Data Project Part 1

Carri Beshears, Samantha Sears, Samuel Johnson, Catalina Perez, and Phuong Tseng

2022/2/12

Questions:

1. [2 marks] What is the problem your are addressing with these data? State the question you are trying to answer and let us know what type of question this is in terms of the PPDAC framework.
2. [2 marks] What is the target population for your project? Why was this target chosen i.e., what was your rationale for wanting to answer this question in this specific population?
3. [2 marks] What is the sampling frame used to collect the data you are using? Describe why you think this sampling strategy is appropriate for your question. To what group(s) would you feel comfortable generalizing the findings of your study and why.

The sampling frame used CDC STI data from 2000-2014. This sampling frame is appropriate to the question because it gives specific information about state, sex, STI, year, and population. Groups to generalize:Sexually active male/female in the US during years of 2000-14. Not comfortable generalizing with non-binary people and possibily MSM groups- because data only specifies M/F And does not differentiate between sexual preferences.

4. [2 marks] Write a brief description (1-4 sentences) of the source and contents of your dataset. Provide a URL to the original data source if applicable. If not (e.g., the data came from your internship), provide 1-2 sentences saying where the data came from. If you completed a web form to access the data and selected a subset, describe these steps (including any options you selected) and the date you accessed the data.

We pulled this dataset from this cdc website: https://wonder.cdc.gov/std.html. This sample data comes from the CDC STI dataset from 2000 - 2014. It was fairly simple to pull this dataset; the link takes the user to a portal after clicking the "Data Request" link where they can fill out the requirements that they want in their dataset such as disease, year, gender, state or region (geography as unit of measurement), and etc. User can also title the dataset, identify the type of measure (e.g., count, population, or rate) then download the file. This sample data was retrieved on February 12, 2021.

Set up the packages To load the libraries that we need without repeating the library function, we used the lapply function in R to load them all at once.

```
sti_data <- read_xlsx("sti_data.xlsx", sheet = "2000-2014")
```

5. [1 mark] Write code below to import your data into R. Assign your dataset to an object. We use the read_xlsx function to read in the data set and specify the excel sheet then assign it to an object called sti_data.

Question 6

6. [3 marks] Use code in R to answer the following questions:

```
dim(sti_data) #returns 23400 rows and 11 columns
```

i) What are the dimensions of the dataset?

```
## [1] 23400    11
```

```
names(sti_data)
```

ii) Provide a list of variable names.

```
##  [1] "Year"         "Year Code"    "State"        "State Code"   "Gender"
##  [6] "Gender Code"  "Disease"      "Disease Code" "Count"        "Population"
## [11] "Rate"
```

```
head(sti_data)
```

iii) Print the first six rows of the dataset.

```
## # A tibble: 6 x 11
##    Year `Year Code` State `State Code` Gender `Gender Code` Disease
##   <dbl>       <dbl> <chr>        <dbl> <chr>  <chr>         <chr>
## 1  2000        2000 Alab~            1 Female F             Chancr~
## 2  2000        2000 Alab~            1 Female F             Chlamy~
## 3  2000        2000 Alab~            1 Female F             Gonorr~
## 4  2000        2000 Alab~            1 Female F             Total ~
## 5  2000        2000 Alab~            1 Female F             Primar~
## 6  2000        2000 Alab~            1 Female F             Primar~
## # ... with 4 more variables: `Disease Code` <dbl>, Count <chr>,
## #   Population <chr>, Rate <chr>
```

Question 7

7. [4 marks] Use the data to demonstrate a statistical concept from Part I of the course. Describe the concept that you are demonstrating and interpret the findings. This should be a combination of code and written explanation.

Explanation