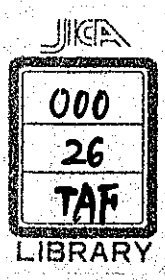
METHODS OF SAMPLE SURVEY

JAPAN INTERNATIONAL COOPERATION AGENCY



CONTENTS

FOREWORD

This booklet was prepared as a text book on the sample survey methods, for the Group Training Course in Agricultural Economic Statistics which is to be held in Tokyo for two and a half months starting from September 1972, sponsored jointly by OTCA and the Statistics and Survey Department of the Ministry of Agriculture and Forestry.

When survey cost per unit is expensive like in the case of Agricultural Economic Survey, application of the sample survey methods is inevitable. And since the number of sample farm households is limited, sampling design of the survey must be made carefully in order to obtain unbiassed and precise enough estimates for usual use.

Chapter 1 discusses the fundamental concepts of the sample survey. Chapter 2 explains as plainly as possible various techniques of the sample survey. A minimum of necessary mathematical expressions are employed to make the discussion clear and accurate. Chapter 3 explains the fundamental theories of the sample survey.

CHAPTER 1
CONCEPTS OF THE SAMPLE SURVEY

1. Complete Enumeration Survey and Sample Survey

Such a survey in which all survey objects in a particular area under consideration are enumerated is called a complete enumeration survey, while such a survey in which only survey objects selected at random are enumerated is called a sample survey. A representative example of the complete enumeration survey is a population census, and that of the sample survey a survey of rice production by crop cutting.

In the sample survey, a clear distinction must be made in mind between the population in a particular area under consideration and the sample selected from the population. The objects to be sampled are called sampling units. A group of all sampling units in a given area under consideration is called a population, and a group of selected sampling units is called a sample.

When the number of survey objects in the population is so large that an enough budget can not be given for conducting a complete enumeration survey, the sample survey becomes the only possible way of survey. The theories and techniques of the sample survey were developed during World War II and came into practical use thereafter. These facts account for the application of the sample survey methods for various statistical surveys conducted by governments of many countries of the world after the War.

2. Random Sampling and Systematic Sampling

The most fundamental characteristics of the sample survey is that the sample is selected at random, not purposely. Selection of the sample at random permits application of the probability theory for evaluating the estimate. That is, the estimate of the sample survey distributes on the normal distribution with the mean equal to the population mean and with the standard deviation equal to $\sigma/\sqrt{n}$, where σ is the population standard deviation and n is the size of sample. In other words, the estimate of the sample survey does not deviate much from the population mean (such an estimate being called an unbiassed estimate), and the extent of error of the

estimate (not enumeration error but sampling error) can be calculated.

In actual surveys, however, the real random sampling using a random number table is not very often used. Instead the systematic sampling is usually used in which sampling is made with a constant interval on the list of sampling units. This is because the control of sampling work is easier and it often happens that a more accurate estimate can be obtained by the systematic sampling than by the random sampling. In this booklet, however, we will assume that the random sampling was applied even if the systematic sampling was actually made.

3. Enumeration Error and Sampling Error

Two types of error are involved in the estimate of the sample survey the enumeration error and the sampling error. The enumeration error is the error which comes from observation error of the enumeration of individual sampling units in the sample. The sampling error is the error which occurs from the fact that the estimate is calculated not on the basis of all the data of the population but on the basis of only the data of the sample. The complete enumeration survey is free from the sampling error, but consists of the enumeration error only.

Now, a comparison will be made between the error of the survey result of the complete enumeration survey and the error of the estimate of the sample survey. The enumeration error of the complete enumeration survey is often a big under-bias, and it is very difficult to make it small. On the other hand, the enumeration error and the sampling error of the sample survey can be controlled.

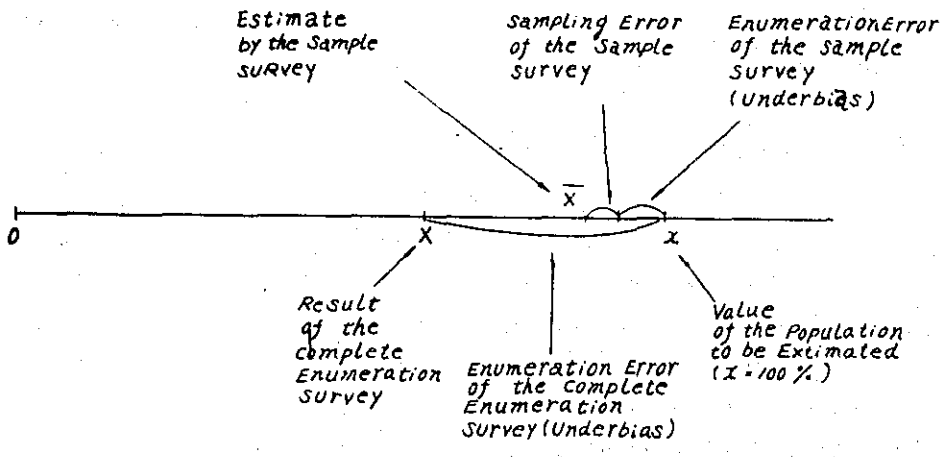
The enumeration error of the sample survey can be reduced by the following ways.

(1) Since the sample survey can be conducted by a smaller number of enumerators, excellent ones can be employed and it is possible to spend more time for training them.

(2) Larger quantities of materials and more time can be employed for enumeration of the individual survey objects in the sample.

The sampling error of the sample survey can be reduced to a desired level at the stage of designing the survey. In other words, since the sampling error of the estimate is approximately $\frac{\sigma}{\sqrt{n}}$ (σ being the standard deviation of the population and n being the size of sample), the sampling error can be reduced to a desired level by enlarging the size of sample, n.

The fact mentioned above can be schematically shown as follows:



4. Advantages of the Sample Survey

The sample survey has the following advantages, as compared with the complete enumeration survey.

(1) It can be conducted with less expense.

Since the enumeration is made on the sample only, less enumerators, less questionnaire forms and less of the other enumeration expenses are required. It should be noted, however, that more expenses may be required for the sample survey than for the complete enumeration survey, in some other aspects of survey. For instance, such works as survey designing, preparation of the list of population, sampling of the sample and calculation of the precision of estimate require additional expenses.

(2) It can be carried out in less time.

Since the enumeration is made on the sample only, the time for enumeration and tabulation can be reduced very much. This is very important when it is taken into consideration that statistics becomes of less value unless it is published early.

(3) Errors can be controlled.

As was mentioned in Paragraph 3, the enumeration error and the sampling error of the sample survey can be controlled to a desired extent at the stage of survey designing.

5. Disadvantages of the Sample Survey

The sample survey has a following disadvantage. Reducing the sampling error to a desired extent can be achieved by increasing the size of sample for a given area. However, when the area is further divided into smaller divisions, a sufficiently large size of sample can no more be obtained. It follows that the precision of the sample survey can be secured only within the area where the sufficient sample size was assigned, but for smaller divisions of that area, it cannot be guaranteed.

This fact means that the sample survey causes inconvenience when statistics are required for various levels of subdivisions like prefectures, cities, towns and villages as in the case of a population census. This is one of the reasons why the sample survey is not employed in the population census. It is also necessary to make sure whether or not the stratified estimates are accurate enough, when the estimates by strata are to be utilized.

6. Complete Enumeration Survey or Sample Survey

In designing a survey, selection of either the complete enumeration survey or the sample survey should be made taking into account advantages and disadvantages of the sample survey discussed above.

Generally speaking, the following factors govern the selection.

(1) The population census and the like are conducted by means of the complete enumeration survey.

(2) When it is practically impossible to conduct a complete enumeration survey, there is no alternative but to resort to the sample survey. For instance, when there is no other way of enumeration than crop cutting for getting good crop data, it is practically impossible to conduct a complete enumeration survey with crop cutting.

(3) When the budget available is too limited for conducting the complete enumeration survey but barely affords the sample survey, the sample survey becomes the only possible method. There are a great number of instances of this sort.

(4) When the budget available affords both the complete enumeration survey and the sample survey, one of them should be chosen by making a careful comparison between advantages and disadvantages of the sample survey.

In either case a deep understanding and rich experience concerning the sample survey are required. It is therefore recommended that specialists of the sample survey be consulted in making a choice.

7. Notes on the Sample Survey

(1) Frame of Sampling Units of the Population

It is necessary to obtain a list of sampling units belonging to the population in order to select a random sample from it. This list is called a frame.

Registration documents, cadastral documents and survey results of various censuses serve as valuable information for preparation of the frame. An effort should be made to obtain an appropriate frame by examining those data carefully. In this case, some means should be contrived to update them, since they were prepared in the past.

No information is, in some cases, available at all for preparation of a frame. In such cases, it becomes necessary to conduct a preliminary

survey to prepare a frame prior to carrying out the sample survey. Although this preliminary survey requires much time and labor, it is of a great value since it not only furnishes the information indispensable for the preparation of the frame but also produces very valuable basic statistics.

(2) Enumeration of Individual Sampling Units in the Sample

The fact that a limited budget can intensively be used for enumeration of a limited number of survey objects, thereby reducing the enumeration errors of individual sample data to a minimum is an important advantage of the sample survey. It is, however, not necessarily easy to obtain accurate individual data for various reasons. There are some methods of enumeration, like the objective measurement, the interview method. The method of enumeration should be decided carefully, since an underbias produced by enumeration is reflected on the estimates directly.

CHAPTER 2
METHODS OF THE SAMPLE SURVEY

This chapter deals with various techniques of the sample survey in as plain terms as possible. Section 1, as a preparatory step, explains the steps of the sample survey, furnishing information as to where those various techniques are positioned in the sample survey. Section 2 introduces a minimum of mathematical expressions required to ensure accuracy of our discussion.

1. Steps of the Sample Survey

(1) Designing of the Sample Survey

At this step, the designer of the sample survey examines the surrounding conditions of the survey as carefully as possible, decides on the type and size of the survey suitable for the purpose of the survey, and formulates a program of implementation of the survey.

He proceeds with his designing work, taking into account all sorts of problems which may arise at various steps of the sample survey. This is why the designer of the sample survey requires a deep understanding of and rich experience in the sample survey.

The main works of a sample survey designer are as follows;

a. Description of the purpose of survey

b. Determination of survey items

c. Selection of the enumeration method

d. Searching for any document which may be utilized for the preparation of the frame (list of sampling units of the population)

e. Determination of the type of the sample survey (The type of sample survey is a combination of selections, that is, whether single-stage sampling or multi-stage sampling, unstratified sampling or stratified sampling, objective measurement or

interview survey or survey by mail, simple estimation or ratio estimation, etc.)

f. Determination of the size of the sample (The size of the sample is determined in such a way that the sampling error of estimate may fall within 3%, 5% or 10% of the estimate. Such an percentage is called "aimed precision")

g. Programming of the sample survey (The implementation program of the survey and its schedule are prepared.)

(2) Preparations for the Survey

a. Designing of the form of frame, sampling form, questionnaires and tabulation forms.

b. Preparation of a survey manual

c. Acquisition of materials and instruments required for the survey (e.g. instruments for objective measurement)

d. Nomination of enumerators

e. Training of enumerators

f. Request for cooperation to the organizations concerned

(3) Preparation of Frame

The first step of the implementation of the sample survey is the preparation of the lists of sampling units in the population, that is, the frame.

(4) Sampling

Sampling is carried out on the frame. The sampling method is to be decided on by selecting a suitable combination of the following alternatives.

a. Unstratified sampling or stratified sampling

b. Single-stage sampling or multi-stage sampling

c. Random sampling, systematic sampling or proportional probability sampling

(5) Enumeration

The enumerators conduct enumeration of the selected samples. Enumeration may be carried out either by objective measurement, interviewing or mail enquiery.

(6) Calculation of Estimates

Estimates are calculated on the basis of sample data. The method of estimation may be either simple estimation or ratio estimation.

(7) Calculation of Precision

The precision of estimate is calculated on the basis of sample data to make sure whether or not the aimed precision computed at the stage of survey designing could be achieved. The precision of estimate computed on the basis of sample data is called achieved precision.

(8) Publication of Survey Results

The estimates are published. The estimates of the principal survey items should carry their achieved precision when published.

2. Mathematical Expressions Used in the Sample Survey

(1) Population and Sample

Population ..... Group of all sampling units
Sample ..... Group of selected sampling units

The sampling unit is a unit to be sampled. For instance, when we are going to estimate the average agricultural income of farm household in a certain region during the previous year by selecting sample farm households, the sampling units are individual farm households of the region. The population is the group of all farm households in the region. The sample is the group of selected farm households.

$x_1, x_2, \ldots\ldots, x_N$ Value of a certain survey item of each sampling unit of the population

$X_1, X_2, \ldots\ldots, X_n$ Values of a certain survey item of each sampling unit of the sample

The symbol N here indicates the number of sampling units in the population, and the symbol n the number of sampling units in the sample. N and n are called the size of population and the size of sample, respectively.

Take a specific example, that is, the estimation of the average agricultural income per farm household in a certain region cited above. $x_1, x_2, \ldots\ldots, x_N$ respectively represent the agricultural income in the previous year of the first farm household, that of the second farm household, ..., that of the last farm household in the list of all farm households. $X_1, X_2, \ldots\ldots, X_n$ respectively represent the agricultural income in the previous year of the first farm household, that of the second farm household, ..., that of the last farm household in the list of sample farm households.

Whenever the sample survey is discussed, the population and the sample must always be distinctly contrasted to each other.

(2) Summation Symbol

x_i ..... Value of the ith sampling unit in the population

X_i ..... Value of the ith sampling unit in the sample

The symbol i in x_i and X_i is called an affix. In the case of x_i, varies from 1 to N, and in the case of X_i, varies from 1 to n.

$$\sum_{i=1}^{N} x_i = x_1 + x_2 + \cdots\cdots + x_N \ldots\ldots \text{Population total}$$

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots\cdots + X_n \ldots\ldots \text{Sample total}$$

Σ is a capital letter of Greek alphabet which means summation, corresponding to an English letter S. $\sum_{i=1}^{N} x_i$ signifies "Sam of x_i varying i from 1 to N." Suppose, for instance, that a small village has ten farm

households, their respective agricultural income in the previous year being as follows:

Serial Number of Farm Household in the Sample	Agricultural Income		
(i)	(x_i)		
	¥1,000		
1	724	=	x_1
2	1,235	=	x_2
3	693	=	x_3
4	2,340	=	x_4
5	463	=	x_5
6	981	=	x_6
7	1,320	=	x_7
8	348	=	x_8
9	1,046	=	x_9
10	833	=	x_{10}
Total	9,983	=	$\sum_{i=1}^{10} x_i$

The average agricultural income of the population is $\sum_{i=1}^{10} x_i \div 10$ = ¥998,300. If the third, sixth and ninth farm households are selected as a sample, then

Serial Number of Farm Household in the Sample	Agricultural Income		
(i)	¥1,000		(X_i)
1	693	=	X_1
2	981	=	X_2
3	1,046	=	X_3
Total	2,720	=	$\sum_{i=1}^{3} X_i$

the average agricultural income of the sample is $\sum_{i=1}^{3} X_i \div 3$ = ¥906,667.

The summation symbol has the following nature.

$$\sum_{i=1}^{N} (x_i + y_i) = \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} y_i$$

The following example will help the understanding of this nature of the summation symbol. Suppose that x_i signifies the income from rice production of the ith farm household and y_i signifies the income from other agricultural production than rice, $x_i + y_i$ therefore means the agricultural income of the ith farm household. In the same context, $\sum_{i=1}^{N} (x_i + y_i)$ signifies the total of agricultural income of all the farm households in the population. While $\sum_{i=1}^{N} x_i$ signifies the total of income from rice production of all the farm households, and $\sum_{i=1}^{N} y_i$ signifies the total of income from other agricultural production, and therefore $\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} y_i$ signifies the total of agricultured income of all the farm households being equal to $\sum_{i=1}^{N} (x_i + y_i)$.

Another characteristics of the summation symbol is expressed by the following equation.

$$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i$$

The following example will also be helpful in understanding this characteristic. If x_i indicates the agricultural income of the ith farm household in 1,000 Yen, multiplying it by 1,000 will give the agricultural income in Yen. Let C = 1,000, and Cx_i will mean the agricultural income of the ith farm household expressed in Yen. The left-hand side $\sum_{i=1}^{n} C\,x_i$ signifies the total of agricultural income in Yen, and the right-hand-side $c\sum_{i=1}^{n} x_i$ the total of agricultural income in 1,000 Yen multiplied by 1,000 Yen. Both sides therefore express the same value.

Lastly, mention should be made of the multi-stage summation symbol which is most frequently used in the discussion of the sample survey. This symbol is used in the discussion of the stratified sampling and multi-stage sampling which will be discussed later. We will now consider, for example, agricultural income of the farm households of two strata, two regions for example. It can be written as follows:

Agricultural income in the previous year of the farm households of the first region

$$x_{11},\ x_{12},\cdots\cdots,\ x_{1N_1}$$

Agricultural income in the previous year of the farm households of the second region

$$x_{21},\ x_{22},\ \cdots\cdots,\ x_{2N_2}$$

The number $_{11}$ in x_{11} is called a double affix. The first $_1$ indicates the first region, and the second $_1$ the first farm household (of the first region). The agricultural income x of the jth farm household in the ith region is generally expressed as x_{ij}. The total of all the agricultural income of all the farm households of both regions is written by the following expression.

$$\sum_{i=1}^{2}\sum_{j=1}^{N_i} x_{ij} = (x_{11} + x_{12} + \cdots\cdots + x_{1N_1}) + (x_{21} + x_{22} + \cdots\cdots + x_{2N_2})$$

$\sum_{j=1}^{N_1} x_{1j}$ indicates the total agricultural income of the first region, and $\sum_{j=1}^{N_2} x_{2j}$ that of the second region. $\sum_{i=1}^{2}\sum_{j=1}^{N_i} x_{ij}$ is therefore the sum total of agricultural income of the two regions.

(3) Mean

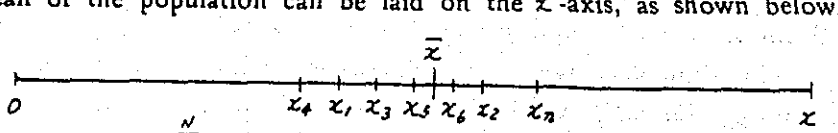
Values which express the nature of the distribution are called characteristics. One of them is the mean or the arithmetic mean.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}(x_1 + x_2 + \cdots\cdots + x_N)$$

This equation gives the mean of the population. The mean of the sample is given by the equation

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots\cdots + X_n)$$

The mean of the population can be laid on the x-axis, as shown below.



Therefore, the total $\sum_{i=1}^{N}(x_i - \bar{x})$, that is the total of the differences of x_i from $\bar{x}$ is zero. This can be proved as follows:

$$\sum_{i=1}^{N}(x - \bar{x}) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N}\bar{x}$$

$$= N\bar{x} - N\bar{x}$$

$$= 0$$

(4) Variance and Standard Deviation

The characteristics which express the degree of variability of values around the mean are variance and standard deviation. The population variance is given by the following formula.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}$$

In other words, the variance is the mean of the square of the differences between each value x_i and the mean $\bar{x}$. It should be noted that N-1 is used in lieu of N, since the expected value (which will be explained later) of the sample variance equals the population variance only when N-1 is used. The variance obtained by dividing by N-1 is called an unbiased variance. The sample variance is given by the following formula.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

The formula can be rewritten as follows to simplify the calculation of, for instance, the population variance:

$$\begin{aligned}
\sigma^2 &= \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 \\
&= \frac{1}{N-1} \sum_{i=1}^{N} (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \frac{1}{N-1} \left(\sum_{i=1}^{N} x_i^2 - 2\bar{x} \sum_{i=1}^{N} x_i + N\bar{x}^2 \right) \\
&= \frac{1}{N-1} \left(\sum_{i=1}^{N} x_i^2 - 2\bar{x} \cdot N\bar{x} + N\bar{x}^2 \right) \\
&= \frac{1}{N-1} \left(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2 \right) \\
&= \frac{1}{N-1} \left\{ \sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N} \right\}
\end{aligned}$$

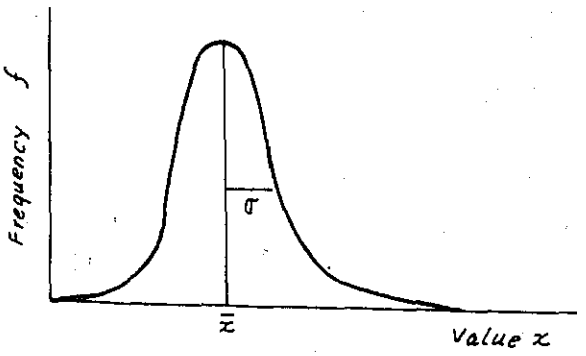
This means that the population variance can be obtained by simply calculating the sum of square of each value x_i and the sum of each value x_i, without taking the trouble of computing the difference between each value x_i and the mean $\bar{x}$ for all x_i's.

The standard deviation is the square root of variance.

Population standard deviation = $\sigma = \sqrt{\sigma^2}$

Sample standard deviation = $s = \sqrt{s^2}$

When the population distributes in the normal distribution, its standard deviation can be expressed in a distribution graph as follows:



(5) Correlation

Suppose that a survey is made on the area of paddy field and the rice production of every sample farm household in a certain region. Generally speaking, the larger (or the smaller) the area of paddy field is, the larger (or smaller) the rice production becomes. The characteristics which serves to express the degree to which the two variables of each object are co-related to each other is called the correlation coefficient. In this example, the object is a farm household, the two variables being the area of paddy field and the rice production.

Lets the area of paddy field and the rice production of the ith farm household in the population be x_i and y_i, respectively. The correlation coefficient can be given by the following formula.

$$\rho = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \, \sigma_y^2}}$$

where,

$$\sigma_x^2 = \frac{\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2 / N}{N-1} \quad \cdots\cdots\cdots \text{ variance of } x_i$$

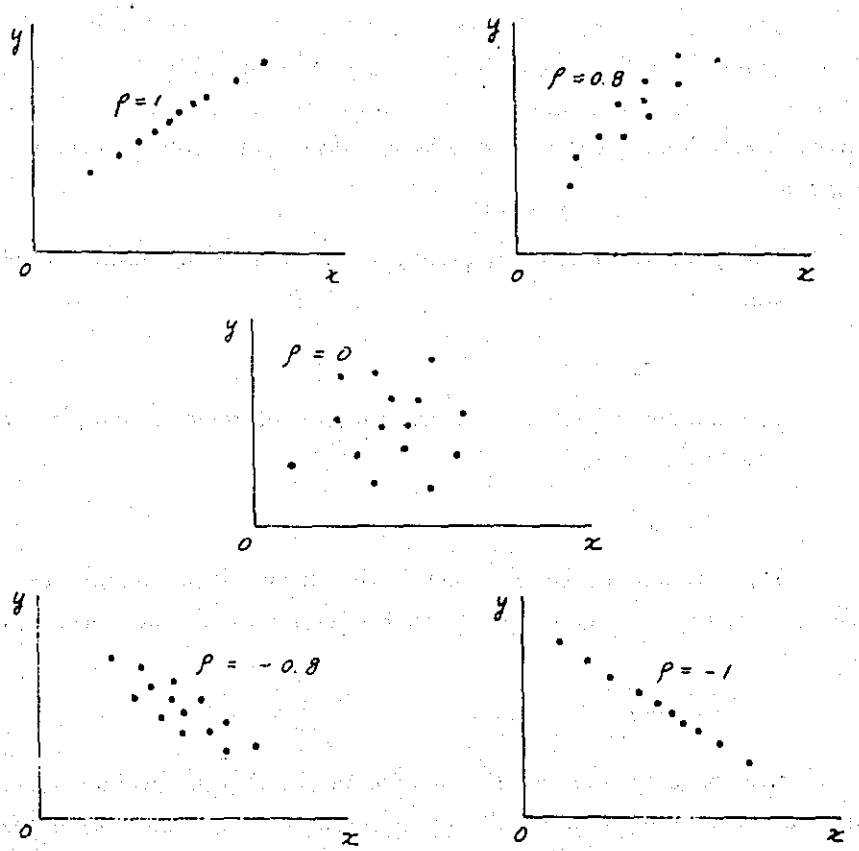
$$\sigma_y^2 = \frac{\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2 / N}{N-1} \quad \cdots\cdots\cdots \text{ variance of } y_i$$

$$\sigma_{xy} = \frac{\sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)/N}{N-1}$$ co-valiance of x_i and y_i

When the sample values are used, the formula is expressed in the following symbols:

$$r = \frac{S_{xy}}{\sqrt{S_x^2 \; S_y^2}}$$

The value of the correlation coefficient varies as shown in the following correlation graphs:



As can be seen from these graphs, ρ ranges from -1 to $+1$. When $\rho = 1$, the relation between x_i and y_i is perfectly on a line with positive direction. When $\rho = 0$, they vary independently of each other. When $\rho = -1$, the relation between x_i and y_i is perfectly on a line with negative direction.

(6) Mathematical Description of the Sample Survey

The sample survey is often called a scientific method, because it is based on mathematical theories. The study of the sample survey therefore requires familiarity with a minimum of necessary mathematical expressions. The simplest type of sample survey will now be discussed in mathematical terms using the mathematical symbols which we have already studied in the preceeding chapters.

Let's here estimate the mean (e.g. average agricultural income per farm household of all the farm households) of a population (e.g. all the households in a certain region) on the basis of the mean of the sample (a group of selected sample farm households) drawn at random from the population.

The population values (agricultural income of all farm households) are;

$$x_1, x_2 \cdots\cdots, x_N$$

The sample values (agricultural income of selected sample farm households) are;

$$X_1, X_2 \cdots\cdots, X_n$$

The estimate to be obtained is the mean of the population, that is, average agricultural income per farm household of all the farm households,

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

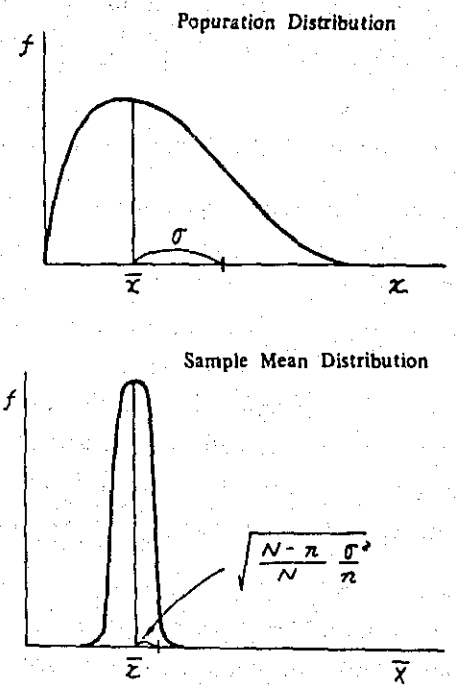
This will now be estimated by the sample mean, that is, average agricultural income per farm household of all the selected sample farm households.

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{n} x_i$$

It should be noted here that, although the mean of the population can never be known by conducting a sample survey, the fact remains that there is a given value of it. On the other hand, the mean of the sample represents nothing but one value calculated from a specific sample drawn from the population. If another sample is drawn from the population, a different sample mean will be obtained. The number of ways by which a set of n units selected out of N units is expressed as NC_n. Therefore, the number of sample means is equal to or less than NC_n (because some values may turn out to be the same by chance). The mean of the sample calculated in this sample survey is therefore one of NC_n sample means.

As will be discussed in detail in Chapter 3 "THEORIES OF THE SAMPLE SURVEY", the NC_n sample means exhibit normal distribution, with the mean which is equal to the population mean $\bar{x}$, and the variance which is equal to $\frac{N-n}{N}\frac{\sigma^2}{n}$ (σ^2 being the population variance).

This relationship is shown by the following graphs.



The standard deviation $\sqrt{\frac{N-n}{N}\frac{\sigma^2}{n}}$ of an estimate (sample mean $\bar{X}$ in this example) is called a standard error. The ratio of the standard error to the value to be estimated is called the precision of the estimate.

In the case of estimation of the population mean by the sample mean, the standard error is as follows.

$$\sqrt{\frac{N-n}{N}\quad\frac{1}{n}}\cdot\sigma$$

And the precision of the estimate is a follows.

$$\sqrt{\frac{N-n}{N}\quad\frac{1}{n}}:\frac{\sigma}{\bar{x}}$$

The rate of the population standard deviation to the population mean $\frac{\sigma}{\bar{x}}$ is called the coefficient of variance of the population and is often expressed in C_x

$$C_x = \frac{\sigma}{\bar{x}}$$

Since $\frac{N-n}{N}$ approximates 1 when the size of a population is by far larger than that of the sample, it may become almost equal to 1, and can be eliminated from the formula. The precision of estimation is often expressed in CV (Coefficient of Variance of the estimate), and its formula can therefore be rewritten as

(Precision of Estimate) $CV \doteqdot \frac{C_x}{\sqrt{n}}$

When the size of the population is very large.

3. Preparation of Frame

(1) Importance of the Frame

The preparation of a frame, that is, the list of all sampling units existing in the population, must precede all steps of the sample survey. Since random sampling is possible only when a frame is available, the preparation of a frame is the foundation of the sample survey and is of major importance.

As a matter of fact, designers of the sample survey often devote the most of their effort and time to looking for any document which may be utilized for the preparation of the frame. It is also usual that much manpower spent for actually compiling the list.

(2) Documents Utilized for Preparing the List

Cadastral documents, registration documents, membership registration documents of various cooperative associations and results of various types of censuses are the principal data which serve as documents for use in the preparation of the frame.

When no document is available for preparing the frame, a preliminary survey is carried out to compile it. Although this preliminary survey often requires a vast amount of manpower, we should think that the effort will be amply rewarded, since the frame itself will furnish us with valuable basic statistics.

(3) Updating of Information in the Document

As has been discussed above, the information which may be utilized in the preparation of a frame is the past one and must be updated. For instance, those sampling units which were contained in the population at the time of the past survey but should not be included for the present survey must be excluded from the list. Those sampling units will be called extinct sampling units. On the other hand, there may be such units as not existing in the past but should be included in the present survey. Those will be called appeared units.

It is possible that the extinct sampling units are included in the sample if not excluded from the list, and the ratio of extinction can be calculated with the sample data. However, the appeared sampling units must be included in the list without fail. Otherwise, a serious error will occur, since the appeared sampling units are totally left out of the frame.

(4) Processing of Information in the Document

It sometimes occur that even if a kind of documents is available for use in the preparation of the frame, it does not constitute the desired frame as it is. When it is desired, for instance, that the population containing only those sampling units of a certain size or larger, the sampling units of smaller size than that size must be excluded. In another case, when it is desired that a group of a certain number of units for

which information has been made available will constitute sampling units, it becomes necessary to combine those units into groups.

(5) Auxiliary Variable for Stratification or Ratio Estimate

When stratification or ratio estimate which will be discussed later is to be carried out, it is necessary to add an auxiliary variable to the name of each sampling unit in making the list of sampling units, or the frame.

(6) Frame for Multi-Stage Sampling

When multi-stage sampling which will be discussed later is to be carried out, it is desirable to obtain a list of secondary sampling units. If it is not available, it is also possible to prepare a list of secondary sampling units only for primary sampling units selected in the sample, at the stage of implementation of the survey.

4. Determination of the Size of Sample

The size of sample is determined at the stage of sample survey designing. It is determined with reference to the budget allocated and the required precision of the estimate. In the case of the most simple type of sample survey, that is, simple estimate with single-stage sampling without stratification, the following equation may used to approximate the size of sample provided that the population is of a large size.

$$n \doteqdot \left[\frac{Cx}{CV}\right]^2$$

where,

n = size of sample

Cx = coefficient of variance of the population

$$C_x = \frac{\sigma}{\bar{x}}$$

σ = population standard deviation

$\bar{x}$ = population mean

CV = Precision of estimate

Let the coefficient of variance of the population be 0.5 and the desired precision of estimate be 5% or 0.05, the size of sample can be determined by approximation as follows:

$$n = [\frac{0.5}{0.05}]^2 = 10^2 = 100$$

However, in the case of stratified sampling or multi-stage sampling which will be discussed later, it is necessary to distribute the size of the whole sample between the strata or stages of sampling. This distribution is usually carried out after the frame has been completed, and its process is called the allocation of sample size. The discussion of the allocation of sample size is included in that of stratification and multi-stage sampling which will follow.

5. Sampling

When the frame has been completed and the size of the sample has been determined, the sample is then drawn from the population. It is selected at random or systematically.

(1) Random Sampling

To select a sample at random means drawing it from the population by a certain method regardless of the values of sampling units. Following is the established sampling procedure.

(a) Assign serial numbers to all the sampling units contained in the list.

(b) Select one random number out of the numbers equal to or less than the size of the population N, using a table of random numbers. Such a sampling units in the frame that its serial number coincides with this random number becomes the first sampling unit of the sample.

(c) Repeat (b) until the number of selected units arrives at the size of the sample n. When the sampling is finished, the

serial numbers and names written in the frame list of those sampling units included in the sample are to be transcribed into the list of sample to complete the sampling work.

(2) Systematic Sampling

If we apply the random sampling, the amount of work becomes large when size of the sample grows. Besides, the systematic sampling has advantages mentioned later. Therefore, systematic sampling is often used in practice to select a sample.

Following is the established method of the systematic sampling.

(a) Calculate the sampling interval by dividing the size of population N by the size of sample n. If that is an aliquant part of the other, round up to one or two digits below decimal point.

$$h = \frac{N}{n}$$

(b) Select one random number from the numbers less than the sampling interval, may be, by opening one page of any book at your hand. This integer a is called the random start.

$$0 < a < h$$

(c) Add sampling interval h to random start a. Add again h to the sum of a and h. Repeat this cycle of calculation until the number of sampled units including the random start becomes n.

$$a$$
$$a + h$$
$$a + 2h$$
$$\vdots$$
$$a + (n - 1)h$$

If h is not an integer, it should be rounded into an integer. The sampling units in the frame which have those serial numbers are selected to prepare the list of sample.

(Example)

There are a hundred farm households in a village, and they are listed on the list of farm households. Twenty-two farm households will now be selected as a sample by systematic sampling.

$$h = \frac{N}{n} = \frac{100}{22} = 4.55 \ldots\ldots \text{ sampling interval}$$

A random number is then selected out of numbers 1, 2, 3 and 4.

Suppose that 3 is selected as a random number.

$$a = 3 \ldots\ldots\ldots\ldots\ldots \text{ random start}$$

Now calculate, a, a + h, a + 2h, ... a + 21h.

Values Calculated	Integers Rounded	
3	3	
7.55	8	
12.10	12	
16.65	17	
21.20	21	
25.75	26	
30.30	30	
34.85	35	The sampling units in the frame which have these serial numbers constitute the sample.
39.40	39	
43.95	44	
48.50	49	
53.05	53	
57.60	58	
62.15	62	
66.70	67	
71.25	71	
75.80	76	
80.35	80	
84.90	85	
89.45	89	
94.00	94	
98.55	99	

The systematic sampling method has the following advantages, as compared with the random sampling method.

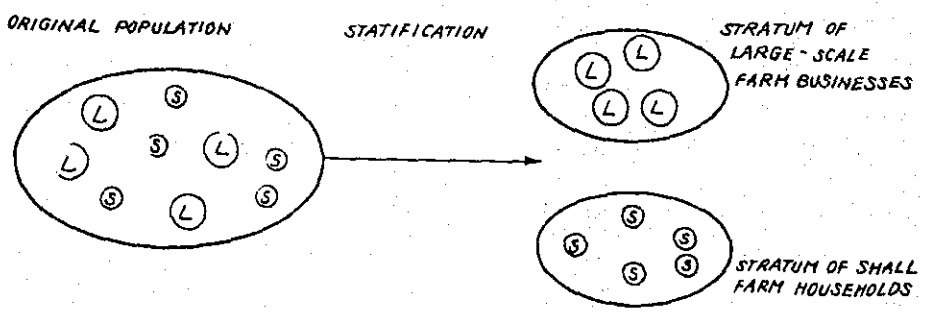
(a) Sampling is easier.

(b) It will not permit purposive sampling.

(c) As the sample is evenly distributed in the frame, a higher precision of estimation can be obtained. If the sampling units of the frame are arranged in the order of their values, the precision of estimation will be furthermore improved.

6. Stratification

Suppose there coexist large-scale farm businesses and small farm households in a village. If, the total agricultural income of this village is to be estimated by selecting a sample from the population which is including both of large farm businesses and small farm households, it will not be difficult to imagine that a high precision of estimation cannot be obtained. A better method of estimation is to select two separate samples from the group of the large farm business and that of small farm house-holds, respectively. Dividing the sampling units of a population into several groups is termed "stratification", and those groups are called strata. A characteristics or variable of each sampling unit, which serves as a criterion for stratification, is called the auxiliary variable for stratification. Information whether each sampling unit is large scale or small scale is an auxiliary variable for stratification in our example.



Large (L) } Auxiliary Variable
Small (S) } for Stratification

(1) Purpose of Stratification

The purpose of stratification, as mentioned above, is to improve the precision of estimate. However, in a practical application of stratification, the sample is often stratified by regions so that statistics (estimates) by regions (strata) is to be published. Accordingly, the stratification is made for either of the following two purposes.

(a) To compile statistics by regions or by strata.

(b) To improve the precision of estimate, even if statistics by stratum are not required.

The establishment of strata for the purpose of (a) depends on the division of statistics required.

The stratification for the purpose of (b) gives rise to the following alternatives.

(i) Whether or not stratification is to be made.

(ii) What kind of auxiliary variable should be employed for stratification.

(iii) How many strata should be established and by what criteria

In coping with these problems, an increase of the precision of estimate which may be attained by stratification and an increase in labor which may be caused by stratification must be taken into consideration. Generally speaking, it is recommended that not more than a few strata is established. The reason is that a large number of strata do not improve the precision of estimate much, while, it increase the amount of calculation to such an extent as not being practical.

(2) Variance within Stratum and Variance between Strata

The stratification which is made for the purpose of improving the precision of estimate should be such that the variance within stratum is small and the variance between strata is large.

The variance within strata may be considered as the weighted

mean of the variances within each stratum. Let the value of the jth sampling unit in the ith stratum be expressed as x_{ij}, and let there be Ni sampling units in ith stratum. The variance within ith stratum is then given as follows.

$$\sigma_i^2 = \frac{\sum_{j=1}^{N_i} (x_{ij} - \bar{x}_i)^2}{N_i - 1}$$

Then, the variance within stratum is given by,

$$\sigma_w^2 = \sum_{i=1}^{R} \frac{N_i}{N} \sigma_i^2$$

The variance between strata is the variance between the means in each stratum, and is given by,

$$\sigma_b^2 = \frac{\sum_{i=1}^{R} N_i (\bar{x}_i - \bar{x})^2}{N}$$

(3) Allocation of Sample between Strata

First, the size of the sample for the whole population is decided upon in relation to the precision of estimate required and the budget allocated. The allocation of sample between strata is carried out in one of the three different ways mentioned below.

(a) Proportional Allocation

The sample is allocated in proportion to the size of each stratum Ni.

$$Ni = \frac{Ni}{N} \quad n$$

(b) Neyman's Allocation

The sample is distributed in proportion to the product of the size of each stratum and the standard deviation.
The mean $\bar{x}_i$ may also be used instead of the standard deviation σ_i.

$$n_i = \frac{N_i \bar{x}_i}{\sum_{i=1}^{R} N \bar{x}_i} \cdot n = \frac{T_i}{\sum_{i=1}^{R} T_i} n$$

where,

Ti = total of values in the ith stratum

(c) Deming's Optimum Allocation

The sample is allocated in proportion to the product of the size of each stratum, the standard deviation and inverse ratio of square root of the survey expense per sampling unit.

$$n_i = \frac{N_i \sigma_i / \sqrt{c_i}}{\sum_{i=1}^{k} N_i \sigma_i / \sqrt{c_i}} n$$

The most widely used method of sample allocation is (a) or (b).

7. Multi-Stage Sampling

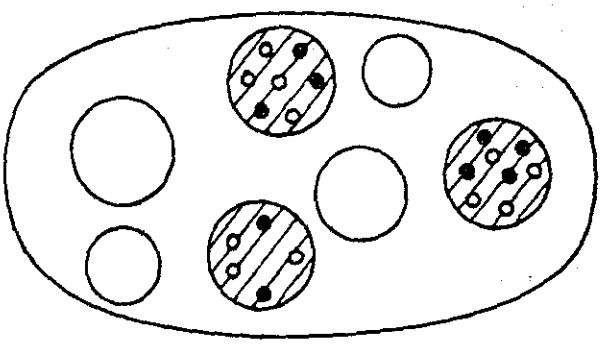
In estimating the average agricultural income of all the farm households in a region, a sample can be selected directly from the list of all the farm households in the region. This is the sampling method which has been discussed up to now, aside from the statification.

Another sampling technique will now be introduced.

(a) A sample of villages are first selected, and then a sample of farm households is selected in each sample village. The average agricultural income of each sample village is estimated on the basis of data of the sample households. Likewise, the average agricultural income in the region is estimated on the basis of the estimated average agricultural income of the sample villages.

(b) It is natural that the distance between farm households in a village is shorter than that of between villages. If sample farm households are selected only in sample villages as in the case of (a) above, the total travel expenses will be much less.

This type of sampling method is called two-stage sampling.



Let the agricultural income of the jth farm household in the ith village be τ_{ij}, and the agricultural income of the jth sample farm household in the ith sample village be Xij, respectively.

The average agricultural income in the region will then be given by the equation

$$\bar{\tau} = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \tau_{ij} \Big/ N$$

where,

M = Number of villages in the region

Ni = Number of farm households in the ith village

N = $\sum_{i=1}^{M} N_i$, total number of farm households in the region

The average agricultural income in the region is estimated by the equation

$$\bar{x} = \left(\frac{M}{m} \sum_{i=1}^{m} \frac{N_i}{n_i} \sum_{j=1}^{n_i} x_{ij} \right) \Big/ N$$

where,

m = Number of the sample villages

ni = Number of sample farm households in the ith sample village

A village in our example is called the primary sampling unit, and a farm household the sub sampling unit.

8. Enumeration

As have already been mentioned, the estimate of the sample survey contains two kinds of errors: enumerating error and sampling error. The sampling error can be controlled (in such a way that the size of the sample is enlarged until the error is reduced to a desired level) at the stage of survey designing. The enumerating error is, however, a different story. To improve the accuracy of enumeration, more time and labor must be used. It is no exaggeration to say that quality of the estimate mainly depends on accuracy of the enumeration.

Agricultural income can be enumerated in the following ways.

(1) Book-keeping

(2) Interview

For adopting the book-keeping method, literacy and willingness of sample farmers are essential. On the other hand, a simple interview method may not give accurate enough data for such a difficult item as agricultural income.

9. Estimation

(1) Simple Estimation and Ratio Estimation

We have already discussed two techniques of sampling, that is, stratification and multi-stage sampling, they have a bearing on the method of estimation. However, the fundamental variations of estimation are the following two:

(a) Simple estimation

(b) Ratio estimation

The simple estimation method is that which has already appeared in the preceding pages. The ratio estimation is a method in which an auxiliary variable which is closely correlated to the variable to be estimated is used at estimation stage to attain higher precision. For example, the area of paddy field planted and the rice production of farmers are roughly

proportional to each other. In other words, they are closely correlated to each other. Suppose that data on the area of paddy field planted of all the farmers in a certain village are available, and we want to estimate the total production of rice in that village with a very limited survey cost. A method will be as follows. First, we select sample farmers of an appropriate number. Second, enumerators will visit those sample farmers to make interview survey by asking their rice production. Then the total rice production in the village can be estimated by the following formula.

$$\hat{T} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} Y_i} \cdot \sum_{j=1}^{N} y_j$$

$\hat{T}$ ..... Estimated total rice production in the village

x_i ..... Rice production of the ith sample farmer

Y_i ..... Area of paddy field of the ith sample farmer

y_j ..... Area of paddy field of the jth farmer in the population

n ..... Number of sample farmer

N ..... Number of all farmers in the village

The ratio used in this equation, that is, $\sum_{i=1}^{n} x_i / \sum_{i=1}^{n} Y_i$, signifies the average rice yield per unit area in sample. The total area of paddy field planted is expressed by $\sum_{j=1}^{N} y_j$. This method of ratio estimation can thus estimate the total rice production by multiplying the average rice yield per unit area obtained from sample data by the total area of paddy field in the village.

Since the ratio is characteristic of the equation, it is called the method of ratio estimation. y is termed as an auxiliary variable. The equation used in the method of ratio estimation can be rewritten to read

$$\hat{T}_R = y \cdot \bar{R}$$

where,

$$y = \sum_{j=1}^{N} y_j$$

$$\bar{R} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} Y_i}$$

As ratio estimation employs more information (auxiliary variable) than simple estimation, the former gives a higher precision of estimate than the latter. Generally speaking, the method of ratio estimation may by used to advantage when the correlation coefficient ρ of variable x and y is 0.8 and over.

(2) Estimation of Variables and Estimation of Composition

Discussions up to this point were concerned with the estimation of variables. It is sometimes required to estimate the composition percentage of specifically characteristic sampling units to the population by means of the sample survey. Suppose, for example, it is necessary to estimate the number of farm households with a certain level of agricultural income, say, 1 million yens or more in a certain village by selecting a sample from the population. As can easily be supposed, an accurate estimation will be very difficult to obtain if the percentage of those farm households is very low. For instance, in a case where there farm households with 1 million yen or more of agricultural income constitute only 3% of the total number of farm households in the village, a great number of sample households are to be sampled for getting a reasonable estimate.

In order to estimate the percentage of specifically characteristic farm households to the population, the following method may be used. Each of the specifically characteristic farm households is assigned with a number of one and the other farm households zero. This method therefore deals with a population comprising of 0 and 1, as in 0, 0, 1, 0, ..., 0. This is called binominal distribution, whose mean is P (percentage of specifically characteristic sampling units). The size of population is N. Let the number of specifically characteristic sampling units be N1, P = N1/N. The variance is $p(1-p)$. Let $1-p=q$, $p(1-p)=pq$.

Take a specific example. The ratio of the number of the farm households with 1 million yens or more of agricultural income to the number of the farm households in the village is 10% precision, the number of necessary sampling units will be given as follows:

Variance: $V(\bar{P}) \fallingdotseq \frac{\sigma^2}{n} = \frac{P(1-P)}{n} = \frac{Pq}{n}$

$$CV(\bar{x}) = \frac{\sigma}{\bar{x}}$$

Precision: $CV(\bar{P}) = \sqrt{V(\bar{P})} / \bar{x} = \sqrt{\frac{PQ}{n}} / P = \sqrt{\frac{0.1 \times 0.9}{n}} / 0.1$

Let the precision be 10%,

$$0.1 = \sqrt{\frac{0.09}{n}} / 0.1$$

$$\therefore \quad 0.01 = \sqrt{\frac{0.09}{n}}$$

$$\therefore \quad 0.0001 = 0.09/n$$

$$\therefore \quad n = 900$$

The size of the sample should be 900.

10. Aimed Precision and Achieved Precision

At the stage of survey designing, the degree of precision desired and the size of sample which will satisfy this requirement must be determined. This precision aimed at is called aimed precision. For the calculation in this connection, existing data are looked for or a preliminary survey is conducted to gather necessary information. And the population variance is computed by approximation.

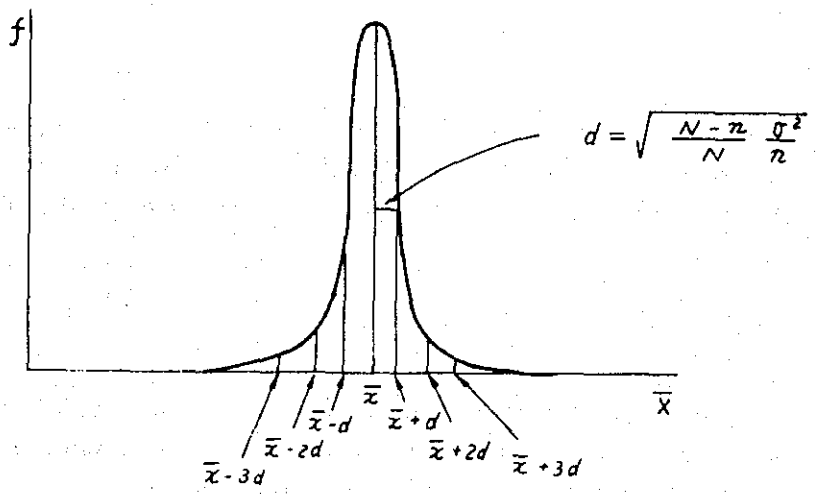
When the estimate is computed after a sample survey was implemented, the calculation of precision must be carried out using sample data in order to make sure the aimed precision was actually achieved. The calculated precision at this stage is called achieved precision. This achieved precision is to accompany with the estimate when published so that users of the statistics can use the data correctly.

CHAPTER 3
THEORIES OF THE SAMPLE SURVEY

1. Fundamental Theorem of the Sample Survey

The most fundamental of the theories of the sample survey is the following theorem. "Whatever the distribution of the population is, the distribution of a sample mean X, as the size of the sample becomes larger (50 or more), approaches a normal distribution in which the mean value is the population mean and the variance is $\frac{N-n}{N}\ \frac{\sigma^2}{n}$ (σ^2 being the population variance)."

Considering, therefore, that the sample mean $\bar{X}$ follows normal distribution, following characteristics of the normal distribution can be utilized.



The probability of $\bar{X}$'s falling between $\bar{x} - d$ and $\bar{x} + d$ is 67%.

The probability of $\bar{X}$'s falling between $\bar{x} - 2d$ and $\bar{x} + 2d$ is 95%.

The probability of $\bar{X}$'s falling between $\bar{x} - 3d$ and $\bar{x} + 3d$ is more than 99%.

The meaning of the precision of estimate, 3% for example, in estimating the population mean on the basis of the sample mean is as follows.

(Precision Coefficient of Variance = $d/\bar{X}$, $0.03 = d/\bar{X}$, and therefore, $d = 0.03\bar{X}$)

"To say that the population mean $\bar{X}$ is between $\bar{X} - 0.03\bar{X}$ and $\bar{X} + 0.03\bar{X}$ holds true in 67 times out of 100 times of declaration. To say that the population mean $\bar{X}$ is between $\bar{X} - 2 \times 0.03\bar{X}$ and $\bar{X} + 2 \times 0.03\bar{X}$ holds true in 95 times out of 100 times. To say that the population mean is between $\bar{X} - 3 \times 0.03\bar{X}$ and $\bar{X} + 3 \times 0.03\bar{X}$ holds true in 99 times out of 100 times." The expression "hold true in a certain number of times out of 100 times" is used here on the assumption that an estimates has been calculated for each sample, of NCn different samples. In reality, however, the sampling is carried out only once.

The same discussion holds true of the estimation of the population total, because the population total is simply obtained by multiplying the population mean by the size of population N.

2. Expectation

Suppose that a variable X takes values of $x_1, x_2, \ldots\ldots, x_N$ and probability of taking these values are $P_1, P_2, \ldots\ldots, P_N$, $(P_1 + P_2 + \ldots + P_N = 1)$. Then the expectation of X, E(X), is computed by the following formula.

$$E(X) = p_1 x_1 + p_2 x_2 + \cdots\cdots\cdots\cdots + p_N x_N$$
$$= \sum_{i=1}^{N} p_i x_i$$

Let's here compute the expectation of a sample value X, when only one sampling unit was sampled. The value which X can assume is any one of the values of sampling units in the population, that is, $x_1, x_2 \ldots\ldots, x_N$. The probability that X assumes each of these values is $\frac{1}{N}$, since all sampling units are selected at random from the population. The expectation of X of a sample value is therefore given as follows.

$$E(X) = \frac{1}{N} x_1 + \frac{1}{N} x_2 + \cdots\cdots\cdots\cdots + \frac{1}{N} x_N$$
$$= \frac{1}{N} \sum_{i=1}^{N} x_i = \bar{x}$$

It is therefore the population mean.

There are the following theorems regarding the expectation.

$$E(X+Y) = E(X) + E(Y)$$

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

$$E(CX) = CE(X)$$

$$E(C) = C$$

where,

C = constant

3. Expectation and Variance of the Sample Mean

In Section 1 of this chapter it was mentioned that the distribution of a sample mean X approaches a normal distribution with mean $\bar{x}$ and variance $\frac{N-n}{N}\frac{\sigma^2}{n}$. To prove the fact of approaching a normal distribution requires knowledge of advanced mathematics, and will not be discussed here. Suffice it to demonstrate here that expectation of sample mean $\bar{X}$ is $\bar{x}$ and that the variance (variance defined by expectation E, that is $E\{\bar{X} - E(\bar{X})\}^2$ is $\frac{N-n}{N}\frac{\sigma^2}{n}$

Since the sample mean variance $\frac{N-n}{N}\frac{\sigma^2}{n}$ is the most fundamental of variances of all sorts of estimates, it will be discussed here in detail.

(1) Expectation of the Sample Mean

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \cdots\cdots + X_n}{n}\right)$$

$$= \frac{1}{n}\left\{E(X_1) + E(X_2) + \cdots\cdots + E(X_n)\right\}$$

$$= \frac{1}{n}\left\{\sum_{i=1}^{N}\frac{1}{N}X_i + \sum_{i=1}^{N}\frac{1}{N}X_i + \quad + \sum_{i=1}^{N}\frac{1}{N}X_i\right\}$$

$$= \frac{1}{n}\left\{n \cdot \bar{x}\right.$$

$$= \bar{x}$$

It now can be said that the expectation of the sample mean is the population mean. It may also be said that the sample mean is an unbiased estimate of the population mean.

(2) Variance of the Sample Mean

The variance of a variable X is defined by using expectation E as

$$V(x) = E\left\{ x - E(X) \right\}^2$$

Hence, the variance of sample mean $\bar{X}$ is as follows.

$$\begin{aligned}
V(\bar{x}) &= E\left\{ \bar{x} - E(\bar{x}) \right\}^2 \\
&= E\left\{ \bar{x} - \bar{z} \right\}^2 \\
&= E\left\{ \bar{x}^2 - 2\bar{x}\bar{x} + \bar{z}^2 \right\} \\
&= E(\bar{x}^2) - 2E(\bar{x})\bar{z} + \bar{z}^2 \\
&= E(\bar{x}^2) - 2\bar{z}^2 + \bar{z}^2 \\
&= \underline{E(\bar{x}^2) - \bar{z}^2} \\
E(\bar{x}^2) &= E\left\{ \left(\frac{\sum_{i=1}^{n} x_i}{n} \right)^2 \right\} \\
&= \frac{1}{n^2} E\left\{ \left(\sum_{i=1}^{n} x_i \right)^2 \right\} \\
&= \frac{1}{n^2} E\left\{ (x_1 + x_2 + \cdots + x_n)(x_1 + x_2 + \cdots + x_n) \right\} \\
&= \frac{1}{n^2} E\left\{ \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} \sum_{j \neq i}^{n} x_i x_j \right\} \\
&= \frac{1}{n^2} \left\{ \underline{\sum_{i=1}^{n} E(x_i^2)} + \underline{\sum_{i=1}^{n} \sum_{j \neq i}^{n} E(x_i x_j)} \right\}
\end{aligned}$$

HERE,

$$E(x_i^2) = \sum_{k=1}^{N} \frac{1}{N} x_k^2 = \frac{1}{N} \sum_{k=1}^{N} x_k^2$$

$$E(x_i x_j) = \sum_{k=1}^{N} \sum_{l \neq k}^{N} \frac{1}{N(N-1)} x_k x_l = \frac{1}{N(N-1)} \sum_{k=1}^{N} \sum_{l \neq k}^{N} x_k x_l$$

THEREFORE,

$$\begin{aligned}
E(\bar{x}^2) &= \frac{1}{n^2} \left\{ \sum_{i=1}^{n} \left(\frac{1}{N} \sum_{k=1}^{N} x_k^2 \right) + \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left(\frac{1}{N(N-1)} \sum_{k=1}^{N} \sum_{l \neq k}^{N} x_k x_l \right) \right\} \\
&= \frac{1}{n^2} \left\{ \frac{n}{N} \sum_{k=1}^{N} x_k^2 + \underline{\frac{n(n-1)}{N(N-1)} \sum_{k=1}^{N} \sum_{l \neq k}^{N} x_k x_l} \right\}
\end{aligned}$$

HERE,

$$\sum_{k=1}^{N} \sum_{l \neq k}^{N} x_k x_l = \left(\sum_{k=1}^{N} x_k \right)^2 - \sum_{k=1}^{N} x_k^2 = N^2 \bar{z}^2 - \sum_{k=1}^{N} x_k^2$$

THEREFORE,

$$E(\bar{x}^2) = \frac{1}{n^2}\left\{\frac{n}{N}\sum_{k=1}^{N} x_k^2 + \frac{n(n-1)}{N(N-1)}\left(N^2\bar{x}^2 - \sum_{k=1}^{N} x_k^2\right)\right\}$$

$$= \left\{\frac{1}{nN} - \frac{(n-1)}{nN(N-1)}\right\}\sum_{k=1}^{N} x_k^2 + \frac{(n-1)N}{n(N-1)}\bar{x}^2$$

HERE,

$$\frac{1}{nN} - \frac{(n-1)}{nN(N-1)} = \frac{(N-1)-(n-1)}{nN(N-1)} = \frac{(N-n)}{nN(N-1)}$$

THEREFORE,

$$E(\bar{x}^2) = \frac{(N-n)}{nN(N-1)}\sum_{k=1}^{N} x_k^2 + \frac{(n-1)N}{n(N-1)}\bar{x}^2$$

THEN,

$$V(\bar{x}) = E(\bar{x}^2) - \bar{x}^2$$

$$= \frac{(N-n)}{nN(N-1)}\sum_{k=1}^{N} x_k^2 + \frac{(n-1)N}{n(N-1)}\bar{x}^2 - \bar{x}^2$$

$$= \frac{(N-n)}{nN(N-1)}\sum_{k=1}^{N} x_k^2 + \left(\frac{(n-1)N}{n(N-1)} - 1\right)\bar{x}^2$$

HERE,

$$\frac{(n-1)N}{n(N-1)} - 1 = \frac{(n-1)N - n(N-1)}{n(N-1)} = -\frac{(N-n)}{n(N-1)}$$

THEREFORE,

$$V(\bar{x}) = \frac{(N-n)}{nN(N-1)}\sum_{k=1}^{N} x_k^2 - \frac{(N-n)}{n(N-1)}\bar{x}^2$$

$$= \frac{N-n}{N}\,\frac{1}{n}\left\{\frac{\sum_{k=1}^{N} x_k^2 - N\bar{x}^2}{N-1}\right\}$$

HERE,

$$\frac{\sum_{k=1}^{N} x_k^2 - N\bar{x}^2}{N-1} = \sigma^2$$

THEREFORE,

$$V(\bar{x}) = \frac{N-n}{N}\,\frac{\sigma^2}{n}$$

4. Precision of Estimate

When the population mean $\bar{x}$ is estimated on the basis of the sample mean $\bar{X}$, $\bar{X}$ is called an estimate. The variance of $\bar{X}$ is written as $V(\bar{X})$. (In the foregoing section it has been proved that $V(\bar{X}) = \frac{N-n}{N}\frac{\sigma^2}{n}$.) The square root of $V(\bar{X})$ is the standard error. This standard error is the sampling error which has been explained before in this booklet.

However, the standard error varies with the absolute value of an estimate and the unit used for measurement. The relative amount of standard error to an estimate is not clear. The ratio of the standard error $\sqrt{V(\bar{X})}$ to the expectation of estimate $E(\bar{X})$, that is, $\sqrt{V(\bar{X})}/E(\bar{X})$ $\left(= \sqrt{\frac{N-n}{N}\frac{1}{n}}\,\sigma/\bar{x} \right)$, is called the coefficient of variance and written as $CV(\bar{X})$. This coefficient of variance of an estimate is called the precision of estimate.

Precision of Estimate = Coefficient of Variance of the Estimate $CV(\bar{X})$

$$= \sqrt{V(\bar{x})}\,/E(\bar{x}) = \sqrt{\frac{N-n}{N}\frac{1}{n}}\cdot\frac{\sigma}{\bar{x}}$$

The formula of the precision of estimate may seem to be complicated, but it can be rewritten in more simple form.

$$CV(\bar{X}) = \sqrt{\frac{N-n}{N}}\cdot\frac{1}{\sqrt{n}}\cdot\frac{\sigma}{\bar{x}}$$

When the size of population N is by far larger than the size of sample n, $\frac{N-n}{N}$ approximates 1.

$$\frac{N-n}{N} = 1 - \frac{n}{N} = 1 \quad \left(\frac{n}{N} \longrightarrow 0\right)$$

The $\frac{\sigma}{\bar{x}}$ indicates the relative value of the population standard deviation to the population mean, and is called a coefficient of variance of the population, being written as C_x.

$$C_x = \frac{\sigma}{\bar{x}}$$

Therefore,

$$CV(\bar{X}) \fallingdotseq \frac{C_x}{\sqrt{n}}$$

The precision of estimate $\bar{X}$ approximates the coefficient of variance of the population C_x divided by the square root of the size of sample.

5. Simple Estimation

The estimation of population mean $\bar{x}$ on the basis of sample mean $\bar{X}$ and the estimation of population total $T = \sum_{i=1}^{N} x_i = N\bar{x}$ on the basis of $N\bar{X}$ are called the method of simple estimation. Their estimation formula and variance formula are as follows:

(1) Simple estimation of the population mean

Estimation $\quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Variance $\quad V(\bar{x}) = \frac{N-n}{N} \frac{\sigma^2}{n}$

(2) Simple estimation of the population total

Estimation $\quad \hat{T} = \frac{N}{n} \sum_{i=1}^{n} x_i$

Variance $\quad V(\hat{T}) = N^2 \frac{N-n}{N} \frac{\sigma^2}{n}$

6. Ratio Estimation

Ratio estimation is a method by which auxiliary variables which are closely correlated to the variable to be estimated is used to achieve higher precision of the estimate. Application of the ratio estimation method requires availability of the population mean or the population total of the auxiliary variable.

(1) Ratio estimation of population mean

Estimation $\quad \bar{X} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i} \bar{y}$

Variance $\quad V(\bar{x}) = \frac{N-n}{N} \frac{1}{n} (\sigma_x^2 - 2R\rho\sigma_x\sigma_y + R^2\sigma_y^2)$

where, $R = \bar{x}/\bar{y}$

(2) Ratio estimation of population total

Estimation $$\hat{T} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i} N\bar{y}$$

Variance $$V(\hat{T}) = M^2 \frac{N-n}{N} \frac{1}{n} (\sigma_x^2 - 2R\rho\sigma_x\sigma_y + R^2\sigma_y^2)$$

WHERE, $R = \bar{x}/\bar{y}$

7. Stratification

When the population is stratified, estimation is made in each stratum. When estimation of the population mean is required, the weighted arithmetic mean of the population mean by stratum with weights of the respective sizes of strata gives an estimate of the population mean. An estimate of each stratum is simply added, when estimation of the population total is required. The variance of the population mean estimate is obtained by the weighted arithmetic mean of the variance of each stratum with a weight of the size of each stratum. The variance of the population total estimate is given by simply adding the variance of each stratum. Here is given an example of the stratified simple estimation of a population total.

Estimation $$\hat{T} = \sum_{i=1}^{\ell} \hat{T}_i = \sum_{i=1}^{\ell} \frac{N_i}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

Variance

$$V(\hat{T}) = \sum_{i=1}^{\ell} V(\hat{T}_i) = \sum_{i=1}^{\ell} N_i^2 \frac{N_i - n_i}{N_i} \frac{\sigma_i^2}{n_i}$$

where,

$\hat{T}_i$ = Total estimate of ith stratum

N_i = size of population of ith stratum

n_i = size of sample of ith stratum

X_{ij} = value of jth sampling unit in ith stratum

σ_i^2 = population variance of ith stratum

ℓ = number of strata

8. Multi-Stage Sampling

When multi-stage sampling is used, the mean or total of each primary sampling unit selected in the sample is estimated on the basis of the sub sampling units, and then the mean or total of the population is estimated on the basis of estimated data of those primary sampling units. The variance of a multi-stage estimate comprises the variance between the primary sampling units and the variance within the primary sampling unit (variance between the sub sampling units).

Hereunder is given an example of the two-stage simple estimation of the population total.

Estimation $$\hat{T} = \frac{M}{m} \sum_{i=1}^{m} \frac{N_i}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

Variance $$V(\hat{T}) = M^2 \frac{M-m}{M} \frac{1}{m} \sigma_b^2 + \frac{M}{m} \sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \frac{\sigma_i^2}{n_i}$$

where, $$\sigma_b^2 = \frac{\sum_{i=1}^{M} (\bar{X}_i - \bar{\bar{X}})^2}{M-1}$$

$$\sigma_i^2 = \frac{\sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i)^2}{N_i - 1}$$

M = Number of primary sampling units of the population

m = Number of primary sampling units of the sample

Ni = Number of sub sampling units in ith primary sampling unit of the population

ni = Number of sub sampling units in ith primary sampling unit of the sample

X_{ij} = Value of variable of jth sub sampling unit of ith primary sampling unit of the population

X_{ij} = Value of variable of jth sub sampling unit of ith primary sampling unit of the sample

$\bar{X}_i$ = Mean of variable within ith primary sampling unit of the population

$\bar{\bar{X}}$ = Mean of population

9. Sampling with Probability Proportional to Size of Unit

As a variation of two-stage sampling there is the method of sampling with probability proportional to size of unit. In two-stage sampling the calculation of estimate requires much work, because the estimates must be calculated within each primary sampling unit. There are two ways to avoid this inconvenience. One is a method by which to allocate the sub sampling units in proportion to the number of sub sampling units within each primary sampling unit, although both primary and sub sampling units are selected at random. When this method is used, the ratio $fi = \frac{ni}{Ni}$, that is, the rate of sampling of sub sampling units, is the same in all primary sampling units (= f). Accordingly, multiplying the sum total of all sample data of sub sampling units by a reciprocal $\frac{M}{m}\frac{1}{f}$ of the total sampling ratio $\frac{m}{M}f$ gives the estimate of the population total.

Another method is the sampling with probability proportional to size of unit. The process by this method consists in selecting primary sampling units according to probability proportional to the number of sub sampling units within each primary sampling unit or the size of the primary sampling unit, and selecting a fixed number of sub sampling units as sample from each selected primary sampling unit. The probability that a primary sampling unit may be selected as sample is $\frac{Ni}{N}m$, and the probability that a sub sampling unit may be selected as sample is $\frac{n}{Ni}$. It follows that the probability that a sub sampling unit may be selected as sample from the whole population is $\frac{Ni}{N}m \times \frac{n}{Ni} = \frac{mn}{N}$. It is a fixed value. Accordingly, multiplying the total of all selected sub sampling units in the sample by $\frac{N}{mn}$ gives the estimate of the population total.

As sub sampling units are selected as sample at the same rate through out the whole population by either method, it is not necessary to multiply the estimate in each primary sampling unit by the weight of the size of each primary sampling unit. They are thus called a self-weighted sample.

Hereunder are shown the method of sampling of the sampling with probability proportional to size of unit, formula of estimation and formula of variance.

(1) Method of Sampling

i	Ni	Accumulation of Ni
1	21	21
2	6	27
(3)	33	(60)
4	13	73
5	2	75
6	28	103
7	11	114
(8)	45	(159)
9	19	178
10	31	209
⋮	⋮	⋮
100	15	2,053
M		$N = \sum_{i=1}^{M} N_i$

Sample size m = 25

$f = \frac{N}{m} = \frac{2,053}{25}$

$= 82.12 \fallingdotseq 82$

A random start is selected as 58.

Random Start = 58

Second Sample = 58 + 82 = 140

Third Sample = 140 + 82 = 222

Fourth Sample = 222 + 82 = 304

(2) Formula of Estimation and Formula of Variance

Estimation $\hat{\bar{Y}} = \frac{N}{m\,n} \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}$

where,

N = Total number of sub sampling unit of population

m = Size of sample of primary sampling units

n = Size of sample of sub sampling units in each primary sampling unit

Variance $V(\hat{\bar{Y}}) = M^2 \frac{M-m}{M} \frac{\hat{\sigma}_b^2}{m}$

where, $\hat{\sigma}_b^2 = \frac{\sum_{i=1}^{m} \bar{X}_i^2 - \frac{\left(\sum_{i=1}^{m} \bar{X}_i\right)^2}{m}}{m-1}$

M = Number of primary sampling units of the population

10. Estimation of Estimate Variance by Sample Data and the Unbiased Variance

So far the population variance was used in all the formula of variance of various methods of estimation except for the last one of the sampling with probability proportional to size of unit. Actually, variance of estimate is to be calculated using sample data then the formula of variance must be given some modification. In the case of the sampling with probability proportional to size of unit, use of sample data makes the formula of variance very simple. In this case only, therefore, the formula of variance using sample data was shown.

(1) Calculation of Variance of Simple Estimate by Sample Data

As one of the simplest examples, discussion will be made of the calculation by sample data of the variance of estimate of the population mean obtained by the simple estimation without stratification, single-stage sampling.

The formula for estimation of the population mean by simple estimation and that of variance were written before as follows:

$$\text{Estimation} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\text{Variance} \quad V(\bar{X}) = \frac{N - n}{N} \quad \frac{\sigma^2}{n}$$

$$\text{where,} \quad \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N - 1}$$

The symbol σ^2 in this formula of variance signifies the population variance. The information on x_i's are not actually available. Only available information is sample data Xi's.

What formula will then be an appropriate one to estimate the population variance by sample data? The fact is that the sample variance is the estimate of the population variance. This fact will be proved later in (2).

Therefore, the formula of variance estimation by sample data will be as follows:

$$\hat{V}(\bar{X}) = \frac{N - n}{N} \frac{s^2}{n}$$

where,

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}$$

(2) Unbiased Variance

We will now prove that the sample variance $s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}$ is an unbiased (expectation of an estimate being equal to the value to be estimated) estimate of $\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N - 1}$

The problem which presents itself here is that the denominators of both variances are n − 1 and N − 1, respectively. If the variances are defined by denominators n and N, respectively, the sample variance is no more an unbiased estimate of the population variance. It is because of this characteristic that the variances divided by n − 1 and N − 1, respectively, are called an unbiased variance.

[Proof]

$$E(s^2) = E\left\{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}\right\} = E\left\{\frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n - 1}\right\}$$

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n} E(X_i^2) - n\,E(\bar{X}^2)\right\}$$

The formula of variance of the sample mean is made use of here.

$$V(\bar{X}) = E\left\{\bar{X} - E(\bar{X})\right\}^2 = E(\bar{X})^2 - \bar{X}^2$$

AND, $V(\bar{X}) = \frac{N - n}{N} \frac{\sigma^2}{n}$

THEN, $E(\bar{X})^2 = \bar{X}^2 + \frac{N - n}{N} \frac{\sigma^2}{n}$

Therefore,

$$E(s^2) = \frac{1}{n-1}\left\{\sum_{i=1}^{n} E(X_i^2) - n\left(\bar{X}^2 + \frac{N-n}{N}\frac{\sigma^2}{n}\right)\right\}$$

HERE, $\sum_{i=1}^{n} E(X_i^2) = \sum_{i=1}^{n}\left(\sum_{i=1}^{N} \frac{1}{N} X_i^2\right)$

$$= n \cdot \frac{1}{N} \sum_{i=1}^{N} X_i^2$$

Therefore,

$$E(s^2) = \frac{n}{n-1}\left\{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2 - \frac{N-n}{Nn}\sigma^2\right\}$$

HERE, $\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2 = \frac{N-1}{N}\left(\frac{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}{N-1}\right)$

$$= \frac{N-1}{N}\sigma^2$$

Therefore,

$$E(s^2) = \frac{n}{n-1}\left\{\frac{N-1}{N} - \frac{N-n}{Nn}\right\}\sigma^2$$

$$= \frac{n}{n-1}\quad\frac{n(N-1)-(N-n)}{Nn}\quad\sigma^2$$

$$= \frac{n}{n-1}\quad\frac{N(n-1)}{Nn}\sigma^2 = \sigma^2$$

Statistics & Information Department
Ministry of Agriculture & Forestry
Government of Japan