# DRP Final Project

Mekias Kebede

3/18/2022

## Power of Computing in Statistics

My project for this quarter supervised by Jess Kunke centered around how powerful computing can be for statistics. Specifically I learned a variety of techniques within the R programming language, which is a free software environment designed specifically for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, including Windows and MacOS. The R language can be a powerful tool in virtually all of statistics; for the scope of this project the specific area of statistics I worked on this quarter was survey statistics.

I choose to analyze a 2003 Immunization Survey done by the CDC in the Chicago area. They utilized a sampling method to draw an appropriate sample from the Chicago area (their sampling frame) of 471 children to survey regarding their vaccination status. This information was later used to conduct estimates on the larger population of children in Chicago as a whole and played a role in estimating nationwide estimates. For the purposes of this project I used their sample by pretending it was the sampling frame and thus drew a sample from their survey data because for logistical and practical reasons conducting a true survey where I would physically survey people myself would have complicated this process. The advantage of doing this is that I have the true answers; and so any estimates I would conduct from my sample could easily be compared to the true answers. The sampling and statistical estimations I did from this data set also served as the medium for which I learned how to use R.

## Survey Statistics

Prior to learning how to use R however, I began by learning some basic statistics integral to survey statistics in particular. This includes how to chose the most appropriate sampling method for a data set. Throughout the quarter we learned of 3 sampling methods:

1. `Simple Random Sampling Without Replacement (SRS WOR)`
2. `Auxiliary Information`
3. `Clustering`

In a large population, we as the statisticians can't always study the entire population for logistical reasons and so we take a sample that is as best of a representation of the whole population as we can get and use that to generalize the whole population. Out of the 3 sampling methods I listed above, we have to chose wisely which to use for sampling because each has its own strengths and weaknesses depending on the data set. For the purpose of analyzing the immunization data set, I used simple random sampling (WOR) because there was no additional information I could use to utilize the auxiliary information method and cluster sampling would not be effective because a better representation of our population could be obtained by using simple random sampling.

### Horvitz-Thompson Estimators

All of these sampling methods can be simplified to a general form called the Horvitz-Thompson (HT) estimator. For a given data set we can find the sum of each individual's value divided by what we call a inclusion probability. The inclusion probability is defined as the chance that the individual people will be selected to be in the sample. This probability is always between 0 and 1 and is defined as our sample size divided by our population size. In the context of our immunization data set, $y_i$ would represent the number of shots an individual child from our sample has received. And the $\pi_i$ represents the inclusion probability for each child (Note that this is not the mathematical constant $\pi$ but a variable). Our inclusion probability is dependent on what sampling method was used; in this case because simple random sampling (WOR) was used, each child has an equal probability of being chosen, thus making our inclusion probability the number of children in our sample divided by the total number of children in population. We then sum $\frac{y_i}{\pi_i}$ for $i \in s$, $s$ meaning the total number of respondents in sample.

$$\hat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

Another way to represent this for simple random sampling is to factor out our inclusion probability from the sum and represent $\pi_i$ as $\frac{n}{N}$ where n is our sample size and N is our total population size.

$$\hat{Y} = \frac{N}{n} \sum_{i \in s} y_i$$

## Getting Started with R

After learning about some of the prerequisite mathematics and statistics needed to sample and estimate information from the data I was now ready to use R. Once R Studio was properly downloaded and installed, I opened R Studio up and immediately 4 distinct sections were displayed in each corner of what is called our work space:

1. Data viewing (top left corner)
   - The top left corner is a place where your data can be viewed a organized table format.
2. Environment (top right corner)
   - The top right corner stores tabs listed as environment, history, connections, and tutorial.
3. Console (bottom left corner)
   - The bottom left corner of the work space is where the console, terminal, render, jobs tabs are located. This is where the code one writes is executed. The code you write can be written here, however you can write your code also in the top left corner in a R script and store it, where it can be executed at a later time.
4. Files (bottom right corner)
   - The bottom right corner includes tabs listed as files, plots, packages, help, and viewer.

### Reading in Data

Oftentimes, for new users of R studio, loading a data set from a file into R and getting it set up to use can be a very tedious step that can be confusing. Down below I have listed out the steps in order of code that must executed to properly read in a data set into your work space for use. In this example I am using the immunization data set.

```
getwd() #finds your working directory
```

```
setwd("/Users/mekiaskebede/Desktop/DRPSTATS") # sets the place you're storing your R files
# as your working directory

datadr = "/Users/mekiaskebede/Desktop/DRPSTATS/"

data <- read.csv(paste0(datadr, "nispuf03chi.txt"), sep=" ") #reads in the data
# you want to analyze into R and the paste0 function strings 2 strings together
```

## Understand Your Data Before Using It

Once the data has been "read in" we can now play around with our data in a variety of ways by using what we call commands or functions in R. However prior to working with any data set it is critical know what the data means and what the specific variables within a data set represent. For example, within the immunization data set I worked on, there are over 30 different variables that represent pieces of information collected by researchers on children. These variables include information like:

- Race
- Sex
- Family Income
- Residing State
- Immunization Status (For Variety of Shots)

It is important to know the specific variable names to be able to utilize the variables when writing more detailed code because often times the variables aren't in plain language and are simplified to confusing acronyms that one may not be easily identify the meaning behind. This information can usually be found in a key of some sort that accompanies most data sets to describe the relevance and meaning of each variable collected as well as identifying values to show whether a variable is true of false conditional (Example: Binary 1 & 0). With this in mind, here are some commands I learned that allow me to view and access specific data I want:

```
View(data) # directs you to where data is displayed in R
# (top left corner in a tab where your data file is named)
```

```
head(data) # prints first 6 rows of data in console

imm = subset(data, PDAT==1) # created variable that
# stores subset of data when PDAT is 1
# Data entries where PDAT is 1 means the child has proof of vaccination
# this is what I mean by "Understand your Data" to accurately analyze the correct stuff.
# Analyzing survey data from children whose vaccination status is not verified
# would make for questionable estimates on the greater Chicago area.

subset(data, I_HISP_K == 1) # prints a subset of data to console

nrow(data) # counts number of rows in data
# (In this data set, rows represents the number of children surveyed)

data[1,1] # finds the specific value in data at row "#" & column "#" because data is 2-D

data[,3] # same command but this time by not adding a specific value for rows
# it prints every row in column 3 in the console
```

```
data$EDUC1 # this can also be written as data[,6] since EDUC1 is located in column 6
# this command prints every row in column 6 which represents our EDUC1 variable
```

### Using Sampled Data for Estimates

As I mentioned earlier drawing samples from sampling frames (sampling frame is the whole population you trying to estimate things about by using a sample from it) is a highly important concept in survey statistics used to estimate things we would like to know about a greater population. In R we can draw samples using the “sample” command, and what it does is basically randomly select an assigned number of integers between zero and your given input value in the parenthesis. As an example, let’s consider trying to estimate the average number of shots every individual child has in Chicago based on a sample survey of children in a segment of the Chicago population. Using the HT estimator and sampling used in simple random sampling, I was able to estimate a variety of things about the immunization data, one being the average number of shots children have had.

```
sample(1:471, 5, replace = FALSE) # prints out a random set of 5 integers between 1 and 471
# "replace = FALSE" means that no number can be repeated; this is the default

sample(1:10, 10, replace = TRUE) # prints out a random set of 10 integers between 1 and 10
# this time "replace = TRUE" meaning a number can be repeated and isn't thrown out after first use

N = nrow(imm) # number of rows in imm data in context of data
# it means the number of children # in the imm data set

n = 50 # variable created to represents arbitrary sample size

imm$P_NUMDTP # displays P_NUMDTP variable entries in imm data
# P_NUMDTP represents the number of DTP shots a child has been given so far

sum(imm$P_NUMDTP) # sums all entries in P_NUMDTP meaning it gives the total
# number of shots all the children combined have had

(N/n)*sum(imm$P_NUMDTP) # prior sum multiplied by total respondents and
# inclusion probability

imm$P_NUMDTP[1:50] # displays P_NUMDTP data again but only the first 50 entries from the full set

imm$P_NUMDTP[sample(1:277, 50, replace = FALSE)] # displays P_NUMDTP data but randomly
# selected entries from a sample created of size 50

(N/n)*sum(imm$P_NUMDTP[sample(1:277, 50, replace = FALSE)]) # outputs an estimated number of
# DTP shots given to whole imm population

(1/n)*sum(imm$P_NUMDTP[sample(1:277, 50, replace = FALSE)]) # prints estimation of
# the number of DTP shots per child in Chicago
```

## Conclusion

Sampling and estimating things using R was the primary focus this quarter but I also learned a variety of other things in R, including how to plot graphs in R, utilize logical conditionals in code, and began learning about LaTeX. The process of learning the basics of R and using statistics in R was a tedious process that left me struggling at times; however with the help of my mentor Jess, clarity was restored and I was able to learn more than I thought I could this quarter about R and statistics with a full load of other course work. Thank

you to the DRP STATS program for this opportunity and Jess for guiding me throughout this quarter.