(Study report 6)

Accuracy calculation on area sample

Shoji KIMURA

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

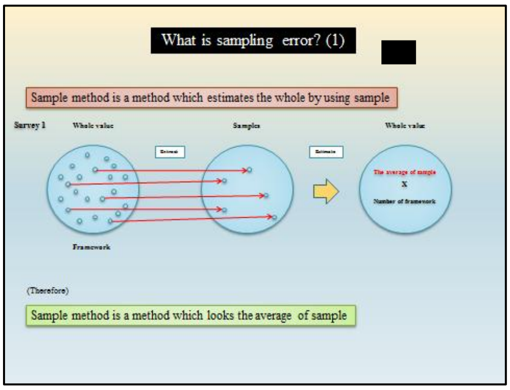
This study report is written for a working level statistician who responses on conducting an area sample survey by ALIS. In order to respond a request he or she should understand that "What is known as the proper number of sample?" Therefore, this report has used the actual data in the result of ALIS activity to verify the data accuracy. However, we should recognize that the accuracy is the just theoretical indication of data not for ensuring of the data value itself.

1. What is accuracy?

First, let clear about "what is accuracy?" The staffs of statistical institution like you are requested to collect or make the data with high accuracy. The high data accuracy means this data is in no or low data error. And as you know, in case of sample survey, data error divides into sampling error and non-sampling error. The sampling error means the differences between the sample value and the whole value. And the non-sampling error means the errors out of sampling error. More specifically, the sampling error shows the differences between the average value of samples by sample survey and the average value of sample framework. Please note that sampling error can be considered the differences of sample data average and whole data average. On the other hand, specifically, we can indicate an insufficient framework, survey miss, arbitrariness and miscalculation as the non-sampling error.

In fact, the sampling error is a statistical error and a calculable error. And the non-sampling error is a human error and an incalculable error. Which are more difficulty problems for data accuracy? It will be the non-sampling error. However, this study subject is an accuracy calculation on area sample survey. Here we are going to focus to the sampling error.

Let consider on "What is the sampling error?" once again. As you know, "a sample method is a method which estimates a whole by using sample". Now we have to understand the concept of a whole value in figure 1.



What is sampling error? (1)

Sample method is a method which estimates the whole by using sample

(Therefore)

Sample method is a method which looks the average of sample

Figure 1

For this purpose, we extract some samples from framework and check the value of each sample. And although we estimate the whole value, the method for whole value estimation becomes the average of sample multiply number of framework. Therefore, "*the sample method is a method which looks the average of sample*"

So, we can make the formula

$$y = \frac{\sum_{n=1}^{n} y_i}{n} N$$

to seek the average of sample data.

N = Number of framework

n = Number of sample

y_i = Value in "i" number sample

On the one survey, it aggregates the value of each sample and divides by number of sample. As the result, we will get the average of sample. If we continue the survey infinity for same framework by same number of sample, we can get the infinity average data of sample.

In this case we take an appearance frequency on a y-axis and the average of sample data on an x-axis so that, we can make a normal distribution graph by a central limit theorem. Some persons who in charge statistics may be misunderstanding about this central limit theorem.

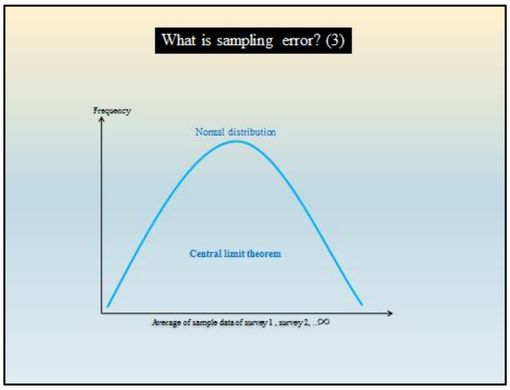
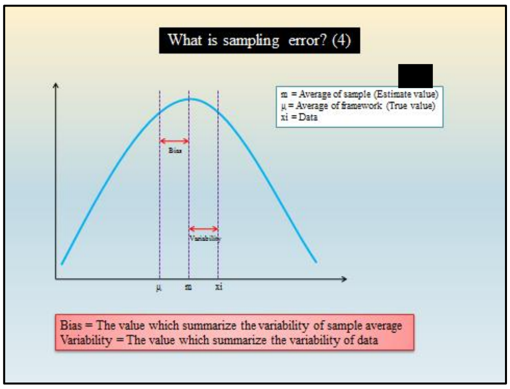

What is sampling error? (3)

Figure 2

Please note that the normal distribution is made by the average of sample data not, by the data itself extracted from framework.

Now we can explain "What is the sampling error" by using this normal distribution graph.


What is sampling error? (4)

m = Average of sample (Estimate value)
μ = Average of framework (True value)
xi = Data

Bias = The value which summarize the variability of sample average
Variability = The value which summarize the variability of data

Figure 3

In fact, we can define "m" as average of sample at the center of this normal distribution graph. This average of sample is the estimated value as the average of framework. However, the true value "μ" exists on an x-axis as the true average of framework. The difference of "m" and "μ" is the sampling error. We call this difference "bias". The definition of bias, Bias = *the value which summarize the variability of sample average*.

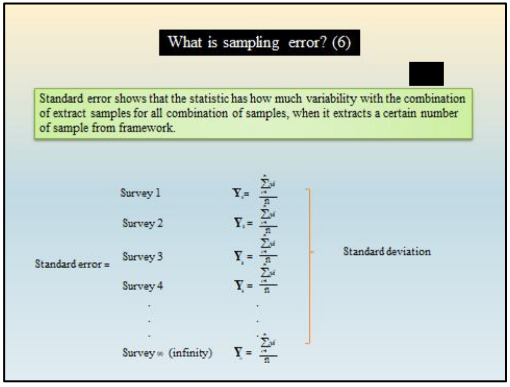
On the other hand, it appears the difference of "m" and an optional data "xi". As you know, this difference is called "variability". The definition of variability, Variability = *the value which summarize the variability of data*. Therefore, we can explain the optional data "xi" by next formula,

Xi = μ + (m - μ) + (xi-m)

So we can make clear rules of statistics. The degree of bias means the error in sample average. And we can indicate this bias by standard error (SE). By the way, the degree of bias is called as accuracy. For comparison, the degree of variability means the variability of data. And we can indicate this variability by standard deviation (SD) or coefficient of variation (CV). The degree of variability is called precision.

2. Standard error

On this study, mainly focus on standard error which indicates the sampling error. "Standard error shows that "the statistics have how much variability with the combination of extract samples for all combination of samples, when it extracts a certain number of samples from framework." It becomes this description is the most difficult description on this report. However you do not need to understand this description itself.



What is sampling error? (6)

[redacted]

Standard error shows that the statistic has how much variability with the combination of extract samples for all combination of samples, when it extracts a certain number of sample from framework.

Figure 4

If we continue the survey infinity for a same framework by a same number of samples, we can get the infinity the average data of sample. The standard deviation of these average data of sample becomes the standard error. As the theoretical conclusion, we can consider below;

・The distribution of the sample average makes the normal distribution.

・The sampling error means an error in sample average (bias).

・The error in sample average is shown by the standard error.

Based on the past study, let move to a deeply explanation of standard error formula.

Here I explain the standard error by using a real data of ALIS in Lao PDR. The standard error is calculated by below formula when it extracts "n" sample from framework which has an element count "N" and the standard deviation "σ".

$$\mathbf{SE} = \sqrt{\frac{N-n}{N-1}}\,\frac{\sigma}{\sqrt{n}} \text{ at that time } \sigma = \sqrt{\frac{\sum (yi-\overline{y})^2}{n-1}}$$

Let pay attention to this part of formula $\sqrt{\frac{N-n}{N-1}}$. It shows the relationship between a framework size and the number of sample. Here we can consider that if number of N is enough large, the solution of $\sqrt{\frac{N-n}{N-1}}$ approaches limitable to "1". For example, in case, the number of framework is 100,000 and the number of sample is 1,000 become 0.994. So we can consider conveniently, SE= $\frac{\sigma}{\sqrt{n}}$

However, in case of real calculation in ALIS, you should calculate this formula's part as well. Because ALIS counts automatically N and n, and prints these number to the result sheet.

Let input the real data to this formula. In case of the first sample survey in Khammouane province, n=3,383 and σ=299.09a, SE becomes 5.14a

$$SE = \frac{299.09}{\sqrt{3.383}} = 5.14a$$

Please remember that the standard error is standard deviation of sample average by extracted optionally.

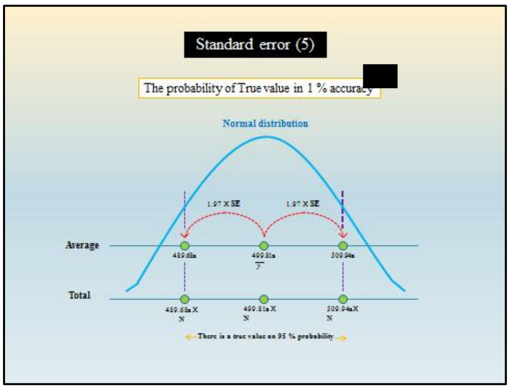
Next let consider the standard error rate.

$$SER = \frac{SE}{\bar{y}} \times 100$$

This is the formula for the standard error rate. It inputs SE=5.14a and the average of first sample data=499.81a. We can get an answer of about 1% accuracy as the standard error rate by input real data.

$$SER = \frac{5.14}{499.81} \times 100 = 1.03\%$$

Figure 5 shows the meaning of this 1% accuracy rate.



Standard error (5)

The probability of True value in 1 % accuracy

[redacted]

Figure 5

I think everybody knows this confidence interval theory. Generally, the theory of confidential interval is explained for the variability as the standard deviation by “σ”. However, we can explain this theory for bias by standard error. In fact, there is a true value during about ±2SE with a center on average of sample on 95% probability. Because as you checked, the average of sample by extracted infinity becomes the normal distribution graph.

So we can consider that a true agricultural land area in this province is during 489.68a × N and 509.94a X N on 95% probability. This consideration method is a proper approach for the result of sample survey.

3. Relationship between the accuracy and the number of sample

On this area sample survey in this province, Department of Planning in Lao PDR has extracted 3,383 samples as the first sample. And they got 1.03% accuracy rate. Here I give you one proposition. How many samples does it need, if you want to get 5% accuracy rate (error rate)?

Please remember this standard error formula.

$$SE = \frac{\sigma}{\sqrt{n}}$$

This formula can solve as follows.

$$SE\sqrt{n} = \sigma \,, \sqrt{n} = \frac{\sigma}{SE} \,, n = \left(\frac{\sigma}{SE}\right)^2$$

Therefore, it is necessary to clear two undefined values of “σ” and SE in order to seek the number of “n.” On the other hand, the formula of SER is;

$$SER = \frac{SE}{\bar{y}} \times 100$$

Probably we can say that the standard error rate is target accuracy. Therefore, if you want to get “a%” accuracy, it solves as follows.

$$a\% = \frac{SE}{\bar{y}} \times 100 \,, \;\; a\% \times \bar{y} = SE \times 100,$$

$$SE = \frac{a\,\% \times \bar{y}}{100}$$

We can get the formulas below to seek “n” which responds to target accuracy.

$$n = \left(\frac{\sigma}{SE}\right)^2 \,, SE = \frac{a\% \times \bar{y}}{100}$$

Here, “σ” and “$\bar{y}$” use the result of feasibility study. $(\sigma = 299.09a \;, \bar{y} = 499.81a)$ Because we can consider that the sample is a reduced figure of framework. In fact, we can assume that *“σ” and “$\bar{y}$” will be hardly affected by number of sample*. As a real story, if this pre-condition doesn’t establish, the sampling theory itself will not be established.

In case of 5% target accuracy, we can calculate as follows.

$$SE = \frac{5\% \times 499.81a}{100}$$

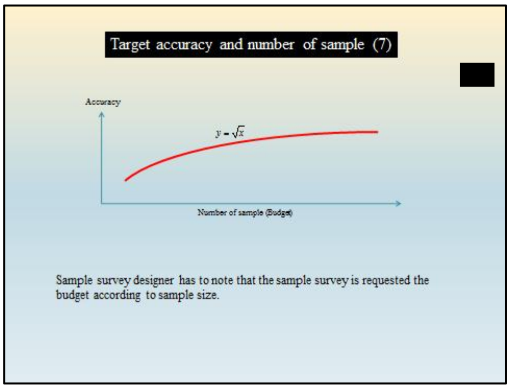
$$SE = \;\; 24.99a$$

$$n = \left(\frac{299.09a}{24.99a}\right)^2 \quad , n = 143$$

In fact, it is required that the approximately 143 samples are surveyed from the framework in order to get the 5% accuracy. Let consider about the relationship between the accuracy and the number of sample based on the above result. To take 1% accuracy need 3,383 samples, and to take 5 % accuracy need only 143 samples. Why in order to get 5 times accuracy needs to prepare about 24 times samples? We can clear this mystery by standard error formula. In fact, *the accuracy is inversely a proportional to the square of number of sample*.

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

Figure 6 shows the relationship between accuracy and number of sample.



Target accuracy and number of sample (7)

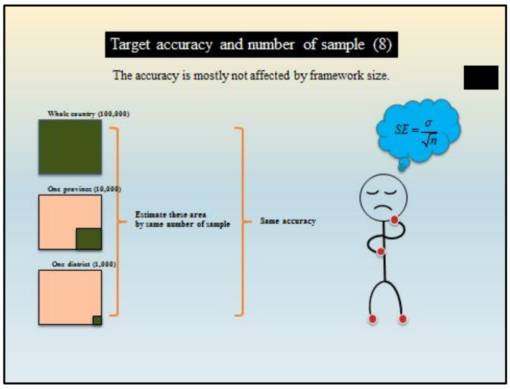
[redacted]

Sample survey designer has to note that the sample survey is requested the budget according to sample size.

Figure 6

Administrative officer trends may want to take many samples for he or she survey. However, we have to know that in order to increase the accuracy by the number of sample has limitations. And a sample survey designer has to note that the sample survey is requested a budget according to the sample size.

In addition, this standard error formula gives one suggestion for us. That is *the accuracy is mostly not affected by the framework size*. Generally speaking, the standard error formula can be established without sigh "N" number of framework. In fact, in case of 100,000 samples frame like whole country 10,000 samples frame like one province, 1,000 samples frame like one district, if you estimate these areas by the same number of sample, these accuracies become almost the same. In other words, the accuracy is mostly affected by the data variability (σ) and the number of sample (n).



Target accuracy and number of sample (8)

The accuracy is mostly not affected by framework size.

[redacted]

Figure 7

4. Conclusion

The practical statistician has to note that the calculated accuracy is a theoretical value absolutely and it is requested to correspond to the non-sampling error in order to secure the accuracy of data itself. Most of part of our effort has to be spent to clear the non-sampling error as a practical statistician.

Your organization may introduce ALIS. ALIS will make the framework and support sample survey. Staffs will be able to operate ALIS easily and will conduct sample survey. However, you just only stand at the start line for the purpose getting high accuracy data if you conduct sample survey by ALIS. You remain many task mainly task in order to reduce non-sampling error like survey technique, reliable field survey and exclusion arbitrariness to get high accuracy data.