A survey to estimate the proportion of Southeast Sulawesi elementary school students reading fluently

Gusti N.A. Wibawa1*, Wayan Somayasa1, Irma Yahya1, Ahid Hidayat2 and Bahriddin Abapihi1

1 Department of Mathematics, Halu Oleo University, Kendari, Indonesia
2 Department of Elementary School Teacher’s Education, Halu Oleo University

gnawibawa@gmail.com (corresponding author)

Abstract. We conducted a survey to estimate the proportion of elementary school students in Southeast Sulawesi who have been able to read fluently while sitting at second grade. The information is needed to contribute to the policy towards enhancement of elementary student ability to read in their early stage of schools. A two-stage cluster sampling method had been applied in order to maintain the representativeness of selected samples over the whole population. At the first stage, we randomly chose the districts and, then, randomly selected the sub-districts. Finally, all elementary schools in each selected sub-district were taken as well as all their second grade students as the sampling units. Doing these, we got 230 respondents randomly selected from 78 elementary schools of 20 sub-districts in all 6 selected districts or regencies. Based on the selected samples, it is approximately only 32% of students were able to read fluently, while the rest of them were unable or less ability to read.

1. Introduction
Southeast Sulawesi is among the five provinces in Indonesia that the development of its education program is grouped into the category of inadequate, especially at primary education level [3]. This is also in accordance with the results of the Asesmen Kompetensi Siswa Indonesia (AKSI, Assessment of Indonesian Student Competencies) in 2016 that 63 – 73% of fourth grade elementary school students in Southeast Sulawesi have poor reading skills [4]. In addition, EGRA (Early Grade Reading Assessment) in 2014 conducted a study on 4,812 second-grade students throughout Indonesia by assessing their abilities to read and to understand reading. The study showed that only about 50% of students can read fluently and understand reading. It also found that only one third of students in Kalimantan and Sulawesi to have the ability to read fluently and to understand reading[5]. These results need to be seriously paid attention because the students who are not able to read in the early grades, especially up to the second grade of elementary school, are more likely to repeat classes and eventually drop out of school. If there is no intervention, then the gap between students who can and cannot read will definitely increase over time. Consequently, a study with a proper sampling method should be conducted in order to precisely obtain better estimation of the proportion. This information can be used by stakeholders in their policies to tackle such a problem.

Some investigations have been conducted on elementary student abilities to read such as in [3], [4], and [5]. However, most of them just simply applied simple random sampling technique as tool of their surveys. Some others conducted sample selection without knowing the sampling frames, and sometimes the samplings were not fully randomized. This paper aims to estimate the proportion or percentages of elementary school second grade students who have been reading fluently based on samples obtained using the two-stage cluster sampling method. By applying this method, it is expected
that the sampling would have good coverage and high precision, especially when the characteristics of districts and sub-districts as clusters may vary from a cluster to another [2].

2. Literature Review

2.1. Sampling

Sampling technique is purposed to draw sample efficiently in term of its cost and representativeness. By applying a proper sampling theory, one can get the samples at the lowest cost in budget, time and personnel. It also brings statistical precision to parameter estimation of the population [6]. A sampling technique is considered efficient if it requires the smallest cost in data collection and provides representative sample to be used in estimation of population parameters [7]. A good sample produced by good sampling technique will provide representative conclusion on population.

In many statistical textbooks, we found some classical sampling techniques that can be used in conducting a survey or drawing samples from a population. Generally, there are two categories of sampling techniques, namely probability and non-probability samplings. In probability sampling, all the members of population have the same opportunity to be chosen as a sample. In this case, all elements of population are usually known and can be listed for random selection. Contrarily, non-probability sampling selects sample without sampling frame or list of population elements. Consequently, the members of population have unequal opportunity to be selected and, even, some members have no chance to be drawn as a sample.

2.2. Two-Stage Cluster Sampling

This method is the development of the classical cluster sampling method. In this technique, it is considered that within a cluster there are also some clusters. As a result, we select randomly some clusters in the first stage and, then within these clusters in the second stage, we do the same by randomly drawing clusters again. Lastly, the sampling units in the second stage clusters will be drawn all or only some members by random selection [2].

After conducting sample selection using two-stage cluster sampling, we can estimate the parameters of the population. In this paper, we will only be discussing about one parameter, population proportion.

Let \( \hat{p}_i \) denotes the proportion of sampled elements from cluster \( i \) that fall into the category of interest. Estimator of a population proportion \( p \) is

\[
\hat{p} = \frac{\sum_{i=1}^{n} M_i \hat{p}_i}{\sum_{i=1}^{n} M_i}
\]

and estimated variance of \( p \) is

\[
\hat{V}(\hat{p}) = \left( \frac{N-n}{N} \right) \left( \frac{1}{nM^2} \right) S_r^2 + \frac{1}{nN M^2} \sum_{i=1}^{n} M_i \left( \frac{M_i - m_i}{M_i} \right) \left( \frac{\hat{p}_i(1-\hat{p}_i)}{m_i - 1} \right)
\]

where

\[
S_r^2 = \frac{\sum_{i=1}^{n} M_i^2 (\hat{p}_i - \hat{p})^2}{n-1} = \frac{\sum_{i=1}^{n} (M_i \hat{p}_i - M_i \hat{p})^2}{n-1}.
\]

and the 95% confidence interval estimation of the population proportion is

\[
\hat{p} \pm 1.96 \sqrt{\hat{V}(\hat{p})}
\]

with
N  = number of clusters in the population
n  = number of cluster samples
M  = number of elements in the population
\( \bar{M} = \frac{M}{N} \) = cluster size average of the population
\( M_i \) = number of elements in \( i \)-th cluster
\( m_i \) = number of samples drawn from \( i \)-th cluster.

3. Research Methodology

We conducted a survey based on the scheme of two-stage cluster sampling. There are 17 districts in Southeast Sulawesi. We considered these districts as main clusters and randomly chose six clusters (districts). Within these districts, there are some sub-districts considered as sub-clusters. The numbers of sub-districts vary from one district to another. Then within a district, some sub-districts were randomly selected and all elementary schools in a selected sub-district were selected for the survey. We got 78 schools in total for all selected sub-district over six selected districts in Southeast Sulawesi. The distributions of samples are presented in Table 1.

| Districts      | Population #Sub-district | Population # Elementary school | Sample #Sub-district | Sample # Elementary school |
|----------------|--------------------------|--------------------------------|----------------------|---------------------------|
| North Buton    | 6                        | 75                             | 3                    | 10                        |
| East Kolaka    | 12                       | 140                            | 3                    | 12                        |
| North Kolaka   | 15                       | 109                            | 3                    | 12                        |
| Konawe         | 29                       | 275                            | 5                    | 20                        |
| North Konawe   | 13                       | 102                            | 3                    | 12                        |
| West Muna      | 11                       | 95                             | 3                    | 12                        |
| **Total**      | **86**                   | **796**                        | **20**               | **78**                    |

In each school, we randomly selected some students of second grade to participate in our survey. Here, the schools are considered as the sampling units, while the students are as subunits. We chose 2 to 3 students in each school. It resulted in 230 respondents of second grade elementary school students in total.

4. Result and Discussion

After the data collection, we calculated some statistics based on the samples and the formulas as presented in Section 2. The results are shown in Table 2.
Table 2. Statistic summaries for the data obtained from two-stage cluster sampling survey

| Sub-district | M_i | m_i | \(\hat{p}_i\) | M_i\(\hat{p}_i\) | M_i\(\hat{p}_i\)-M_i\(\hat{p}\) |
|--------------|-----|-----|------------|----------------|-------------------------|
| 1            | 81  | 9   | 0.33       | 27.00          | 1.11                    |
| 2            | 39  | 9   | 0.11       | 4.33           | -8.13                   |
| 3            | 84  | 12  | 0.33       | 28.00          | 1.15                    |
| 4            | 51  | 12  | 0.50       | 25.50          | 9.20                    |
| 5            | 62  | 12  | 0.42       | 25.83          | 6.02                    |
| 6            | 85  | 12  | 0.33       | 28.33          | 1.16                    |
| 7            | 92  | 12  | 0.42       | 38.33          | 8.93                    |
| 8            | 87  | 12  | 0.50       | 43.50          | 15.69                   |
| 9            | 96  | 12  | 0.25       | 24.00          | -6.68                   |
| 10           | 78  | 12  | 0.33       | 26.00          | 1.07                    |
| 11           | 79  | 12  | 0.25       | 19.75          | -5.50                   |
| 12           | 57  | 12  | 0.58       | 33.25          | 15.03                   |
| 13           | 82  | 12  | 0.33       | 27.33          | 1.12                    |
| 14           | 82  | 12  | 0.42       | 34.17          | 7.96                    |
| 15           | 45  | 15  | 0.13       | 6.00           | -8.38                   |
| 16           | 27  | 9   | 0.33       | 9.00           | 0.37                    |
| 17           | 71  | 8   | 0.13       | 8.88           | -13.82                  |
| 18           | 63  | 12  | 0.08       | 5.25           | -14.89                  |
| 19           | 55  | 12  | 0.17       | 9.17           | -8.41                   |
| 20           | 43  | 12  | 0.25       | 10.75          | -2.99                   |
| Total        | 1359| 230 | 434.38     |                |                         |

\[ \hat{p} = \frac{\sum_{i=1}^{20} M_i \hat{p}_i}{\sum_{i=1}^{20} M_i} = \frac{434.38}{1359} = 0.32 \]

We calculated the estimate of variance of \(\hat{p}\), as follows:

\[ S_r^2 = \frac{\sum_{i=1}^{20} M_i^2 (\hat{p}_i - \hat{p})^2}{n - 1} = 76.11 \]

and

\[ \sum_{i=1}^{n} M_i^2 \left( \frac{M_i - m_i}{M_i} \right) \left( \frac{\hat{p}_i (1 - \hat{p}_i)}{m_i - 1} \right) = 1503.74 \]

So the variance of \(\hat{p}\) with \(\bar{M}\) is estimated by the sample mean, 67.95, is

\[ \hat{V} (\hat{p}) = \frac{N - n}{N} \left( \frac{1}{nM^2} \right) S_r^2 + \frac{1}{nNM^2} \sum_{i=1}^{n} M_i^2 \left( \frac{M_i - m_i}{M_i} \right) \left( \frac{\hat{p}_i (1 - \hat{p}_i)}{m_i - 1} \right) \]

\[ \hat{V} (\hat{p}) = \left( \frac{86 - 20}{86} \right) \left( \frac{1}{20(67.95)^2} \right) 76.11 + \frac{1}{(20)(86)(67.95)^2} 1503.74 \]

\[ \hat{V} (\hat{p}) = 0.00082 \]
The estimation of the proportion of the students who are able to read fluently is

\[ \hat{p} = 0.32 \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.32 \pm 1.96 \sqrt{0.00082} = 0.32 \pm 0.056. \]

We estimated that the proportion of the students reading fluently is 0.32, with a bound of 0.056 on the error of estimation. In addition, the estimate of proportion confidence interval of the students reading fluently lies between 0.263 and 0.376.

5. Conclusion
We have demonstrated how to estimate the proportion based on samples drawn from a survey using two-stage cluster sampling method. Our estimate showed that most of second grade elementary school students in Southeast Sulawesi, Indonesia, were unable or less ability to read with the proportion of 0.68 or 68%. It was a small proportion of 0.32 or only about 32% of them are reading fluently with the estimate of confident interval of proportion between 0.263 and 0.376. This result is confirming the study of EGRA in 2014, and vice versa.

References
[1] Aliaga A and R. Ren. 2006. Optimal Sample Sizes for Two-stage Cluster Sampling in Demographic and Health Surveys. https://dhsprogram.com/pubs/pdf/WP30/WP30.pdf. [14 September 2018]
[2] Scheaffer, R.L, W. Mendenhall, and R.L. Ott. 2006. Elementary Survey Sampling. Sixth edition. (Canada: Thomson Brooks/Cole) p 312
[3] Kemendikbud. 2016a. Keberhasilan Program Pembangunan Pendidikan Dasar Dan Menengah Tahun 2015/2016. Jakarta: Pusat Data dan Statistik Pendidikan dan Kebudayaan (PDSPK). http://publikasi.data.kemdikbud.go.id/uploadDir/isi_C3999016-EE40-429C-B4A6-CBA416DC3776_.pdf. [15 Juni 2017]
[4] Kemendikbud. 2016b. Asesmen Kompetensi siswa Indonesia (AKSI) http://puspendik.kemdikbud.go.id/inap-sd. [12 Juni 2017]
[5] USAID. 2014. Indonesia 2014: The National Early Grade Reading Assessment (EGRA) and Snapshot of School Management Effectiveness (SSME) Survey.
[6] Cochran, W.G. 2001. Teknik Penarikan Sampel. Alih bahasa oleh Rudiansyah (Jakarta: Universitas Indonesia)
[7] Levy P.S. and S. Lemeshow. 1999. Sampling of Populations (New York: A Wiley Latherscience Publication)
[8] Scheaffer, R.L, W. Mendenhall, and R.L. Ott. 2006. Elementary Survey Sampling. Sixth edition. (Canada: Thomson Brooks/Cole) p 312