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.

```
n <- nrow(video_data)
# POINT ESTIMATE:
# We can count the number of students who played a video game in the week prior to the survey by counti
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 popul
# 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 - 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.

Conclusion

The point estimate gives the best estimate for the proportion of students who played a video game in the week prior to taking the survey through a single value, while the interval estimate estimates this proportion through a range of values. However, this range is not guaranteed to have all sample proportions within this range if we repeated the sampling process many times. The higher our confidence level, the wider the interval range since there is a higher likelihood of the sampling proportions falling into the confidence level. Similarly, the lower our confidence level, the smaller the confidence interval.

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.