Problem set 4

Vivian Yeh

Scope: C sets

Question 1: Scope

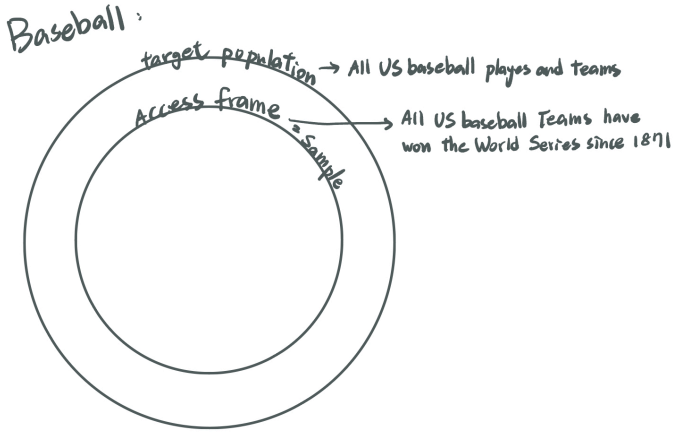
Baseball:

Target population: All US baseball players and teams

access frame: All US baseball teams have won the World Series since 1871

sample: All US baseball players and teams have won the World Series since 1871

unit of observation: A team in a given year



National Youth Tobacco Study:

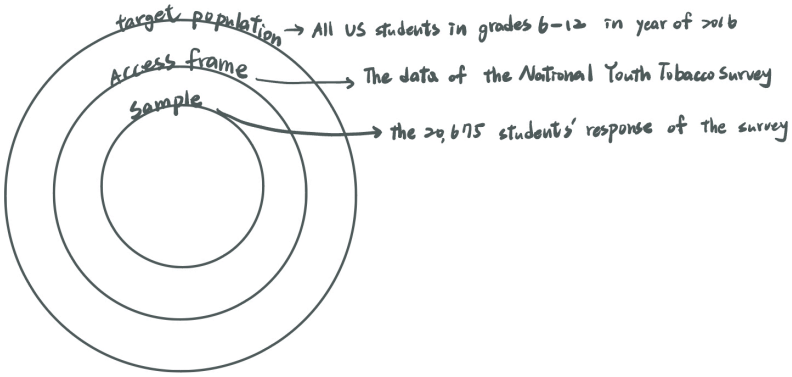
Target population: All US students in grades 6-12 in year of 2016

access frame: The data of The National Youth Tobacco Survey

sample: the 20,675 student's response of the survey

unit of observation: a student

Tobacco :



Question 2: Data

Baseball: The unites are in the target population and also in the access frame. Sean Lahman collected the data because he wanted to figure out the relationship between team payrolls and winning the World Series. The data were collected in 2021. Luhman collected the data on Lahman Baseball Database.

National Youth Tobacco Study: The unites are in the target population and also in the access frame. CDC and the Food and Drug Administration collected the data in order to address the reasons middle and high school students used e-cigarettes. The data were collected in 2016, and they collected the data on the National Youth Tobacco Survey.

Question 3: Question

Refine the question, if needed, to one that is answerable by the data. To do this: Consider whether the granularity implied by the question matches that of the sample. Consider whether the question is too vague or too ambitious.

Baseball: What is the relationship between the before-taxed income of a player and the top three teams which have won the World Series.

National Youth Tobacco Study: What is the most common reason that motivates a 6-12 grade student starting to use an e-cigarette.

Question 4: Bias

Baseball: non-response bias - There may be some teams that are the winner of the World Series do not want to share the payrolls of players, as the result, we can not access the income of the players from the data.

National Youth Tobacco Study: non-response bias - Since this is a questionnaire, there may be many students that did not answer all not questions completely or properly. Coverage bias - Some private schools may refuse to take the surveys, so their data are missing.

Question 5: Variation

Determine whether a chance mechanism was used in the data collection process. If so, identify the type of mechanism: sampling, assignment, measurement, or some combination of them.

Baseball: The data collector did not use a chance mechanism to collect this database.

National Youth Tobacco Study: The data collector did not use a chance mechanism to collect this database.

Question 6: Comparison

The scope diagrams of baseball data is different from the National Youth Tobacco Study because the access frame of baseball data is same as its sample range, and the access frame and sample of Tobacco is different. Baseball data is collected by a person, and the data is about the performance of the players and teams. The data is hard to be incorrect, but there are still some missing data. For the tobacco data, it is more unreliable than the baseball data because students may be not honest or non-response. Moreover, some private school may reject to participate in the survey test. Therefore, the data of baseball is more reliably generalizable than the other.

Based on the names alone, I would classify the baseball data as an example of an administrative sources. And, the Tobacco data is the survey sample.

Box Model

A town has 1,000 people in it. They live in 300 households. The number of people in a household ranges from 1 to 6, inclusive. About 400 of the 1,000 people in the town are children.

For each of the sampling schemes below, describe a box model that imitates the chance process.

You may use multiple boxes, if needed. For each box, describe:

1. The box and tickets (what tickets go in the box, and their values)

2. The number of draws

3. Whether or not the tickets are replaced between draws

Question 7

Select 50 households at random and report the average number of people in the household.

1. The tickets will be the household, and each household contains different numbers of people. There are total 300 tickets in the box.

2. For this question, it has to draws 50 times.

3. The tickets are not replaced between draws.

Question 8

Visit the elementary, middle, and high schools (there is one of each in the town) and take a sample of 25 from each school. Report the average household size across the 75 students sampled.

1. There are total 400 students in this town, and we have to classify which of them are in elementary, middle, and high schools, and then we create three boxes: one for each school. Each ticket represents a student, and the value of the ticket is the household size of the student.

2. We draw 25 times for each boxes, so the total draws will be 75 times.

3. The tickets are not replaced between draws.

Question 9 BONUS

There are at least three problems regarding sampling bias with the scenario in question 6. Can you identify them?

There are coverage, selection, and non-response biases.