Chapter I
INTRODUCTION

This study is a review of the sampling methodology of the Rural Manpower Utilization Project, initiated by the National Economic Development Board (NEDB), Kingdom of Thailand. The main purpose of the study is to investigate the quality of the intensively studied samples as an aid to interpretation of the data from them. A secondary but related purpose is to indicate possible improvements in sampling methodology for similar studies in the future.

A. The Parent Project

The NEDB Rural Manpower Utilization Project [1] was started in 1969 by the Manpower Planning Division of the National Economic Development Board, and conducted jointly by the Chulalongkorn University Social Science Research Institute and the Department of Agricultural Economics, Kasetsart University. Its principal objectives[2] were to collect accurate data on rural labour utilization, on the extend and nature of rural unemployment, and on farm incomes and expenditures. As the utilization of the rural labour force is closely related to many social and economic factors, such as population size and characteristics, land tenure, opportunities for and training in

---

1 For a brief discussion of preliminary results from the project, see Fuhs, Friedrich W., and Vingerhoets, Jan. Rural Manpower, Rural Institutions and Rural Employment in Thailand, National Economic Development Board, Bangkok, 1972. Chapter III, pp. 53 - 105.

2 Ibid, pp. 53.

agricultural and other rural occupations, marketing facilities, credit facilities, the introduction of innovations, etc..., the desirability of studying these was also indicated.

The approach - intensive and extensive

The simultaneous interest in accurate and intensive data on labour utilization, income and expenditures, and data on a range of related socio-economic factors lent itself readily to the administrative and fieldwork structure which evolved. The project was divided into two parts, the intensive study of labour utilization, incomes, and expenditures, and the extensive study of related socio-economic variables. Dr. J. Amyot, of the Chulalongkorn University Social Science Research Institute, (CUSSRI) was the research director for the extensive socio-economic study, while the intensive study was directed by Mr. Arb Nakajud, head of the Department of Agricultural Economics, Faculty of Economics and Business Administration, Kasetsart University. Another staff member, Dr. Sopin Tongpan served as senior researcher.

The overall planning for and supervision of the project was done by a research committee consisting of the research directors, and advisors from the Northeastern Economic Development (NEED) team of the National Economic Development Board, the Population Council, and UNICEF. This committee was chaired by the project co-ordinator, Dr. F.H. Fuhs, I.L.O. expert working as an advisor to the National Economic Development Board.

Fieldwork was carried out by Research Assistants of the Chulalongkorn University Social Science Research Institute, who were assigned to

Project Areas in teams of two persons. The fieldwork, which took one year for each round of the project, consisted of day-by-day enumeration of the households selected for intensive study, completion of CUSSRI's "Socio-Economic Profile Schedule" for all households in the villages studied, and various case studies.

The data collected from the intensively studied sample consisted of the following :

(1) Labour Utilization. The number of hours worked by type of work for every member of the household labour force, each day of the year. Types of work were divided into 30 different activities.

(2) Income and Expenditures. Again, data were recorded daily and in considerable detail, some of the main subdivisions being income from crops and from livestock, other farm income, farm costs, non-farm income, household expenditures, and changes (increases and decreases) in stock.

The "Socio-Economic Profile Schedule" [3] is an instrument developed by CUSSRI out of its previous research experience. It collects the following data: household census, information on children of household head permanently away from home, farm size and land use, land tenure (including source of owned land and inheritance practices), details on off-farm employment and own-account non-farm work, use of hired

---

[3] The two versions of the Socio-Economic Profile Schedule used for this project are available at CUSSRI. A list of the variables for the first round which are on computer cards is in Appendix B (page 182)

labour and participation in labour exchange, crops planted (including yields, amounts sold, and marketing practices), use of modern equipment, use of modern farm inputs, livestock inventory, savings and debts, innovations, and influences leading to the adoption of innovations.

The case studies done included: Village Social Organization; Groups and Associations; Descriptive Account of Life in a Typical Household; Innovations; and Land Holding History.

Villages studied to date :

At the present time, two rounds of the project have been carried out, and a third is in the planning stage. The villages of the first two rounds were :

Round I (fieldwork carried out June 15, 1969 - June 15, 1970)

Project Area I - 3 villages in Changwat[4] Ayutthaya, of which 2 were in Amphoe Bang Pahan, and 1 was in Amphoe Nakhon Luang.

Project Area II - 3 villages in Amphoe San Kamphaeng, Changwat Chiang Mai.

---

[4] Explanations of the Thai terms used in this paper are to be found in a glossary, Appendix C.

Project Area III - 2 villages within the Nong Wai Irrigation Project area, Changwat Khon Kaen: 1 village in Amphoe Muang, and 1 village in Amphoe Nam Pong.

Project Area IV - 2 villages in Amphoe Phuvieng, Changwat Khon Kaen.

Round II (fieldwork carried out June 15, 1970 - June 15, 1971)

Project Area V - 2 villages in Amphoe Yang Talat, Changwat Kalasin.

Project Area VI - 2 villages in Changwat Prachinburi, 1 in Amphoe Kok Pip and 1 in Amphoe Si Mahapot.

Project Area VII - 2 villages in Amphoe Tung Song, Changwat Nakhon Si Thammarat.

Project Area VIII - 2 villages in Changwat Songkhla: 1 in Amphoe Tungyai which was Thai-Buddhist, and 1 in Amphoe Hat Yai, which was Thai-Muslim.

Selection of the Villages Studied.

Statistically-based selection of villages was not advisable, for the interest was in studying villages which were different socio-economically, and to date there is a lack of information on the types of villages that do exist in rural Thailand, and the frequency with which different types occur. However, the collective experience of the research committee made the identification of some common types of villages in each region possible, and villages were selected which appeared to represent some of these types.

The selection of the particular provinces was not based on an assumption that these particular provinces as such were necessarily "typical" of the region in which they were situated, but rather with the knowledge that village types which were common in the region could be found in these provinces. One constraint on the selection of research field sites was that 2 or 3 villages of different types had to be close enough together (in a transportation sense, not just in physical distance) to permit one team to visit each village regularly (daily or every other day). With these two criteria taken into consideration, the selection of provinces was largely on a pragmatic basis. For instance, in the case of Ayutthaya, it was because CUSSRI had considerable knowledge of conditions there due to its previous research. Chiang Mai was chosen in the North because (in addition to satisfying the two criteria already discussed) it provided the opportunity to link up with a similar project of Chiang Mai University which provided professional guidance for the research assistants as well as training in the form of regular joint seminars. In the northeast, two project areas were desired,one in and one outside an irrigation area. Khon Kaen offered this in the same province, as well as professional guidance from personnel at Khon Kaen University.

The three villages in Ayutthaya may serve as an illustration of the way in which the villages selected represent different types of villages. One of them is dependent almost entirely on agriculture, its main crop being rice grown by the transplant method. In addition, most households grow fruit and vegetables for their own consumption. A second village presents quite a different picture, although quite a

common one in the central plains. It broadcasts rice, but of greater economic importance is its home industry - brick-making. The fact that this activity can be carried on almost year round means that the villagers do not take advantage of their opportunities to diversify agriculture or to engage in other kinds of occupations (with the exception of civil service, which is related mainly to ease of communication with the provincial capital). The third village in Changwat Ayutthaya also broadcasts rice. However, poor yields, (the result of either drought or flooding at the wrong times in the growing cycle) have forced the villagers to diversify. They do grow a few other crops with results as uncertain as they are for rice. Their main substitute is livestock raising. Poorer households concentrate on ducks, and those better off on pigs and cattle. As will be discussed more fully later, these activities influence the pattern of labour utilization, for pigs and cattle require labour input year round, and duck-raising takes up much of the period between rice harvesting one year and rice planting the next.

In a similar way, the villages selected in other regions of the country represent the different kinds of economic and social conditions existing there. The three Project Areas (III, IV, and V) in the Northeast consist of 6 villages, three of which are located in irrigation areas, and 3 of which are not. One of those in an irrigation area is very close to the Lam Pao irrigation project, which has provided opportunities for employment to a large section of the labour force. Two of the three not in an irrigation area are fairly recently settled

villages located on higher ground with poor soil. They represent a type of village common in the Northeast. The third village outside an irrigation area is an old village located in the valley of a small stream which supplies the water for its long-established village-level irrigation system.

A few more notes could be made along these lines: the selection of one Thai-Muslim and one Thai-Buddhist village in Songkhla, the inclusion of two villages in Changwat Prachinburi, and Project Area II in the North, which consisted of two villages developing diversified cropping patterns, and one heavily dependent on cottage industry (cloth-weaving).

Sampling for the Intensively Studied Households

In order to take full advantage of both the intensive data on labour utilization, incomes and expenditures, and the extensive socio-economic data for each village, it was desirable that the sample of intensively studied households in each be as representative of the whole village as possible. However, the factors affecting labour utilization, incomes and expenditures are many. Moreover, at the time when selection was made, accurate data on most of these were not available.

Two variables were used in the selection of the intensively studied households in all villages, namely farm-size, and the willingness of the selected households to co-operate with the research assistants regularly over a long period of time.

Information on farm sizes and frequency of their occurrence was obtained from the headman of each village. In Ayutthaya, information on land

tenure was also obtained, and the sample was stratified on both size of holding, and tenure status. In Chiang Mai, there was no need to worry about different tenurial situations (as most farmers own their land) so the individual making the selection stratified on size of holding. However, within this stratification, he selected only households for which the average number of household members was close to the village average. In most of the other villages, sampling was done strictly on a basis of farm-sizes.

The necessity of including the factor of co-operation may have seriously biased the sampling. However, conducting a research project of this type, which requires day-to-day contact of the research assistants with the selected households, is impossible without it. Willingness to co-operate (or the lack of it) is a source of bias that can probably never be eliminated.

A further source of bias may have been introduced by the village headmen, as the result of their opinions of the households selected. How much this factor played a part depends on three things: the extent to which each headman offered his advice; the extent to which each was biased for or against the households selected; and the extent to which the research assistants listened to the advice of the village headman.

B. The problem of study, its purpose and scope

As is clear from the project description above, the NEDB Rural Manpower Utilization Project is a large project with extensive data that can be

analysed in many different ways, for many different purposes. Substantial reports have been[3] and are being drafted. The present study does not in any way attempt to overlap the considerable effort which has been and is still being put into analysing the data and writing the reports from the project. Rather, it attempts to fill in one gap not yet considered - how representative were the samples of intensively studied households of their villages, with particular reference to labour utilization. This problem of the representativeness of the samples for the labour utilization data was chosen for three reasons. The first was the relevance of this problem to the interpretation of the labour utilization data from the intensively studied samples, and the extent to which the results can be generalized to the villages. The second is the hope that such an inquiry will indicate ways of avoiding past sampling pit-falls if further rounds of the project are done. The last is the fact that the present investigator was interested, and had the skills necessary to do the task, and at the same time did not see anyone else working on the project who had both the interest in and the needed skills for investigating this particular problem.

The fact that sampling for this project was not a straight-foreward task has already been indicated above. Many variables are related

---

3 Amyot, Jacques. Village Ayutthaya: Social and Economic Conditions of a Rural Population in Central Thailand. (Bangkok: CUSSRI, 1974). Mimeographed; available at the Chulalongkorn University Social Science Research Institute, Faculty of Political Science, Bangkok.

to labour utilization. Sampling had to be done on a basis of one or two variables that seemed most important, and on which prior data were available. Farm-size was always used, and in most project areas, was the only variable used. Obviously, farm-size is more closely related to labour utilization in a village dependent entirely on crops than it is in a village with lucrative non-agricultural work, or one with extensive livestock raising. The use of other variables, namely tenure status and household size, provided refinements of sampling over what was possible using only farm-size, and should have helped improve sample representativenes.

Given the various types of villages selected, and the sampling procedure used, it can be predicted that the quality of the samples varied from village to village. In some villages, the variables used in stratified sampling may well have resulted in little if any bias. In other villages, particularly those in which many factors were involved in the use of labour, the samples may have considerable bias.

The procedure for the present study will be to determine for some of the villages studied the extent of sample bias on the variables covered by the Socio-Economic Profile Schedule, and then to consider the more important problem of what these biases imply for the interpretation of the labour utilization data from the samples. (It might be noted that income and expenditure data were also collected for the sample. The present investigator will not consider this aspect because she lacks an adequate background in economics. However, it is hoped that given the results of this study, a better qualified

person could fairly quickly and easily cover that aspect.) A final section will be devoted to suggestions for improving sampling for future work on the same project or similar projects.

Related Work

The present study is related principally to the larger study already described, of which it is a part. Its methodology, which is quite different from the larger study, is based on fundamental principles of statistical sampling.

Methodology

The main purpose of this study, once again, is to determine the extent to which the household samples in the NEDB Rural Manpower Utilization Project were representative of the villages from which they were drawn, in order to facilitate the interpretation of the labour utilization data. This is possible because of the fact that the Socio-Economic Profile Schedule was administered to all households in each village under study.

This instrument has provided a wealth of information about the households in each village, and has made it possible to compare the socio-economic data for the households studied intensively with that for all households in each village, to determine the extent and nature of biases existing in the sample, if any. Of the more than 130 variables coded for the Socio-Economic Profile Schedule (see Appendix B for a complete list), 75 were considered to be fairly directly

related to the utilization of the labour force. These variables were used for the purpose of identifying sample biases.

The statistics used are somewhat unusual in the repertoire of social science investigations, for they apply to data available for both the sample and the population. Most statistical procedures are based on the fact that only sample data are available, and inferences must be made on that basis. The present investigation makes use of data available for both sample and population as a basis for determining how representative the samples were.

Two different types of variables, quantitative and discrete (or descriptive), were considered, making it necessary to use different types of statistical treatment in the comparison of sample and population. The first type, "quantitative" consists of variables such as the number of persons in the household, the number of rai of land owned, etc..., which are coded as a quantity from zero up. The second "discrete" on "descriptive" type can be exemplified by "occupation" or "education", or "place of birth" for which numerical codes are given to represent categories such as "farming", "home industry","cannot read or write","completed Prathom 4", or "village of present residence". For these variables, the numbers are only codes for categories which do not have my numerical relationships to each other.

Quantitative variables were treated statistically in two ways, one to determine the probability for differences in means, and the other for differences in variances between population and sample. The

variance test used was Chi-square ($\chi^2 = s^2/\sigma^2$). This statistic, in fact, tests "goodness of fit" of two distributions (Popham, 1967). A Chi-square value of "zero" indicates a perfect fit. The less the two distributions are alike, the greater the value of χ^2. A probability for χ^2 means the probability that s^2 (the sample variance) did not originate from a population with variance σ^2. Thus, a probability of p = .05 means there is 1 chance in 20 that the sample did not originate from the population with variance σ^2, and 19 chances in 20 that this amount of deviation was just the result of random error. A probability of p = .10 means there is 1 chance in 10 that the sample did not originate from the population with variance σ^2, or, in the present study, where the sample was chosen from the population it was intended to represent (with one exception), the greater the sample bias. For the purposes of the present study, the significance level for χ^2 was set at .05. A value lower than that was interpreted as meaning that the variances were essentially equal. A value larger than .05 was taken to mean that the variances were different, and thus the sample was biased. A fairly stringent significance level was chosen for this test, as the second test (Z test) used on quantitative variables was dependent for its validity on the sample and population variances being equal.

The test used for sample and population means was the Z test ($Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$) This test is based on the assumptions that the population data are disbributed normally, that the population variance is known, and that the sample and population variances are equal. In this study, the population variance is always available for the variables under consideration. The last condition, that sample and population variance

be equal is covered by the Chi-square test already described. The assumption that the variables be normally distributed was not checked out, as only radical departures from normality seriously affect the power of the test. For the purposes of the present study, the significance level for Z was set at .10. A probability of P = .10 means that there are 10 chances in 100 (1 in 10) that the difference in means is due to random variation, and 90 in 100 that the difference is real, indicating a sample bias. Similarly, a probability of .50 means there are 5 chances in 100 that the difference is due to random error, or 95 in 100 that the difference is real. Thus, for the Z test, the lower the value of P, the more significant the difference between population and sample.

The 45 quantitative variables to which the χ^2 and Z tests were applied are to be found in Appendix B, where they are marked with asterisks.

Means and variances are meaningless for discrete variables. The procedure for comparing samples and populations on these variables was again to use the Z test, but this time using the formula for proportions:

$$Z = \frac{P - p}{pq/N} \qquad \text{or} \qquad Z = \frac{X - Np}{Npq}$$

where : P is the proportion of successes in the sample

p is the proportion of successes in the population

q = 1 - p

N is the number of cases in the sample

X is the number of successes in the sample

Again, for the discrete variables, the significance level for Z was set at .10. This statistic was calculated for every category of the 30 discrete variables marked with asterisks in the list in Appendix B.

For the most part, the discussion is limited to variables which showed significant differences on the χ^2 or Z tests (the significance levels being .05 and .10 respectively, as discussed above). Occasionally, where it is necessary to fill in on a point, variables without significant differences are brought in. Otherwise, variables for which the differences were not significant by the above definitions do not appear in either the text or the tables. Summarized tables for the quantitative variables with significant differences are an integral part of the text. The full tables for quantitative variables appear in Appendix D. The discrete variables present a problem, as there are many of them. To prevent unwieldiness of the text, tables for the discrete variables for which significant differences between the proportions of occurances in different categories exist are presented in appendix A. Discrete variables for which no significant differences in category proportions appeared are not included.

In Chapter IV, in the discussion of relationships among variables showing sample biases, a t-test is occasionally used. The procedure for the t-test is:

1. Calculate an F ratio

$$F = \frac{s_1^2}{s_2^2} \qquad \text{d.f.} = n_1 - 1,\ n_2 - 1$$

to decide whether $s_1^2 = s_2^2$ or $s_1^2 \neq s_2^2$

2. a. If $s_1^2 = s_2^2$ (and $n_1 \neq n_2$)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{\Sigma X_1^2 + \Sigma X_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$df = n_1 + n_2 - 2$$

b. If $s_1^2 \neq s_2^2$ (and $n_1 \neq n_2$)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

To find the value of t_α when using formula b., first find the values of t_α at df = n_1-1 and at df = n_2-2. Average these values of t_α at df = n_1-1 and df = n_2-2. Add the average to the value of t_α for the lower df (n_1-1 or n_2-2). This is the value of t which the computed value must be greater than or equal to in order for a significant difference to exist.

For the F ratio, the significance level used was .05. If the value of F at the required degree of freedom (d.f.) was lower than the value at .05 (i.e. if $P > .05$), the two variances were considered equal. If the value of F was larger than the value at .05 (i.e. if $P < .05$), the variances were considered unequal. The appropriate formula, (a. or b.) for the t-test was then chosen.

The significance level chosen for the t-test was also .05. As in the case of the Z test (to which t is closely related) the higher the value of the statistic t, the lower the probability of its occurance and the greater the significance of the difference between means.