

Lab 9

Lab Overview

This lab will look at cluster sampling using another variation of the bike dataset. Note new stations that have less than 10 total rentals are filtered out of this dataset. The `bikes_final` dataset contains the starting station and the trip duration in minutes.

```
bikes <- read_csv('http://www.math.montana.edu/ahoegh/teaching/stat446/Bike_Trips.csv')
trip_time <- bikes %>% dplyr::select(Start.station, interval_time)

#remove stations with less than 10 rentals
new_stations <- trip_time %>% group_by(Start.station) %>% tally() %>% arrange(n) %>% filter(n < 11)%>%
  pull(Sum)

bikes_final <- trip_time %>% filter(!Start.station %in% new_stations)

all_stations <- bikes_final %>% group_by(Start.station) %>% tally()
```

Using the `bikes_final` dataset, design a cluster sampling procedure. You will be allowed to sample a total of 100 clusters.

1. (4 points)

What are the PSUs and SSUs in this scenario?

2. (2 points)

How many PSUs are there in the sampling frame?

3. (4 points)

Implement and describe a sampling procedure to select 100 PSUs. You can sample all SSUs within a PSU.

4. (4 points)

Based on your sample, what is your estimate of the average bike trip time?

5. (4 points)

Create a confidence interval for the average bike trip time.

6. (2 points)

If the trip times were similar within the each station, but different across stations, would cluster sampling lead to a high or low variance estimator?