STAT3003 Problem Sheet 3

ZHENG Weijia (William, 1155124322)

April 11, 2021

1 Q1

1. (Adapted from Exercise 7.3 of Scheaffer et al. (2012)) A retail store with four departments has charge accounts arranged by department, with past-due accounts at the top of each departmental list. Suppose the departments average around ten accounts each, with approximately 40% past due. On a given day the accounts might appear as shown in the accompanying table (with account numbers 1 through 40). The store wishes to estimate the proportion of past-due accounts by systematic sampling.

Account numbers	1-11	12-20	21-28	29-40
Delinquent accounts	1,2,3,4	12, 13, 14	21, 22, 23, 24, 25	30, 31, 32, 33

Table 1: Retail Store Accounts Study Data

   (a) List all possible 1-in-10 systematic samples and compute the exact variance of the sample proportion.

   (b) List all possible 1-in-5 systematic samples and compute the exact variance of the sample proportion.

   (c) Compare the result in part (a) with an approximate variance obtained in a SRS sample of size $n = 4$ from this population. Similarly, compare the result in part (b) with that obtained from a SRS with $n = 8$. Can you explain what is going on?

1.1 (a)

We need to fix the starting point, when start at 1: the number of deliquent accounts is 3, the sample proportion is $\frac{3}{4}$.

When start at 2: the number of deliquent accounts is 4. The sample proportion is 1.

When start at 3: the number of deliquent accounts is 4. The sample proportion is 1.

When start at 4: the number of deliquent accounts is 3. The sample proportion is $\frac{3}{4}$.

When start at 5: the number of deliquent accounts is 1. The sample proportion is $\frac{1}{4}$.

When start at 6: the number of deliquent accounts is 0. The sample proportion is 0.

When start at 7: the number of deliquent accounts is 0. The sample proportion is 0.
When start at 8: the number of deliquent accounts is 0. The sample proportion is 0.
When start at 9: the number of deliquent accounts is 0. The sample proportion is 0.
When start at 10: the number of deliquent accounts is 1. The sample proportion is $\frac{1}{4}$.
Therefore the exact variance of sample proportion is 0.165.

1.2 (b)

When start at 1: the number of deliquent accounts is 3. The sample proportion is 0.375.
When start at 2: the number of deliquent accounts is 4. The sample proportion is 0.5.
When start at 3: the number of deliquent accounts is 4. The sample proportion is 0.5.
When start at 4: the number of deliquent accounts is 3. The sample proportion is 0.375.
When start at 5: the number of deliquent accounts is 2. The sample proportion is 0.25.
Hence the exact variance of sample proportion is 8.75×10^{-3}.

1.3 (c)

Using the formulae

$$Var(\hat{p}) = \frac{N-n}{N-1}p(1-p).$$

And we have $p = 0.4, N = 40$, therefore $Var(\hat{p}) = 0.0554$.
For $n = 8$, $Var(\hat{p}) = 0.0246$.
When n=4, SRS's variance is smaller and when n=8, SRS's is larger.
Because when $n = 4$, the ICC (intracluster coorelation coefficient) is tend to be larger, because ICC represents how things are similar inside a cluster. Therefore $Var(\hat{p})$ tends to be underestimating.
And when $n = 8$, the ICC is smaller, then $Var(\hat{p})$ tends to be overestimating.

2 Q2

2. (Adapted from Exercise 7.19 of Scheaffer et al. (2012)) A farmer wishes to estimate the total weight of fruit to be produced in a field of pumpkins by sampling just prior to harvest. The plot consists 20 rows with 400 plants per row. The manufacturer of the seeds says that each plant can yield up to 10 pounds of fruit. Outline an appropriate systematic sampling plan for this problem so as to estimate the total weight of fruit to within 2000 pounds.

Note that we can regard every plant's condition is the same. Therefore we can apply the formulae from SRS:

$$n = \frac{1}{\frac{1}{N} + \frac{d^2}{N^2 v^2 z^2_{1-\frac{\alpha}{2}}}(1 - \frac{1}{N})}.$$

The v is the "pre-sample estimate". We calculate it by Range Rule of Thumb, which is we calculate the variance by range divide by 4: $v = (10 - 0)/4 = 2.5$

And we have $N = 20 \times 400 = 8000$, $d = 2000$, $z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$.

Hence we can calculate $n = 366.6016$. Therefore we take $n = 367$.

And by $8000/366.6 = 21.8221$. So take k (the stepsize) be 21.

To summarize, we should perform a 1-in-21 systematic sampling.

3 Q3

3. (Adapted from Exercise 8.6 of Scheaffer et al. (2012)) A political scientist developed a test designed to measure the degree of awareness of current events. She wants to estimate the average score that would be achieved on this test by all students in a certain high school. The administration at the school will not allow the experimenter to randomly select students out of classes in session, but it will allow her to interrupt a small number of classes for the purpose of giving the test to every member of the class. Thus, the experimenter selects 25 classes at random from the 108 classes in session at a particular hour. In total there are 3500 students in class at this hour. The test is given to each member of the sampled classes, with results as shown in Table 2. Use a biased estimator to estimate the average score that would be achieved on this test by students in the school and provide a 95% confidence interval. Compare your answer to that from Question 3 from Problem Sheet 2, which used an unbiased estimator. Which answer do you like better?

Class	# students	Total Score	Class	# students	Total Score
1	31	1590	14	40	1980
2	29	1510	15	38	1990
3	25	1490	16	28	1420
4	35	1610	17	17	900
5	15	800	18	22	1080
6	31	1720	19	41	2010
7	22	1310	20	32	1740
8	27	1427	21	35	1750
9	25	1290	22	19	890
10	19	860	23	29	1470
11	30	1620	24	18	910
12	18	710	25	31	1740
13	21	1140			

Table 2: Current Event Test Study Data

For sure we will apply the "using auxiliary data" method.
Define X_i is the number of students in sampled class i. (E.g. $X_1 = 31$.)
Define Y_i is the total score of sampled class i. (E.g. $Y_1 = 1590$.)
Hence $\bar{Y} = \frac{1}{25}\sum_{i=1}^{25} Y_i = 1398.28$. And $\bar{X} = \frac{1}{25}\sum_{i=1}^{25} X_i = 27.12$. Then

$$\hat{R} = \frac{\bar{Y}}{\bar{X}} = 51.5590.$$

Hence $\hat{\mu} = 51.5590$.
Then using the formulae $\widehat{Var(\hat{\mu})} = \frac{N}{M^2}(N-n)\frac{1}{n}\hat{\sigma_r^2}$. Where $\hat{\sigma_r^2} = 10808.7251$.
We can gain that $\widehat{Var(\hat{\mu})} = 0.3164$.
Therefore, an 95% C.I. can be $(51.5590 \pm t_{24,0.975}\sqrt{0.3164}\ ) = (50.3981, 52.7199)$.
Recall that the result from Problem Set 2 is $(43, 147 \pm 4.3081)$.
I prefer the biased approach, because this has a much smaller variance.

4 Q4

(Adapted from Exercises 9.14 and 9.15 of Scheaffer et al. (2012)) Suppose (another) sociologist wants to estimate the total number of retired people residing in a certain city. She decides to sample blocks and then sample households within blocks. (Block statistics from the Census Bureau aid in determining the number of households in each block.) Four blocks are randomly selected (by SRS) from the 300 of the city. From the data in Table 3, estimate the average number of retired residents per household in the city using a ratio estimator and provide an estimate for the standard error.

Block	# households	# households samples	# retired residents per household
1	18	3	1, 0, 2
2	12	3	0, 3, 0
3	9	3	1, 1, 2
4	14	3	0, 1, 1

Table 3: Retired People Study Data

Make it clear that we want to estimate μ the average number of retired residents per household.
Before beginning, we need to make sure that $n = 4, N = 300$.
We are in the sampling method of "ratio estimation and two-stage cluster sampling."
And adopting an point estimate of μ is

$$\hat{\mu} = \frac{\sum_{i=1}^{n} M_i \bar{Y}_i}{\sum_{i=1}^{n} M_i} = \frac{18 * 1 + 12 * 1 + 9 * (4/3) + 14 * (2/3)}{18 + 12 + 9 + 14} = 0.9686.$$

And we then need to determine

$$\widehat{Var(\hat{\mu})} = \frac{1}{M^2} N(N-n) \frac{1}{n} \hat{\sigma_r^2} + \frac{1}{M^2} \frac{N}{n} \sum_{i=1}^{n} M_i (M_i - m_i) \frac{1}{m_i} \hat{\sigma_i}^2.$$

And $\hat{\sigma_r}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{Y_i} - M_i \hat{\mu_r}) = 9.7017$.
And $\hat{\sigma_i}^2$'s are variance inside each cluster i. Hence the 4 values are 1, 3, 0.3333, 0.3333.
And we do not have M, so we need to estimate $\hat{M} = \frac{N}{n} \sum_{i=1}^{n} M_i = 3975$.
Hence $\widehat{Var(\hat{\mu})} = 0.0136 + 1.0495 \times 10^3 = 0.01465 = 0.1210^2$.
Therefore the standard error is 0.1210.

5 Q5

(Adapted from Exercise 2.21 of Scheaffer et al. (2012)) Readers of the magazine *Popular Science* (August 1990) were asked to phone in their responses to the following question: "Should the U.S. build more fossil-fuel generating plants or the new so-called safe nuclear generators to meet the energy crisis of the 90s?". Of the total call-ins, 86% chose the nuclear option. What do you think about the way the poll was conducted? What do you think about the way the question was worded? Do you the results are a good estimate of the prevailing mood of the country?

Note that the survey is conducted in a way that without any sampling. Hence they just let readers to state their response voluntarily.

And in the wording of the question, the "so-called safe" is not needed. This implies that the question is not balanced.

Hence I don't think the results are a good estimate of prevailing mood of the country. If one wants to get good estimate of prevailing mood of the country, one needs to sample the country by adopting some proper sampling method.

6 Q6

(Adapted from Exercise 2.7 of Scheaffer et al. (2012)) Discuss the relative merits of using personal interviews, telephone interviews, and mailed questionnaires as methods of data collection for each of the following situations:

(a) A television executive wants to estimate the proportion of viewers in the country who are watching her network at a certain hour.

(b) A newspaper editor wants to survey the attitudes of the public toward the type of news coverage offered by his paper.

(c) A district councillor is interested in determining how homeowners feel about a proposed zoning change (from "industrial" to "residential", say).

(d) The Department of Health wants to estimate the proportion of dogs that have had vaccinations for rabies within the last year.

6.1 (a)

Telephone interview is a feasible way to conduct the survey. Because the pool to be sampled is the whole country, and the survey needs to be conducted soon after the TV programme starts.

6.2 (b)

Suppose the "public" means the readers (subscriber) of his paper, then since the paper company already has the frame of those readers, then mailed questionnaires with small reward is good.

6.3 (c)

I think the councillor should have a frame, since he can adopt the mailed questionnaires/ personal interview method. Since the problem seems like cannot be discussed through a simple telephone call.

6.4 (d)

Personal interview. Because in this situation, the survey is officially conducted, I suppose the Department of Health has a frame of registration institutes of pet hospitals, etc..

So there are many certificates or documents (e.g. the registration documents of dogs, etc.) needed to be checked. In this sense, telephone and mailed questionnaires are not that proper.

If the Department can assure the document and certificates problem remotely, then telephone interviews /mailed questionnaires are also feasible.

7 Q7

(Adapted from Exercise 2.10 of Scheaffer et al. (2012)) Give an example of a question that could force a response in a certain direction because of its strong wording. *Note: please do not use an example already given in the notes or exercises or previous solutions...*

"Do you for or against the execution of Lantau Tomorrow Plan of Hong Kong government?" comparing with "Do you for or against the execution of Lantau Tomorrow Plan at any cost?"

8 Q8

(Adapted from Exercise 2.31 of Scheaffer et al. (2012)) Balance of questions makes logical sense but does not always have a strong impact on the results. Its impact is increased by the strength of a counterargument. The following two comparisons were drawn in one study reported by Schuman and Presser (1996):

> A1: If there is a serious fuel shortage this winter, do you think there should be a law requiring people to lower the heat in their homes?
>
> B1: If there is a serious fuel shortage this winter, do you think there should be a law requiring people to lower the heat in their homes, or do you oppose such a law?

Another study in a later year compared these forms:

> A2: If there is a serious fuel shortage this winter, do you think there should be a law requiring people to lower the heat in their homes, or do you oppose such a law? (Same as B1 form).
>
> B2: If there is a serious fuel shortage this winter, do you think there should be a law requiring people to lower the heat in their homes, or do you oppose such a law because it would be too difficult to enforce?

Within each pair of questions, which one do you think received the lower percentage of responses favouring the law? Explain your reasoning.

In the A1B1 pair, the B1.

Because in A1B1 pair, because B1 reminds the respondents of a negative attitude. Hence B1 is expected to have a lower percentage of favouring the law. And A1 is not as balanced as B1.

In the A2B2 pair, I cannot do a confident guess.

It is proper to say A2 will have a lower percentage, because B2 only include one possibility of opposing the law, while there are many other reasons to object the law.

But it is also reasonable to expect that B2 will have a lower percentage. Because A2 does not give the respondents a concrete and vivid situation (argument), or the feeling of difficuility of enforcing the law. While the B2 gives a vivid situation of hindering the law's enforcing.

Hence I cannot make a confident guess.