TO DO NEXT . . .

SM 625: Week 8 Sampling Project Notes

Assume that you will decide to allocate your final computed n_{opt} number of clusters to each of the nine project strata based on the proportions of the total number of students in the population in each stratum (i.e., if 20% of the population of students comes from Region 1, you would sample 20% of your clusters from that region). Describe the first-stage sampling fractions for each stratum, where the total number of schools to sample at the first stage in each stratum is defined by your proportionate allocation of the n_{opt} clusters.

Next your team should extend your design to consider stratified PPeS selection of schools from each of the nine strata at the first stage of your sample design.

You have been provided with a sampling frame that lists the schools within each region. Given the information on the sampling frame, how might you sort this list to achieve implicit stratification within the regions? You can treat the overall student count from a previous year (tot_all) as the measure of size for the PPeS sampling. Given this information, compute your zone size for systematic PPeS sampling within each of the nine strata (regions), and proceed with systematic selection based on fractional intervals to select the allocated number of schools within each stratum using PPeS sampling. What is your first-stage sampling fraction within each of the nine strata?

Now, suppose that you have the entire list of sampled schools from a given stratum. As outlined in the project description, the client wishes to work under a paired selection model for estimation of sampling variance. How would you form pseudo-strata to meet this request, based on your systematic sampling within each first-stage stratum? We aren't ready to write the estimation methods section yet, but you can describe this general process as a part of the description of your first-stage sampling.

In future weeks, we will work on determining the sampling rates within each of the sampled schools to maintain epsem across the nine strata. For now, your focus should be on refining the description of the first-stage sampling process.

```
# Load necessary libraries
library(tidyverse)
library(sampling)

# Given values
total_N <- 830138
n_opt <- 87
m_opt <- 53

# Compute proportional allocation of clusters to each stratum
region_counts <- school_frame %>%
  group_by(Region) %>%
  summarise(pop_count = sum(tot_all), .groups = "drop") %>%
  mutate(prop_allocation = pop_count / sum(pop_count),
         clusters_allocated = pmax(1, round(n_opt * prop_allocation)))  # Ensure at least 1 cluster per
region_counts
```

```
## # A tibble: 9 x 4
##   Region pop_count prop_allocation clusters_allocated
##    <dbl>     <dbl>           <dbl>              <dbl>
## 1      1      3561         0.00429                  1
## 2      2      5474         0.00659                  1
## 3      3      8631         0.0104                   1
## 4      4      4855         0.00585                  1
```

```
## 5      5     18907         0.0228                  2
## 6      6     33133         0.0399                  3
## 7      7    191992         0.231                  20
## 8      8    188830         0.227                  20
## 9      9    374755         0.451                  39
```

```r
# Implement systematic PPeS sampling
set.seed(123)
school_frame$MOS <- school_frame$tot_all
total_MOS <- sum(school_frame$MOS)
school_frame$pik <- pmin(1, (n_opt * school_frame$MOS) / total_MOS)

# Implement systematic PPS sampling
s <- UPsystematic(school_frame$pik)
sampled_schools <- school_frame[s == 1, ]

# Generate pseudo-strata for paired selection
sampled_schools <- sampled_schools[order(sampled_schools$MOS), ]
num_pairs <- floor(nrow(sampled_schools) / 2)
paired_strata <- split(sampled_schools, rep(1:num_pairs, each = 2, length.out = nrow(sampled_schools)))

# Results
sampled_schools
```

```
## # A tibble: 87 x 17
##    BCODE BNAME                  DCODE District_Name Region County_ID County_Name
##    <chr> <chr>                  <chr> <chr>          <dbl> <chr>     <chr>
##  1 07809 Caro Alternative Educ~ 79020 Caro Communi~      9 79        Tuscola
##  2 08867 Presque Isle Academy ~ 71902 Presque Isle~      5 71        Presque Is~
##  3 09446 Eastern Washtenaw Mul~ 81908 Eastern Wash~      8 81        Washtenaw
##  4 00964 Harbor High School     63220 Huron Valley~      9 63        Oakland
##  5 08372 Vanderbilt Charter Ac~ 70905 Vanderbilt C~      7 70        Ottawa
##  6 09606 Hanley International ~ 82986 Hanley Inter~      9 82        Wayne
##  7 08601 Will Carleton Charter~ 30902 Will Carleto~      8 30        Hillsdale
##  8 03939 ST MONICA SCHOOL       39010 <NA>               9 63        Oakland
##  9 08210 Concord Academy-Petos~ 24901 Concord Acad~      6 24        Emmet
## 10 03209 River Valley Middle S~ 11033 River Valley~      7 11        Berrien
## # i 77 more rows
## # i 10 more variables: `Public/nonpublic` <chr>, g7_totl <dbl>, g8_totl <dbl>,
## #   g9_totl <dbl>, g10_totl <dbl>, g11_totl <dbl>, g12_totl <dbl>,
## #   tot_all <dbl>, MOS <dbl>, pik <dbl>
```