SURV625 Applied Sampling

2025-02-18

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 build a function that allows us to estimate *roh* from varying design effects, average cluster size, and cluster/element costs to find the optimum element size.

We 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, we can use the average sample cluster size of $m = 7,500/150 = 50$.

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

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

To compute the optimum m size we use the following equation:

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

Finally, we compute the sampling cost as $n \times c_n + n \times m \times c_m$ which we defined these terms above.

```
# function to estimate roh from deff, and optimum m plus sample cost.
optimum_m_sample_fun <- function(deff, m, n, c_n, c_m, C) {

  roh = (deff - 1) / (m - 1)

  m_opt = sqrt( (c_n / c_m) * ( (1-roh)/roh) )


  # is total larger than 500000?
  # m * cost + n * cost
  projected_cost = (c_n * n) + (c_m * n * m_opt) # cost per element

  cat(
    str_c(
      "With a design effect of ", deff, " and an average cluster size of ", m,
      "\nour estimated ICC/roh is ", round(roh, 5),
      "\nand our optimum m is ", round(m_opt),
      "\n\nWe woudl expect our cost to be ", scales::dollar(projected_cost),
      "\nWhich we are left with ", scales::dollar(C-projected_cost),
      " from our total budget.\n\n\n\n")
  )
}
```

We can loop the function through the three given design effects we save in a vector.

```
deff_vector <- c(2.5, 2.0, 1.7)


nm = 7500 # sample total elements
n = 150 # sample number of cluster


# pass function we created above
walk(deff_vector, function(deff_value) {

  optimum_m_sample_fun(
  deff = deff_value,
  m = nm/n, # sample cluster size
  n = n, # sample number of clusters,
  c_n = 3000, # cost per cluster
  c_m = 50, # cost per element within cluster
  C = 500000 # total budget
  )

  }
)
```

```
With a design effect of 2.5 and an average cluster size of 50
our estimated ICC/roh is 0.03061
and our optimum m is 44

We woudl expect our cost to be $776,917
Which we are left with -$276,917 from our total budget.
```

```
With a design effect of 2 and an average cluster size of 50
our estimated ICC/roh is 0.02041
and our optimum m is 54

We woudl expect our cost to be $852,492
Which we are left with -$352,492 from our total budget.
```

```
With a design effect of 1.7 and an average cluster size of 50
our estimated ICC/roh is 0.01429
and our optimum m is 64

We woudl expect our cost to be $932,571
Which we are left with -$432,571 from our total budget.
```

We are clearly above our budget, and if we wanted to optimize cluster m size within our total budget constraint we can assume that we want $m = 64$ cluster size and solve for how many clusters n we would need to sample to stay within our budget. We can do this by calculating $\$500,000 = n(3000 + 64 * 50)$ which we arrange as $\$500,000 = 6200n = 80.6$, and if we round down this estimate of the number of cluster our final cost would be \$n ×c_n + n ×m ×c_m = \$ $80 \times 3000 + 80 \times 64 \times 50 = \$496,000$. Therefore, with a sample design of 80 clusters of size 64 elements each, we would be within our budget constraint while optimizing element size.

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")
```

```
# A tibble: 9 x 2
  Region County_ID
   <dbl> <chr>
1      1 07, 31, 66
2      2 22, 27, 36, 55
3      3 02, 21, 52
4      4 17, 48, 49, 77
5      5 01, 04, 06, 16, 20, 26, 35, 60, 65, 68, 69, 71, 72
6      6 05, 10, 15, 18, 24, 28, 40, 43, 45, 51, 53, 57, 67, 83
7      7 03, 08, 11, 12, 13, 14, 34, 39, 41, 54, 59, 61, 62, 64, 70, 75, 80
8      8 09, 19, 23, 25, 29, 30, 33, 37, 38, 46, 47, 56, 73, 78, 81
9      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_allo = n/sum(n)) |>
  rename(pop_count = n)
```

```
# A tibble: 9 x 3
  Region pop_count prop_allo
   <dbl>     <dbl>     <dbl>
1      1      3561   0.00429
2      2      5474   0.00659
3      3      8631   0.0104
4      4      4855   0.00585
5      5     18907   0.0228
6      6     33133   0.0399
7      7    191992   0.231
8      8    188830   0.227
9      9    374755   0.451
```

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

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.

Desired sample size for variables of interest:

For the first variable ever smoked one cigarette (expected proportion = 0.25), we estimate the

```
# total student population in Michigan
N <- school_frame |> tally(tot_all) |> pull()

# desired coefficient variance
CV <- 0.05
```

The estimated element variance for this proportion is given as:

```
p_hat <- 0.25
```

$$S^2 = \frac{N}{N-1}p(1-p)$$

```
s_2 <- (N / (N-1)) * p_hat*(1-p_hat) ;  s_2
```

```
[1] 0.1875002
```

We can estimate the desired standard error using the desired CV and the estimated proportion for this variable.

```
se <- CV * p_hat ; se
```

```
[1] 0.0125
```

```
se
```

```
[1] 0.0125
```

Our desired variance for this proportion is than:

$$var(\hat{p}) = se(\hat{p})^2, \text{ where } se(\hat{p}) = \sqrt{var(\hat{p})}$$

```
V <- se^2; V
```

```
[1] 0.00015625
```

Now we can compute the desired sample size with a desired CV = .05 as:

```
n_0 <- s_2 / V
n <- n_0 / (1+(n_0/N)) ; round(n)
```

```
[1] 1198
```

For the next variable **ever smoked marijuana** (expected proportion = 0.15), we do the same procedure as before, but in one code chunk as:

```
p_hat <- 0.15
s_2 <- (N / (N-1)) * p_hat*(1-p_hat) ; s_2 # element variance
```

```
[1] 0.1275002
```

```
se <- CV * p_hat
V <- se^2
n_0 <- s_2 / V
n <- n_0 / (1+(n_0/N)) ; round(n)
```

```
[1] 2260
```

In working with a sample mean, the formulas are slightly different. For the third variable, **age when first approach to smoke cigarettes or marijuana** (expected mean = 12, expected SD = 1), we compute the desired sample size for a desired CV=0.05 using the given expected mean and standard deviation as:

```
y_bar <- 12
sd <- 1
s_sqrd <- sd^2

#  critical value
z <- 1.96
# build equation to calculate n_0
n_0 <- s_sqrd / (CV*y_bar)^2
n <- n_0 / ( 1 + (n_0/N) )
round(n)
```

```
[1] 3
```