R Coding Demonstration Week 8: Probabilities and Reopening Campus

Matthew Blackwell

Gov 51 (Harvard)

Introduction

- Reopening campuses under COVID is very tricky!
- Roughly 2,400 students have returned to campus this Fall.
- What is the probability of transmission under various scenarios?
    - Useful for planning gather size limitations, etc.
- Probability: what kinds of data/outcomes should we expect under certain assumptions about how the data is generated.
    - Soon, inference: what can we learn about true probabilities from data.

Data

- Fictitious data on each returning student:

```
covid_data <- read.csv(”data/covid_data.csv”)
```

Variable Name	Description
id	Student ID number
covid	Does the student currently have COVID (1) or not (0)
house	What House is the student living in?

```
head(covid_data, 3)
```

```
##   id covid    house
## 1  1     0 Kirkland
## 2  2     0  Currier
## 3  3     0 Winthrop
```

Question 1

Suppose there are 2 students currently infected with COVID among the 2400 students living on campus. A student randomly chooses 20 students from the 2,400 **with replacement** to an indoor party. What is the probability that at least one of the infected students is invited?

Use the rules of probability to analytically calculate this probability. Then, use the `covid_data.csv` data to simulate the probability.

Answer 1 (Analytic)

- First, we want at least one, but this is easier to work with the complement (no one has covid).

$$\mathbb{P}(\text{At least one COVID case at party}) = 1 - \mathbb{P}(\text{No COVID cases at party})$$

- Sampling with replacement means all 20 invitees have equal probability of having COVID and are independent.

$$\mathbb{P}(\text{No COVID cases at party}) = \left(1 - \frac{2}{2400}\right)^{20} = 0.9835$$

- Putting these together:

$$\mathbb{P}(\text{At least one COVID case at party}) = 1 - \left(1 - \frac{2}{2400}\right)^{20} = 0.0165$$

Answer 1 (Simulation)

```r
n_sims <- 1000
party_with_result <- rep(NA, times = n_sims)

for (i in 1:n_sims) {
  party <- sample(covid_data$covid, size = 20, replace = TRUE)
  party_with_result[i] <- sum(party) > 0
}
mean(party_with_result)
```

```
## [1] 0.021
```

Question 2

With the same setup as the last time, now assume that a random sample **with replacement** of 1,000 students are invited to a CS 50 lecture in Sanders theater. What is the probability that one of the infected students is invited to the lecture?

Use the rules of probability to analytically calculate this probability. Then, use the `covid_data.csv` data to simulate the probability.

Answer 2 (Analytic)

Using the same approach as the last problem:

$$\mathbb{P}(\text{At least one COVID case at party}) = 1 - \left(1 - \frac{2}{2400}\right)^{1000} = 0.5656$$

Answer 2 (Simulation)

```r
n_sims <- 1000
cs50_with_result <- rep(NA, times = n_sims)

for (i in 1:n_sims) {
  cs50 <- sample(covid_data$covid, size = 1000, replace = TRUE)
  cs50_with_result[i] <- sum(cs50) > 0
}
mean(cs50_with_result)
```

```
## [1] 0.55
```

Question 3

Suppose there are 2 students currently infected with COVID among the 2000 students living on campus. A student randomly chooses 20 students from the 2,400 **without replacement** to an indoor party. What is the probability that at least one of the infected students is invited?

Use the `covid_data.csv` data to perform Monte Carlo simulation to calculate this probability.

Answer 3

```r
n_sims <- 1000
party_wo_result <- rep(NA, times = n_sims)

for (i in 1:n_sims) {
  party <- sample(covid_data$covid, size = 20, replace = FALSE)
  party_wo_result[i] <- sum(party) > 0
}
mean(party_wo_result)
```

```
## [1] 0.016
```

Question 4

With the same setup as the last time, now assume that a random sample **without replacement** of 1,000 students are invited to a CS 50 lecture in Sanders theater. What is the probability that at least one of the infected students is invited to the lecture?

Use the `covid_data.csv` data to perform Monte Carlo simulation to calculate this probability.

Answer 4

```r
cs50_wo_result <- rep(NA, times = n_sims)

for (i in 1:n_sims) {
  cs50 <- sample(covid_data$covid, size = 1000, replace = FALSE)
  cs50_wo_result[i] <- sum(cs50) > 0
}
mean(cs50_wo_result)
```

```
## [1] 0.656
```

Question 5

Now, suppose Professor Blackwell is part of a mentoring program that randomly pairs an on-campus student with a professor. The randomly chosen student insists on meeting in person. What is the probability that the student has COVID?

Now, suppose the Crimson has reported that both of the COVID cases are in Adams House and the student lives in Adams. What is the probability that Prof. Blackwell has a COVID mentee? Assume that the 2400 students have been equally divided into the 12 houses. Use the definition of conditional probability to calculate this quantity and verify it via simulation using the `covid_data.csv` data.

Answer 5 (Analytic)

- Randomly sampling a COVID: student: 2/2400.
- Knowing Adams House: use the definition of conditional probability/Bayes' Rule:

$$\mathbb{P}(\text{COVID} \mid \text{Adams}) = \frac{\mathbb{P}(\text{Adams} \mid \text{COVID})\mathbb{P}(\text{COVID})}{\mathbb{P}(\text{Adams})}$$

- We know that both of the COVID cases are in Adams, so $\mathbb{P}(\text{Adams} \mid \text{COVID}) = 1$ and equal division of the students means that $\mathbb{P}(\text{Adams}) = 1/12$. Thus:

$$\mathbb{P}(\text{COVID} \mid \text{Adams}) = \frac{1 \times (2/2400)}{1/12} = \frac{1}{100} = 0.01$$

Answer 5 (Simulation)

```
adams <- subset(covid_data, house == ”Adams”)

mentor_result <- sample(adams$covid, size = n_sims, replace =
mean(mentor_result)
```

```
## [1] 0.013
```