[PROC. OF JSCE No. 347/IV-1 (Infrastructure Planning and Management) July 1984]

THE EFFECT OF CLUSTER SAMPLING ON THE DATA PRECISION OF PERSON TRIP SURVEY

*By Koichi YAMAGATA**

Person trip survey is a fundamental data source for transportation studies, but its characteristics in data precision is rather vague. The author points out that the survey makes up the sample by cluster sampling and that the effect of the sampling method on data precision cannot be ignored. Hitherto, the standard error of estimates obtained by the survey has been estimated on the assumption that the sample from the survey is random. This assumtion was examined by means of a re-sampling experiment and shown to be incorrect. Then, a method to estimate the standard error was developed, taking the effect of cluster sampling into consideration. Finally, the advantages of using a household and a person as a sampling unit were compared from the viewpoints of data precision and survey cost.

1. INTRODUCTION

Person trip survey is recognized as the main means to collect the data for comprehensive transportation studies in most metropolitan areas having a population of more than 300 thousand in Japan. As person trip survey is a kind of sampling survey, the data contain sampling error whose magnitude is affected by the sampling ratio, sampling unit and so on. The data of trip-volume obtained by person trip survey are used for determining the problems and travel characteristics in the study area and building appropriate models to make prediction. Then, the error of data is brought into the stage of analysis and prediction. Accordingly, it is important to know the characteristics of the data collected by person trip survey for the following reasons : to design the survey in a precise way so that it satisfies the requirement of accuracy from the viewpoint of analysis and forecast ; to design study process to correspond in precision with available data ; and to evaluate the confidence of forecasts as a information for decision making[9].

Person trip survey is composed of home interview survey, external-cordon survey and transport-enterprise interview survey. Each survey collects subsets of the sample in the study area. In many cases of person trip surveys in Japan, more than 95 percent of the trips in the sample are obtained by home interview survey. Thus, the accuracy of the data of person trip survey is thougth to depend on the home interview survey.

There are few reports on the data precision of person trip survey. Works in this field have concentrated on developing practical information to execute the survey. The Chicago Area Transportation Study showed in a table the relative error in estimating a trip-volume in relation to the value of the trip-volume to be estimated[1]. The National Committee on Urban Transportation illustrated curves which showed the

* Professor of Construction Engineering, Ibaragi University
(4-12-1, Nakanarusawa-cho, Hitachi-shi, 316)

relationship between the magnitude of relative error in estimating a trip-volume and the sampling ratio, by the value of the trip-volume[2]. Nevertheless, both reports provided little information on the theoretical background to deduce the figure or the table. However, it appears that the following assumptions were used : (1) a population of trips in the study area is assumed, and a sample of which element is a trip is drawn by the home interview survey; (2) the elements in the sample, i. e., trips, are mutually independent, that is, the sample is random; (3) the size of the population is infinite. The Tokyo Area Transportation Study changed the third assumption so that the size of the population is finite, supposing that the travel flow is rather stationary with days[3].

However, home interview survey observes the behaviour of individual persons, and trips in the sample are collected as a chain of sequential trips being made by one person in a day. This means that the elements of sample, i.e., trips are collected by cluster sampling in which the sampling unit is a person. Furthermore, the person whose behaviour is observed is also drawn by cluster sampling in which the sampling unit is a household. Therefore, home interview survey draws clusters of trips which is bundled both by a person and a household. Sampling theory shows that the estimator from cluster sampling has a larger variance than that from simple random sampling, when the measures of elements are homogeneous in a cluster.

This paper aims at developing a method to estimate the data precision of person trip survey from the viewpoint that the survey draws the sample by cluster sampling. The following three subjects were studied :

Firstly, the assumption that the sample is random was examined to determine if it causes significant error in evaluating the data precision of person trip survey. The standard error of estimator by an actual person trip survey was estimated through a re-sampling experiment, and this standard error was compared with the one which was theoretically deduced from the assumption. The result showed that the assumption is inaccurate, in general, to estimate the data precision of actual survey.

Secondly, in this research it was found that the standard error by cluster sampling is proportional to that by simple random sampling. Then, a method for estimating the standard error of data collected by person trip survey was developed according to this finding. The availability of this method was confirmed by statistical test.

Thirdly, as it is impossible to draw individual trips directly by simple random sampling, the efficiencies of a person and a household as a sampling unit were compared from the viewpoint of data precision and cost.

2. STANDARD ERROR OF ESTIMATOR BY CLUSTER SAMPLING

Cluster sampling is a method which makes up a sample in such a way that after dividing the population into subsets, subsets are drawn by simple random sampling, then the sample is composed of all the elements belonging to the drawn subsets. Here, the element whose characteristics are observed is called observation unit, and the subset is called sampling unit. Observation unit equals sampling unit in the case of simple random sampling. When we think of the population of trips in the survey area, the population can be divided into subsets so that a chain of sequential trips made by one traveler in a day corresponds to one subset. Then, to draw a person and to observe the behaviour of the person means to select a subset of trips in the population. Home interview survey makes up the sample by collecting these subset of trips obtained through observation of the drawn persons. Here, the sampling unit is a person.

A household is another subset of trips in the population. As a household is a set of persons, a household also makes a set of trips. Then, to draw a household and to observe the behaviour of all members belonging to the drawn household means to draw the subset of trips in the population, that is, to employ cluster sampling. Here, the sampling unit is a household.

Through the discussion above, it is pointed out that home interview survey collects the sample by cluster sampling. This fact should be taken into considertion, when the precision of estimator is discussed.

Let the size of the population be N, and suppose that each element in the population has the measure y_i respectively. Now, we are concerned with the mean of y_i in the population, which is denoted by $\overline{Y}$. When $\overline{Y}$ is estimated from a sample in which elements are drawn by simple random sampling, the estimator for $\overline{Y}$ is,

$$\overline{y}_r=\sum_{i=1}^{n} y_i/n \quad \cdots\cdots (1)$$

where, n : the size of the sample, r denotes the sample to be random.

The sample variance of $\overline{y}_r$ around $\overline{Y}$ is,

$$V_r=(S^2/n)(N-n)/N \quad \cdots\cdots (2)$$

where, S^2 : population variance of y_i.

In the case of cluster sampling, we suppose that the population is composed of L subsets, i.e., clusters, which contain, respectively, M_j elements. That is, $N=\sum_{j=1}^{L} M_j$.
The measure of k th element in j th cluster is denoted by y_{ik}.

Some kinds of estimators for $\overline{Y}$ corresponding to cluster sampling have been proposed. One of the most popular estimators is

$$\overline{y}_{cl}=\sum_{j=1}^{L} y_j/\sum_{j=1}^{l} M_j, \quad y_j=\sum_{k=1}^{Mj} y_{jk} \quad \cdots\cdots (3)$$

y_{cl} is a ratio estimator, and the variance of $\overline{y}_{cl}$ is affected by the correlation coefficient between y_j and M_j. Subscript $c\ell$ denotes cluster sampling. The variance of estimator $\overline{y}_{cl}$ is shown below, when the size of each cluster is equal, that is, $M_j=\overline{M}=N/L$.

$$V_{cl}=\frac{L-l}{L}\frac{1}{l\overline{M}}S^2\left\{\frac{\overline{M}\,l-1}{(L-1)\overline{M}}-\frac{L(\overline{M}-1)}{L-1}\rho\right\} \quad \cdots\cdots (4)$$

where

$$\rho=\frac{\sum_{j=1}^{L}\sum_{k+k'}^{\overline{M}}(y_{jk}-\overline{Y})(y_{jk}-\overline{Y})}{L\overline{M}(\overline{M}-1)S^2} \quad \cdots\cdots (5)$$

ρ is called intra-class correlation coefficient. When we consider Eq. (5), we know that ρ is positive when most of y_{jk} values in a cluster are larger $\overline{Y}$, or when most of y_{jk} values are smaller than $\overline{Y}$, and negative when the values in each cluster map unbiasly around $\overline{Y}$. In other words, ρ is positive when cluster is composed of homogeneous elements, and negative when it is heterogeneously composed[4].

Now, the ratio of the variance of estimator by cluster sampling against that by simple random sampling has been expressed as Eq. (6) when the sizes of both samples equal[5].

$$V_{cl}/V_r=1+(\overline{M}-1)\rho \quad \cdots\cdots (6)$$

Eq. (6) teaches that the estimator obtained by cluster sampling is less precise than that by simple random sampling, when ρ is positive.

It is not plausible to assume that the numbers of members in a household or the numbers of trips made by a person are equal. When the sizes of clusters are not equal, formulations corresponding to Eqs. (4) ~ (6) has not been developed, except when each cluster is drawn with the probability in proportion to its size[5]. It is one of the purposes of this paper to determine if Eqs. (4) ~ (6) might be meaningful as approximations, when the sizes of clusters are not equal.

From the consideration of the characteristics of cluster sampling, the following points are clarified. Firstly, the precision of trip-volume estimated by home interview survey might be different from the one which is theoretically derived on the assumption that trip in the sample is random. It is empirically

known that the value of ρ tends to be positive in many surveys, because the cluster is often defined as a set of elements which share some homogeneousness in some sense. Then, the former precision tends to be inferior to the latter one, and the magnitude of inferiority depends on ρ. Secondly, it is possibly foreseen that the variance of estimator by home interview survey could be expressed in the form of product of the variance of random sampling by a certain value which is a function of ρ and $\overline{M}$. Thirdly, the second term of Eq. (6) describes the trade-off relationship between cost and precision. That is, when the size of cluster $\overline{M}$ is large, the cost of survey usually decreases, while Eq. (6) shows that the larger $\overline{M}$ is, the less precise is the estimator

3. DIFFERENCE OF PRECISION BETWEEN THE ESTIMATOR BY CLUSTER SAMPLING AND THE ONE BY SIMPLE RANDOM SAMPLING

This chapter addresses to determine the difference of precision between the estimator by cluster sampling and that by simple random sampling. When we consider the trips made by the same person, we find that the measure of a considered characteristic on these trips tend to be homogeneous. For instance, the mode of trip is the characteristic in which we are interested. We know that some persons tend to use a car frequently and others rarely. Now, we look at j th cluster, which contains M_j trips i.e., y_{jk}, $k=1\cdots M_j$. Suppose y_{jk} has the mark C when the trip corresponding to y_{jk} is a car-using trip, and N when it is otherwise. Observing all the marks of y_{jk} in j th cluster, we see that the most of marks on y_{jk} are C in the cases that j th person uses a car frequently, and that marks on y_{jk} are almost uniformly N in the other cases. That is, the marks tend to be unified into C or N in each cluster. This kind of homogeneousness can be found in characteristics regarding origin, purpose and mode of trip.

The homogeneousness of behaviour in a house hold is expected regarding above items, especially when a person is accompanied by members of his family. Thus, the homogeneousness in the cluster makes the variance of the estimator larger than that by simple random sampling in two steps, that is, the step in which trips are bundled by a person, and the step in which persons are bundled by a household.

(1) Calculation of the standard error of estimator by cluster sampling by means of a re-sampling experiment

A sample of trips which is the actual result of home interview survey is postulated to be a supposed population of trips. Hereafter, this is called the supposed population. Samples are drawn from the supposed population through a method that simulates an actual survey procedure. This procedure is called re-sampling experiment. An estimate for an item from the supposed population such as trip generation of a zone by purpose, is postulated to be a true value for the estimates from these samples. Then, the estimates from the samples map around this supposed true value. So, when we draw many samples through the fixed sampling rule, we can get standard error by aggregating the deviation of estimates from each sample, for a defined case in terms of sample ratio and sampling unit. When the sampling unit is a person, the standard error includes the increase of variance due to clustering trips by a person. Hence, this standard error is understood to be an experimental value for the square root of variance by cluster sampling, i.e., $\sqrt{V_{cl}}$ in Eq. (4).

On the other hand, the standard error of estimator by simple random sampling is easily calculated by substituting the size, variance and true value of the supposed population and the size of the sample corresponding to the defined cluster sampling into Eq. (2). Then, the difference between these two kinds of standard errors reflects the effect of cluster sampling.

The Tokyo Metropolitan Area Person Trip Survev completed in 1968 was employed as the supposed population. This survey covered the Tokyo Metropolis, Kanagawa Prefecture, and the most of Chiba and Saitama Prefectures which had a population of 19 million people. Two percent of the households in

this area were drawn randomly, and all members belonging to the drawn households were selected as respondents. Therefore, the supposed population contained the records of approximately 100 thousand households, 320 thousand persons and 800 thousand trips.

In order to simulate an actual home interview procedure, the drawn clusters, i.e., a person or a household were specified by pseudo-random number generating programme, and all the records of drawn cluster were transferred to make the sample. In order to examine the difference between cluster samplings in which sampling unit is a person and a household, two series of re-sampling experiments corresponding to these sampling unit were provided. In each series, re-sampling experiment was completed at four levels of sampling ratio, that is, 2%, 5%, 10% and 20% so that the effect of sampling ratio could be found. Hereafter, i denotes sampling unit and j denotes sampling ratio. The combination of sampling unit i and sampling ratio j is called case in this paper. That is, eight cases of sampling experiment were carried out in each series. In each case, five samples which were mutually stochastic independent were drawn from the supposed population.

As person trip survey is main data source of trip-volume in a transportation study, it is important to know the effect of cluster sampling on the data of trip-volume. The comparison was accomplished for the following items of trip-volume : (a) trip generation and attraction of zones by purpose; (b) trip generation and attraction of zones by mode; (c) trips between zones. The size of zone is approximately a Ward in the central part of the study area, and a city or town in the outskirts.

Now, we focus on an item of trip-volume, for example, trip generation of zone l for a certain purpose, and denote it by x_l. The true value of x_l in the supposed population is assumed to be μ_l, and an estimate for x_l from the k th sample of i th sampling unit and j th sampling ratio is denoted by x_l^{ijk}. Then, x_l^{ijk} is a probability variable and its expectation is μ_l.

One of the purposes of the re-sampling experiment is to determine the standard error of x_l^{ijk} around μ_l, that is, σ_l^{ij}. Confidence interval for estimating x_l through ij th sampling method is expressed by $(x_l^{ijk}-z_{1-\alpha/2}\,\sigma_l^{ij},\ x_l^{ijk}-z_{1-\alpha/2}\,\sigma_l^{ij})$, when confidence coefficient is $1-\alpha$. $z_{1-\alpha/2}$ denotes the point at which the intergration of standard normal distribution from $-\infty$ is $1-\alpha/2$. Therefore, the precision of estimation is indicated by standard error.

The standard error of estimator from ij th case of sampling σ_l^{ij} is estimated as SE_l^{ij} in Eq. (6).

$$SE_l^{ij}=\sqrt{\sum_{K=1}^{m}(x_l^{ijk}-\mu_l)^2/m} \quad \cdots\cdots (7)$$

where, m : the number of samples.

SE_l^{ij} is also a probability variable and fluctuates remarkably when m is small. Notice that

$$\sum_{j=1}^{m}\{(x_l^{ijk}-\mu_l)/\sigma_l^{ij}\}^2=m(SE_l^{ij}/\sigma_l^{ij})^2 \quad \cdots\cdots (8)$$

has chi-spuare distribution whose degree of freedom is m.

In the re-sampling experiment, m is not enough to estimate σ_l^{ij} accurately. In order to supplement the shortcoming, zones are gathered together into some categories. As the value of μ_l affects the value of SE_l^{ij}, the zones whose μ_l are equal have same property from the stochastic point of view. Zones are sorted according to their values so that each category is composed of zones whose μ_l are approximately equal. Then, standard error of category SE_l^{ij} is estimated instead of that of SE_l^{ij}. Notice that subscript is converted from l to c which denotes number of category.

$$SE_c^{ij}=\sqrt{\sum_{l\in c}\sum_{K=1}^{m}(x_l^{ijk}-\mu_l)^2/m^*} \quad \cdots\cdots (9)$$

where, m^* : product of m by the number of zones which belonge to category c.

m^* varies with categories and items, but 30~60 measures are used to calculate SE_c^{ij}.

The standard error to estimate x_l from the simple random sample whose size is equal to ij th case of the cluster sample, namely TSE_l^{ij}, is calculated from Eq. (10)

$$TSE_l^{ij}=\sqrt{\mu_l(N-\mu_l)/n_{ij}(N-n_{ij})/(N-1)} \quad \cdots\cdots (10)$$

Eq. (10) is easily deduced from Eq. (2).

Corresponding to the categolization of zones, the standard error of zones need to be converted to the one of categories. The mean μ_c of μ_l's where l th zone belongs to c th category is employed as the representative of category, and the standard error of categories TSE_c^{ij} is calculated by substituting μ_c into Eq. (10) instead of μ_l.

(2) Criteria for comparison

If cluster sampling of trip by a person or a household has no effect on the precision of estimator, standard error estimated by means of the re-sampling experiment SE_c^{ij} is expected to match with the theoretical one TSE_c^{ij} deduced on the assumption of simple random sampling. That is, the ratio,

$$\varphi_c^{ij}=SE_c^{ij}/TSE_c^{ij} \quad \cdots\cdots (11)$$

is expected to be 1 and to map around 1 in the range of eventuality in sampling. Then, we can determine the effect of cluster sampling by scanning φ_c^{ij}.

Reviewing Eq. (6), we find that φ_c^{ij} is an estimate of $\sqrt{V_{cl}/V_r}$ when we estimate x_c from the ij case of sampling. Furthrmore, the size of sample n and the value of $\overline{Y}$ are not contained in Eq. (6). Therefore, the value of φ_c^{ij} does not depend on the sample size and the value of μ_c. Then, φ_c^{ij} has commonly an expected value φ_i regardless of sampling ratio and category. But, its variance varies, because m^* is not equal for each φ_c^{ij}.

In order to evaluate the map of φ_c^{ij}, whose number is the product of the number of categories C by the number of cases of sampling ratio J, the following criteria were adopted:

① The weighted mean of φ_c^{ij}

$$\overline{\varphi}_i=\sum_{j=1}^{J}\sum_{c=1}^{C} w_c^j\,\varphi_c^{ij}/\sum_{j=1}^{J}\sum_{c=1}^{C} w_c^j \quad \cdots\cdots (12)$$

The weighting factors were determined in such a way that the difference of confidence interval to estimate φ_c^{ij} was reflected in w_c^j. The weigted means resulted in a very small difference from the mean without weight.

② Sign test

Under the null hypothesis that the sample by cluster sampling could be assumed to be a simple random sample, the expected value of φ_c^{ij} is equal to 1. Sigh test was employed so that this hypothesis is examined. If this hypothesis was rejected, it was inferred that the expected value of φ_c^{ij} differed from 1.

③ The number of φ_c^{ij} values which are realized in the confidence interval.

Under the hypothesis stated above, the variance of x_c^{ijk} is equal to $(TSE_c^{ij})^2$. Then, the interval where the sample variance $(SE_c^{ij})^2$ will be realized with the probability $1-\alpha$ is expressed by Eq. (13).

$$\left(\frac{X^2_{\frac{\alpha}{2}}(m_c^*)}{m_c^*}(TSE_c^{ij})^2,\ \frac{X^2_{1-\frac{\alpha}{2}}}{m_c^*}(TSE_c^{ij})^2\right) \quad \cdots\cdots (13)$$

where, $X^2_{\frac{\alpha}{2}}(m_c^*)$: the point where the integration of chi-square distribution whose degree of freedom is m_c^* is $\alpha/2$.

Then, the confidence interval for φ_c^{ij} is

$$\left(\sqrt{X^2_{\frac{\alpha}{2}}(m_c^*)/m_c^*},\ \sqrt{X^2_{1-\frac{\alpha}{2}}(m_c^*)/m_c^*}\right) \quad \cdots\cdots (14)$$

Whether a value of φ_c^{ij} is realized in the interval specified by its m_c^* makes independent test for each combination j and c to examine the hypothesis that the expected value of a φ_c^{ij} is 1. Then, the number of φ_c^{ij} which realized in the interval was counted, and if it was large, the hypothesis was rejected.

(3) Consideration of results

The criteria for some items of trip-volume which are often used in transportation studies, are summarized in Table 1 for both series of sampling units. Columns ①, ② and ③ correspond to the

number of criteria stated in 3 (2), respectively.

It is pointed out that for most of the items the mean $\overline{\varphi}_i$ are different from 1 and that the differences are meaningful from statistics point of view. That is, standard error of estimator by cluster sampling differs from that by simple random sampling. In most cases, the former exceeds the latter and the deterioration in precision is caused due to cluster sampling. It is certified that the effect of cluster sampling cannot be ignored from the viewpoint of data precision. The magnitude of effect varies with the sampling unit and the items concerned.

In the case that the sampling unit is a household, deteriorations in precision are found for such items as trip generation and attraction for all purposes, business purpose by mode, inner-zonal trip distribution and so on. As to the cases in which the sampling unit is a person, for the items listed above the deteriorations in precision are also found.

The characteristics found in Table 1 are in harmony with the empirical knowledge on the characteristics of the behaviour of a person and a household. That is, the intra-class correlation coefficient ρ is positive for such items of trip that are exclusively made by limited persons with a certain attribute. In other words, clusters are classified into two groups in such a way that the clusters in one group contain many elements whose measure is a certain value, and that clusters in the other group contain few. As to business trip, it is clear that some persons working in business fields such as management, sales and so on make business trip frequently, while others make them rarely. That is, the homogeneousness in a cluster is plausible. On the other hand, as to items in regards to commuting trip, the precision by cluster sampling is superior, or even, to that by simple random sampling. It is understood that the cluster is rather heterogeneous regarding this item, because a person makes at most one commuting trip and a large proportion of people in the population have the possibility to commute.

Moreover, it is pointed out that the precision of estimator by household sampling is affected not only by the intra-class correlationship of trip attributes within a person, but also by the one of personal attributes within a household. The household sampling makes lagerer deterioration, in general, because of the above reason. But, as to business trip, cluster sampling makes estimators more precise. As mentioned earlier, from occupational point of view, only limited attributes of persons make business trips, and the occupation of each member in a household is rather heterogeneous. Above discussion demonstrates that the method of analysis employed in this research is correct.

(4) A method to estimate the data precision by cluster sampling and its test

Through the discussion in 3 (3), it has been made clear that the effect of cluster sampling cannot be ignored. Next task is to develop a method to estimate the standard error taking the effect of cluster sampling into consideration.

Eq. (6) indicates that the product of the standard error by simple random sampling by a constant migth be appropriate to estimate the standard error by cluster sampling. Eq. (15) is proposed as a method to estimate the standard error.

$$SE_{ci}=\varphi^{*i}\cdot SE_r \quad \cdots\cdots (15)$$

φ^{*i} corresponds to square root of $1+(\overline{M}-1)\rho$ in Eq. (6) and varies with the items concerned and the difference of the sampling unit, but is constant against the sampling ratio. Referring to Eqs. (11) and (12), the mean $\overline{\varphi}^i$ is understood to be an estimate for φ^{*i} obtained through the re-sampling experiment.

According to Eq. (15), the standard error for a certain item c by cluster sampling in which the sampling unit is i and the sampling ratio is j, is expressed by Eq. (16).

$$SE_c^{*ij}=\varphi^{*i}\cdot TSE_c^{ij} \quad \cdots\cdots (16)$$

This standard error is hypothetical and is denoted by SE_c^{*ij} in order to distinguish from the experimental standard error SE_c^{ij}. Now, it was tested whether Eq. (17) might be appropriate as a method to estimate the standard error by cluster sampling. Under the hypothesis that true standard

Table 1. Criteria to examine the difference between cluster sampling and simple random sampling by means of re-sampling experiments.

Item \ Criterion	Sampling Unit	Person (1) Mean of $\mathcal{S}$	Person (2) Sign Test	Person (3) Mapping around 1	Person (4) Mapping around $\mathcal{S}^*$	Household (1) Mean of $\mathcal{S}$	Household (2) Sign Test	Household (3) Mapping around 1	Household (4) Mapping around $\mathcal{S}^*$	(5) $\bar{\mathcal{S}}$ (household) / $\bar{\mathcal{S}}$ (person)
Trip Generation & Attraction by Purpose										
All purpose	G	1.31	R	0/ 8	7/ 8	1.52	R	0/12	8/12	1.16
	A	1.31	R	1/ 8	7/ 8	1.48	R	0/ 8	6/ 8	1.13
Commuting	G	.87	R	10/20	17/20	.91	R	15/20	18/20	1.05
	A	1.02	A	16/16	15/16	1.00	A	15/16	15/16	.98
Business	G	1.50	R	0/20	16/20	1.48	R	0/16	14/16	.99
	A	1.50	R	0/20	18/20	1.51	R	0/16	14/16	1.01
Private	G	1.02	A	14/16	15/16	1.23	R	6/16	16/16	1.21
	A	1.06	R	14/16	16/16	1.22	R	5/16	15/16	1.15
Trip Generation & Attraction by Mode										
Railway	G	.96	R	18/20	18/20	1.08	R	13/20	16/20	1.13
	A	1.00	A	15/16	15/16	1.12	R	10/16	15/16	1.12
Bus	G	1.22	R	1/12	12/12	1.30	R	0/16	14/16	1.07
	A	1.21	R	1/12	12/12	1.29	R	2/16	11/16	1.07
Car	G	1.57	R	0/16	14/16	1.78	R	0/16	15/16	1.13
	A	1.56	R	0/16	14/16	1.77	R	0/16	13/16	1.13
Bicycle	G	1.65	R	0/12	11/12	1.72	R	0/16	16/16	1.04
	A	1.62	R	0/16	14/16	1.71	R	0/16	15/16	1.06
Foot	G	1.42	R	1/16	14/16	1.64	R	0/16	13/16	1.15
	A	1.43	R	0/16	15/16	1.64	R	0/16	13/16	1.15
Trip Generation & Attraction for Commuting Purpose by Mode										
Railway	G	.91	R	12/20	19/20	.94	A	10/16	12/16	1.03
	A	1.02	A	20/20	20/20	.98	A	13/20	20/20	.96
Bus	G	.96	R	12/12	12/12	1.02	A	12/12	12/12	1.06
	A	.98	A	12/12	11/12	1.06	R	7/12	11/12	1.08
Car	G	.97	A	12/12	12/12	1.06	R	10/12	11/12	1.09
	A	1.01	A	12/12	12/12	1.07	R	8/12	12/12	1.06
Foot	G	.98	A	15/16	16/16	1.08	R	3/16	16/16	1.10
	A	1.00	A	12/12	12/12	1.10	R	2/12	12/12	1.10
Trip Generation & Attraction for Business Purpose by Mode										
Railway	G	1.12	R	2/12	12/12	1.11	R	2/12	12/12	1.00
	A	1.08	A	1/12	11/12	1.01	A	6/12	11/12	.94
Car	G	1.46	R	2/20	16/20	1.45	R	0/16	14/16	1.00
	A	1.38	R	4/20	17/20	1.46	R	0/20	17/20	1.06
Foot	G	1.37	R	2/ 8	6/ 8	1.22	R	3/ 8	8/ 8	.89
	A	1.32	R	3/ 8	5/ 8	1.25	R	0/ 8	8/ 8	.95
Trip Distribution for all Purpose										
Inner Zonal		1.50	R	2/16	16/16	1.62	R	0/16	16/16	1.08
Inter Zonal		1.03	R	19/24	21/24	1.06	R	12/24	22/24	1.03
Rate of Trip Frequency										
All Purpose		not calculated because rate is defined on a person				1.11	R	9/20	20/20	1.11
Commuting						.99	A	1/20	19/20	.99

Level of significance; 5%, A; Accept, R; Reject

error is SE_c^{*ij}, $m^*(SE_c^{ij}/SE_c^{*ij})^2$ has chi-square distribution whose degree of freedom is m_c^*. Then, the interval where SE_c^{ij} will map with probability $1-\alpha$ is as follows.

$$[(SE_c^{*ij})^2 \cdot X^2_{\frac{\alpha}{2}}(m_c^*)/m_c^*, (SE_c^{*ij})^2 \cdot X^2_{1-\frac{\alpha}{2}}(m_c^*)/m_c^*] \quad \cdots\cdots (17)$$

Therefore, when the experimental value of SE_c^{ij} is realized in this interval, the hypothesis is acceptable. This test was repeated by $J \times C$ times for an item. The number of SE_c^{ij}'s realized in the interval is shown in column ④ in Table 1 with the number of trials. This test is not perfectly

independent, because each SE_c^{ij} is used to calculate φ^{*i}. But, it is accurate enough to evaluate the status of the distribution of SE_c^{ij} around SE_c^{*ij}.

It was pointed out that more than three fourths of SE_c^{ij}'s maped in the interval for almost every item and that SE_c^{ij} concentrated sharply around SE_c^{*ij}. Therefore, SE_c^{*ij} was acceptable as an expected value of SE_c^{ij} and Eq. (16) was reliable as a method to estimate the standard error by cluster sampling.

4. COMPARISON BETWEEN A HOUSEHOLD AND A PERSON AS A SAMPLING UNIT

(1) Comparison of the precision of estimator with respect to trip-volume

As it is impossible to make a random sample of trips directly through home interview survey, the sample of trips has to be collected by observing the transportation behaviour of persons. There are two posssible sampling unit to collect trips, i.e., a person and a household. This chapter compares the household sampling, which home interview survey usually employs, with the person sampling from the viewpoint of data precision and cost.

The comparison was accomplished by comparing standard errors. These standard errors have already been measured in the form of $\overline{\varphi}_i$, that is, the ratio of increase of standard error by cluster sampling in comparison with simple random sampling. Then, the ratio of $\overline{\varphi}_1$ by household sampling against $\overline{\varphi}_2$ by person sampling indicates the increase of standard error due to drawing persons by cluster sampling in which sampling unit is a household. The measures of the ratio are shown in Table 1, column ⑤.

Table 1 shows that the standard error by household sampling tends to be larger than that by person sampling, and that the increase of standard error is up to 20 % at maximum. The magnitude of deteriortion in precision varies with the items concerned, depending on the magnitude in intra-class correlationship of attributes of members in a household. The magnitudes range from 5 % to 15 % for most items. Remarkable deterioration is found for such items as trip generation and attraction for all purpose, trips by mode, and trips for commuting purpose by mode. The deterioration regarding trip generation is understandable because the quantities of activity of each member in a household have a correlationship to a certain extent. The reason for the deterioration in respect with mode is related to the dwelling place of the members. Each member has the same access condition to modes, which tends to make them choose a mode in common.

(2) Comparison of the precision of estimator with respect to attributes of a person

The rate of trip frequency is another important statistic estimated from home interview survey. The rate is characteristic which is defined on a person. That is, observation unit is a person for this characteristic, while observation unit is a trip for the statistics in regard with trip-volume. Then, the problem concerned is the effect of drawing persons by a cluster of a household.

In order to measure the effect, a method estimating the intra-class correlation coefficient ρ directly from the sample is adopted. Let y_{jk} be trip frequency of k th member of j th household, and ρ is calculated by inputting y_{jk} into Eq. (5). Then, the ratio of variance by cluster sampling against that by simple random sampling, i.e., V_{cl}/V_r is estimated by Eq. (6). Now, Eq. (6) is formulated under the condition that the number of elements in each cluster is equal. So, the estimated V_{cl}/V_r is an approximation, because the number of members in a household is not equal. The square root of V_{cl}/V_r is the ratio of standard error and expresses the increase of standard error caused by cluster sampling.

$$R=SE_{cl}/SE_r=\sqrt{V_{cl}/V_r}=\sqrt{1+(\overline{M}-1)\rho} \quad \cdots\cdots (18)$$

Table 2 shows the value of R for rates of trip frequency by purpose and by mode.

The re-sampling experiment is also employed for the purpose to test the result in Table 2. The same procedure that is described in Chapter 3 is employed. By this method ρ is calculate only for the rates of trip frequency for all purposes and the one for commuting purpose. The values are 1.11 and 0.99,

Table 2. Intra-class Correlation of Rate of Trip Frequency in a Household by Means of Direct Calculation of ρ.

Item	Calculated ρ	SE_{cl}/SE_r
Rate by purpose		
All purpose	.1376	1.11
Commuting	.0278	1.02
School	- .0231	0.98
Business	.0208	1.02
Shopping	- .0358	0.97
Social	.0775	1.07
To home	.1290	1.11
Rate by mode		
Railway	.1097	1.09
Bus	.0672	1.06
Car	.0708	1.06
Bicycle	.1144	1.10
Foot	.0742	1.06

respectively. These fit the values in Table 2, demonstrating their reliability.

Table 2 indicated that clustering persons by a household caused about 10 % deterioration in precsion at maximum. The magnitude of deterioration was remarkable for the rate of trip frequency by mode. This tendency was common with the one mentioned in 4 (1).

(3) Data precision under a budget constraint

The results in 4 (1) and (2) makes it clear that household sampling is inferior to person sampling from the viewpoint of data precision, if the sizes of each sample are equal. Now, we take the size of sample under a cost constraint into consideration.

Let C_h and C_p be the cost to investigate a person's behaviour by household sampling and person sampling, respectively, and let C_t be the amount budget allowed to the survey. Then, the sample sizes collected by each sampling unit are $n_h=C_t/C_h$ and $n_p=C_t/C_p$. Referring to Eqs. (10) and (16), the standard error to estimate a certain item of which the true value is μ, from the sample whose size is n_h, and sampling unit is a household, is expressed by Eq. (19).

$$SE_h=\varphi_h^*\sqrt{\mu(N-\mu)/n_h}\sqrt{(N-n_h)/(N-1)} \quad \cdots\cdots (19)$$

Similarly, the standard error from the sample whose sampling unit is a person is expressed by Eq. (20).

$$SE_p=\varphi_p^*\sqrt{\mu(N-\mu)/n_p}\sqrt{(N-n_p)/(N-1)} \quad \cdots\cdots (20)$$

φ_h^* and φ_p^* denote the rate of increase of standard error due to cluster sampling. Then, the condition for $SE_h<SE_p$ is

$$\varphi_h^*/\varphi_p^*<\sqrt{N-n_p)/n_p}\,/\sqrt{N-n_h)/n_h} \quad \cdots\cdots (21)$$

Substituting costs into sample size, let R^* denote φ_h^*/φ_p^*, then Eq. (21) is converted into

$$(R^*)^2C_h-C_p<|(R^*)^2-1|C_t/N \quad \cdots\cdots (22)$$

Roughly, Eq. (22) is satisfied regardless of N and C_t, when

$$(R^*)^2<C_p/C_h \quad \cdots\cdots (23)$$

It is inferred from the analyses in 4 (1) and (2) that R^* is approximately 1.2 at maximum, but cost is vague. We assume that survey cost is composed of field survey costs such as interviewer cost, which are proportional to the number of visits and processing cost such as coding which are proportional to the number of persons investigated. Most home interview surveys in Japan employ the so-called home questionnaire method. In this case, the cost to investigate all the members in a household is not more expensive than the cost to investigate a specified person in a household. Thus, the field survey cost for household sampling is assumed to be roughly equal to that for person sampling.

Now, we assume that the field survey cost is ¥ 600 per a household and that the processing cost is ¥ 400 per a person. This assumption is probably not very different from actual cost. This makes C_h be ¥ 600 and C_p be ¥ 1000. Substituting these values into Eq. (23), it is clarified that household sampling has a greater advantage in data precision than person sampling when the allocated costs are equal.

5. CONCLUSION

Person tip survey is a fundamental data source for transportation studies, but the characteristics of

data obtained from the survey are rather vague from the viewpoint of precision. The author pointed out that the survey made up the sample by cluster sampling, and investigated the effect of cluster sampling on data precision. Hitherto, the standard error of estimators by the survey has been estimated on the assumption that the sample from the survey is random. Firstly, this assumption was examined and shown to be incorrect by means of a resampling experiment. Then, a method to estimate the standard error was developed successfully. Finally, the characteristics of a household and a person as a sampling unit were considered from the viewpoints of data precision and survey cost.

The main results are summarized as follows;

(1) It was clarified that the standard error of estimators from the sample collected by home interview survey differed from that which was deduced on the assumption that the sample was random. It is necessary to take into consideration the effect of drawing the elements of a sample by cluster sampling.

(2) The magnitude of the effect varies with the items concerned. A remarkable effect is found for items concerning mode and business purpose. The deterioration in precision attains up to 80 % at maximum in comparison with the simple random sample.

(3) The standard error of estimator by cluster sampling were formulated as a product of the standard error of estimator from the simple random sample by a constant. The constant varied with the items and the sampling unit. The constants for the items important for transportation studies were identified in Table 1 for the case of Tokyo Metropolitan Area Transportation Study.

(4) In order to evaluate the effect of clustering persons by a household, a method to estimate the intra-class correlation coefficient directly was applied. This method produced a very good approximation and turned out to be a simplified way to inspect the effect of cluster sampling.

(5) A household and a person are possible sampling units to collect trips. When the size of the samples equal, the precision of estimators from the sample whose sampling unit is a person is better than the other by about 20 %. On the other hand, when the costs being allocated to the survey equal, the sample whose sampling unit is a household gives more precise estimates than the other.

The intra-class correlationship of trip attributes within a person or a household is thought to be affected by the pattern of transport behaviour, which varies with areas and times. Further studies are needed on the stability of the intra-class correlationship over time and area.

The auther wishes to thank Professor Hideo Igarashi of the Hokkaido University for his valuable suggestions.

REFERENCES

1) State of Illinois ;County of Cook ; City of Chicago : Chicago Area Transportation Study, Final Reports, Vol. 1 Appendix, p. 107, 1962.
2) Bureau of public Roads and National Committee on Urban Transportation : Origin-Destination and Land Use, Rrocedure Manual 2 A ; pp. 13~15, Public Administration Service, 1958.
3) Committee on Tokyo Metropolitan Area Person Trip Survey : Reports on Tokyo Metropolitan Area Comprehensive Transportation Study, p. 230, Ministry of Construction et al., 1970, (In Japanese).
4) Miyazawa, K. : Modern Statistics, pp. 211~213, Baifu-kan, 1962, (In Japanese).
5) Nakayama, I. et al. : Modern Statistics Dictionary, pp. 451~455, Toyo Keizai Sinposha, 1962, (In Japanese).
6) Saito, K, and Asai, A. : Design of Sampling Survey, Baifu-kan, 1970, (In Japanese).
7) Wilks, S. S. : Mathematical Statistics, John Wiley & Sons, Inc. 1963.
8) Stopher, P. R. and Meyburg, A. H. : Urban Transportation Modeling and Planning, Lexington Book, 1975.
9) Yamagata, K. : The precision of Data and its Effect on the Accuracy of Transport Forecasts, Annual Papers of Regional Science Association of Japan, Vol. 12 pp. 85~103, 1981. (In Japanese).
10) Yamagata, K. : Data precision of person Trip Survey Taking the Effect of Sampling Unit into Consideration, Annual Papers of Civil Engineering Planning, Vol 5 pp. 366~370, 1983, (In Japanese).

(Received January 12 1984)