Exercise ark #1.
Probability sampling. Simple random sampling

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

October 14, 2020

Exercise 0

1. Explain in your own words the reasons why it may be an advantage to select a sample instead of doing a census

2. Discuss the properties of a good sample

Exercise 1

For each survey further, describe the target population, sampling frame, sampling unit and observation unit. Discuss any possible sources of selection bias or inaccuracy of responses:

1. Potential jurors in some jurisdictions are chosen from a list of county residents who are registered voters or licensed drivers over age 18. In the fourth quarter of 1994, 100 300 jury summons were mailed to Maricopa County, Arizona, residents. Approximately 23 000 of those were returned from the post office as undeliverable. Approximately 7 000 persons were unqualified for service because they were not citizens, were under 18, were convicted felons, or other reason that disqualified them from serving on a jury. An additional 22 000 were excused from jury service because of illness, financial hardship, military service, or other acceptable reason. The final sample consists of persons who appear for jury duty; some unexcused jurors fail to appear. (Lohr, 2019, p.20)

2. A survey is conducted to find the average weight of cows in a region. A list of all farms is available for the region, and 50 farms are selected at random. Then the weight of each cow at the 50 selected farms recorded. (Lohr, 2019, p.20)

3. Kripke et al. (2002) claim that persons who sleep 8 or more hours per night have a higher mortality risk than persons who sleep 6 or 7 hours. They analyzed data from the 1982 Cancer Prevention Study II of the American Cancer Society, a national survey taken by about 1.1 million people. The survival or date of death was determined for about 98% of the sample six years later. Most of the respondents were friends and relatives of American Cancer Society volunteers; the purpose of the original survey was to explore factors associated with the development of cancer, but the survey also contained a few questions about sleep and insomnia. (Lohr, 2019, p.21)

Exercise 2

Let $N = 6$ and $n = 3$. For purposes of studying sampling distributions, assume that all population values are known.

$$y_1 = 98 \quad y_2 = 102 \quad y_3 = 154$$
$$y_4 = 133 \quad y_5 = 190 \quad y_6 = 175$$

We are interested in $\bar{y}_U$, the population mean. Two sampling plans are proposed.
• Plan 1. Eight possible samples may be chosen.

Sample number	Sample, S	P(S)
1	{1,3,5}	1/8
2	{1,3,6}	1/8
3	{1,4,5}	1/8
4	{1,4,6}	1/8
5	{2,3,5}	1/8
6	{2,3,6}	1/8
7	{2,4,5}	1/8
8	{2,4,6}	1/8

• Plan 2. Three possible samples may be chosen.

Sample number	Sample, S	P(S)
1	{1,4,6}	1/4
2	{2,3,6}	1/2
3	{1,3,5}	1/4

1. What is the value of $\bar{y}_U$?
2. Let $\bar{y}$ be the mean of the sample values. For each sampling plan, find (i) $E(\bar{y})$; (ii) $V(\bar{y})$; (iii) Bias$(\bar{y})$; (iv) MSE$(\bar{y})$.
3. Which sampling plan do you think is better? Why? (Lohr, 2019, p.61)

Exercise 3

(R code available) Let U be a population of size $N = 8$ with the index set $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$. The values of y_i are

i	1	2	3	4	5	6	7	8
y_i	1	2	4	4	7	7	7	8

For this population, consider sampling scheme

S	$P(S)$
{1,3,5,6}	1/8
{2,3,7,8}	1/4
{1,4,6,8}	1/8
{2,4,6,8}	3/8
{4,5,7,8}	1/8

1. Find the probability of selection π_i for each unit i.
2. What is the sampling distribution of $\hat{t} = 8\bar{y}$?
3. Assume a SRS of size 3 without replacement and find expectation and variance of $\bar{y}$
4. Assume a SRS of size 3 with replacement and find expectation and variance of $\bar{y}$
5. For 3. and 4., draw the histogram of the sampling distribution of $\bar{y}$. Which sampling distribution has the smaller variance, and why? (Lohr, 2019, p.61)

Exercise 4

(R code available) A university has 807 faculty members. For each faculty member, the number of refereed publications was recorded. This number is not directly available on the database, so requires the investigator to examine each record separately. A frequency table for number of refereed publications is given below for an SRS of 50 faculty members.

Refereed Publications	0	1	2	3	4	5	6	7	8	9	10
Faculty Members	28	4	3	4	4	2	1	0	2	1	1

1. Plot the data using a histogram. Describe the shape of the data.
2. Estimate the mean number of publications per faculty member, and give the SE for your estimate.
3. Do you think that $\bar{y}$ from 2. will be approximately normally distributed? Why or why not?
4. Estimate the proportion of faculty members with no publications and give a 95% confidence interval. (Lohr, 2019, pp.62-63)

Exercise 5

A typical opinion poll surveys about 1000 adults. Suppose that the sampling frame contains 100 million adults including yourself, and that an SRS of 1000 adults is chosen from the frame.

1. What is the probability that you are selected to be in the sample?
2. Now suppose that 2000 such samples are selected, each sample selected independently of the others. What is the probability that you will not be in any of the samples?
3. How many samples must be selected for you to have a 0.5 probability of being in at least one sample? (Lohr, 2019, p.68)

Exercise 6

It is important to keep track on unemployment level from month to month. Let's assume that you have been assigned a task to find a size of SRS n. You have no idea about the proportion of unemployed people in the population, but you read in literature that p = 0.5 is often used when you have no idea about the value of proportion. So you decide to use as p = 0.5. Assume that population size is large and, thus, you can ignore fpc.

- What is the value of n if your margins of error is c 0.001, 0.002, 0.005?
- Why proportion of interest p = 0.5 is often used when you have no idea about the value of the proportion?