Verification on number of first sample in ALIS

Shoji KIMURA
ASEAN Food Security Information System, email: wood_v@yahoo.co.jp

-Based on the result of agricultural land area sample survey in Philippines, Nueva Ecija province-

1. Preface

After the operation of making area framework, ALIS operator extracts the first sample. This operation itself is very easy operation with only input number of target sample to ALIS. However, it needs to pay attention before this operation that the operator has to conduct area borderline attaching operation to the extracted all first samples. This operation is not easy. Therefore, we have to consider about proper number of first sample.

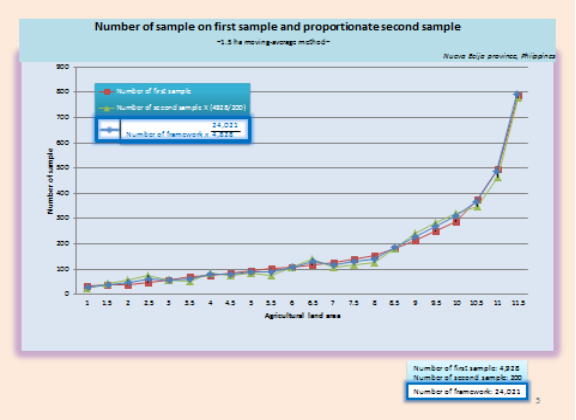
Generally speaking, administrative officer or academic researcher who request the survey conducting to get some data trend to request taking many samples to the statistical organization. Because they consider that the estimated data with high accuracy is gotten by many samples.

On the other hand, the statistician who requested many samples consider about the limitation of budget and labor as a survey designer. Here, we verify the relation between data accuracy and number of sample for the decision of proper number of first sample in ALIS.

2. What is the sample at all?

Let's consider about "what is the sample at all?" for our consideration on the relation between data accuracy and number of sample. We have learned about sample as "sample is a reduction figure of whole value" at a mathematic class. This is a principle of sample for framework. In fact, we can say that sample property (character) and framework property are almost same. By the way, "what's property?" It is verified by using the result of ALIS in Nueva Ecija province, Philippines.

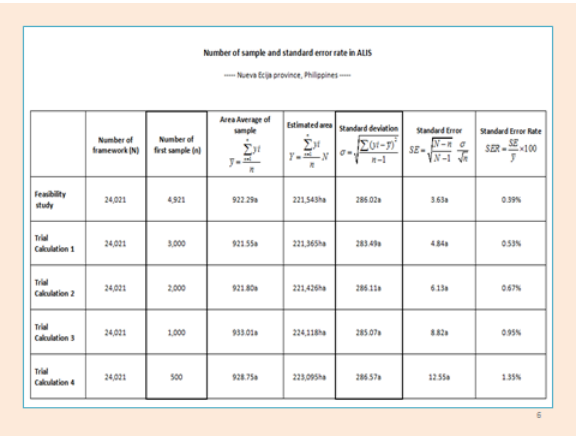
In this graph, the red line shows the appearance number of area mesh by each area category of agricultural land area on first samples in Nueva Ecija province. I have used the "moving-average method" to clear abnormal value for this graph. The green line shows the appearance number of area mesh on second samples which extracted from first samples.



The data of number of second samples are conducted "ratio adjusting calculation" to compare the appearance rate condition of sample with first samples. You can say that the second sample is a reduction figure of first sample. If we measure the agricultural land area of all area mesh in framework, we can suppose this blue line. In this case, we can say that the sample is a reduction figure of framework.

In fact, framework and sample have same property of data variability. Generally, the data variability indicates as the standard deviation which it is positive square root of variance. So we can say that the common property of sample and framework is the standard deviation (σ).

3. Verification of sample size


Number of sample and standard error rate in ALIS

---- Nueva Ecija province, Philippines ----

	Number of framework (N)	Number of first sample (n)	Area Average of sample $\bar{y}=\frac{\sum y_i}{n}$	Estimated area $\bar{Y}=\frac{\sum y_i}{n}N$	Standard deviation $\sigma=\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}$	Standard Error $SE=\sqrt{\frac{N-n}{N-1}}\frac{\sigma}{\sqrt{n}}$	Standard Error Rate $SER=\frac{SE}{\bar{y}}\times100$
Feasibility study	24,021	4,921	922.29a	221,543ha	286.02a	3.63a	0.39%
Trial Calculation 1	24,021	3,000	921.55a	221,365ha	283.49a	4.84a	0.53%
Trial Calculation 2	24,021	2,000	921.80a	221,426ha	286.11a	6.13a	0.67%
Trial Calculation 3	24,021	1,000	933.01a	224,118ha	285.07a	8.82a	0.95%
Trial Calculation 4	24,021	500	928.75a	223,095ha	286.57a	12.55a	1.35%

This table shows the relation between the number of first sample and standard error (accuracy) on Nueva Ecija province. In the feasibility study in ALIS, Bureau of Agricultural Statistics (BAS) in Philippines extracted 4,921 samples as first sample[note1]. They can get the total estimated agricultural land area in this province 221,543ha by simple estimation and this Standard Error Rate becomes 0.39%. In addition, I have calculated the case of random sample size of 3,000, 2000, 1000 and 500 area meshes as the first sample.

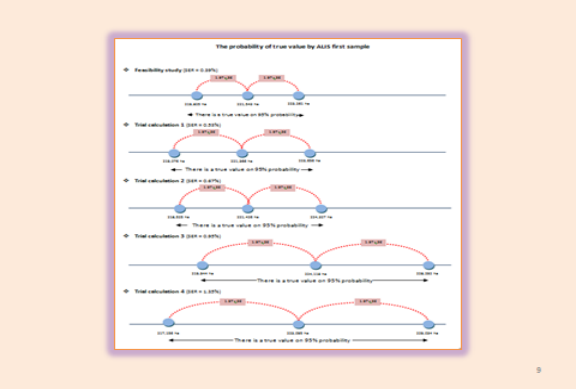
First, you can check that the standard deviation of each sample size is almost the same. In fact, we can say that these sample size have the same property for the estimation of whole value. On the other hand, the difference of sample size emerges as the difference of Standard Error.

In fact, considering the standard error formula, the standard error (accuracy) is influenced by only number of sample due to the standard deviation of each sample size is almost equally. The relation between the sample size and accuracy can explain by "the accuracy is inversely proportional to the square of number of sample".

4. What is the accuracy?

Here, we should consider "what is the accuracy?" again. Many officers who work in the statistical office say the accuracy is important, so it needs to take many samples in order to get high accuracy data. However, it is hard to seek the person who can explain statistically "what is the accuracy".

Accuracy means the theoretical value by sampling method. The difference of Standard Error indicates the difference of the confidential interval of estimated data. In fact, to take many samples (to increase accuracy) is only to narrow the confidential interval. We have to recognize that the accuracy is not the evaluation for the estimated data itself.



This diagram shows the difference of the confidential interval of each sample size on previous table. As you see, to increase the accuracy means to narrow the confidential interval. In fact, the accuracy does not ensure the accuracy of estimated data itself, but only indicates the "reliability" of estimated data.

In addition, we have to review the relation between accuracy and number of sample. The accuracy is inversely proportional to the square of number of sample. It means that we have to do 4 times effort to get 2 times accuracy.

To take many samples in order to increase accuracy has a risk which it becomes "mountain in labor"

5. Conclusion

However, it is difficult to indicate how many samples are proper as the first sample in ALIS. This difficulty occurs because the target accuracy should be considered according to the utilization purpose of data and budget and labor condition in statistical office. It is recommended to simulate proper number of sample by using ALIS data list on feasibility study. The standard deviation would be almost same in any sample size. The simulation method for necessary number of sample on target accuracy would be explained by other study report, however, it can suggest only this, it does not need to take many samples to estimate area data.

note 1 BAS has set 5,000 target samples as the first sample; however, the operator has judged 79 area meshes as the area mesh does not including cultivated land at the operation of the attaching area borderline and registered these meshes without area borderline information. So ALIS has judged the number of first sample is 4,921.