Video Game Preferences Among UC Berkeley Stats Students in Fall 1994

Heidi Tam and Paige Pagaduan

2024-10-19

0. Contribution Statement

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.

```r
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.

```r
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.

```r
z <- 1.96
lower_interval_estimate_fraction <- point_estimate_fraction - z * sqrt(point_estimate_fraction * (1 - p
upper_interval_estimate_fraction <- point_estimate_fraction + z * sqrt(point_estimate_fraction * (1 - p
```

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%).

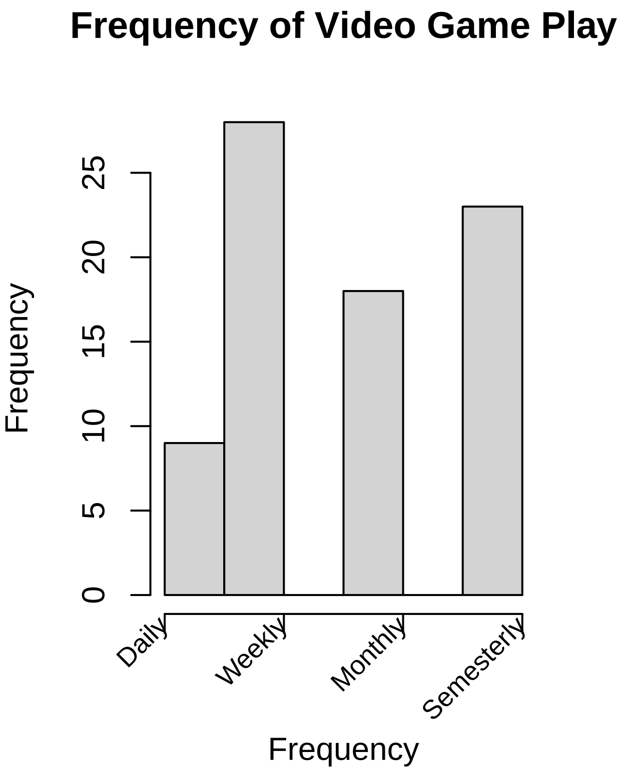
Analysis

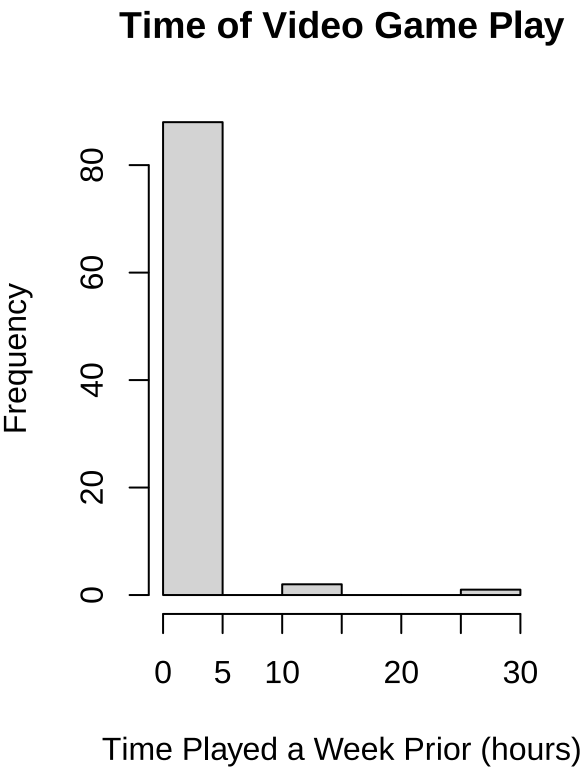
Point estimates are a single numerical value that summarizes the data based on the sample data. In this case, point_estimate_fraction is a point estimate (sample proportion) that estimates the proportion of students who played a video game in the week through a single number, approximately 0.374. Interval 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.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
1	2	3	2.705	4	4	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.

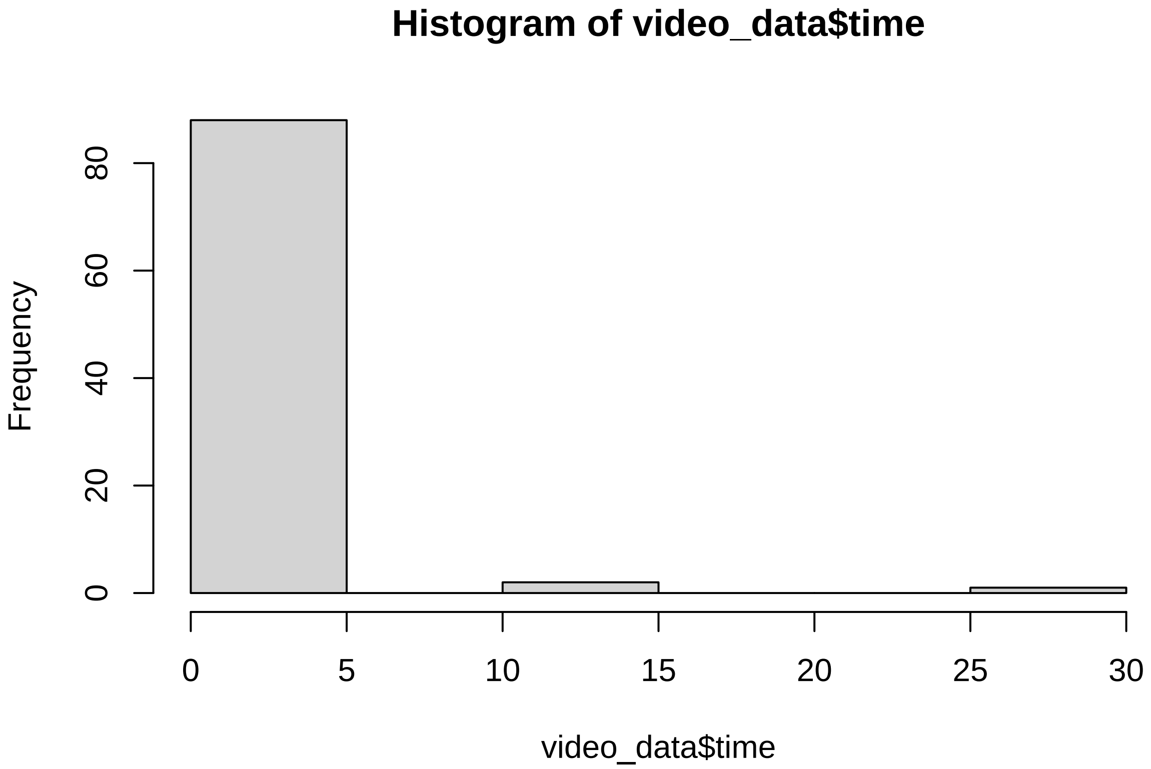
Conclusion

The results of graphing each feature of the survey

Question 03

Methods

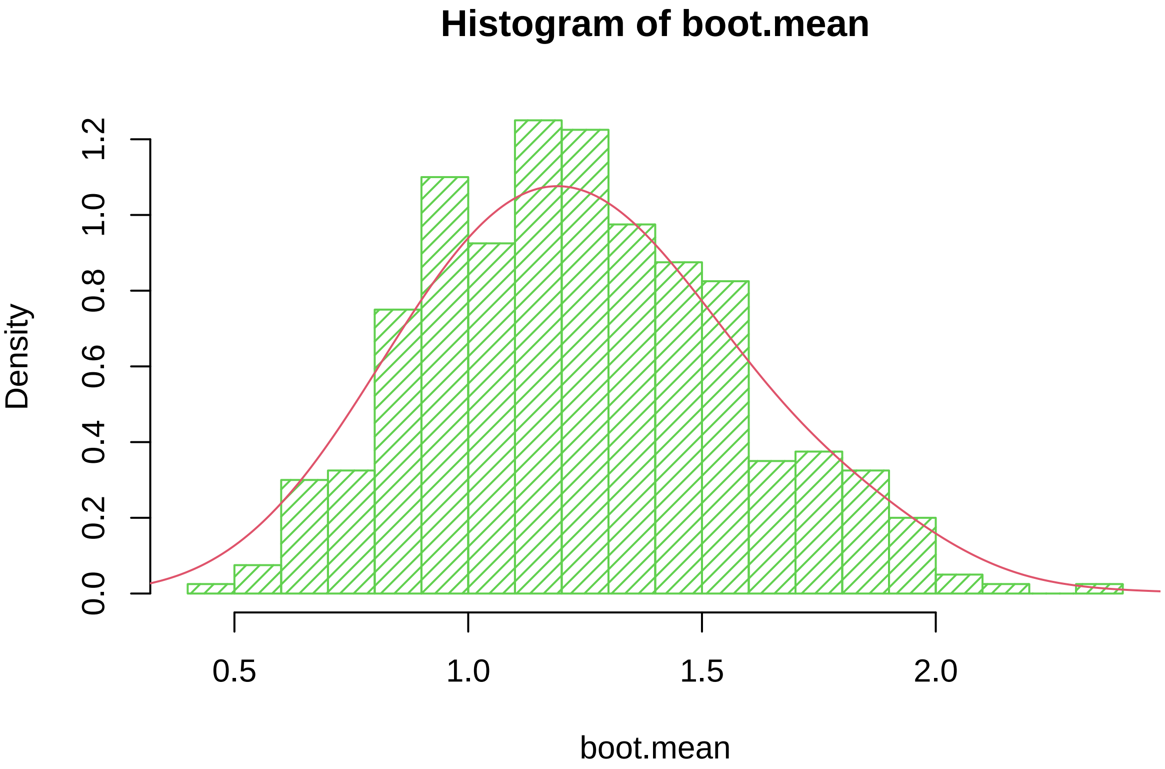
```r
hist(video_data$time)
```



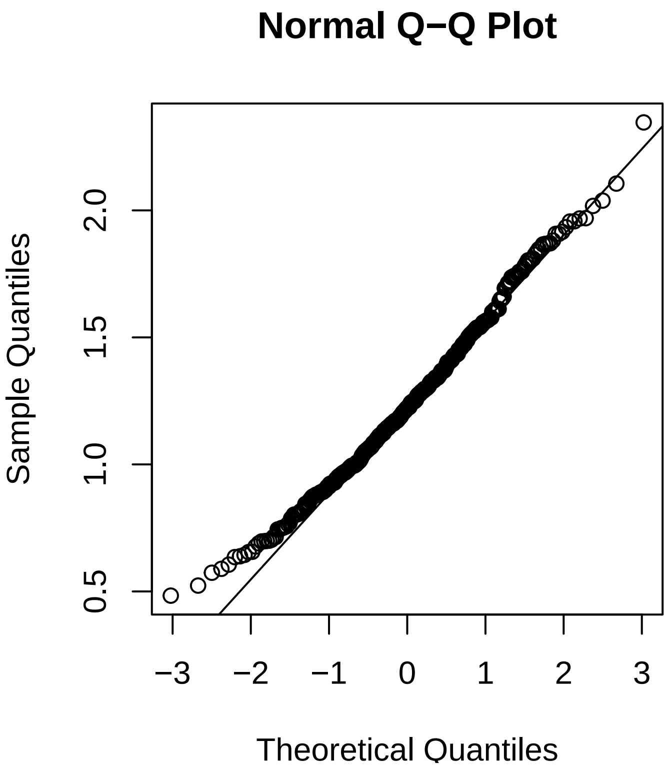
Since the distribution of the amount of time people played video games in the week prior to taking the survey is not approximately normal, we need to create a bootstrap population. We can do this by repeating every sample for 314/ 91 is about 3.45 times since our sample was drawn from a population of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

```r
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
```

```r
hist(boot.mean, breaks = 20, probability = TRUE, density = 20, col = 3, border = 3)
lines(density(boot.mean, adjust = 2), col = 2)
```



```r
# 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

```
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.

```
# 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.