Video Game Preferences Among UC Berkeley Stats Students in Fall 1994

Student 1 and Student 2

2024-10-19

0. Contribution Statement

Student 1

Student 1 wrote the description of the data and worked on questions 1, 3, and 4 (except for the graphs). Student 1 also worked on the extra credit problem (#6). Student 1 also did the regression analysis.

Student 2

Student 2 mainly worked on questions 2 and 5 and provided graphs for question 4. Student 2 also did the conclusion and discussion.

Introduction

Data

The data from videodata.txt and videoMultiple.txt were collected as part of a survey for students at UC Berkeley enrolled in a particular statistics course that had about 3000-4000 students. Students from Statistics 2, Section 1, in Fall 1994, were invited to participate in the survey if they partook in the second exam of the course. Within this section, 314 students were eligible to participate and 95 students were randomly selected through a random number generator. The data from videodata.txt was the first part of the survey and asked about background information of the survey participant. The data was numerical and discrete (e.g., Time, the number of hours played in the week prior to the survey as an integer) or categorical but encoded as a numerical value (e.g., Like to play, rated on a scale of 1 to 5). The data from videoMultiple.txt was the second part of the survey and covered whether the student likes or dislikes playing video games. In these questions, more than one response could be given. The data in this file was recorded as binary values indicating whether the student selected that option. Some columns also contain string values if the student provided a reason that was not given as to why they dislike video games.

Basic Analysis

Question 01

Methods

First, we loaded the data in through R.

To determine a point estimate of the fraction of students we played a video game in the week prior to the survey, we need the total number of students. The video_data table has 91 observations and 15 variables. Since we have no way of proving that each student only filled out the form once, we will assume so for this assignment.

Data Cleaning:

We will replace all values of 99 with NA.

```
video_data[video_data == 99] <- NA
```

POINT ESTIMATE: We can count the number of students who played a video game in the week prior to the survey by counting the number of students who played a video game for more than zero hours in the week prior.

```
n <- nrow(video_data)
played_count <- sum(video_data$time > 0)
point_estimate_fraction <- played_count / n
```

INTERVAL ESTIMATE: We will construct a confidence interval that contains a range of values that likely contain the population parameter, the proportion of all students who answered the survey who played a video game in the week prior to the survey. For a 95% confidence interval, we will use z = 1.96.

```
z <- 1.96
lower_interval_estimate_fraction <- point_estimate_fraction - z *
sqrt(point_estimate_fraction * (1 - point_estimate_fraction) / n)
upper_interval_estimate_fraction <- point_estimate_fraction + z *
sqrt(point_estimate_fraction * (1 - point_estimate_fraction) / n)
```

We constructed the confidence interval using the formula p-hat = z +/- sqrt((p-hat)*(1 - p-hat) / n), where p-hat is the sample proportion of students who played a video game in the past week and n is the sample size. Z is the Z-score for the confidence level we chose (95%).

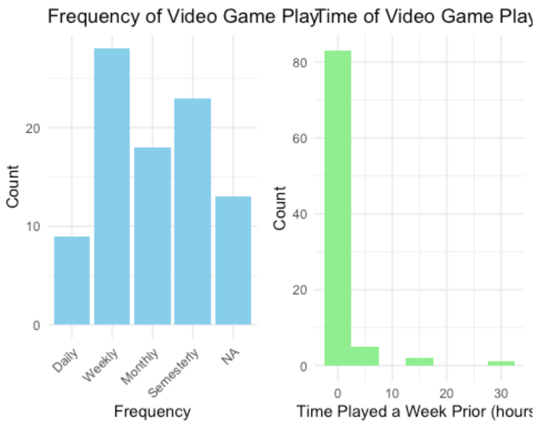
Analysis

Point estimates are a single numerical value that summarizes the data based on the sample data. In this case, point_estimate_fraction is a point estimate (sample proportion) that estimates the proportion of students who played a video game in the week through a single number, approximately 0.374. Interval estimates, on the other hand, provide some leeway in case the point estimate is close but not exactly the same as the population proportion. The interval estimate uses the point estimate and creates a range depending on our chosen confidence level, which in this case, is 95%. In this case, we can say that if we take many independent samples of size n = 91 from the population, about 95% of them would cover the true population proportion (p). It is important to note that it would be incorrect to state we are 95% confident that p is between 0.274 (the lower interval) and 0.473 (the upper interval) because the confidence interval is not measuring our confidence in a particular interval. It is also incorrect to state that there is a 95% chance that p is between 0.274 and 0.473 because p is a fixed value, and we thus should not calculate the probability of it occurring within a range.

Question 02

Methods

To understand the differences in distributions of the hours played per week and the reported frequency of play, we graphed each as a frequency histogram.



Additionally, we examined the basic summary statistics of first the frequency cateogory then time category.

Daily	Weekly	Monthly	Semesterly	NA's
9	28	18	23	13

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	1.243	1.25	30

Analysis

Based on the frequency histograms, most students report their video gaming frequency to be weekly, followed by semesterly, then monthly, and finally daily. A majority of survey submissions report the count of hours spent playing video gamees in the week prior to be between 0 and 5, with some people reporting 10 to 15 hours and fewer 25 to 30 hours.

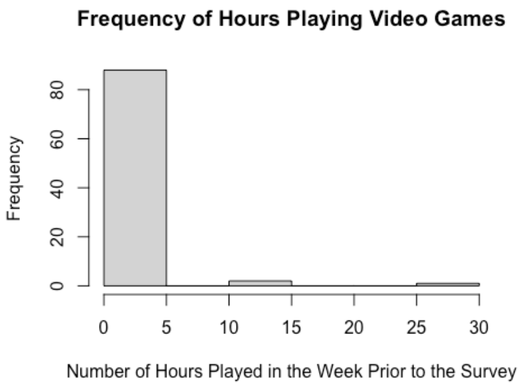
Conclusion

The results of graphing each feature of the survey showed that most students were correctly reporting the frequency of their video game play. The highest frequency of hours played the week prior was 0 to 5 hours, which is logical considering that the highest frequency of video game play was weekly. It is reasonable to assume that a majority of students would play video games only a few hours a week. Additionally, the external factor of an exam during the week prior would most likely have a negative association with the number of hours played for that week.

Question 03

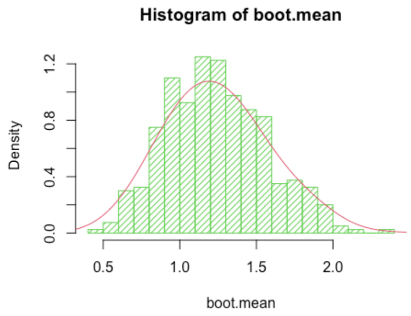
Methods

```
hist(video_data$time, xlab = 'Number of Hours Played in the Week Prior to the Survey', main = 'Frequency of Hours Playing Video Games')
```

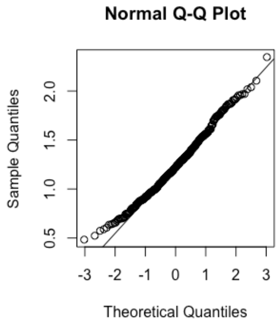


Since the distribution of the amount of time people played video games in the week prior to taking the survey is not approximately normal, we need to create a bootstrap population. We can do this by repeating every sample for 314/ 91 is about 3.45 times since our sample was drawn from a population of N = 314 eligible students.

```
set.seed(371)
shuffle.ind = sample(1:nrow(video_data))
boot.population <- rep(video_data$time[shuffle.ind], length.out = 314)
# Choose our first sample
sample1 <- sample(boot.population, size = 91, replace = FALSE)

# Choose 400 samples from our bootstrap population and store them in a 2D Array
# Each row represents a bootstrap sample of size 91 (we have 400 rows/samples)
# Each column represents an element (we have 91 elements)
set.seed(6653)
B = 400
boot.sample <- array(dim = c(B, 91))
for (i in 1:B) {
  boot.sample[i, ] <- sample(boot.population, size = 91, replace = FALSE)
}

# Calculate the sample mean of each bootstrap sample
boot.mean <- apply(X = boot.sample, MARGIN = 1, FUN = mean)

# Histogram of Bootstrap sample means
hist(boot.mean, breaks = 20, probability = TRUE, density = 20, col = 3, border = 3)
lines(density(boot.mean, adjust = 2), col = 2)
```



```
# QQ plot of Bootstrap sample means
par(pty = 's')
qqnorm(boot.mean)
qqline(boot.mean)
```



Both the histogram and QQ plot (the points closely follow the straight reference line) imply the bootstrap sample means are approximately normally distributed. This means we can construct a 95% confidence interval.

POINT ESTIMATE: the mean of the bootstrapped sample means

```
# Close to the sample average
point_estimate_average <- mean(boot.mean)
```

CONFIDENCE INTERVAL: contains a range of values that likely contain the population parameter, the average amount of time spent playing video games in the week before the survey for UC Berkeley students enrolled in a particular section.

```
# Bootstrap confidence interval
# Find the values between 2.5% to 97.5% to capture the middle 95% of the
data.
interval_estimate <- c(quantile(boot.mean, 0.025), quantile(boot.mean,
0.975))
lower_interval_estimate_average <- interval_estimate[[1]]
upper_interval_estimate_average <- interval_estimate[[2]]
```

Analysis

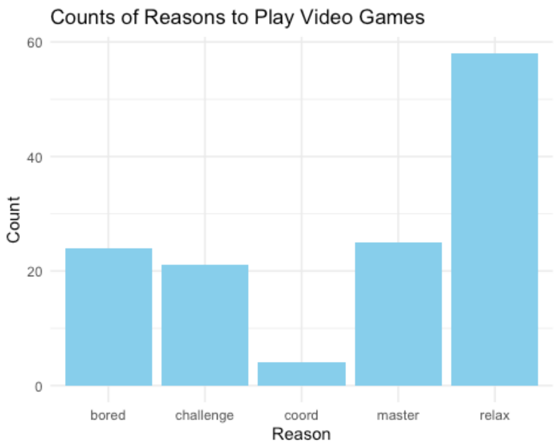
The point estimate average, about 1.242, means that after bootstrapping from the original population, we would expect the average amount of time spent playing video games in the week prior to the survey to be about 1.242 hours. The confidence interval estimate average, from about 0.676 to 1.902, means that if we took many independent samples of size 91 from the population, 95% of the samples would capture the true population parameter within this range.

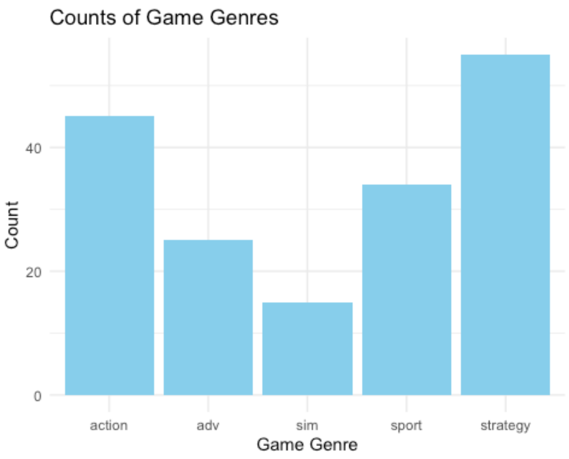
Conclusion

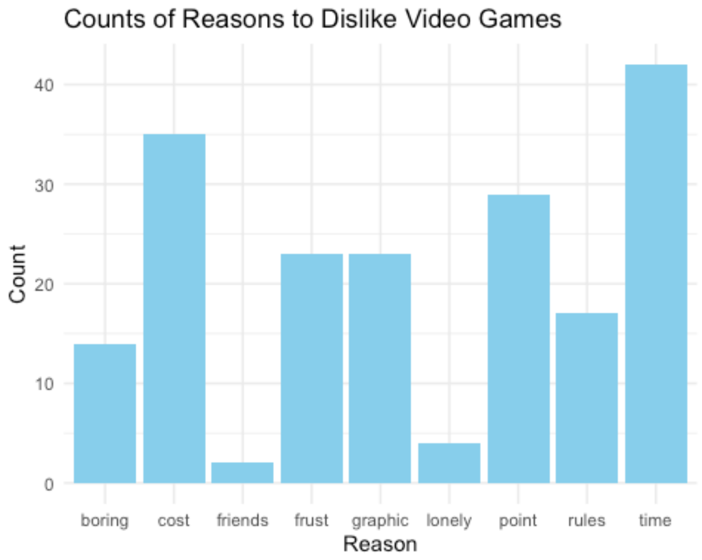
From the histogram and the qq plot, we can see the sample distribution of the number of hours spent playing video games is extremely skewed. Additionally, n is large but n / N = 91/314 is not small. We are unsure about whether the probability distribution of the sample average follows a normal curve, so can bootstrap to estimate the properties of the population. We collected 400 samples, each with a size of 91 since our initial sample was of size 91. We double checked for normality using a histogram and QQ Plot before continuing with bootstrapping. Then, we created the point estimate by taking the mean of the bootstrapped samples. Our point estimate was about 1.242, which would be the average amount of time we would expect to be spent playing video games in the week prior to the survey if we had to summarize this in a single value. We also created a 95% confidence interval, which captures 95% of the samples within the range of 0.676 to 1.902 hours.

Question 04

Methods







The second question of the Attributes section keeps track of why people play the games checked above. This section was marked with binary values, where 1 means that the person selected that option and 0 means that the person didn't select the option. We made a list of the mean results for each option in this question, which represents the proportion of people who chose that particular option as why they enjoy playing video games.

```
reasons_played <- list()
for (col in c("graphic", "relax", "coord", "challenge", "master", "bored")){
     mean_value <- mean(video_multiple[[col]], na.rm = TRUE)
     reasons_played[[col]] <- mean_value
}
```

This yields 0.26 for graphic, 0.67 for relax, 0.05 for coordination, 0.24 for challenge, 0.29 for master, and 0.28 for bored. Since two-thirds of gamers listed relaxation as one of their reasons for playing, it’s reasonable to conclude that most people who play video games likely find them enjoyable.

People who do not play video games were told to skip the first question. This means that they marked neither 0 nor 1 in the question options and left them as NA. We can calculate the number of people who skipped the first question by choosing a column and counting the NAs. We can choose a random column because the number of NAs should be the same in each category (Action, Adventure, Simulation, Sports, Strategy) if people followed directions and skipped the question when they have not played video games before.

```
sum(is.na(video_multiple$action))

## [1] 4
```

Four people out of 91 respondents have not played video games before. Since two-thirds of the people who DO play video games play them for relaxation, that means 2/3 * (91 - 4) = 58 people out of 91 who enjoy playing video games. That’s equivalent to approximately 63.74%. This is more than half, so it is fair to say in general, most students enjoy playing video games!

By sorting the reasons_played list in decreasing order by value, we can see the two most popular reasons for playing video games are for relaxation and people enjoy the feeling of mastery.

```
T2_pros <- sort(unlist(reasons_played), decreasing = TRUE)
```

We will repeat this process to determine the main reasons why people dislike playing video games.

```
disadvantages <- list()
for (col in c("time", "frust", "lonely", "rules", "cost", "bored", "friends",
"point")){
  mean_value2 <- mean(video_multiple[[col]], na.rm = TRUE)
  disadvantages[[col]] <- mean_value2
}
T2_cons <- sort(unlist(disadvantages), decreasing = TRUE)
```

The top two reasons why people dislike playing video games are time and cost. We will add relaxation, mastery, time, and cost to the variable reasons_for likeliness.

```
reasons_for_likeliness <- c('relaxation', 'master', 'time', 'cost')
```

Analysis

It is important to note there is a distinction between the first two bullets, which can only be generalized to students who play video games, and the last two bullets, which can be generalized to all students. * About **0.667 of the students who play video games** enjoy them because they are **relaxing**. It is also worth noting that in the ‘Other’ category, people listed other reasons why they enjoy playing video games, and these included ‘excitement’, ‘fun’, and ‘lowers stress’, which supports the idea that many students who play video games find them relaxing and destressing. * About **0.287 of the students who play video games** enjoy them because they enjoy the feeling of **mastery**. * About **0.483 of all students** don't like video games because of the amount of **time** they take. Time is a very critical factor when it comes to playing video games. In fact, of the nine people who provided an additional reason why they don't like video games (denoted in the 'Other2' column), four of them listed that they felt unproductive while playing video games. * About **0.402 of all students** don't like video games because of their **cost**. This number is almost half, showing that cost is also a very significant factor people consider when it comes to playing video games.

Conclusion

In general, students enjoy playing video games. The most important determinants of whether students like or dislike playing video games are relaxation, mastery, time, and cost. However, it is important to note there may be non-response bias because only the people who played video games were able to say what they liked about the games. All students, regardless of whether they played video games or not, had to answer the question of what they disliked about playing video games. For the question of what people disliked, there could be a difference in the answers between those who have played video games and those who have not. Without considering the breakdown of whether an individual plays video games, we can simply state that students in general dislike the time and cost when it comes to playing games.

Question 05

Methods

First, we find the indices of each respondent group and checked which answers from the attitude questions had the most responses in order to narrow down the range of our analysis.

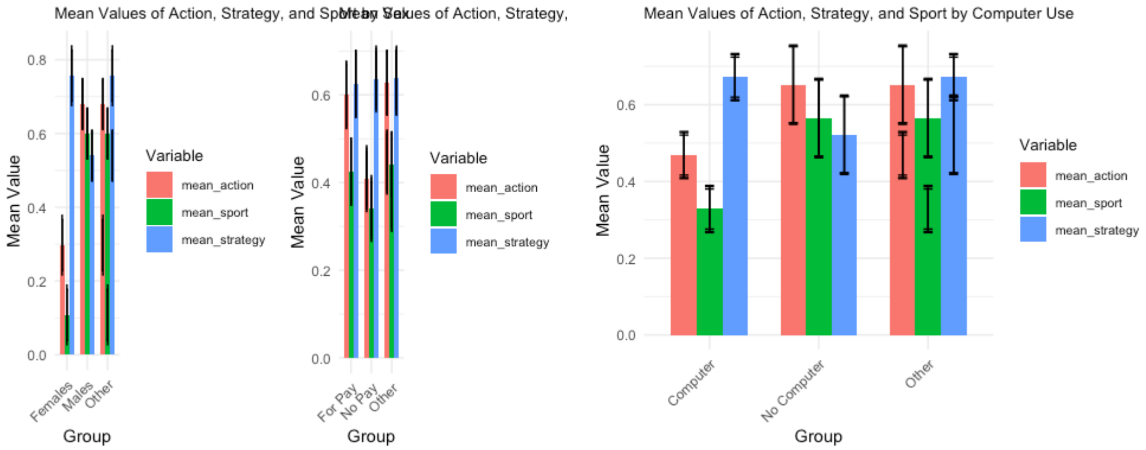
Table continues below

action	adv	sim	sport	strategy	relax	coord	challenge	master
45	25	15	34	55	58	4	21	25

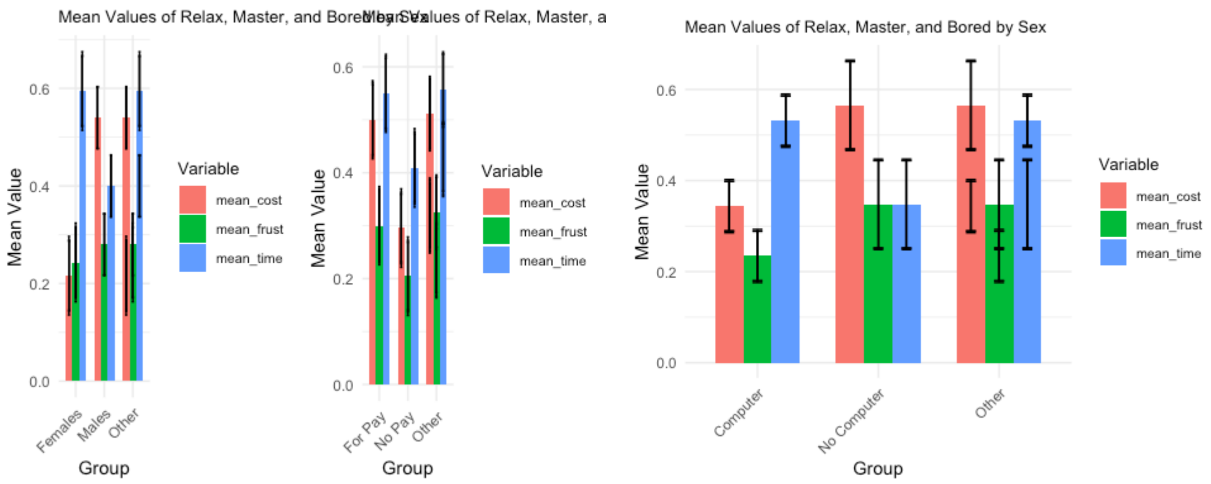
bored	graphic	time	frust	lonely	rules	cost	boring	friends	point
24	23	42	23	4	17	35	14	2	29

Then, we checked the means of the genres Action, Strategy, Sport (answers with the top three highest frequency for this question) by each group.

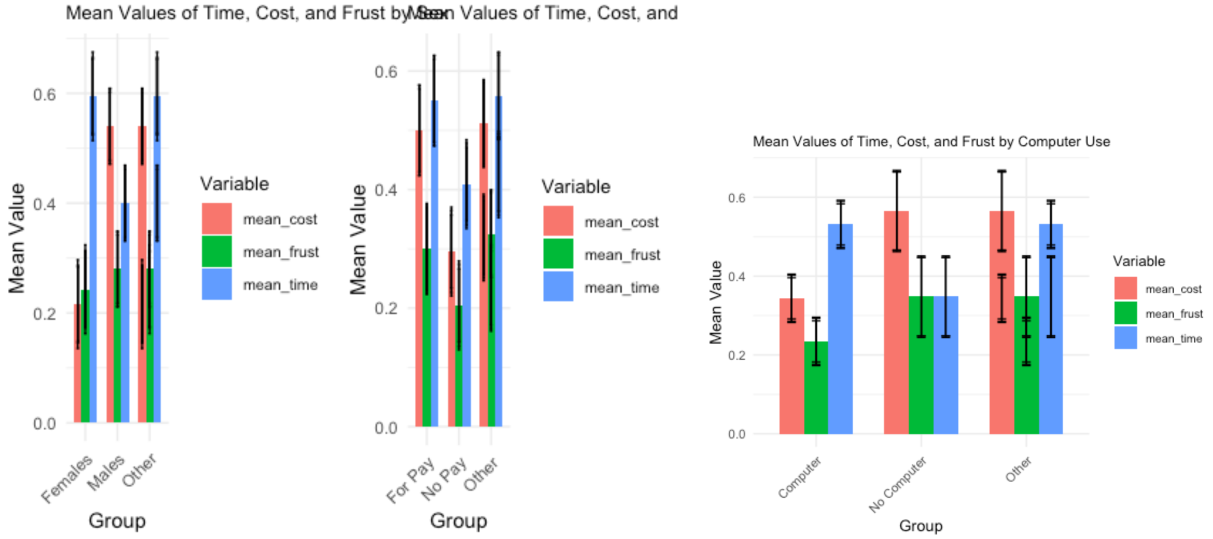
```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```



Next, we analyzed the means of Relax, Master, and Bored (answers with the top three highest frequency for this question) by each group.



Finally, we examined the means of Time, Cost, and Frust (answers with the top three highest frequency for this question) by each group.



Analysis

Based on the plots above, it is clear that there are significant differences in the distribution of each answer by within each group and across all groups. Generally, the plots for sex by group and ownership of computer by group have the similar distributions. More specifically, the distribution for responses from males is similar to the distribution for responses from people who do not own a computer. Additionally, the only answer with the highest frequency across all groups was Relax in response to the question concerning why the respondent played video games, if they did at all.

Conclusion

The graphical representations of the means of the three most popular answers by group show that there significant differences between the distributions of each answer. The differences vary between and among different groups of respondents, and among respondents who like to play video games and those that do not. The greatest difference is among those who own and do not own computers, which makes logical sense. Although video games started with video game consoles, they have recently transitioned to personal computers. Someone who does not own a computer would not have the ability to play most video games, which is why the responses for those without computers is more negative to neutral.

Question 06 (Extra Credit)

Methods

Ho: Our null hypothesis is that the distribution of the grades matches our target distribution. H1: Our alternative hypothesis is that one or more categories of grades differs from the expected.

We created a graph to show the distribution of the grades as a histogram.

```r
library(ggplot2)
distr_of_grades <- ggplot(video_data, aes(x = as.factor(grade))) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(x = "grade", y = "Count", title = "Distribution of Grades") +
  theme_minimal()
```

It looks like no one expected to receive a D (1) in the course. Most (more than 50 out of 91 respondents) expected to receive a B (3) in the course. The next most popular expectation was an A (4), with more than 30 out of 91 respondents expecting to receive that grade. The least number of people (about 8 out of 91) expected to receive a C (2).

Now, we will calculate the proportions of each grade received by creating a proportion table.

```r
grade_prop <- prop.table(table(video_data$grade, useNA = 'no'))
```

From the proportion table, we can see that there were about 34.07% A's, 57.14% B's, 8.79% C's, and 0% D's. This is quite different from the target distribution of 20% A's, 30% B's, 40% C's, and 10% D's.

Analysis

We are interested in whether our findings of the sample proportions significantly differ from the target distribution. To investigate this, we will conduct a chi-squared goodness of fit test.

```r
counts <- table(video_data$grade, useNA = 'no')
# We need to modify the observed counts to have a 0 to account for the
```

```
missing grade of D.
observed_counts <- c(0, 8, 52, 31)
chi_squared_test <- chisq.test(observed_counts, p = c(0.1, 0.4, 0.3, 0.2))
```

With a p-value of 1.629 * 10^(-13) < 0.05, we can reject the null hypothesis. There is a difference in the proportion of grades in the target distribution and the true outcome.

Now, we will investigate the impact of non-response (4 people) having failing grades.

```
observed_counts2 <- c(4, 8, 52, 31)
chi_squared_test2 <- chisq.test(observed_counts, p = c(0.1, 0.4, 0.3, 0.2))
```

With our new chi-squared test, we still have a p-value of 1.223 * 10^(-11) < 0.05, we can reject the null hypothesis. There is a difference in the proportion of grades in the target distribution and the true outcome, and the non-respondents do not change the picture of the grade distribution.

Conclusion

Using a proportion table and chi-squared test, we found that there is a significant difference between the proportion of grades expected to be achieved and the target distribution. However, there is no difference in the grading distribution if the four non-respondents had failing grades.

Advanced Analysis

Methods

For our advanced analysis of this dataset, we decided to do a regression analysis to see how grade varies with the number of hours spent playing video games.

```
library(ggplot2)
model <- lm(grade ~ time, data = video_data)
```

Analysis

From the regression analysis, we are able to see that a line of regression fitting the data would have a slope of about -0.009725 and an intercept at (0, 3.265). This means that for every additional hour someone plays video games in the past week before their exam, their expected grade decreases by the 0.009 in grade points. There is a negative association between the number of hours spent playing video games and the grade achieved. Additionally, the intercept is at (0, 3.265), which tells us that when people did not play any video games in the past week, we would expect their grade to be 3.265, which means most likely to be a B with a smaller chance of receiving an A.

Conclusion

Typically, more hours playing video games per week results in a lower grade in the course. We suspect this is because students neglect their studies when they are focused on their video games, and thus, are less prepared for the exams. However, there may also be confounding variables that cause this relationship in lower grades, such as difficulty of the class, innate ability to problem solve, and previous experience in statistics. Another confounding variable is income and accessibility to resources. For example, a family may provide their child with many video games because they are wealthy, but the stability of the family's wealth may also cause the child to not take school seriously.

Conclusion

Conclusion Summary

Our analysis of this survey data provided important insight into the video gaming habits of UC Berkeley students enrolled in a particular stats class. It is clear that these students were a diverse sample through our analysis of their responses. Additionally, we were able to understand the reasons and motivations behind their video game habits and explored the complexities of their responses.

Discussion

After analyzing the data in depth, it is clear that it has major flaws that prevent further analysis. One flaw that was prominent throughout the data analyzation process was that some questions of the survey included customizable responses that each respondent could write their own response into. This limits the scope of our analysis, as we had no was to effectively assess this category of answers. Another flaw that we noticed was that it was not clear how the two dataframes were related. There was no clear relational key that allowed us to clearly understand which respondent provided which answers.