Exercise ark #2.
Stratified sampling

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

October 9, 2020

Exercise 0

1. Explain in your own words what is stratified sample?
2. What can be reasons to use a stratified sample rather than SRS?

Exercise 1

Consider a population of 6 students. Suppose we know the test scores of the students to be

Student	1	2	3	4	5	6
Score	66	59	70	83	82	71

1. Find the mean $\bar{y}_U$ and variance S^2 of the population.
2. How many SRS's of size 4 are possible?
3. List the possible SRS's. For each, find the sample mean. Find $V(\bar{y})$, using
$$V(\bar{y}) = (1 - n/N)n^{-1}S^2.$$
4. Now let stratum one consist of students 1–3, and stratum two consist of students 4–6. How many stratified random samples of size 4 are possible in which 2 students are selected from each stratum?
5. List the possible stratified random samples. Which of the samples from (3) cannot occur with the stratified design defined in (4)?
6. Find $\bar{y}_{str}$ for each possible stratified random sample. Find $V(\bar{y}_{str})$, and compare it to $V(\bar{y})$. (Lohr, 2019, p.102)

Exercise 2

A stratified sample is being designed to estimate the prevalence p of a rare characteristic, say the proportion of residents in Milwaukee, Wisconsin, who have Lyme disease. Stratum one, with N_1 units, has a high prevalence of the characteristic; stratum two, with N_2 units, has low prevalence. Assume that the cost to sample a unit (for example, the cost to select a person for the sample and determine whether he or she has Lyme disease) is the same for each stratum, and that at most 2 000 units are to be sampled.

1. Let p_1 and p_2 be the proportions in stratum one and stratum two with the rare characteristic. If $p_1 = 0.10$, $p_2 = 0.03$, and $N_1/N = 0.4$, what are n_1 and n_2 under optimal allocation?

2. If $p_1 = 0.10$, $p_2 = 0.03$, and $N_1/N = 0.4$, what is $V(\hat{p}_{str})$ under proportional allocation? Under optimal allocation? What is the variance if you take an SRS of 2 000 units from the population? (Lohr, 2019, p.110)

Exercise 3

(R code available) Consider the same problem as in exercise 4 previous section. In the SRS of 50 faculty members not all the departments were represented. The SRS contained several faculty members from psychology and from chemistry but none from the foreign languages. It was therefore decided to take a stratified simple random sample, using the areas biological sciences, physical sciences, social sciences and humanities as the strata. Proportional allocation was used in this sample. The distribution of the strata for population and sample are given below

Stratum	Number of faculty members in the stratum	Number of faculty members in the samples
1 - Biological sciences	102	7
2 - Physical sciences	310	19
3 - Social sciences	217	13
4 - Humanities	178	11
Total	807	50

The data from the stratified sample turned out to be

Number of faculty members	Number of refereed publications Biological	Physical	Social	Humanities
0	1	10	9	8
1	2	2	0	2
2	0	0	1	0
3	1	1	0	1
4	0	2	2	0
5	2	1	0	0
6	0	1	1	0
7	1	0	0	0
8	0	2	0	0

1. Estimate the total number of refereed publications by faculty members in the college and compute the standard error.

2. How does the result from part (3) compare with the result from the SRS in Exercise 4 previous session?

3. Estimate the proportion of faculty with no refereed publications and compute the standard error and 95% confidence interval.

4. Compare the result in part (3) with the result from the SRS in Exercise 4 previous session.

5. Did the stratification increase precision for the two estimation problems considered in parts (1) and (3) ? Explain why you think it did or did not. (Lohr 2019 pp 103-104)

Exercise 4

A public opinion researcher has a budget of \$20 000 for taking a survey. She knows that 90% of all households have known telephone numbers. Telephone interviews cost \$10 per household; in-person interviews cost \$30 each if all interviews are conducted in person and \$40 each if only households with unknown telephone numbers are interviewed in person (because there will be extra travel costs). Assume that the variances in phone and nonphone strata are similar and that the fixed costs are $c_0 = \$5\,000$.

1. How many households should be interviewed in each group if all households are interviewed in person ?
2. How many households should be interviewed in each stratum if households with a known telephone number are contacted by telephone and the other households are contacted in person?
3. Which of the two data collection methods would you choose? Give a justification for your answer. Lohr, 2019, p.104

Exercise 5

Burnard (1992) sent a questionnaire to a stratified sample of nursing tutors and students in Wales, to study what the tutors and students understood by the term *experiential learning*. The population size and sample size obtained for each of the four strata are given below:

Stratum	Population size	Sample size
General nursing tutors (GT)	150	109
Psychiatric nursing tutors (PT)	34	26
General nursing students (GS)	2 680	222
Psychiatric nursing students (PS)	570	40
Total	3 434	397

Respondents were asked which of the following techniques could be identified as experiential learning methods; the number of students in each group who identified the method as an experiential learning method are given below:

Method	GS	PS	PT	GT
Role play	213	38	26	104
Problem solving activities	182	33	22	95
Simulations	95	20	22	64
Empathy-building exercises	89	25	20	54
Gestalt exercises	24	4	5	12

Estimate the overall percentage of nursing students and tutors who identify "Role play" techniques as "experiential learning". Be sure to give standard errors for your estimate. (Lohr, 2019, p.106)