Chapter–3

Research Methodology

S.K. Acharya, Rubi Kundu and G.C. Mishra

In any research, it is very important for making the concept of the study, methods and techniques which are utilized to design the sampling, collect and accumulate information, find revelation of truth and formulation of theories. This chapter deals with the research methodology, which have been adopted for the purpose of the present study. However, the entire discussion has been made under the following sub-themes :

A. Locale of Research

B. Pilot study

C. Sampling design

D. Variables and their measurements

E. Method of data collection

F. Statistical tools used for analysis of data.

1. SELECTION OF LOCALE

The present study was conducted at Basirhat Block-1(rural) of North 24 Parganas district in the state of West Bengal. The district and block were selected purposively due to the following characteristics :

i. Having abundance of problem traits, components and dynamics under study and as describe in objectives.
ii. It would provide easy access.
iii. Having abundance of causal factors to be correlated with the consequence factors, i.e. school drop-out at primary level.
iv. Heterogeneity of problem content and factors.
v. Having gender discrimination – in terms of educational attainment, at primary level as demanded by the objective of the study.

2. PILOT STUDY

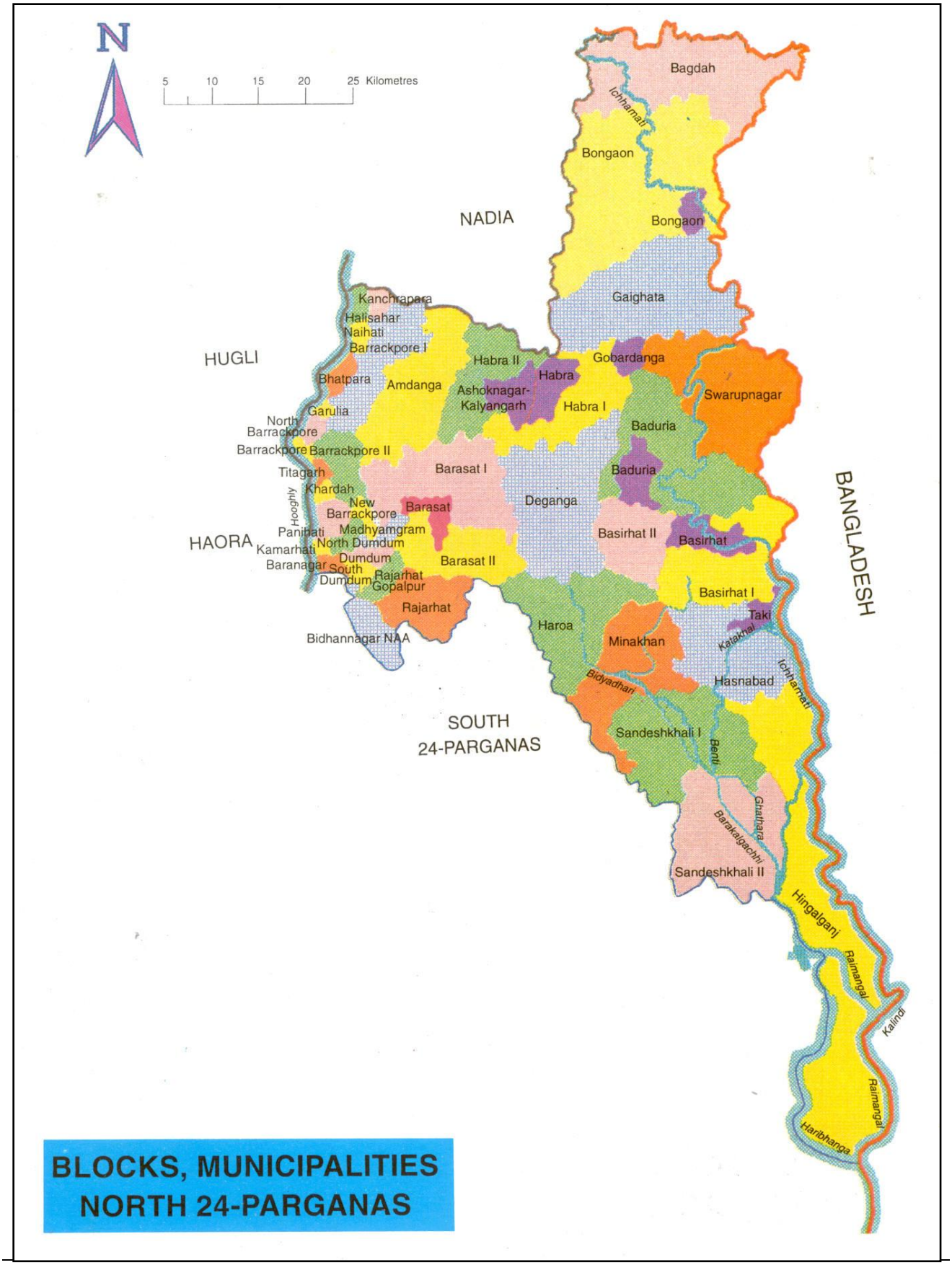
The pilot study was organized to get a primary delineation of the status, intensity and distribution of the researchable problem along with the prospective respondents performing here in.

Obviously, the economic and educational marginalisation of a large section of Indian society should be a matter of concern for all. Today, the situation is being rendered more serious as a result of impact of globalisation and neo-liberal economic policies, that has hit badly that marginalised groups such as peasants, landless labours and artisans.

Again, the situation of Indian children is marked by diversity and persistant disparities. Progress is visible in some developed sectors, yet disparities are striking, with poor status of woman and illiteracy as major barriers. In 1991, 61 per cent of Indian women cannot read or write.

We have selected North 24 Parganas. The district shows huge diversities in the field of nature, financial condition, education, occupation among population. One portion of the district is comprised of Industrial belt. The adjacent area of Kolkata is the place for Middle class educated people. But the rural part of the district across the border of Bangladesh is dominated by marginal farmers and landless labours, most of whom are minority, schedule caste and schedule tribe. The map and location of the district is as follows :

ISBN: 978-81-930585-5-8



ISBN: 978-81-930585-5-8

3. SAMPLING DESIGN : APPROACHES AND PROCESS

Sampling design is the well thought out logical frame work for selective sample representative to the system or population – and thus the selection of sample is used for estimating the parameters of population as a whole.

The purposive as well as simple random sampling and systematic sampling techniques were adopted for the present study. It may be termed as multistage random sampling procedure. The state, district and block are purposively selected for the study.

The table (3.1) furnished which contains data regarding number of drop-outs from different schools that has been obtained through a pilot-study, teacher and parental interaction and verification of attendance register.

ISBN: 978-81-930585-5-8

SAMPLING approach

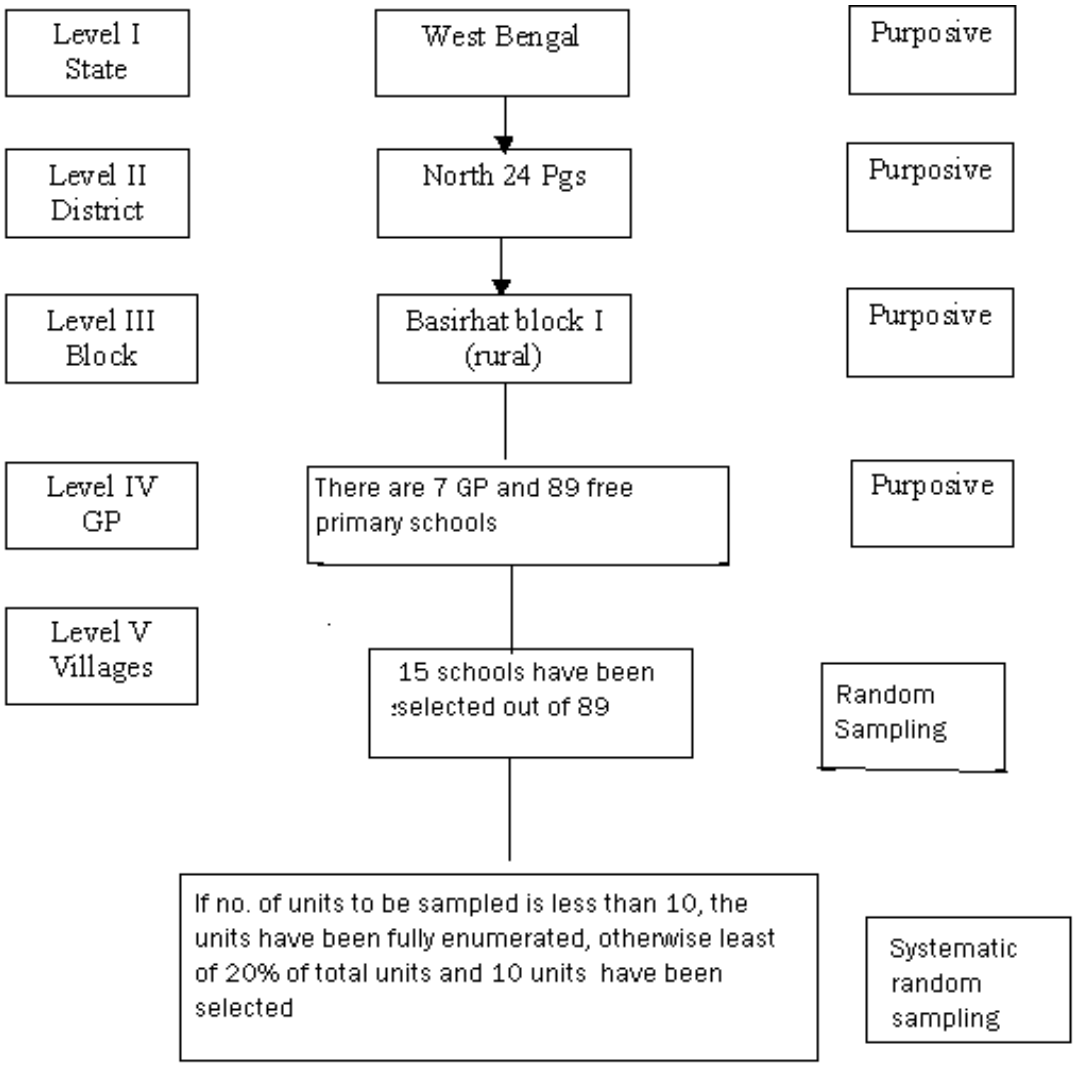


Table 3.1: Table showing the name of the free primary schools and the corresponding Gram Panchayets in Basirhat Block-I (rural) selected as a random sample of 15 schools out of 89; 150 respondents have finally been selected

Sl. No.	Gram Panchayet	School Slected	Dropou /school in last 2 years	Fully enumerated/ sampled	No. of individuais sampled
1	Pifa	Pifa Free Priimary School	60	Sampled	15
2	Pifa	Atkaria Ramnagar Free Priimary School	4	Fully Enumerated	4
3	Pifa	Paikpara Free Priimary School	9	Fully Enumerated	9
4	Sangrampur Shibhati	Biramnagar Free Priimary School	80	Sampled	20
5	Sangrampur Shibhati	Sangrampur Free Priimary School	6	Fully Enumerated	6

ISBN: 978-81-930585-5-8

6	Sangrampur Shibhati	Amarkati Free Priimary School	40	Sampled	10
7	Madhyampur-Olaichandi	Madhyampur Olai Chandi Free Priimary School	2	Fully Enumerated	2
8	Gotra	Anantapur Free Priimary School	92	Sampled	23
9	Shankchura Bagundi	Soladana Free Priimary School	7	Fully Enumerated	7
10	Shankchura Bagundi	Hariharpur Free Priimary School	40	Sampled	10
11	Gacha Aakharpur	Prasannakati Free Priimary School	9	Fully Enumerated	9
12	Gotra	Laxmankati Free Priimary School	48	Sampled	12
13	Gotra	Gotra Free Priimary School	3	Fully Enumerated	3
14	Itinda Panitar	Ghochadanga Free Priimary School	60	Sampled	15
15	Itinda Panitar	Itinda Free Priimary School	5	Fully Enumerated	5

Sampling process

Multistage random sampling process has been followed here to select the sample ultimately. The state, district and the block have been purposively selected based on the following factors.

1) Domicile,
2) Accessibility,
3) Predomination of the problem,
4) Prevalence of the desired variables,
5) Even distribution of the respondents.

Selection of the school catchment area

Initially 89 school catchment areas have been delineated from which 15 school catchment areas have been selected based on proportionate random sampling and total enumeration process.

4. SELECTION OF VARIABLES : VARIABLES AND THEIR MEASUREMENT

The variables were selected based on the following considerations and stages :

a) Consultation with experts, parents of drop-out, teachers, panchayat personnel, school management personnel;

b) Related research works have also been taken care of so that the variable selection could go relevant, precise and truly revealing the problem situation;

c) The variables were primarily selected at a greater number hut subsequently rationalized to fit into the study design.

d) The variables have been catalogued into dependent(Y) and independent categories (X) in such a way as to present a relationship of cause-effect/degree of independence to support the objective of the study. The dependent variables in the present context are; Nature of drop-out(Y), Age at drop-out (Y_1), and level of drop-out (Y_2).

ISBN: 978-81-930585-5-8

Explained variable (Dependent variables)		
1.	Nature of drop out (Y)	Number of months the girls dropped out of the school (Seasonality/Period).
2.	Age at drop out (Y_1)	The chronological age (in years) at which the girl dropped out of the school expressed in years.
3.	Level of drop out (Y_2)	Class (I-IV) in which the girl dropped out
Explanatory variable (Independent Variable)		
1.	Father's Education (X_1)	Education may be operationalised as the amount of formal schooling attained/literacy acquired by the respondent at the time of interview. Education is instrumental in building personality structure and helps in changing one's behaviour in social life. Illiteracy and literacy have been quantified as zero and one. Then educational score has been quantified through number of years of schooling and score of literacy.
2.	Father's Age (X_2)	In all societies, age is one of the most important determinants of social status and social role of the individual. In the present study, the number of years rounded in the nearest whole number of the chronological age the respondents' father, taken as a measure of age of the father.
3.	Total land owned (X_3)	Operationally holding size may be defined as a total size of land possessed by an individual family for the purpose of habitation and cultivation. In the present study homestead land areas in cottah as well as cultivable land has been taken as measure of holding size.
4.	Irrigation Index (X_4)	This is a ratio between land under irrigation to total size of holding expressed in digital form.
5.	Nature of holding (X_5)	Nature of holding has been measured in terms of land ownership status and respective scores. The following has been categorization : owner cultivator (3), share cultivator (2), land less labourer (5).
6.	Cropping intensity (X_6)	Cropping intensity has been operationalised as the proportion of total annually cropped area to the cultivable land area expressed in percentage. The cropping intensity is calculated by the formula, $CI = \frac{\text{Total Annual Cropped Area}}{\text{Net Cultivated Area}} \times 100$
7.	Number of days utilized as family labour by boys in a season (X_7)	To measure the weekly average working days utilized for her family in a year.
8.	Number of days utilized as family labour by girls in a season (x_8)	To measure the average working days utilized for her family in a year.
9.	Expenditure towards Health care (x_9)	Amount in Rupees the family of the girl drop-out incurs towards health-care per month
10.	Expenditure towards Education (x_{10})	Amount in Rupees the family of the girl drop-out incurs towards Education per month
11.	Per capita family expenditure/year (x_{11})	It is obtained by dividing the total yearly expenditure of family.
12.	Monthly family income (x_{12})	It is obtained by dividing the total annual income of the family by twelve (months).
13.	Family size (x_{13})	It is operationalised as the number of members in the individual family.

ISBN: 978-81-930585-5-8

14.	Social interaction value (x_{14})	Frequency of interaction is considered in scores and expressed as sum of total of formal and informal interaction.
15.	Perceived Reason of drop-out (x_{15})	As itemized, there are 18 types of reason. Every perceived reason may be supportive or non-supportive or both. Attach value 1 for non-supportive and 3 for non-supportive statement. So, the maximum possible score is 54. The perceived reasons are as follows:
		Low Family Income
		Agril. and Occupational Engagement(B/G)
		Household Activities(Boys/Girls)
		Health hazards
		Availability of food at school time
		Non-Supportive parental attitudes
		Home Environment
		Lack of Dress Materials
		Non-Supportive Teachers
		School Environment
		Non-Understandability of Text
		Inadequate sitting arrangement
		Relevancy with real life situation
		Meteorological and Geo-climatic features
		Distance of School
		Communication Facility
		Reward & Non-Rewarding School Hours
		Social and Cultural Barrier(taboo)
16.	Fertility status (x_{16})	Fertility status = $\frac{\text{Number of children}}{\text{14-45 (Age of mother)}}$
17.	Frequency of key institutional interaction (x_{17})	Frequency of key institutional interaction that has been expressed as sum of no. of times interacted in a year with different institution like health centre, bank, Primary school, etc.
18.	Distance Matrix (x_{18})	The total linear distance of the house of the girl dropped out from the School, Hospital, Market, and Panchayet separately. They the some of the distance has been divided by the total number of institution the family interacts.
19.	Recreational facility (x_{19})	The frequency (number of times in a month) with which the parents of the girls drop out attended cultural programmes, no. of times the child used to go playground and number of times she attends sports programme.
20.	Mother's age (x_{20})	The variable take age of the mother in years of girl dropped out.
21.	Mother's Education (x_{21})	Number of years of schooling mother of girl drop out finishes.

ISBN: 978-81-930585-5-8

22.	Number of hours mother engaged in hh activity (x_{22})	This variable has been measured by counting the total number of hours the mother of girl drop out engaged herself in household activities in a week has been quantified by the number of household activity.
23.	Girl's age(x_{23})	The variable has been quantified by taking the chronological age of the girl dropout.
24.	Number of hours girl engaged in household activity (x_{24})	$\frac{\text{Total number of hours the girl engaged in household activities/7}}{7}$
25.	Access to text (x_{25})	Access to text is considered for the subject Bengali, Arithmetic, Science, History and Geography. The girl dropped out can meet one of the three levels of access to the subject. The three levels are Poor, Fair and Good. The score corresponding to this three levels are 1, 2, and 3 respectively The variable take average of the scores accessed over the subjects by the girl.
26.	Family Education Score (x_{26})	Family education score (FES) has been quantified by summating up each of the individual family member's education score and then has been divided by the family size. The total score of a family on education has been summated and then divided by the total number of family member of the same family. $FES = \frac{\text{Total Educational Score of a family}}{\text{Family Size}}$
27.	Calorie Intake (x_{27})	It has been the expression of total food values in terms of calorie intake, obtained through the calculation of respective calorie values taken in by the girls per day. The quantification follows the standard nutrition and calorie count inscribed in the table.

Calorie intake is calculated as follows :

Food item	Kcal/ 100gms	Remarks
Rice	334.34	Averaged over different varieties
Roots and tubers	76.81	Averaged over different types
Leafy vegetables	39.79	Averaged over different types
Vegetables	69.83	Averaged over different types
Nuts and Oil seeds	500.33	Averaged over different types
Pulses	378.34	Averaged over different types
Fish	106.59	Averaged over different types
Egg	163	Averaged over different types
Fruit	42	Averaged over different types
Milk	67	Averaged over milk of different types domestic animals like cow, goat.

ISBN: 978-81-930585-5-8

		Condiments and spices	242.56	Averaged over different types

28.	Information use index (x_{28})	There are three levels of a tool for collecting information; sometimes, often, very often and 1, 2, 3 values have been attached respectively to this levels. The different sources as itemized by the responded for gathering information are as follows:			
		a) Radio	Some times (1)	Often (2)	Very often (3)
		b) Newspaper			
		c) Educational film			
		d) Farm publication			
		e) Poster			
		f) Leaflet			
		g) Booklet			
		h) Television			
		i) Field trip			
		j) Demonstration			
		k) Krishimela			
		l) Exhibition			

Construction of Schedule: Prior in framing of the schedule the following activities to be catered

1) **Collection of items :** Items have denoted the factors interplaying in characterizing the dropout phenomen in the research locale. It is done through pilot survey, face to face interaction, brain storming, focus group discussion with the different layers of stakeholder of the local.
2) **Selection of items :** The items presenting not even perceptable variation were discussed
3) **Schedule has been drafted first,** After drafting the schedule, necessary correction, rectification or improvisation have been made where it demands.
4) **Finalization of schedule : After** primary field trial , and subsequent revisions catered there in, for the finalizations of the schedule.

5. METHOD OF DATA COLLECTION

Construction of schedule after pretesting

The draft schedule for collection of data, incorporating the tools and techniques of different variables was presented twice each time on sample respondents. The quantification was done for each and every variable after operationalising them. Before final data collection, entire schedule was pre-tested for elimination, addition and alternation with non-sample respondents of the study area. In pre testing, care was taken not to include respondents who were selected as sample for final interview. On the basis of the experiences in pretesting, appropriate changes in the construction of item and their sequence were made. The schedule was then finalized and multiplicated. The final form of the schedule is given in the appendix.

Field data collection

The data were collected during schedule constructed for the study. The schedule was administered to the respondent in local language and the responses were recorded in English on the schedule. The interview was carried out by the researcher herself.

ISBN: 978-81-930585-5-8

Case study method

The case study method is a very popular form of qualitative analysis and involves a careful and complete observation of a social unit, be that unit a person, a family, an institution, a cultural group or even the entire community. It is a method of study in depth rather than breadth. The case study places more emphasis on the full analysis of a limited number of events or conditions and their interrelations. The case study deals with the processes that take place and their interrelationship. Thus, case study is essentially an intensive investigation of the particular unit under consideration. The object of the case study method is to locate the factors that account for the behaviour patterns of the given unit as an integrated totality.

6. STATISTICAL TOOLS USED

Mean, Standard deviation, Correlation Co-efficient. Path Analysis, Discriminant Analysis, Factor Analysis, Step down regression, Canonical Analysis.

Analysis

Tools	Analysis
Mean	The mean is the arithmetic average and is the result obtained when the sum of values of the individuals in the data is divided by the number of individuals in the data. Mean is the simplest and relatively stable measure of central tendency. We can work it out as under : Mean or ($\overline{X}$) = $\frac{\sum Xi}{n} = \frac{X_1 + X_2 + ... + X_n}{n}$ Where, ($\overline{X}$) = The symbol we use for mean (pronounced as × bar) ∑ = Symbol for summation X_i = Values of the i^{th} item X, i = 1, 2 ……, n n = total number of items
Tools	**Analysis**
	Mean is used in summarizing the essential features of a series and in enabling data to be compared. It is a relatively stable measure of central tendency. But it suffers from some limitation viz., it is unduly affected by extreme, it may not coincide with the actual value of an item in a series, and it may lead to strong impressions, particularly when the item values are not given with the average. However, mean is better than other averages, specially in economic and social studies where direct quantitative measurements are possible.
Standard deviation	Standard deviation is the most widely used measure of dispersion of a series and is commonly denoted by the symbol "σ" (pronounced as sigma). Standard deviation is defined as the square root of the average of squares of deviations, when such deviations for the values of individual items in a series are obtained from the arithmetic average. It is worked out as under. $\text{Standard deviation}(\sigma) = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n}}$

Contd. …

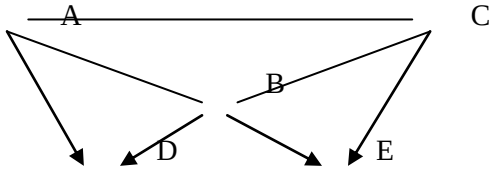
ISBN: 978-81-930585-5-8

Coeffi-cient of varia-tion	When we divide the standard deviation by the arithmetic average of the series, the resulting quantity is known as coefficient of standard deviation, which happens to be a relative measure, and is often used for comparing with similar measure of other series. When this coefficient of standard deviation is multiplied by 100, the resulting figure is known as coefficient of variation. Sometimes, we work out the squares of standard deviation, known as variance, which is frequently used in the context of analysis of variation. $\text{Coefficient of Variation (CV)} = \frac{\text{Standard deviation}}{\text{mean}} \times 100$
Coefficient of correlation	Karl Pearson's coefficient of correlation (or simple correlation) is the most widely used method of measuring the degree of relationship between two variables. This coefficient assumes the following :
Tools	**Analysis**
	That there is linear relationship between the two variables; that the two variables are causally related which means that one of the variable is independent and the other one is dependent; and A large number of independent causes are operating in both variables so as to produce a normal distribution. Karl Pearson's coefficient of correlation $r = \frac{\sum (X_1 - \overline{X})(Y_i - \overline{Y})}{n.a \times \sigma Y}$ Where, X_i = i^{th} value of X variable $\overline{X}$ = mean of X Y_i = i^{th} value of Y variable $\overline{Y}$ = mean of Y n = number of pairs of observations of X and Y σX = Standard deviation of X σY = Standard deviation of Y
	Karl Pearson's coefficient of correlation is also known as the product moment correlation coefficient. The value of 'r' lies between + 1 to –1. Positive values of r indicate positive correlation between the two variables (i.e., changes in both value take place in the same direction), whereas negative values of r indicate negative correlation i.e., changes in the two variables taking place in the opposite direction. A zero value of 'r' indicates that there is no association between the two variables. When r (+) 1, it indicates perfect positive correlation and when it is (-) 1, it indicates perfect negative correlation, meaning thereby that variations in independent variable (X) explain 100 per cent of changes in dependent variable (Y). We can also say that for a unit change in independent variable, if there happens to be a constant change in the dependent variable in the same direction, then correlation will be termed as perfect positive. But if such changes occurs in the opposite direction, the correlation will be termed as perfect negative. The value of 'r' nearer to + 1 or – 1 indicates high degree of correlation between the two

ISBN: 978-81-930585-5-8

Tools	Analysis
Regression Analysis	To assess the regressional effect. $\beta = r^*(\sigma y / \sigma x)$ Regression analysis : Variable selection technique like factor analysis was used to find out the Minimum Data Set (MDS) to identify the key variables (predictor) and the effect of such variables will be tested by various technique like stepwise regression technique and logistic regression technique. Regression is the determination of a statistical relationship between two or more variables. When there are two or more than two independent variables, the analysis concerning relationship is known as multiple correlation and the equation describing such relationship as the multiple regression equation. We here explain multiple correlation and regression taking only two independent variables and one dependent variable. In this situation the results are interpreted as shown below $T = a + b1X1 + b2X2$ Where X_1 and X_2 are two independent variables and y being the dependent variable, and the constants a, b_1 and b_2 can be solved by solving the following three normal equations : $\sum y_i = na + b_1 \sum X_{1i} + b_2 \sum X_{2i}$ $\sum X_{1i}y_i = a \sum X_{1i} + b_1 \sum X^2_{1i} + b_2 \sum X_{1i} X_{2i}$ $\sum X_{2i}y_i = a \sum X_{2i} + b_1 \sum X_{1i} \sum X_{2i} + b_2 \sum X^2_{2i}$ It may be noted that the number of normal equations would depend upon the number of independent variables. If there are 2 independent variables, then 3 erquations, if there are 3 independent variables then 4 eauations and so on, are used. In the present study, three regression equations were extracted for three y variables separately as dependent variables and X_1-X_{28} as predictor variable
Tools	**Analysis**
Stepping Method Criteria	These options apply when stepwise variable selection method has been specified. Variables can be entered or removed from the model depending on either the significance (probability) of the F value or the F value itself. Method selection allows you to specify how independent variables are entered into the analysis. Using different methods, you can construct a variety of regression models from the same set of variables. Stepwise variable entry and removal examines the variables in the block at each step for entry or removal. This is a forward stepwise procedure. The significance values in your output are based on fitting a single model. Therefore, the significance values are generally invalid when a stepwise method (Stepwise, Forward, or Backward) is used. All variables must pass the tolerance criterion to be entered in the equation, regardless of the entry method specified. The default tolerance level is 0.0001. Also, a variable is not entered if it would cause the tolerance of another variable already in the model to drop below the tolerance criterion. All independent variables selected are added to a single regression model.

ISBN: 978-81-930585-5-8

Path Analysis	$r^* \frac{\sigma_y}{\sigma_x}$ Is used to assess the direct, indirect and residual influences of exogenous variables on endogenous variables. Path analysis is an extension of the regression model, used to test the fit of the correlation matrix against two or more causal models which are being compared by the researcher. The model is usually depicted in a circle-and-arrow figure in which single arrows indicate causation. A regression is done for each variable in the model as a dependent on others which the model indicates are causes. The regression weights predicted by the model are compared with the observed correlation matrix for the variables.
Tools	**Analysis**
	# Path model. A path model is a diagram relating independent, intermediary, and dependent variables. Single arrows indicate causation between exogenous or intermediary variables and the dependent(s). Arrows also connect the error terms with their respective endogenous variables. Double arrows indicate correlation between pairs of exogenous variables. Sometimes the width of the arrows in the path model are drawn in a width which is proportional to the absolute magnitude of the corresponding path coefficients (see below). 1. Causal paths to a given variable include (1) the direct paths from arrows leading to it, and (2) correlated paths from endogenous variables correlated with others which have arrows leading to the given variable. Consider this model:  This model has correlated exogenous variables A, B, and C, and endogenous variables D and E. Error terms are not shown. The causal paths relevant to variable D are the paths from A to D, from B to D, and the paths reflecting common anteceding causes -- the paths from B to A to D, from C to A to D, and from C to B to D. Paths involving two correlations (C to B to A to D) are not relevant. Likewise, paths that go backward (E to B to D, or E to B to A to D) reflect common effects and are not relevant. Path coefficient/path weight. A path coefficient is a standardized regression coefficient (beta) showing the direct effect of an independent variable on a dependent variable in the path model. Thus when the model has two or more causal variables, path coefficients are partial regression coefficients which measure the extent of effect of one variable on another in the path model controlling for other prior variables, using standardized data or a correlation matrix as input.
Tools	**Analysis**
	Recall that for bivariate regression, the beta weight (the b coefficient for standardized data) is the same as the correlation coefficient, so for the case of a path model with a variable as a dependent of a single exogenous variable (and an error residual term), the path coefficient in this special case is a zero-order correlation coefficient.

Contd. …

ISBN: 978-81-930585-5-8

Factor Analysis	Many statistical methods are used to study the relation between independent and dependent variables. Factor analysis is different; it is used to study the patterns of relationship among many dependent variables, with the goal of discovering something about the nature of the independent variables that affect them, even though those independent variables were not measured directly. Thus answers obtained by factor analysis are necessarily more hypothetical and tentative than is true when independent variables are observed directly. The inferred independent variables are called factors. A typical factor analysis suggests answers to four major questions: a) How many different factors are needed to explain the pattern of relationships among these variables? b) What is the nature of those factors? c) How well do the hypothesized factors explain the observed data? d) How much purely random or unique variance does each observed variable include?
Discriminant Analysis	The main use of discriminant analysis is to predict group membership from a set of predictors. Discriminant function analysis consists of finding a transform which gives the maximum ratio of difference between a pair of group multivariate means to the multivariate variance within the two groups. Accordingly, an attempt is made to delineate based upon maximizing between group variance while minimizing within group variance. The predictors characteristics are related to form groups based upon
Tools	**Analysis**
	similarities of distribution n-dimensional space which are then compared to groups which are input by the user as truth. This enables the user to test the validity of groups based upon actual data, to test groups which have been created, or to put objects into groups. Discriminant analysis (DA) may act as a univariate regression and is also related to ANOVA (Wesolowsky, 1976). The relationship to ANOVA is such that DA may be considered as a multivariate version of ANOVA. The underlying assumptions of DA are: • the observations are a random sample, • each group is normally distributed, DA is relatively robust to departures from normality, • the variance/covariance matrix for each group is the same, • each of the observations in the initial classification is correctly classified (training data). Mahalanobis Distance A measure of how much a case's values on the independent variables differ from the average of all cases. A large Mahalanobis distance identifies a case as having extreme values on one or more of the independent variables.
Canonical Correlation Analysis	There are several measures of correlation to express the relationship between two or more variables. For example, the standard Pearson product moment correlation coefficient (r) measures the extent to which two variables are related; there are various nonparametric measures of relationships that are based on the similarity of ranks in two variables; Multiple Regression allows one to assess the relationship between a dependent variable and a set of independent variables; Multiple Correspondence Analysis is useful for exploring the relationships between a set of categorical variables. A sociologist may want to investigate the relationship between two predictors of social mobility based on interviews, with actual subsequent social mobility as measured by four different indicators.

Contd. …

ISBN: 978-81-930585-5-8

Tools	Analysis
	A medical researcher may want to study the relationship of various risk factors to the development of a group of symptoms. In all of these cases, the researcher is interested in the relationship between two sets of variables, and Canonical Correlation would be the appropriate method of analysis. Computational Methods and Results Some of the computational issues involved in canonical correlation and the major results that are commonly reported will now be reviewed. Eigenvalues : When extracting the canonical roots, you will compute the eigenvalues. These can be interpreted as the proportion of variance accounted for by the correlation between the respective canonical variates. It is noted that the proportion here is computed relative to the variance of the canonical variates, that is, of the weighted sum scores of the two sets of variables; the eigenvalues do not tell how much variability is explained in either set of variables. It is computed as many eigenvalues as there are canonical roots, that is, as many as the minimum number of variables in either of the two sets. Significance of Roots : The significance test of the canonical correlations is straightforward in principle. Simply stated, the different canonical correlations are tested, one by one, beginning with the largest one. Only those roots that are statistically significant are then retained for subsequent interpretation. Canonical weights : After determining the number of significant canonical roots, the question arises as to how to interpret each (significant) root. In general, the larger the weight (i.e., the absolute value of the weight), the greater is the respective variable's unique positive or negative contribution to the sum. To facilitate comparisons between weights, the canonical weights are usually reported for the standardized variables, that is, for the z transformed variables with a mean of 0 and a standard deviation of 1.
Tools	**Analysis**
	Factor structure : Another way of interpreting the canonical roots is to look at the simple correlations between the canonical variates (or factors) and the variables in each set. These correlations are also called canonical factor loadings. The logic here is that variables that are highly correlated with a canonical variate have more in common with it. Therefore, one should weigh them more heavily when deriving a meaningful interpretation of the respective canonical variate. This method of interpreting canonical variates is identical to the manner in which factors are interpreted in factor analysis. Variance extracted : As discussed earlier, the canonical correlation coefficient refers to the correlation between the weighted sums of the two sets of variables. It tells nothing about how much variability (variance) each canonical root explains in the variables. However, you can infer the proportion of variance extracted from each set of variables by a particular root by looking at the canonical factor loadings. Redundancy : The canonical correlations can be squared to compute the proportion of variance shared by the sum scores (canonical variates) in each set. If you multiply this proportion by the proportion of variance extracted, you arrive at a measure of redundancy, that is, of how redundant one set of variables is, given the other set of

ISBN: 978-81-930585-5-8

	variables. In equation form, you may express the redundancy as: Redundancy left = [(loadingsleft2)/p]*Rc2 Redundancy right = [(loadingsright2)/q]*Rc2 In these equations, p denotes the number of variables in the first (left) set of variables, and q denotes the number of variables in the second (right) set of variables; Rc2 is the respective squared canonical correlation. One can compute the redundancy of the first (left) set of variables given the second (right) set, and the redundancy of the second (right) set of variables, given the first (left) set. Because successively extracted canonical roots are uncorrelated,
Tools	**Analysis**
	you could sum up the redundancies across all (or only the first significant) roots to arrive at a single index of redundancy. Practical significance : The measure of redundancy is also useful for assessing the practical significance of canonical roots. With large sample sizes (see below), canonical correlations of magnitude R = .30 may become statistically significant (see above). If one square this coefficient (R-square = .09) and use it in the redundancy formula shown above, it becomes clear that such canonical roots account for only very little variability in the variables. Of course, the final assessment of what does and does not constitute a finding of practical significance is subjective by nature. However, to maintain a realistic appraisal of how much actual variance (in the variables) is accounted for by a canonical root, it is important to always keep in mind the redundancy measure, that is, how much of the actual variability in one set of variables is explained by the other.

ISBN: 978-81-930585-5-8