing educational group" has higher mean time spent playing.

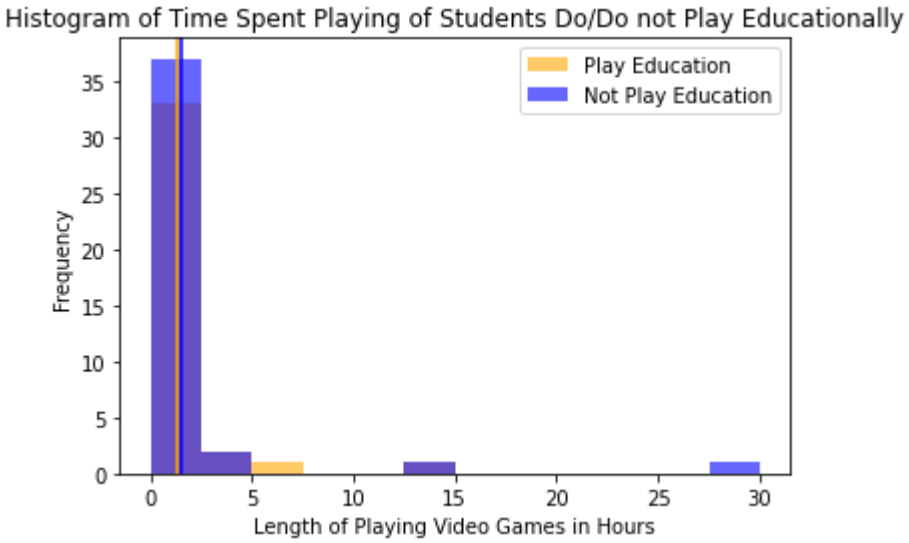

Figure 15. Histogram of the two distributions.

The histogram of time spent playing for two groups is shown above. Specifically, the "not play educational" group has overall higher time spent playing according to the graph. Therefore, we need to explore whether these two groups come from the same distribution.

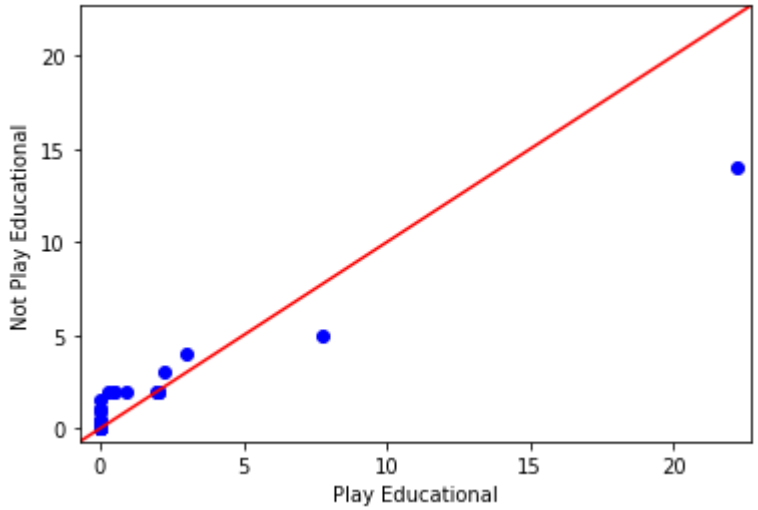

Figure 16. Q-Q plot between the two distributions.

Firstly, the Q-Q plot is used as the empirical test. The graph does not have the slope of 1 and intercept of 0. Therefore, we can conclude that two groups do not come from the same distribution.

Since this is just an empirical test, the chi-square independent test is used as well as a statistical test.

For the chi-square test, we define the time spent playing more than 2.5 hours is long time playing. Thus, we define the contingency table below.

	Play Long Time	Not Play Long Time
Play Educational	33	4
Not Play Educational	37	4

Table 7. Contingency table showing time spent playing video games and whether not games are played for educational purposes.

Then, we use the chi-square independent test to calculate the p-value and use 0.05 as the significance level. Additionally, our test's null hypothesis is: there is no relation between the play educational times and not play educational times. Meanwhile, the alternative hypothesis is: there is a significant relationship between them.

p -value: 0.826

Consequently, we cannot reject the null hypothesis since our p-value, about 0.83, is greater than the significant level, 0.05. In other words, the data is in favor of the null hypothesis. In other words, the times spent playing two groups are independent. Therefore, the data of the two groups are not from the same distribution.

Additional Question II - Where are video games played?

We circle back to Scenario 5, where we observe a non-relationship between computer ownership and whether or not one likes to play video games. Even though there was no relationship, we did observe that about one-third of the computer owners did like video games. Perhaps, this could be attributed to a lesser need to play video games elsewhere as more activities are available on computers.

Refer to the Data variable descriptions table to see how the “where” variable is divided. We will only focus on those who play exclusively on their home systems, home computers, or at the arcade. Thus, only those categories will be included in our graphical summary and statistical test.

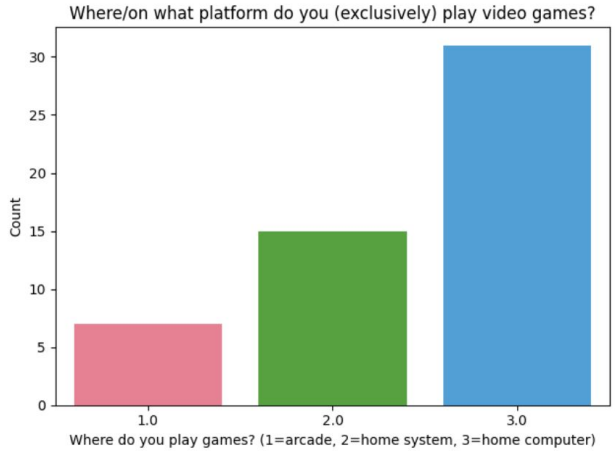

Figure 17. Count plot showing how many students play video games at the arcade, on their home arcade systems, or on their computers.

Our sample size will be smaller for this test: $n = 53$. Displayed below is the heatmap for the aforementioned categories.

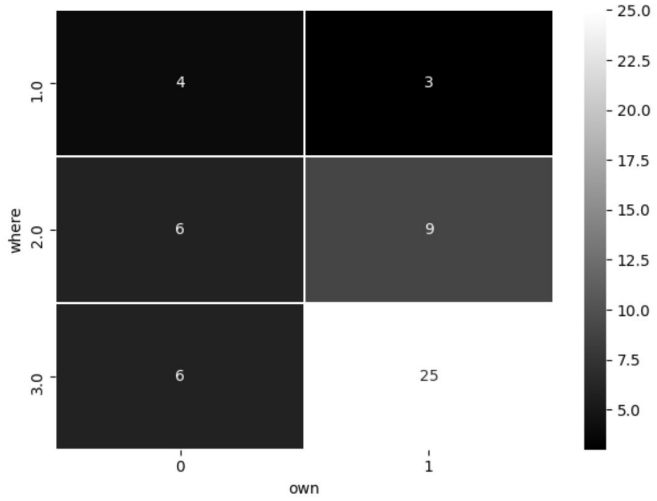

Figure 18. How many students who own a computer or don't own a computer play video games exclusively at the arcade, on their home systems, or their home computers?

The acquired Chi-square test of association results between the two variables and hypotheses are as follows:

H_0 = There is no relationship between computer ownership and where students play video games.

H_0 = There is a relationship between computer ownership and where students play video games.

χ^2 value: 4.825, p-value: 0.09, df: 2

Since the p-value is more than the significance level of 0.05, we will fail to reject the null hypothesis and conclude that there is not enough evidence to show that the students who owned

computers were not more or less likely to play video games through any particular avenue, including their computers.

Furthermore, we observe that there must be other factors that must determine whether or not students like video games. We have seen that there is no relationship between computer ownership and video game liking, and we confirmed that by seeing that the lack of correlation in this section demonstrates that there isn't sufficient evidence that students who own computers at home necessarily avoid other ways to play video games.

A quick note is that this dataset is clearly imbalanced, heeding to the count plot, and continued analysis using data balancing methods and re-sampling techniques, which are beyond the scope of this paper, is necessary.

Theory

Point Estimates

Point estimates involve the use of observed data, sample data, to calculate the unknown population parameters. For example, using sample mean, $\bar{x}$, to estimate the population mean, μ. Because

$$E[x_{I(j)}] = \frac{1}{N}\sum_{i=1}^{N} x_i = \mu \quad E[\bar{x}] = \frac{1}{n}E[x_{I(1)} + x_{I(2)} + \ldots + x_{I(n)}] = \frac{1}{n}(\mu + \mu + \ldots + \mu) = \mu$$

Thus, $E[\bar{x}] = \mu$, we can say that the sample mean is an unbiased estimator of population mean.

Confidence Intervals

Confidence intervals are interval estimates of the population parameter from the statistics of observed data. The confidence level of the confidence interval indicates the probability of containing the real value of the population parameter in the interval.

We can calculate the confidence interval simply by central limit theorem if the data points are independent, identically distributed.

$$CIs = (\bar{x} - Z_{a/2}\frac{s}{\sqrt{n}},\ \bar{x} + Z_{a/2}\frac{s}{\sqrt{n}}) \text{ where } Z_{a/2} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

However, for this project, the data is survey data, and, thus, the data points are not independent. Therefore, we should use a finite population correction factor to calculate the confidence interval or use bootstrap to generate bootstrap confidence intervals. By using finite population correction factor, the confidence interval is:

$$CIs = (\bar{x} - Z_{a/2}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}},\ \bar{x} + Z_{a/2}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}) \text{ where } Z_{a/2} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

Central Limit Theorem

The Central Limit Theorem (CLT) can serve as a confirmation of the normality of a given population distribution. This is because it maintains that as long as the size of random sample drawn from the population is sufficiently large, that sample will follow a normal distribution. Thus, if we are concluding the distributions in this paper are normally distributed, we may surmise that population distribution is Gaussian, and that sample mean is a good estimate.

Analytically, the CLT is written as follows for a sufficiently large n:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \sim \aleph(\mu, \frac{\sigma^2}{n}),$$

where μ is the population mean and σ^2 is the variance. We observe that as n increases, $\bar{x}$ converges to a value equal to μ.

Skewness

Skewness is a measurement of how symmetrical our distribution is, with a perfectly symmetrical dataset having a skewness coefficient of 0. It is the average of the third power of our standardized data. It is defined as:

$$skewness = \frac{1}{n}\sum_{i=1}^{n}(\frac{\bar{x}-x_i}{S})^3,$$

where n is the sample size, $\bar{x}$ is our sample mean, and S is the sample standard deviation

Kurtosis

Kurtosis is a measure of the combined weight of the tails of our distribution, compared to the center of it. A negative kurtosis value indicates that our distribution has less weight in the tails compared to the normal distribution, while a positive kurtosis value indicates that our distribution has more weight in the tails compared to the normal distribution.

It is defined as:

$$kurtosis = \frac{1}{n}\sum_{i=1}^{n}(\frac{\bar{x}-x_i}{S})^4,$$

where n is the sample size, $\bar{x}$ is our sample mean, and S is the sample standard deviation.

Hypothesis Testing

A hypothesis test is a type of statistical inference which uses data collected from a sample in order to make inferences about either the population parameter or the population distribution.

First we make an assumption H_0 about our distribution, which is called the null hypothesis. We then create an alternative hypothesis, H_1, which is the opposite of our null hypothesis. We then use our sample data in order to determine whether H_0 can be rejected. However, there are 2 types of errors we may come across when it comes to hypothesis testing. A type I error is when we reject

our null hypothesis H_0 when H_0 is actually true, denoted by α. A type II error is when we accept a null hypothesis H_0 when H_0 is actually false, denoted by β.

Chi-Squared Test of Independence

We adopt the Chi-squared test of independence to examine the strength of association between the incidence of low-weight newborns and smoking status. To implement this particular test, the data should be arranged in a contingency table. The Chi-square test works by weighing the observed frequencies for a particular categorical variable to the frequencies for that variable expected if the null hypothesis is correct. Thus, as we implemented the Chi-squared test, we assumed that the expected frequencies are all equal.

The Chi-squared statistic is calculated by dividing the squared difference between observed and expected value by the expected value. Then, we sum these results from i to the number of cells in the contingency table.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Bootstrap

Bootstrap helps us generate confidence intervals if we do not know whether the estimator has a normal distribution. To be specific, we can use bootstrap to check whether the data meet the central limit theorem assumptions, and we can get the confidence interval via the central limit theorem. Otherwise, if the data does not meet the central limit theorem assumptions, we can use bootstrap to generate a bootstrap confidence interval, which is also a valid confidence interval.

To be specific, we can use bootstrap to generate the pseudo estimate values and check whether they are normally distributed. If they are normally distributed, the central limit theorem assumptions are satisfied. Otherwise, we can use bootstrap data to generate bootstrap confidence intervals.

Quantile-Quantile Plot

The quantile-quantile plot, or Q-Q plot, is used as our graphical method and empirical test to compare the distributions of two groups. To be specific, a quantile-quantile plot shows how two distributions are similar. The quantile-quantile plot is defined as:

$$C(P) = (F_x^{-1}(P),\ (F_y^{-1}(P)), \text{ where } P \in (0, 1)$$

We can get the curve of the plot by plotting the two groups of data, all x_i and all y_i, into the plot. As a result, if we get a straight line with intercept 0 and slope 1, the two distributions are identical. Otherwise, they are not the same distribution. Specifically, a non-zero intercept means there is a shift for two distributions, and not one slope indicates there is scale change among two distributions.

In our analysis, we use the Q-Q plot as the empirical test for normality. To be specific, we use x-axis as the ideal normal distribution, and y-axis as the distribution which we want to check. Therefore, if the graph has approximately slope 1 and intercept 0, the distribution is approximately normal.

Simulation Study

We can use simulation studies to check the normality of a distribution. Generally, we can generate pseudo random values from a given distribution, and then check the similarity of the simulated distribution with that of the observed data. To be specific, for checking normality, we calculate the kurtosis and skewness coefficients of observations, and simulate with data from a normal distribution. The size of simulations is equal to the size of observations. Then, we repeat the simulation many times, and we plot the histogram of simulations' kurtosis and skewness coefficients. Eventually, we determine whether the observed distribution is normally distributed according to the histogram.

Contributions

Jared Dishman: Data cleaning
Alison Camille Dunning: Data (Survey Variable Descriptions, connecting variables to the main question, other data tables), Investigation (Scenario 1, 5, Additional Question II), Theory (CLT, Chi-Square Test of Independence), paper editing, table descriptions, LaTeX conversion
Ruotian Gao: Investigation (Scenario 4, 6)
Yaoxin Li: Investigation (Scenario 2, 3, Additional Question I), Theory (Point Estimates, Confidence Intervals, Bootstrap, Quantile-Quantile Plot, Simulation Study)
Bryan Talavera: Introduction, Data (Survey Variable Descriptions), Theory (Hypothesis Testing, Skewness, Kurtosis)
Abdiaziz Weheliye: Introduction
Ye Yint Win: Background

References

1. Simmons, A. E. (2019, March 2). *The Disadvantages of a Small Sample Size*. Sciencing. https://sciencing.com/disadvantages-small-sample-size-8448532.html.
2. *1.4 - Confidence Intervals and the Central Limit Theorem: STAT 506*. PennState: Statistics Online Courses. https://online.stat.psu.edu/stat506/lesson/1/1.4.
3. Flexjobs. (2020, December 29). *How Many Hours Is a Part-Time Job?: FlexJobs*. FlexJobs Job Search Tips and Blog. https://www.flexjobs.com/blog/post/how-many-hours-will-you-work-in-a-part-time-position/.

4. Muir-Herzig, Rozalind G. (February 2004). *Technology and Its Impact in the Classroom.* Computers and Education, v42 n2 p111-131
5. Hainey, Thomas (2013, July 23). *Students' Attitudes Toward Playing Games and Using Games in Education: Comparing Scotland and the Netherlands,* v69 p474-484