Chapter - III

3.1. Introduction

Since the first objective of this study is to identify inclusion and exclusion errors in the identification of the poor of Cachar district of Assam, this involved an exhaustive study of the standard of living, economic condition and socio-health status of the households in the study area. Second objective is the estimation of the extent and magnitude of inclusion and exclusion errors in the identification of the poor. This was done by analyzing the result of the assessment made between the Government identification and the present study of identification. Third, the significance of various factors on the poverty estimate in the study area was studied through regression analysis. The research methodology has to be robust in order to minimize errors in the data collection and analysis. Owing to this, various methods, viz., survey interview (face to face method), were chosen for data collection. Analytical tools and techniques employed to best assess, address and understand the objectives of the study, including their limitations are presented. This chapter contains sampling design and size of the sample, database, a brief discussion of the schedule, procedure of data collection, and analytical tools employed.

3.2. Sample Design and Size of the Sample:

To test the first hypothesis, it is important to know the status of the two types of households. First those households who have a BPL/ AAY card but do not deserve

to fall under the category of poor households and Secondly, other households who do not have BPL / AAY card but deserve to fall under the category of poor household. Therefore, we are required to collect the data from both Government identified poor and non-poor households. For this, we collect data from both these types of households. The population in the present study is the households of Cachar district. The households of Cachar district amount to a huge number. It has 3, 79,955 households which is too vast to tackle in a research work of present nature[1]. In view of this, the present study has adopted a multistage random sampling technique. There are three stages which were followed in the present study to collect data. In the first two stages, units of interest (Revenue circles and Wards/ Villages) have been selected randomly using lottery method. However, in the third stage, units (households) have been selected using systematic random sampling.

At the first stage, we have randomly selected the five revenue circles of Cachar district. These are -- Silchar, Lakhipur, Katigora, Sonai and Udharbond. To retain the true representativeness of the entire area, the sites of samples are selected from all the five revenue circles. Among the revenue circles, Silchar and Lakhipur have the Municipal Boards but the other revenue circles have no such boards. So, there are two segments of revenue circle. These are – urban segment and rural segment. Among these, there are two urban segments, viz., Silchar and Lakhipur and three rural segments, viz., Katigora, Sonai and Udharbond. In the second stage, from each urban segment, two wards have been selected randomly, viz., Ward 4 and Ward 8 from Silchar and Ward 5 and Ward 6 from Lakhipur as well as two villages have been selected from each rural segment, viz., Tarapur and Kusiarkul from Katigora, Sonabarighat Part - I and Narsinghpur Part-IV from Sonai, Durganagar Part-VI and

[1] Census of India, 2011

Doyapore Part-II from Udharbond. In the third stage, sample size has been selected from each of the selected wards or villages using a scientific formula originally developed by Cochran (1977)[2]. The formula used in our sample size is:--

$$SS = \frac{Z^2 * (p) * (1 - p)}{c^2}$$

Where,

SS = Sample Size

Z = Z value (e.g. 1.96 for 95% confidence level)

p = percentage picking a choice, expressed as decimal (.1 used for sample size needed)

c =confidence interval, expressed as decimal (e.g., .06 = ± 6)

Correction for finite population:

$$\text{New SS} = \frac{SS}{1 + \frac{SS\text{-}1}{pop}}$$

Where,

pop = Population

The ward/village wise sample sizes so determined are: 87 out of 927 from Ward 4 and 77 out of 376 from Ward 8 of Silchar Revenue circle, 77 out of 394 from ward 5 and 63 out of 179 from Ward 6 of Lakhipur circle, 58 out of 144 households from Kusiarkul and 83 out of 622 from Tarapur of Katigora, 77 out of 387 households from Durganagar Part VI and 77 out of 583 households from Doyapore Part II of Udharbond and 56 out of 132 from Narsinghpur Part IV and 83 out of 594 households from Sonabarighat Part I of Sonai. This sample selection procedure resulted in 738 households out of 4338 households. Once the sample size is

[2] Cochran, W.G. (1977). Sampling technique (3rd edition) New York: John Willey & Sons

determined, the individual households (ultimate sampling unit) have been selected using systematic random sampling. The field survey has been taken during September, 2014 to April, 2015. The whole story of sample design is shown in Table 3.1.

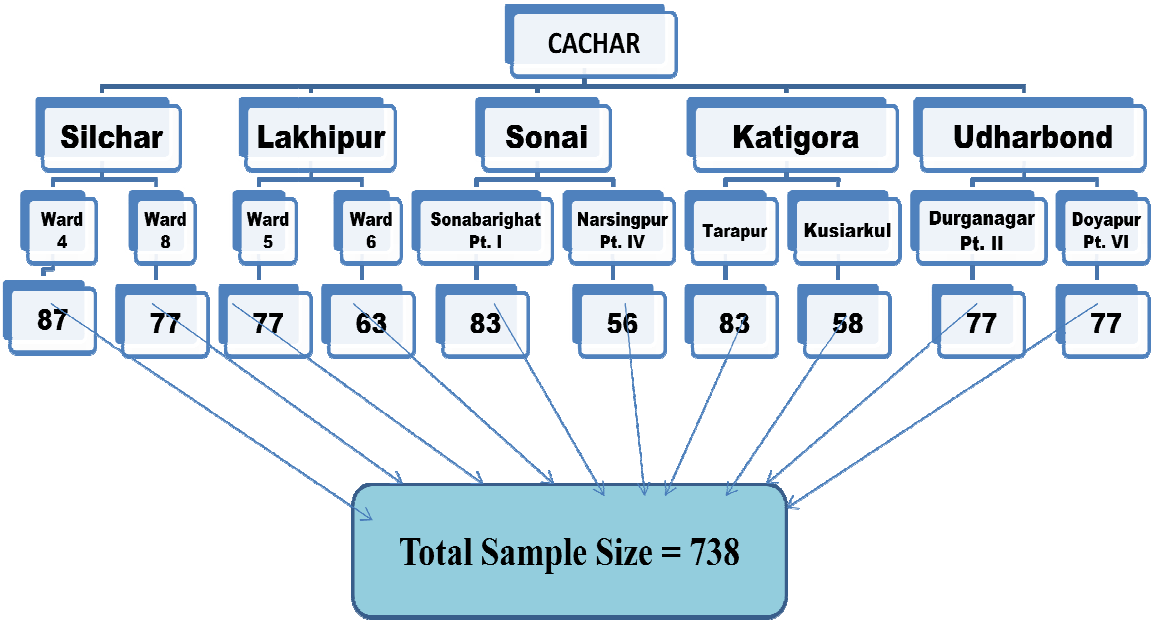
Table 3.1: Statistics of Total Sample Households in the study area

Revenue Circle	Rural / Urban	Villages/ Wards	Total number of households of the Sample villages /Wards	Total number of sample households
Silchar	Urban	Ward – 4	927	87
		Ward – 8	376	77
Lakhipur	Urban	Ward – 5	394	77
		Ward –6	179	63
Katigora	Rural	Tarapur	622	83
		Kusiarkul	144	58
Udharbond	Rural	Durganagar Part VI	387	77
		Doyapore Part II	583	77
Sonai	Rural	Narsinghpur Part IV	132	56
		Sonabarighat Part I	594	83
Cachar District			4338	738

Source: Compiled from Field survey, 2014-15

The actual picture of sample design and size of the sample is presented with the help of a chart which is shown in the figure 3.1.:

Figure 3.1.: Sample design and Sample size



3.3. Database:

The present study has relied both on the primary and on the secondary sources of data. The primary data were collected from both types of households with and without a BPL/AAY card. The data were collected with the help of an interview schedule, having questions both structured and open-ended. In order to understand the socio-economic conditions of the households and to elicit certain information which was difficult to get from the households by asking questions, like the type of sanitation, the observation method was also adopted. The researcher has tried to cross-check the information through discussions with the other members of the households as well as with the local persons and Ward Commissioners of the area.

The sources of secondary data include census reports, Government reports, research papers, published and unpublished works on poverty and Government BPL lists. The secondary data were also collected from the publications of government agencies like, the Office of the Department of Food and Civil Supply, Cachar and the Department of Economics and Statistics, the Government of Assam. Besides, different articles, journals and websites are also referred for the purpose.

3.4. A Brief Description of the Schedule:

In order to collect suitable and desirable information from the household, the schedule has been broadly divided into three parts. These are:

(i) Standard of living of the households,

(ii) Households' economic conditions,

(iii) Information regarding households' social and health conditions.

Accordingly, the interview schedule of the study has been designed and divided into three sections, viz.,

Section I: Standard of Living Status,

Section II: Economical Assessment,

Section III: Social and Health Profile.

Section I: Standard of Living Status:-- This section contains the information about the residential status of the respondents, enquiring about the type of house (where five categories has been given – houseless, kachha, assam-type, semi-pucca and pucca), Household with electricity (in terms of 'yes' or 'no'), Sanitation (which has been categorized as no facility, dry toilet, pit latrine—without slab, pit latrine –

with slab, flush), Type of Cooking fuel (categorization has been made as electricity, bio-gas, coal, wood, LPG, kerosene, animal dung and straw), Source of drinking water.

Section II: Economical Assessment: -- This section is designed to grab the economical condition of the households. It enquires on whether or not the households demonstrated the poor economic condition. Here, six questions are provided. These questions are mainly related to means of livelihood of the economic head of the household, number of working and non-working members in the households, assets of the household (which included the options of TV, electric fan, sound system, radio, mobile fan, refrigerator, two/three/four wheeler, computer, washing machine, air conditioner, inverter), household's expenditure, bank account and holding of BPL/AAY card.

Section III: Social and Health Profile: -- Here, the information about the socio-health background of the households are included. The households were investigated about the family size, religion, caste, gender of the head of the household, status of the household (where information regarding adult male, adult female, teen male, teen female, child male and child female) were collected. They were asked about whether all eligible members of the household have completed the primary education (in terms of 'yes' or 'no') as well as whether any members suffer from chronic disease during the last six months (in terms of 'yes' or 'no'). It is also asked about whether any child aged (5-14) of the household does not attend school (in terms of 'yes' or 'no'). The condition of women whether they participate in decision

making has also been asked by categorizing as – own health care, visit to her family, decision on child education and child health and participate in political affairs.

3.5. Procedure of data collection:

Administration of tools in collecting data is the next important step in any research survey. In the present study, a face to face interview has been conducted through the interview schedule, rather than the self-fill questionnaire survey, telephonic or mail survey etc. Using face to face survey method, the exact scenario can be defined and explained more thoroughly and thus, non-response is minimized.

At first a pilot survey of the study area has been made, by visiting the revenue circles, and establishing contact with the Ward Commissioners, Gaon Panchayat members and local people. Then the interview schedule has been pre-tested in the field, questioning 32 households in order to see how well it serves the purpose of obtaining the needed data. Accordingly, slight necessary changes were made in the schedule, like, minor changes in literacy and health status to make them easily understandable in view of the type of households exist in Cachar district. In order to know the dependency ratio, questions like number of working and non-working members have also been included.

In the final survey, data collection has been done in the eight months period during September, 2014 to April, 2015. Almost all the data were collected by the researcher having accompanied by local people. To begin with, the households were explained about the research study and were convinced to spare some few minutes to give his/her opinion. The standard of living status, economical assessment, social and health profile were presented and asked in a simple and usual manner.

3.5.1. Processing of data:

Raw data collected from the field cannot be analysed straight way; data are to be processed so as make them amenable to analysis. In the present study, processing of data included editing, coding and tabulation.

Editing: Firstly, data collected from the field were edited and examined one by one carefully. Only the completed schedules are accepted. This assures the reliability and accuracy of the data. Keeping in view the editing part, while collecting data in the final survey a little more collection of sample was done than the targeted sample size, so that after sorting and editing the targeted sample size of 10% of the total population does not get affected. Hence, for analyzing the data we ended up with the fixed sample size, i.e., the total of 738 respondents.

Coding: The qualitative data which cannot be measured numerically are to be converted into a quantitative form by assigning some numbers or symbols. In the present study, some attributes are assigned number coding to measure them quantitatively. The detail is given in the logistic regression section.

Tabulation: In order to organize the data systematically, it is preferred to represent the data in tabular format. This helps to facilitate the process of comparison, detection of errors and omission, and to express data in the least possible space. Moreover, it provides a basis for various statistical computations and helps to make the purpose of enquiry more clear.[3]

[3] Verma, (2005): "Practical Approach to Research Methodology", Akansha Publishing House, New Delhi

In this study, some tabular interpretation of data has been done while describing the inclusion and exclusion errors of the identification of the poor and in analyzing various statistical calculations and analysis.

Diagrammatic representation: With a view to present data in an attractive way, it is suitable to present the data diagrammatically. This helps to understand data easily, helps to compare quickly and most importantly helps to make the purpose of study fruitful. This study used mainly bar-diagrams and pie-diagrams.

3.6. Analytical tools:

3.6.1. Identification of the poor household: the Methodological Issues

The popular multidimensional method developed by Sabina Alkire and James Foster has been used in the present study to identify the poor. According to this method, a household is identified as poor either by the Union Approach or by the Intersection Approach. According to *the Union Approach*, a household is identified as poor if the household is deprived in at least one dimension. On the other hand, a household is identified as poor according to *the Intersection Approach* if the household is deprived in all dimensions.[4]

According to Sabina Alkire and James Foster, no one indicator can capture the multiple aspects that constitute poverty, well being or empowerment. And for analysis, it is essential to track the multiple and interconnected disadvantages that poor people experience. The multidimensional poverty measurement method is a flexible technique that can incorporate several different 'dimensions' of poverty or

[4] Alkire and Seth (2008): "Measuring *Poverty in India: A New Proposal*", OPHI Working Paper No. 15.

well-being. This method identifies 'who is poor' by considering the range of deprivations they suffer. It aggregates that information to reflect societal poverty in a way that is robust and can be easily broken down to reveal how people are poor. Measures constructed using the Alkire Foster method can identify interconnections among deprivations and improve policy design. The method captures the percentage of people who are poor (incidence) and the intensity of the poverty experienced by the poor. It is flexible and can incorporate a wide range of dimensions, indicators and cut-offs.

The set of dimensions and the set of indicators which are used for the poverty measurement in the present study are summarized in Table 3.2.

Table 3.2: Dimensions and Indicators for the Poverty Measure

Dimensions	Indicators
1. Standard of Living	1. Type of House
	2.Household with Electricity
	3.Sanitation
	4.Type of Cooking Fuel
	5.Source of Drinking Water
2. Occupational Status	6. Means of Livelihood
	7.Household with Assets
3. Social and Health Status	8. Literacy Status
	9. Health Status
	10. Status of the Women

From the table 3.2, it is clear that the study has used three dimensions and ten indicators. The first dimension, i.e., Standard of living includes five indicators. These are: type of house, household with electricity, sanitation, type of cooking fuel, source of drinking water. Occupational status, the second dimension, has two indicators, viz., means of livelihood and household with assets. Another dimension of this method is social and health status which contains literacy status, health status and status of the women as its indicators.

The detailed description of the indicators and the cut offs can be found in Table 3.3 which is given below:

Table 3.3.: Dimensions, Indicators and cut offs for the Poverty Measure

Dimensions	Indicators	Poverty Cut Off	Status
1. Standard of living	1. Type of House	Live in a Kachha House	Yes= Poor No = Non-Poor
	2.Household with Electricity	No access of Electricity	Yes= Poor No = Non-Poor
	3. Sanitation	Uses no facility/uses bush /field, Composting Toilet or Dry Toilet, Pit Latrine– without slab.	Yes= Poor No = Non-Poor
	4.Type of Cooking Fuel	Uses coal, animal dung, wood, straw/ shrubs/grass.	Yes= Poor No = Non-Poor
	5.Source of drinking water	Uses Unprotected Well and Spring, River, Dam, Lake /Pond, Tanker Truck.	Yes= Poor No = Non-Poor

2. Occupational Status	6.Means of livelihood	Agricultural / Plantation labourers, Casual laborer, Vendor, Fisherman and Driver.	Yes= Poor No = Non-Poor
	7. Household with assets	Owns (any one of the) a TV, an Electric Fan, a Radio and a Mobile Phone but at the same time does not own (any one of the) a Refrigerator, Two Wheelers, Four Wheelers, Computer, Air Conditioner.	Yes= Poor No = Non-Poor
3. Social and Health Status	8. Literacy Status	Maximum year of education completed by any member (eligible) is less than five years	Yes= Poor No = Non-Poor
	9. Health Status	Any member in the household suffers from chronic disease during the last six months.	Yes= Poor No = Non-Poor
	10. Status of the Women	No women in the household has the right to take decision alone at least in one of the following fields: Own healthcare, Purchases	Yes= Poor No = Non-Poor

		of daily household needs, Visits to her family or relatives, Decision on child education, Decision on child health, Participate in political affairs.	

3.6.2. . Normalization of Poverty Indicators: Why and How?

Normalization of values means adjusting values measured on different scales to a notionally common scale, often prior to averaging. The simplest method is rescaling the range of features to scale the range in (0, 1). In some statistical problems, since the range of values of raw data varies widely, objective functions do not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes proportionately to the final distance.

The present study uses three dimensions and ten indicators in the identification of poor. We collected data on all these dimensions and indicators through field survey. It has been observed that, for most variables, the range of values is not uniform. Besides, the unit by which those variables are measured is also different. Thus, it becomes technically necessary to normalize the variables before we analyze them.

In order to normalize the dimensional variables, we have used the popular method which is followed in calculating the Human Development Index (HDI). This is symbolically represented here:

$$Zi = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

Where,

Zi = normalised value of i^{th} dimension.

x_i = actual value of i^{th} dimension.

min x_i = minimum value of i^{th} dimension.

max x_i = maximum value of i^{th} dimension.

3.6.3. Construction of Relative Poverty Index:

Once we have normalized the indicators, in the next step, we have used the following formula to calculate the value of Relative Poverty.

$$\text{RPI} = \frac{\text{SOL} + \text{OS} + \text{SH}}{3}$$

Where, RPI is the Relative Poverty Index, SOL is the normalized value of Standard of Living; OS is the normalized value of Occupational Status and SH is the normalized value of Social and Health status. The value of RPI ranges from 0 to 1. A value of RPI close the 0 indicates extreme poverty while a value close to 1 indicates insignificant poverty.

It may be noted here that in recent time, UNDP has replaced the Arithmetic Mean (AM) formula with Geometric Mean (GM) formula to calculate Human Development Index (HDI) for a number of reasons. However, despite the advantages of GM over AM, the use of GM is not feasible in the present study. This is because

use of GM is condition by the fact that none of the variables can take value equal to zero for any observation. Since in the present study, we do have values for some observations equal to zero, therefore, the application of old method, i.e., the arithmetic mean method is thought to be more suitable than the geometric mean method.

3.6.4. Construction of Severity of Poverty Index:

In order to construct Severity of Poverty Index, in the first step, we have identified the poor household on the basis of the multidimensional criteria. Once we have identified the poor households, we have again followed the above mentioned procedure to normalize the dimensional indicators. In the third step, we have used the following formula to calculate severity of poverty.

$$\text{SPI} = \frac{\text{SOLp} + \text{OSp} + \text{SHp}}{3}$$

Where, SPI is the Severity of Poverty Index, SOLp is the normalized value of Standard of Living of the poor household; OSp is the normalized value of Occupational Status of the poor household and SHp is the normalized value of Social and Health status of the poor household. The value of SPI ranges from 0 to 1. A value of SPI close to 0 indicates extreme poverty while a value close to 1 indicates insignificant poverty.

3.6.5. Descriptive Statistics:

Descriptive statistics is the discipline of quantitatively describing the main characteristics of a collection of information, or the quantitative description itself. Some measures that are commonly used to describe a data set are measures of central

tendency and measures of variability or dispersion. These measures are widely used in the present study.

3.6.6. Test of Mean difference:

In simple terms, the t-test compares the actual difference between two means in relation to the variation in the data (expressed as the standard deviation of the difference between the means). A t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution if the null hypothesis is supported. It can be used to determine if two sets of data are significantly different from each other, and is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistic (under certain conditions) follows a Student's *t* distribution. The t-statistics was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland ("Student" was his pen name).

Most *t*-test statistics have the form $t = Z/s$, where Z and s are functions of the data. Typically, Z is designed to be sensitive to the alternative hypothesis (i.e., its magnitude tends to be larger when the alternative hypothesis is true), whereas s is a scaling parameter that allows the distribution of t to be determined.

$$t = \frac{Z}{(s/\sqrt{n})} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{(s/\sqrt{n})}$$

As an example, in the one-sample *t*-test, where $\bar{X}$ is the sample mean from a sample $X_1, X_2, \ldots, X_n$, of size n, s is the ratio of sample standard deviation over

population standard deviation, σ is the population standard deviation of the data, and μ is the population mean.[5]

The assumptions underlying a *t*-test are that

- X follows a normal distribution with mean μ and variance σ^2
- s^2 follows a $\chi 2$ distribution with p degrees of freedom under the null hypothesis, where p is a positive constant
- Z and s are independent.

Two-sample *t*-tests for a difference in mean involve independent samples or unpaired samples. Paired *t*-tests are a form of blocking, and have greater power than unpaired tests when the paired units are similar with respect to "noise factors" that are independent of membership in the two groups being compared. In a different context, paired *t*-tests can be used to reduce the effects of confounding factors in an observational study. The independent samples *t*-test is used when two separate sets of independent and identically distributed samples are obtained, one from each of the two populations being compared.

Paired samples *t*-tests typically consist of a sample of matched pairs of similar units, or one group of units that has been tested twice (a “repeated measures” *t*-test). A paired samples *t*-test based on a “matched-pairs sample” results from an unpaired sample that is subsequently used to form a paired sample, by using additional variables that were measured along with the variable of interest. The matching is carried out by identifying pairs of values consisting of one observation from each of the two samples, where the pair is similar in terms of other measured variables. This

[5] Gupta (2013): “Fundamental of Statistics”, Seventh revised and enlarged edtition, Himalaya Publishing House.

approach is sometimes used in observational studies to reduce or eliminate the effects of confounding factors. Paired samples *t*-tests are often referred to as "dependent samples *t*-tests".

In each case, the formula for a test statistic that either exactly follows or closely approximates a *t*-distribution under the null hypothesis is given. Also, the appropriate degrees of freedom are given in each case. Each of these statistics can be used to carry out either a one-tailed or two-tailed test. Once a *t* value is determined, a p-value can be found using a table of values from Student's *t*-distribution.

In testing the null hypothesis that the population means is equal to a specified value μ_0, one uses the statistic

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

where $\overline{x}$ is the sample mean, *s* is the sample standard deviation of the sample and *n* is the sample size. The degrees of freedom used in this test are $n - 1$. Although the parent population does not need to be normally distributed, the distribution of the population of sample means, $\overline{x}$, is assumed to be normal. By the central limit theorem, if the sampling of the parent population is independent then the sample means will be approximately normal. (The degree of approximation will depend on how close the parent population is to a normal distribution and the sample size, *n*.)

The *t* statistic to test whether the means are different can be calculated as follows:

$$t = \frac{X_1 - X_2}{s_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where

$$s_{X_1X_2} = \sqrt{\frac{(n_1-1)s_{X_1}^2 + (n_2-1)s_{X_2}^2}{n_1+n_2-2}}.$$

Note that the formulae above are generalizations of the case where both samples have equal sizes (substitute n for n_1 and n_2). $s_{X_1X_2}$ is an estimator of the common standard deviation of the two samples: it is defined in this way so that its square is an unbiased estimator of the common variance whether or not the population means are the same. In these formulae, n = number of participants, 1 = group one, 2 = group two. $n - 1$ is the number of degrees of freedom for either group, and the total sample size minus two (that is, $n_1 + n_2 - 2$) is the total number of degrees of freedom, which is used in significance testing.

The t-test has a number of applications in statistics. Some of these applications are: t-test for significance of single mean, population variance being unknown; t-test for the significance of the different between two sample means, the population variances being equal but unknown; t-test for significance of an observed sample correlation co-efficient.

t- test for two population means with unknown but equal variance is used in the present study. The t values are calculated to capture a part of the first objective i.e. to find whether there is an inclusion and exclusion error in the identification of the poor.

The test of significance based on t-distribution is an adequate procedure only for testing the significance of the difference between two sample means. In a situation when we have three or more samples to consider at a time, an alternate procedure is needed for testing the hypothesis that all the samples are drawn from the populations with the same mean. The basic purpose of the analysis of variance is to test the

homogeneity of several means. By this technique the total variation in the sample data is expressed as the sum of its non-negative components where each of these components is a measure of the variation due to some specific independent source or factor or cause. ANOVA test is based on the test statistic F (or Variation Ratio). It is classified as one-way analysis of variance and two-way analysis of variance. This technique enables us to compare several population means simultaneously and thus results in lot of savings in terms of time and money as compared to several experiments required for comparing two populations means at a time. Moreover, it is now frequently applied in testing the linearity of the fitted regression line or the significance of the correlation ratio.

In this study, we used one-way analysis of variance. Let us suppose that n sample observations of a random variable X are divided into k classes on the basis of some criterion or factor of classification. Let the ith class of n_i observations and let X_{ij} = jth member of the i^{th} class; j= 1, 2,….., n_i ; i = 1,2, ...,k

$$n= n_1+n_2+ .....+n_k = \sum_{i=1}^{k} n_i$$

Such scheme of classification according to a single criterion is called one-way classification and its analysis of variance is known as one-way analysis of variance. [6]

An F-test is any statistical test in which the test statistics has an F-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled. Exact "F-tests" mainly arise when the models have been fitted to the data using least squares. The name was coined by

[6] Gupta (2013): "Fundamental of Statistics", Seventh revised and enlarged edtition, Himalaya Publishing House.

George W. Snedecor, in honor of Sir Ronald A. Fisher. Fisher initially developed the statistic as the variance ratio in the 1920s.

The F-test in one-way analysis of variance is used to assess whether the expected values of a quantitative variable within several pre-defined groups differ from each other. For example, suppose that a medical trial compares four treatments. The ANOVA F-test can be used to assess whether any of the treatments is on average superior, or inferior, to the others versus the null hypothesis that all four treatments yield the same mean response. This is an example of an "omnibus" test, meaning that a single test is performed to detect any of several possible differences. Alternatively, we could carry out pair wise tests among the treatments (for instance, in the medical trial example with four treatments we could carry out six tests among pairs of treatments). The advantage of the ANOVA F-test is that we do not need to pre-specify which treatments are to be compared, and we do not need to adjust for making multiple comparisons. The disadvantage of the ANOVA F-test is that if we reject the null hypothesis, we do not know which treatments can be said to be significantly different from the others, nor, if the F-test is performed at level α we can state that the treatment pair with the greatest mean difference is significantly different at level α.

The formula for the one-way ANOVA F-test statistics is:--

$$F = \frac{\text{explained variance}}{\text{unexplained variance}},$$

or

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}.$$

The "explained variance", or "between-group variability" is

$$\sum_i n_i(\bar{Y}_{i\cdot} - \bar{Y})^2/(K-1)$$

where $\bar{Y}_{i\cdot}$ denotes the sample mean in the i^{th} group, n_i is the number of observations in the i^{th} group, $\bar{Y}$ denotes the overall mean of the data, and K denotes the number of groups.

The "unexplained variance'' or "within-group variability" is

$$\sum_{ij}(Y_{ij} - \bar{Y}_{i\cdot})^2/(N - K),$$

Where Yij is the j^{th} observation in the i^{th} out of K groups and N is the overall sample size. This F-statistic follows the F-distribution with K−1, N −K degrees of freedom under the null hypothesis. The statistic will be large if the between-group variability is large relative to the within-group variability, which is unlikely to happen if the population means of the groups all have the same value.

We have used One Way ANOVA to find out variation in inclusion and exclusion errors across revenue circles and to determine whether there is a significant differences in case inclusion and exclusion errors among the revenue circles.

3.6.7. Regression analysis to identify factors affecting Poverty status of the households:

Regression analysis is concerned with describing and evaluating the relationship between the given variable called the dependent variable and one or more other variable called explanatory variable or independent variable. Generally, the dependent variable is denoted by the symbol Y and the explanatory variable X's.

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function. Logistic regression can be seen as a special case of

generalized linear model and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions (about the relationship between dependent and independent variables) from those of linear regression. In particular the key differences of these two models can be seen in the following two features of logistic regression. First, the conditional distribution $y \mid x$ is a Bernoulli distribution rather than a Gaussian distribution, because the dependent variable is binary. Second, the predicted values are probabilities and are therefore restricted to [0, 1] through the logistic distribution function because logistic regression predicts the probability of particular outcomes. Logistic regression is used when the response or dependent variable is dichotomous (i.e. binary or 0-1), the predictor variables may be quantitative, categorical or a mixture of the two.

The present study adopted the Logistic Regression Analysis considering both the Poor and Non-Poor households, to evaluate the variables influencing poverty status. For this purpose, we have estimated two models. In the first model, we have considered five variables. In the second model, eight variables including four revenue circles have been included. Here, the revenue circle 'Silchar' is not included in the model and it works as bench mark circle. The set of variables which are used to identify the factors influencing the poverty status of the households are discussed and defined below.

Here, the independent variables are five socio-economic variables of the households, viz., literacy, religion, residence, caste, ratio of the working member, and also the five other dummy variables to represent five revenue circles, viz., Silchar, Lakhipur, Sonai, Katigora and Udharbond. It is however noted that out of this five

dummy variables we have used only four while estimating the regression model to avoid perfect multicollinearity problem.

In this study, the above mentioned variables have been identified on the basis of review of literatures and field experience of the researcher. In order to measure the variable, proper conceptualization of the variables is very important task and here, from the Interview Schedule, some variables / attributes are coded with number such as –

Literacy: Literate is defined as minimum attainment of education level that is up to primary education. In the present study, a household is literate if every member in the household has completed primary education. Here, literacy has been categorized and coded as:

- Literate household coded as "1"
- "0" for illiterate household.

Residence: The variable residence, in this study, means the household resides either in urban or in rural areas. Since the variable is a dichotomous one and hence, numerical code has been assigned:

- "Urban" is coded as "1" and
- "Rural" is coded as "0"

Religion: It is a popular one to be considered as one of the major factors that affect people's lifestyle. In our study, the household's religion has been categorised and coded as follows:

- "Hindu" is coded as "1" and
- "Others" is coded as "0"

Caste: In India, caste system is a system of social stratification. In our study, it is simply referred as different classes, viz., General, SC, ST an OBC. For examining influence of caste on the poverty status, we have been categorized the caste system and coded as follows:

- "General" is coded as "1" and
- "Others" is coded as "0"

Ratio of Working Member: In the present study, it simply implies ratio of working members to total members of the household.

Lakhipur: Lakhipur is one of the urban revenue circles. In order to evaluate the poverty status of the households of this circle, we have categorized and coded as below:

- "Household resides in Lakhipur" is coded as "1" and
- "Household resides in other circle" is coded as "0"

Katigora: Katigora is one of the rural revenue circles. In this study, for finding out the poverty status of the households of this circle, the circle is coded as below*:*

- "Household resides in Katigora" is coded as "1" and
- "Household resides in other circle" is coded as "0"

Sonai: Sonai is one of the backward revenue circles as it is not only far away from the Sadar (Silchar) of the Cachar district but also deprived from modern facilities. In this study, we have categorized and coded this circle as below*:*

- "Household resides in Sonai" is coded as "1" and
- "Household resides in other circle" is coded as "0"

Udharbond: Udharbond is the nearest revenue circle of the sadar (Silchar) of the Cachar district. It avails a little more modern facilities compared to other rural revenue circles. In this study, we have categorized and coded this circle as below*:*

- "Household resides in Udharbond" is coded as "1" and
- "Household resides in other circle" is coded as "0"

To form the model in the next step, the dependent variable needs to be specified. The dependent variable of the study is poverty status, i.e., if "a household is non-poor or not".

The foregoing discussion is presented in Table 3.4.

Table 3.4: Description of Variables

Variable	Description
Religion	ith Household's religion, 0=Others, 1=Hindu
Residence	ith Household's residence, 0=Others, 1=Urban
Caste	ith Household's caste, 0=Others, 1= General
Literacy	ith Household's literacy, 0=Others, 1=Literate
Ratio of Working Member	Ratio of working members to total members
Lakhipur	ith Household resides in Lakhipur, 0= Others, 1 = Lakhipur
Katigora	ith Household resides in Katigora, 0= Others, 1 = Katigora
Sonai	ith Household resides in Sonai, 0= Others, 1 = Sonai
Udharbond	ith Household resides in Udharbond, 0= Others, 1 = Udharbond

In case of logit model, generally, the dependent dichotomous variable is coded for the two possible answers. The logit regression estimates the probability of an event happening. In the current context, the 'event' is the non-poor household. Thus, the probability of non-poor household, i.e. the favourable case takes a value of "1" and the poor household i.e. the unfavourable case is coded as "0" as given in table 3.5.

Table 3.5: Dependent Variable Encoding

Original Value	Internal Value
Poor	0
Non-Poor	1

The logistic regression equation used in this study, based on the cumulative logistic probability function and it is specified as follows:

For poverty status without revenue circles:

$$Pi = \frac{1}{1+e^{-Zi}}$$

; $Z_i = a + b_1 (REL) + b_2 (LIT) + b_3 (CAST) + b_4 (RWM) + b_5 (RES) + Ui$

; i = 1, 2 ...5

Where,

P_i = 1 if ith household is non-poor household and 0 for poor household;

Religion (REL) = 1 if ith household is Hindu, 0 otherwise;

Literacy (LIT) = 1 if ith household is literate, 0 otherwise;

Caste (CAST) = 1if ith household is General, 0 otherwise;

Ratio of working member (RWM) = Ratio of working members to total members of the households.

Residence (RES) = 1 if ith household resides in Urban area, 0 otherwise;

b_1 to b_5 = Coefficient of the independent variables;

U_i = Stochastic error.

For poverty status with revenue circles:

$$Pi = \frac{1}{1+e^{-Zi}}$$

$$; Z_i = a + b_1 (REL) + b_2 (LIT) + b_3 (CAST) + b_4 (RWM) + b_5 (LAKH) + b_6 (KATI) + b_7 (SON) + b_8 (UD) + Ui$$

$$; i = 1, 2 \ldots 8$$

Where,

P_i = 1 if ith household is non-poor household and 0 for poor household;

Religion (REL) = 1 if ith household is Hindu, 0 otherwise;

Literacy (LIT) = 1 if ith household is literate, 0 otherwise;

Caste (CAST) = 1if ith household is General, 0 otherwise;

Ratio of working member (RWM) = Ratio of working members to total members of the households.

Lakhipur (LAKH), = 1 if ith household resides in Lakhipur, 0 otherwise;

Katigora (KATI) = 1 if ith household resides in Katigora, 0 otherwise;

Sonai (SON) =1 if ith household resides in Sonai, 0 otherwise;

Udharbond (UD). =1 if ith household resides in Udharbond, 0 otherwise;

b_1 to b_8 = Coefficient of the independent variables;

U_i = Stochastic error.

For poverty status of a particular revenue circle:

$$Pi = \frac{1}{1+e^{-Zi}}$$

$$; Z_i = a + b_1 (REL) + b_2 (LIT) + b_3 (CAST) + b_4 (RWM) + Ui$$

$$; i = 1, 2 ...4$$

Where,

P_i = 1 if ith household is non-poor household and 0 for poor household;

Religion (REL) = 1 if ith household is Hindu, 0 otherwise;

Literacy (LIT) = 1 if ith household is literate, 0 otherwise;

Caste (CAST) = 1if ith household is General, 0 otherwise;

Ratio of working member (RWM) = Ratio of working members to total members of the households.

b_1 to b_4 = Coefficient of the independent variables;

U_i = Stochastic error.