SURV625 Applied Sampling

2025-03-13

SM 625: Week 4 Sampling Project Notes

For each of the three variables that will be the focus of the final course project, the Department of Education would like to generate estimates of means and proportions having a coefficient of variation of no more than 0.05. Using the numbers provided to you in the description of the final project, compute estimates of the element variances for each variable. Given these estimates, compute the desired level of precision (the desired sampling variance) for each estimate that corresponds to the desired coefficient of variation.

Now, given the desired levels of precision for each estimate, compute estimates of the necessary sample sizes for each of these three estimates (assuming simple random sampling), ignoring the finite population correction. These will be starting points for the eventual two-stage cluster sample design.

We first build a table to store our results for each week's assignments.

- We also add the expected averages for each outcome variable.

```
# build dataframe with inputs
MI_school_samples <- tibble(
  Outcome = c("smoked_cig", "smoked_mj", "age_approached_to_smoke"),
  type = c("prop", "prop", "mean"),
  desire_cv = rep(.05, 3),
  expect_mean = c(.25, .15, 12)
)
  # calculate element
MI_school_samples |> kable()
```

Outcome	type	desire_cv	expect_mean
smoked_cig	prop	0.05	0.25
smoked_mj	prop	0.05	0.15
age_approached_to_smoke	mean	0.05	12.00

Our process is to:

- 1st, calculate the estimated element variance.
  - For a proportion, to get the element variance we use $\hat{p}(1-\hat{p})$.
  - For a mean, to get the element variance we simply just square the estimated standard deviation $v(\bar{y}) = \sigma^2$.
- 2nd, we calculate the estimated standard error as $se(\hat{p}) = CV \times \hat{p}$.
- 3rd, we compute the desired sampling variance as: $var(\hat{p}) = se(\hat{p})^2$, where $se(\hat{p}) = \sqrt{var(\hat{p})}$

```
MI_school_samples <- MI_school_samples |>
  mutate(
    # compute element variance
    s_sqrd = if_else(type=="prop", # for proportions
                               expect_mean * (1 - expect_mean),
                               if_else(type=="mean",# for means
                                       1^2, NA)),
    # compute standard error
    se = desire_cv * expect_mean,
    # compute variance
    V = se^2
    )

MI_school_samples |> select(-type) |> kable()
```

Outcome	desire_cv	expect_mean	s_sqrd	se	V
smoked_cig	0.05	0.25	0.1875	0.0125	0.0001563
smoked_mj	0.05	0.15	0.1275	0.0075	0.0000562
age_approached_to_smoke	0.05	12.00	1.0000	0.6000	0.3600000

We now estimate the desired sample sizes when we desire a CV =.05 as $n = \frac{s^2}{se^2}$

```
MI_school_samples <- MI_school_samples |>
  mutate(SRS_n = s_sqrd / V)

MI_school_samples |> select(1, SRS_n) |> kable()
```

Outcome	SRS_n
smoked_cig	1200.000000
smoked_mj	2266.666667
age_approached_to_smoke	2.777778

SM 625: Week 5 Sampling Project Notes

For this week, we will consider the information available for stratified sampling of students. Eventually you are going to design a stratified cluster sample of students, where the clusters (or PSUs) are schools, but we aren't there yet.

Recall the regions of interest in the sampling project description:

```
school_frame <- read_xls(
  "~/repos/SURV625project/data/MI_school_frame_head_counts.xls")
```

Region	County_ID
1	07, 31, 66
2	22, 27, 36, 55
3	02, 21, 52
4	17, 48, 49, 77
5	01, 04, 06, 16, 20, 26, 35, 60, 65, 68, 69, 71, 72
6	05, 10, 15, 18, 24, 28, 40, 43, 45, 51, 53, 57, 67, 83
7	03, 08, 11, 12, 13, 14, 34, 39, 41, 54, 59, 61, 62, 64, 70, 75, 80
8	09, 19, 23, 25, 29, 30, 33, 37, 38, 46, 47, 56, 73, 78, 81
9	32, 44, 50, 58, 63, 74, 76, 79, 82

As "State officials are interested in providing, if at all possible, separate estimates for each of nine education regions in the state, where the regions are defined by groups of counties", we will use these nine regions as strata.

Prepare a table that includes the:

- Overall population counts in each of these nine strata (the total count of students in the target population at each school is in the tot_all column on the sampling frame).
- Given these counts, once you have the working overall sample size (unknown for now and will be decided by your team next week), what is the proportionate allocation plan of that sample of students across these nine strata?

```
# we will use Region, County_ID, and tot_all

# region counts
school_frame |>
  group_by(Region) |>
```

```
  tally(tot_all) |>
  mutate(prop_allocation = n/sum(n)) |>
  rename(pop_count = n) |>
  kable()
```

Region	pop_count	prop_allocation
1	3561	0.0042896
2	5474	0.0065941
3	8631	0.0103971
4	4855	0.0058484
5	18907	0.0227757
6	33133	0.0399126
7	191992	0.2312772
8	188830	0.2274682
9	374755	0.4514370

```
# what is the proportionate allocation plan of that sample
## of students across these nine strata?
```

SM 625: Week 6 Sampling Project Notes

From a previous study, you obtain estimates of the following design effects for each of these three estimates:

- proportion ever smoked one cigarette = 2.5;
- proportion ever smoked marijuana = 2.0; and
- mean age when first asked to smoke = 1.7.

This previous study featured a sample of size n = 7,500 students between the ages of 13 and 19, selected from a total of a = 150 clusters. Using this information, compute a synthetic estimate of roh for each of the three variables. These synthetic estimates of roh will be used to consider alternative cluster sample designs as you continue with your project work. Finally, budget and cost information is now available. The total budget for data collection for this project will be $500,000. The client and the data collection organization estimate that the data collection will cost $3,000 per primary stage cluster (school), and $50 per completed questionnaire within a cluster. We will use this cost information moving forward for optimal subsample size calculations.

We can estimate the sample ICC or roh from the given design effect estimate as:

$$\hat{roh} = \frac{deff - 1}{m - 1}$$

We now that the sample total is $nm = 7500$ and the sample number of cluster is $n = 150$, which we can take the mean cluster size as $m = nm/n = 7500/150 = 50$ and use it to calculate roh.

```
nm <- 7500
n <- 150
m <- nm / n

MI_school_samples <- MI_school_samples |>
  # add deff and roh to our table
  mutate(desire_deff = c(2.5, 2.0, 1.7),
         # compute roh
         roh = (desire_deff - 1) / (m - 1)
```

```
        )

MI_school_samples |> select(Outcome, desire_deff, roh) |>  kable()
```

Outcome	desire_deff	roh
smoked_cig	2.5	0.0306122
smoked_mj	2.0	0.0204082
age_approached_to_smoke	1.7	0.0142857

SM 625: Week 7 Sampling Project Notes

Recall that the client and the data collection organization estimated that the data collection would cost $3,000 per primary stage cluster (school), and $50 per completed questionnaire within a cluster. We will now use this information for optimum subsample size calculations. Recall that the total budget for data collection will be $500,000.

Given this cost information and your estimates of roh for the three different variables of primary interest from last week, compute the optimum subsample size (and the corresponding optimal number of first stage clusters, given the total budget above) for each of the variables.

- We now have budget constraints and denote the cost per cluster as $c_n = \$3,000$ and cost per element as $c_m = \$50$, with a total budget constraint of $C = \$500,000$. Since we know there are $n = 150$ clusters and a total sample size of $7,500$ students.
- To compute the optimum m size we use the following equation:

$$m_{opt} = \sqrt{\frac{c_n}{c_m}\frac{1-roh}{roh}}$$

```
c_n = 3000 # cost per cluster
c_m = 50 # cost per element within cluster
C = 500000 # total budget

MI_school_samples <- MI_school_samples |>
  mutate(
    # compute optimum m size
    m_opt = sqrt( (c_n / c_m) * ( (1-roh)/roh) ),
    n_opt = C / (c_n + m_opt * c_m),
    # compute new deff
    deff_new = 1 + (m_opt-1) * roh,
     # compute total SSU
    total_nm = m_opt * n_opt)

MI_school_samples |> select(Outcome, roh, m_opt, n_opt) |>
  kable()
```

Outcome	roh	m_opt	n_opt
smoked_cig	0.0306122	43.58899	96.53536
smoked_mj	0.0204082	53.66563	87.97734
age_approached_to_smoke	0.0142857	64.34283	80.42281

How will you decide on a single overall optimum subsample size to use in your design?

- Above we estimated the new design effects which range from 2.3 to 1.9, which are almost in line with our desired design effects of 2.5, 1.7. Below we print the new design effects, optimum number of cluster and cluster size, total sample size `total_nm`for our projected $500,000 budget for all three outcome variables.
  - Finally, we compute the sampling cost as $n \times c_n + n \times m \times c_m$ which we defined these terms above.

```
MI_school_samples |>
  select(Outcome, roh, m_opt, n_opt, deff_new, total_nm) |>
  mutate(projected_cost = (c_n * n_opt) + (c_m * n_opt * m_opt),
         projected_cost = scales::dollar(projected_cost)) |>
  mutate_at(2:6, round, 4) |>
  kable()
```

Outcome	roh	m_opt	n_opt	deff_new	total_nm	projected_cost
smoked_cig	0.0306	43.5890	96.5354	2.3037	4207.879	$500,000
smoked_mj	0.0204	53.6656	87.9773	2.0748	4721.359	$500,000
age_approached_to_smoke	0.0143	64.3428	80.4228	1.9049	5174.631	$500,000

Think about a comparison of alternative cluster sample designs: under a fixed cost constraint, how would we decide which design would be best? What will be your overall sample size (n) under this new optimum subsample size?

As you make progress in writing up what you have done so far, provide some discussion of the rationale for your choices in this regard.

Next, given this optimum subsample size and treating the values of roh as portable, compute the new expected DEFF for each estimate given the new design (this can be specific to each variable / estimate, given the different optimum subsample sizes). In addition, compute a new expected SRS variance for each variable under the new design, using the new "optimum" overall sample size (remember that you can treat the element variances for each variable estimated last week as portable). Finally, compute the new expected sampling variance for each estimate under this new cluster sample design. Are you still meeting the client's precision requirements?

- Given that we have three outcome variables, we also have three optimum number of clusters and cluster size estimates. That is, we can design and examine three options of different optimum number of clusters and cluster sizes.
-

We will use the portable roh estimate and calculate new design effects, SRS variance, and complex design variance for each outcome variable.

```
map(seq(1,3), function(x){


  MI_school_samples |>
  select(Outcome, s_sqrd, roh, m_opt, n_opt, total_nm) |>
  # we can print projected total cost for n=50
  mutate(m_opt = m_opt[x], # optimum m from first row
         n_opt = n_opt[x],
         total_nm = m_opt*n_opt,
         # calculate new deff
         deff_new = 1 + (m_opt-1) * roh,
         # calcualte SRS variance
         var_srs = s_sqrd / (total_nm - 1),
         # calculate complex design variance
         var_crs = var_srs * deff_new,
         #m_opt = round(m_opt)
         ) |>
    mutate_at(4:6, floor) |>
    mutate_at(7:9, round, 5) |>
  select(-roh, -s_sqrd) |>
  kable()

}) |>
  set_names(str_c(rep("Option ", 3), seq(1,3)))
```

```
$`Option 1`


|Outcome                | m_opt| n_opt| total_nm| deff_new| var_srs| var_crs|
|:----------------------|-----:|-----:|--------:|--------:|-------:|-------:|
|smoked_cig             |    43|    96|     4207|  2.30374| 0.00004| 0.00010|
|smoked_mj              |    43|    96|     4207|  1.86916| 0.00003| 0.00006|
```

```
|age_approached_to_smoke |    43|    96|     4207|  1.60841| 0.00024| 0.00038|

$`Option 2`


|Outcome                 | m_opt| n_opt| total_nm| deff_new| var_srs| var_crs|
|:-----------------------|-----:|-----:|--------:|--------:|-------:|-------:|
|smoked_cig              |    53|    87|     4721|  2.61221| 0.00004| 0.00010|
|smoked_mj               |    53|    87|     4721|  2.07481| 0.00003| 0.00006|
|age_approached_to_smoke |    53|    87|     4721|  1.75237| 0.00021| 0.00037|

$`Option 3`


|Outcome                 | m_opt| n_opt| total_nm| deff_new| var_srs| var_crs|
|:-----------------------|-----:|-----:|--------:|--------:|-------:|-------:|
|smoked_cig              |    64|    80|     5174|  2.93907| 0.00004| 0.00011|
|smoked_mj               |    64|    80|     5174|  2.29271| 0.00002| 0.00006|
|age_approached_to_smoke |    64|    80|     5174|  1.90490| 0.00019| 0.00037|
```

We print standard error for the complex design with 95% confidence intervals.

```r
map(seq(1,3), function(x){


  MI_school_samples |>
  select(Outcome, expect_mean, s_sqrd, roh, m_opt, n_opt, total_nm) |>
  # we can print projected total cost for n=50
  mutate(m_opt = m_opt[x], # optimum m from first row
         n_opt = n_opt[x],
         total_nm = m_opt*n_opt,
         # calculate new deff
         deff_new = 1 + (m_opt-1) * roh,
         # calcualte SRS variance
         var_srs = s_sqrd / (total_nm - 1),
         # calculate complex design variance
         var_crs = var_srs * deff_new,
         # compute confidence intervals
         se = sqrt(var_crs),
         lower = expect_mean  - 1.96*se,
         upper = expect_mean + 1.96*se) |>
    mutate_at(3:5, round, 4) |>
```

```
  select(Outcome, expect_mean, se, lower, upper) |>
  kable()

}) |>
  set_names(str_c(rep("Option ", 3), seq(1,3)))
```

$`Option 1`

Outcome	expect_mean	se	lower	upper
smoked_cig	0.25	0.0101330	0.2301393	0.2698607
smoked_mj	0.15	0.0075266	0.1352479	0.1647521
age_approached_to_smoke	12.00	0.0195532	11.9616756	12.0383244

$`Option 2`

Outcome	expect_mean	se	lower	upper
smoked_cig	0.25	0.0101863	0.2300348	0.2699652
smoked_mj	0.15	0.0074861	0.1353272	0.1646728
age_approached_to_smoke	12.00	0.0192675	11.9622357	12.0377643

$`Option 3`

Outcome	expect_mean	se	lower	upper
smoked_cig	0.25	0.0103207	0.2297715	0.2702285
smoked_mj	0.15	0.0075168	0.1352671	0.1647329
age_approached_to_smoke	12.00	0.0191884	11.9623908	12.0376092

- Option 2 with a number of cluster of 87 and cluster size of 53 is the design we will choose given that the total sample size of 4,721 is within the allocated budget ($491,550). We prefer this model because it stays close to the desired design effects we received from the customer. Additionally, the standard errors we estimate for this second option overall are the smallest resulting in tighter 95% confidence intervals for the expected estimates we were provided. This design in close to option 3, yet we prefer having a

slightly smaller SSU if we can increase the number of PSUs sampled since this gives us a cost efficiency.

The client has also provided other new information: the estimated size of the target population is N = 830,138. Given this population size and your overall sample size (n) under the new optimum subsample size computed above, what is your overall working sampling fraction (f)? Does it seem like finite population corrections will be necessary in your sampling variances if you choose to perform SRSWOR at some point?

```
# total pop
N <- 830138
# optimum n
total_nm <- 4721
samp_frac <- total_nm / N; samp_frac
```

```
[1] 0.005687006
```

```
MI_school_samples |>  select(Outcome, expect_mean, s_sqrd, roh, m_opt, n_opt, total_nm) |>
  # we can print projected total cost for n=50
  mutate(m_opt = m_opt[2], # optimum m from first row
         n_opt = n_opt[2],
         total_nm = m_opt*n_opt,
         # calculate new deff
         deff_new = 1 + (m_opt-1) * roh,
         # calcualte SRS variance
         var_srs =(1 - samp_frac) * s_sqrd / (total_nm - 1),
         # calculate complex design variance
         var_crs = var_srs * deff_new,
         # compute confidence intervals
         se = sqrt(var_crs),
         lower = expect_mean  - 1.96*se,
         upper = expect_mean + 1.96*se) |>
  select(Outcome, var_srs, var_crs, se, lower, upper) |>
  mutate_at(2:6, round, 5) |>
  kable()
```

Outcome	var_srs	var_crs	se	lower	upper
smoked_cig	0.00004	0.00010	0.01016	0.23009	0.26991
smoked_mj	0.00003	0.00006	0.00746	0.13537	0.16463
age_approached_to_smoke	0.00021	0.00037	0.01921	11.96234	12.03766

Outcome	var_srs	var_crs	se	lower	upper

Our overall sampling fraction is .0057. In examining the SRS and complex design variances, and recalculating the expected standard error and 95% confidence interval, it does not appear that accounting for a population correction makes a huge impact, and we suggest it will not be necessary in our sampling variance for an SRSWOR design.

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.