Exercise ark #4.
Cluster sampling (one-stage, two-stage)

Oğuz–Alper, Melike & Pekarskaya, Tatsiana, Statistics Norway

October 9, 2020

Exercise 0

Discuss:

- What is the difference between stratified and cluster samples?

- What is the difference between one-stage and two-stage sampling?

Exercise 1

Kleppel et al. (2004) report on a study of wetlands in upstate New York. Four wetlands were selected for the study: Two of the wetlands drain watersheds from small towns and the other two drain suburban watersheds. Quantities such as pH were measured at two to four randomly selected sites within each of the four wetlands.

1. Describe why this is a cluster sample. What are the psus? The ssus? How would you estimate the average pH in the suburban wetlands?

2. The authors used Student's two-sample t test to compare the average pH from the sites in the suburban wetlands with the average pH from the sites in the small town wetlands, treating all sites as independent. Is this analysis appropriate? Why, or why not? (Lohr, 2019, p.208)

Exercise 2

A city council of a small city wants to know the proportion of eligible voters that oppose having a incinerator of Phoenix garbage opened just outside of the city limits. They randomly select 100 residential numbers from the city's telephone book that contains 3 000 such numbers. Each selected residence is then called and asked for (a) the total number of eligible voters and (b) the number of voters opposed to the incinerator. A total of 157 voters were surveyed; of these, 23 refused to answer the question. Of the remaining 134 voters, 112 opposed the incinerator, so the council estimates the proportion by

$$\hat{p} = 112/134 = 0.83582$$

with

$$\hat{V}(\hat{p}) = 0.83582(1 - 0.83582)/134 = 0.00102$$

Are these estimates valid? Why, or why not? (Lohr, 2019, pp.207-208)

Exercise 3

(R code available) The new candy Green Globules is being test-marketed in an area of upstate New York. The market research firm decided to sample 6 cities from the 45 cities in the area and then to sample supermarkets within cities, wanting to know the number of cases of Green Globules sold.

City	Number of supermarkets	Number of cases sold
1	52	146, 180, 251, 152, 72, 181, 171, 361, 73, 186
2	19	99, 101, 52, 121
3	37	199, 179, 98, 63, 126, 87, 62
4	39	226, 129, 57, 46, 86, 43, 85, 165
5	8	12, 23
6	14	87, 43, 59

Obtain summary statistics for each cluster. Plot the data, and estimate the total number of cases sold, and the average number sold per supermarket, along with the standard errors of your estimates. (Lohr, 2019, p.209)

Exercise 4

An accounting firm is interested in estimating the error rate in a compliance audit it is conducting. The population contains 828 claims, and the firm audits an SRS of 85 of those claims. In each of the 85 sampled claims, 215 fields are checked for errors. One claim has errors in 4 of the 215 fields, 1 claim has 3 errors, 4 claims have 2 errors, 22 claims have 1 error, and the remaining 57 claims have no errors. (Data courtesy of Fritz Scheuren.)

1. Treating the claims as psus and the observations for each field as ssus, estimate the error rate, defined to be the average number of errors per field, along with the standard error for your estimate.

2. Estimate (with standard error) the total number of errors in the 828 claims.

3. Suppose that instead of taking a cluster sample, the firm had taken an SRS of $85 \times 215 = 18\,275$ fields from the 178 020 fields in the population. If the estimated error rate from the SRS had been the same as in (1), what would the estimated variance $\hat{V}(\hat{p}_{srs})$ be? How does this compare with the estimated variance from (1)? (Lohr, 2019, p.210)

Exercise 5

(R code available) The file measles.dat contains data consistent with that obtained in a survey of parents whose children had not been immunized for measles during a recent campaign to immunize all children between the ages of 11 and 15. During the campaign, 7 633 children from the 46 schools in the area were immunized; 9 962 children whose records showed no previous immunization were not immunized. In a follow-up survey to explore why the children had not been immunized during the campaign, Roberts et al. (1995) sent questionnaires to the parents of a cluster sample of the 9 962 children. Ten schools were randomly selected, then a sample of the m_i nonimmunized children from each school was selected and the parents of those children were sent a questionnaire.

1. Estimate, separately for each school, the percentage of parents who returned a consent form (variable *returnf*). For this exercise, treat the "no answer" responses (value 9) as not returned.

2. Using the number of respondents in school i as m_i, construct the sampling weight for each observation.

3. Estimate the overall percentage of parents who received a consent form along with a 95% CI.

4. How do your estimate and interval in part (3) compare with the results you would have obtained if you had ignored the clustering and analyzed the data as an SRS? Find the ratio:

$$\frac{\text{estimated variance from (3)}}{\text{estimated variance if the data were analyzed as an SRS}}$$

What is the effect of clustering? (Lohr, 2019, p.212)