Sampling Project Overview

Project Overview

The project will require a written report and 5 minute oral discussion.

Regardless of whether you choose option 1 (collect your own data), option 2 (use your own dataset), or option 3 (use one of the datasets below), the following components need to be included:

1. Introduction: Concise statement of research question
2. Sample Size Calculations: Sample size calculation
3. Sample Design: Description of process for selecting sampling units – (include code)
4. Estimation and Results: Mathematical description of estimation procedure
5. Conclusion: summary of your results from part 3 and 4

Project Rubric

The projects will be evaluated from the following rubric.

Component	Expectation	Points
Introduction		**10**
	Provide thorough overview and background information about the problem you are studying.	5
	Clearly articulate the research questions you are trying to answer.	5
Sample Size Calculations		**10**
	Provide thorough discussion of sample size calculations, including discussion on choosing d the desired margin of error and how S^2 is estimated.	10
Sample Design		**10**
	Discuss how the data will be collected (e.g. survey, visual observation, sampling from existing data file, ect...)	5
	Describe and justify your sampling design. Why is this the most appropriate procedure to answer your research questions?	5
Estimation		**15**
	Describe your estimation procedure, (e.g. SRS, regression, ect...) and justify why this is the best method for estimation.	5
	Compute point estimates and confidence intervals for your quantities of interest. Detail the confidence interval procedure and discuss any implicit assumptions.	10
Summary of Findings		**10**
	Describe your results of your estimation procedure and the inferences you can make in regards to your defined research question. What is the target audience Note any problems with scope of inference of limitations in your results.	10
Other		**25**
	Include code for reproducible research?	5
	Presentation is well rehearsed and engaging, stayed within specified time limits.	10
	Written paper is easy to read, contains proper structure, and grammar.	10

Potential Data Sets

This section contains a set of curated datasets for students to explore with the course project. Specific research questions still need to be identified.

Cluster Sampling: Washington School Immunization

This dataset contains a set of schools in Washington. Each row in the dataset is a school, so this would be a cluster sampling scheme with the primary sampling unit being a school and within a school all secondary sampling units (the students) are selected. The dataset also contains information about the total number of students at the school, proportion that have been immunized along with additional school information and characteristics.

```
WA_Imm <- read_csv('http://www.math.montana.edu/ahoegh/teaching/stat446/WA_Immunization.csv')
WA_Imm
```

```
## # A tibble: 2,365 x 10
##    School_Name School_year Reported K_12_enrollment Percent_complet~
##    <chr>       <chr>       <chr>               <dbl>            <dbl>
##  1 SKY VALLEY~ 2016-17     Y                      21             81
##  2 LACROSSE H~ 2016-17     Y                      21             85.7
##  3 CALVARY CH~ 2016-17     Y                      21             52.4
##  4 RIVERDAY S~ 2016-17     Y                      22             81.8
##  5 MONTESSORI~ 2016-17     Y                      22             95.5
##  6 PARAMOUNT ~ 2016-17     Y                      22             95.5
##  7 GRACE LUTH~ 2016-17     Y                      22            100
##  8 QUEETS-CLE~ 2016-17     Y                      22              0
##  9 COLUMBIA E~ 2016-17     Y                      22             90.9
## 10 MUKILTEO A~ 2016-17     Y                      23            100
## # ... with 2,355 more rows, and 5 more variables: School_District <chr>,
## #   County <chr>, ESD <chr>, Has_kindergarten <chr>, Has_6thGrade <chr>
```

Allegheny County Dog Licenses

The contains records on dogs registered in Allegheny County, PA (near Pittsburgh). The data set contains information on gender of the dog, breed, and dogs name. This dataset would best answer questions about population proportions, or totals, (such as proportion of dogs that have been spayed or neutered or the proportion of dogs that are a labrador retriever / golden retriever.

Note: This will require a little more coding expertise to organize the data.

```
dogs <- read_csv('http://www.math.montana.edu/ahoegh/teaching/stat446/Dogs.csv')
dogs
```

```
## # A tibble: 2,544 x 3
##    LicenseType                Breed              DogName
##    <chr>                      <chr>              <chr>
##  1 Dog Lifetime Neutered Male MIXED              BANDIT
##  2 Dog Lifetime Spayed Female CHES BAY RETRIEVER DAISY
##  3 Dog Lifetime Neutered Male CHIHUAHUA MIX      MCGEE
##  4 Dog Lifetime Male          BICHON FRISE       ROCKY
##  5 Dog Lifetime Spayed Female AM PITT BULL MIX   HARLEY
##  6 Dog Lifetime Spayed Female GER SHEPHERD MIX   EMMA
##  7 Dog Lifetime Spayed Female BOXER              ADRIAN
```

```
##  8 Dog Lifetime Spayed Female BEAGLE MIX         LOLA
##  9 Dog Lifetime Spayed Female COCKAPOO           MACY ANN
## 10 Dog Lifetime Spayed Female GOLDENDOODLE       BIANCA
## # ... with 2,534 more rows
```

Pittsburgh 311 Calls

311 calls are non-emergent police calls (in contrast to 911). The dataset contains a list of ~ 70,000 requests. This will be somewhat similar to the previous example, where it might be best to think about estimating population proportion or total of some characteristics.

```
pitt <- read_csv('http://www.math.montana.edu/ahoegh/teaching/stat446/Pitt311.csv')
pitt
```

```
## # A tibble: 70,029 x 4
##    CREATED_ON          REQUEST_TYPE          REQUEST_ORIGIN NEIGHBORHOOD
##    <dttm>              <chr>                 <chr>          <chr>
##  1 2016-12-07 12:02:00 Leak                  Call Center    Brookline
##  2 2016-12-12 07:32:00 Leak                  Call Center    Beechview
##  3 2016-08-12 10:25:00 Dead Tree (3TB)       Call Center    Larimer
##  4 2016-07-25 14:20:00 Park Shelter          Call Center    Sheraden
##  5 2016-06-01 09:17:00 Traffic Shop          Control Panel  Troy Hill
##  6 2016-09-06 15:03:00 Abandoned Vehicle (p~ Call Center    Central Norths~
##  7 2016-02-25 13:23:00 Landslide             Call Center    Spring Hill-Ci~
##  8 2016-08-29 10:05:00 Landslide             Call Center    South Side Slo~
##  9 2016-08-31 08:51:00 Weeds/Debris          Call Center    Manchester
## 10 2016-12-06 16:01:00 Abandoned Vehicle (p~ Control Panel  Brighton Heigh~
## # ... with 70,019 more rows
```

Bike Trips 2017

The bike trips data contains information for nearly 250,000 bike rentals during March 2017. For each bike rental, the data contains: start station, end station, rental length (minutes), hour (the rental started), and day of the week.

```
bikes <- read_csv('http://www.math.montana.edu/ahoegh/teaching/stat446/Bike_Trips.csv')
bikes
```

```
## # A tibble: 244,687 x 8
##    Start.station End.station Member.Type start_time
##    <chr>         <chr>       <chr>       <dttm>
##  1 14th & Irvin~ 15th & Euc~ Registered  2017-03-31 23:55:00
##  2 15th & P St ~ 17th St & ~ Registered  2017-03-31 23:52:00
##  3 5th St & Mas~ 17th St & ~ Registered  2017-03-31 23:51:00
##  4 Columbus Cir~ 8th & F St~ Registered  2017-03-31 23:50:00
##  5 Columbia Rd ~ 16th & Har~ Registered  2017-03-31 23:50:00
##  6 Columbia Rd ~ 9th & Upsh~ Registered  2017-03-31 23:49:00
##  7 1st & D St SE 1st & K St~ Registered  2017-03-31 23:47:00
##  8 Wilson Blvd ~ Key Blvd &~ Casual      2017-03-31 23:46:00
##  9 Eastern Mark~ 15th St & ~ Registered  2017-03-31 23:46:00
## 10 Eastern Mark~ Potomac & ~ Registered  2017-03-31 23:45:00
## # ... with 244,677 more rows, and 4 more variables: end_time <dttm>,
## #   hour <dbl>, interval_time <dbl>, week_day <dbl>
```