Sampling Design in the Social Survey of Language at the City of Shirakawa

By Chikio HAYASHI

(Received August 10, 1950)

Research workers were organized of the members of the below three Institutes: (i) The National Language Research Institute. (ii) The Institute of Statistical Mathematics. (iii) The Institute of Folkore. This survey was performed in Octover-December, 1949 by about 15 surveyers.

This short report, that on sampling design, is a part of the works done by the members of the Institute of Statistical Mathematics, C. Hayashi M. D. Ishida, F. Maruyama, S. Nishihira.

The purpose of the survey is to make clear the linguistic factors which disturb rationalization of national social life. Then, at a particular spot we endeavour to research the language in relation to cultural circumstances.

As a particular spot, the city of Sirakawa, Hukusima prefecture was selected by the following reasons: (1) A cultural centre in a region. (2) The dialect which is not difficult for us to understand and different from common language. (3) The spot which is convenient for running to and from Tokyo, and where we can get special co-operation.

This survey needs delicate techniques that aims at finding out the dynamic patterning of language in social life, so interviewing method by well trained interviewers must be used.

Samples (testees) are inquired by interviewers concerning with the prepared items: socio-cultural factors, social attitude, the degree of consciousness of common language, accent, phonemic characteristics, morphological characteristics.

These reactions of them are analized from the various points of view.

Thus sample size will be naturally limited, because there are a few interviewers. Considering the several conditions, sample size is

decided to be about 500 from the point of view of administration. In many cases, this size probably secures the sufficient confidence level in the estimates we make. For example, if the coefficient of variation of the label in population is 0.05 or 0.03, the coefficient of variation of unbiased estimate for the population mean, i.e. the sample mean, is about 2.2% or 1.3% respectively. If the population ratio is estimated, the confidence interval is about 5%, 4% and 3.5% under the confidence level 95%, at the population ratio 50%, 40% and 30% respectively. In order to secure as high level of confidence as possible in the analyses of the results obtained from the samples of this size, stractification must be made.

First, we make clear the universe. The universe consists of the citizens to 69 from 15 in age who lead a normal social life at this city in November 1949.

Population is constructed from the universe by giving equal sampling probability to every element (individual), when its label is the reaction type of every item tested in this survey. That is to say sampling unit is indvidual.

Simple stratified sampling methods has to be adopted. But in the real restrictions, (time, money and labour), stratisfication of about 20,000 (the population size) is difficult to do.

So we adopt a kind of double sampling. The charactiristics of stratification are sex, age, occupation, and residence. Other characteristics of influencing a person's language in social life are considered in several directions but unable to use without knowing previously them. In fact they are previously unknown. First we make the strata by using sex and age. Fortunately the population size of every stratum is known. Then we can reap the fruits of stratified sampling by using a simple sampling technique without stratifying the whole population, so far as the population means about certain labels are unbiasedly estimated in the linear form. Suppose that first sample size m is drawn unrestrictedly from the total population and these samples are divided into the strata mentioned above. Let n be the final sample size. Let p_i $(i=1, 2, \ldots R)$ be the ratio of the size of the i-th stratum to the total and n_i be the sample size allocated to the i-th stratum, m_i the first sample size belonging to the i-th stratum in the first sample.

If $m_i \geqq n_i$ $(i=1, 2, \ldots, R)$ and p_i is known, the sampling method

from every stratum in the 1st sample, where the sample size of the i-th stratum is n_i, has the same effect as the stratified sampling from the population when the linear unbiased estimates of the population means are considered. This proof is easily obtained by using the idea of two stage sampling system.

Let N be the population size the elements of which have the label $X_1, X_2, \ldots, X_N$, and equal sampling probabilities, i. e. $\frac{1}{N}$.

Let $\bar{X}$ and σ^2 be the population mean and variance

Suppose that the first samples of size m are drawn, and next the samples of size n are drawn from the sample of size m. Then as the estimate of the population mean $\bar{X}$, sample mean $\bar{x}$ is made.

Let $\sigma_{\bar{x}}^2$ be the variance of $\bar{x}$. Then we obtain

$$\sigma_{\bar{x}}^2 = \frac{m-n}{m-1}\frac{\sigma_w^2}{n} + \sigma_b^2$$

where

$$\sigma^2 = \sigma_w^2 + \sigma_b^2$$

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

so $\sigma_{\bar{x}}^2$ turns out to be equal to the following

$$\sigma_{\bar{x}}^2 = \frac{N-n}{N-1}\frac{\sigma^2}{n}$$

So far as he relation $m>n$ holds, $\sigma_{\bar{x}}^2$ is independent of m. As a results, the proposition mentioned above holds. In order to secure the relation $m_i \geqq n_i$, the first sample size m must be decided following the consideration as below. In our case, the population ratioes concerning with sex and age are known. They are as followings.

TABLES of p_i (%)

Age	15—19	20—24	25—29	30—34	35—39	40—44	45—49	50—54	55—59	60—64	65—69
Male	9.52	6.84	5.31	4.28	4.06	4.28	3.89	3.18	2.89	1.97	1.28
Female	8.90	7.75	6.74	5.24	5.25	4.76	4.04	3.24	2.70	2.13	1.75

In our design, samples are allocated to every stratum in proportion to the size, because many tables necessary for complicated conclusions must be easily and speedily tabulated at the lowest cost.

As $n_i = np_i$, $m_i \geqq np_i$ ($n=500$ in our case) must hold. m_i is a random variable, the mean and variance of which are respectively mp_i and $\sigma^2_{m_i} = \frac{N-m}{N-1} \frac{p_i(1-p_i)}{m}$ where N is population size. In order that the relation $m_i \geqq np_i$ may hold under the confidence level 99.7% or 99.9%, the first sample size m must be decided to secure the following relation,

$$m\left\{p_i - 3(4)\sqrt{\frac{N-m}{N-1}\frac{p_i(1-p_i)}{m}}\right\} \geqq np_i \qquad (n=500)$$

If the sample size m is about 2000, the above relation holds when p_i is more than 1%.

From the point of view, it is decided that the samples of size about 2000 are unrestictedly drawn from the population. This is easy. Then systematic sampling method from the rationing records was adopted. Of course the sampling interval was desided not to synchronize with the average number of the members in a household.

The results of this sampling was satisfactory, we proceed to the next step. Thus the obtained first samples of size 2148 are divided into strata using the labels of sex and age. And 500 p_i samples are drawn from the i-stratum with equal sampling probabilities. Here the idea of so-called double sampling method is used. Every stratum is divided into strata using the characteristics of residence and occupation of every element in it. From every stratum made by this procedure samples of size 500 $p_i q_{ij}'$ are drawn, where q_{ij}' is the sample ratio of the size of the j sub-stratum in the i-stratum. Now suppose that the population mean $\bar{X}$ of lable X_i ($i=1, 2, \ldots N$, N is population size) will be estimated by sampling.

$$\bar{x} = \sum_{i=1}^{R} p_i \bar{x}_i$$

where $\bar{x}$ is sample mean,

N_i is the size of the i stratum,

$$p_i=\frac{N_i}{\sum_{i=1}^{R} N_i},$$

$\bar{x}_i$ is the mean of the samples from i stratum,

R is the number of strata by sex and age.

$\bar{x}_i$ is obtained by the procedure mentioned above, that is to say,

$$\bar{x}_i=\sum_{j=1}^{R_i} q'_{ij}\,\bar{x}'_{ij}$$

where $\bar{x}'_{ij}$ is the sample mean of j sub-straum in i stratum,

R'_i is the number of sub-strata in i straum.

Of course

$E(\bar{x}_i)=\bar{X}_i$ where $\bar{X}_i$ is the mean of i stratum. So $E(\bar{x})=\bar{X}$, i.e. $\bar{x}$ is the unbiased estimate. We consider the variance of $\bar{x}$, $\sigma_{\bar{x}}^2$,

$$\sigma_{\bar{x}}^2=E\left(\sum_{i=1}^{R} p_i\bar{x}_i-\sum_{i=1}^{R} p_i\bar{X}_i\right)^2$$

$$=\sum_{i=1}^{R} p_i^2\,E\,(\bar{x}_i-\bar{X}_i)^2$$

Calculating the $E(\bar{x}_i-\bar{X}_i)^2$ by using the theory of conditionary probability, $\sigma_{\bar{x}}^2$ is approximately as below, where population size is large, and the ratio of sample size to population size in every stratum is small

$$\sigma_{\bar{x}}^2\fallingdotseq\frac{\sigma_2}{n}-\frac{1}{n}\left(\sum_{i=1}^{R} p_i\,(\bar{X}-\bar{X}_i)^2+\sum_{i=1}^{R}\left(1-\frac{n_i}{mp_i}\right)\sigma_{b_i}^2\,p_i\right)$$

where $\sigma_{b_i}^2$ is the variance between sub-strata in i stratum. (This relation is almost equal and the sign $\sigma_{\bar{x}}^2\leqq\cdots$holds.) Because the relation $\frac{n_i}{mp_i}<1$ always holds, the second term in the right hand of the above is positive. The effect of stratification is expressed by this. In our case $\frac{n_i}{mp_i}$ is about $\frac{1}{4}$, n is 500, R is 22,

$$\sigma_{\bar{x}}^2=\frac{\sigma^2}{500}-\frac{1}{500}\left(\sum_{i=1}^{22} p_i\,(\bar{X}-\bar{X}_i)^2+\frac{3}{4}\sum_{i=1}^{22}\sigma_{b_i}^2\,p_i\right)$$

In sampling procedure, more than 500 samples are drawn, because there are non-responses (by inevitable reasons) the ratio of which is about 10%.

Non-responses are substituted by inflation samples.

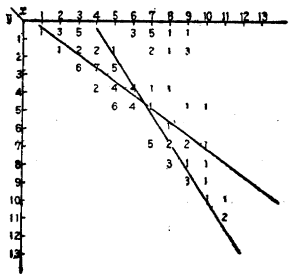
But follow up of samples must not be neglected. Non-responses without inevitable reasons are, of course, in question.

APPENDIX

The difficulties of this survey consists in the techniques of interviewing and in call back of samples.

As to the former, the interviewers (well-trained) were trained and adjusted to agree in receptions of reactions of testees in pre-test.

As to the latter the proportion of the absentees when interviewed must be estimated. If this is not so high, it is easy to call back them. In order to estimate this ratio, the following survey was performed in pre-test.



$\rho = 0.684$

$\bar{y} = 0.763x - 0.453$

$\bar{x} = 0.613x + 3.509$

From 0900 to 1600, random samples of households (size 99) were interviewed, and inqired about the number of members in their households and the number of those at home at this time interviewed. These results are shown is the left graph.

It may be recognized from this pattern that the several conditions in the method of regression estimate are fulfilled.

Moreover we know the average value, in the city of Shirakawa, of the number of members in a household to be 4.9.

Estimated correlation coefficient is 0.684. So the estimated number of these at home is 3.3, in the sense of mean with respect to time during 0900-1600, its variance V^2 is approximately

$$V^2 \fallingdotseq \frac{\sigma_y^2(1-\rho^2)}{n} = 0.042$$

where n is the size of sample, 99.

σ_y^2 is the variance of the numbers of those at home in population (actually, sample).

Thus the proportion of those at home, A is estimated as followings

$$A=\frac{3.3\pm0.4}{4.9}=0.67\pm0.08$$

where ±0.08 is the confidence interval under the confidence level 95%. As the proportion of those at home is about 60%, call back of samples seems not to be so difficult.

Under these considerations, the main test was preformed. In it, special difficulties were not found out and the plans was, on the whole, performed as expected.

In order to weave rich patternings of language phenomena in social life besides this survey, the following researches were done, about

(i) Reading and writing abilities and activities.

(ii) 24 hours' language research.
Observation and recording, of a perticular person's language in daily life in constant attendance to a testee from early morning till late night.

(iii) Lingustic assimilation of evacuated pupils from Tokyo.

(iv) Linguistic structure.

(v) Social circumstance and folkore.

Institute of Statiscal Mathematics