B.Sc. GEOGRAPHY LAB MANUAL

3rd Semester

MIDNAPORE CITY COLLEGE

MCC

ESTD. - 2017

Prepared By

Pure & Applied Science Dept.

Geography

MIDNAPORE CITY COLLEGE

MIDNAPORE CITY COLLEGE

DEPARTMENT OF PURE AND APPLIED SCIENCES

LABORATORY MANUAL FOR BACHELOR OF SCIENCE (HONOURS)

IN

GEOGRAPHY

SEMESTER - III

PREFACE TO THE FIRST EDITION

This is the first edition of Lab Manual for BSc Honours in Geography (Third Semester). Hope this edition will help you during practical. This edition mainly tried to cover the whole syllabus. Some hard topics are not present here that will be guided by responsive teachers at the time of practical.

ACKNOWLEDGEMENT

We are really thankful to our students, teachers, and non-teaching staffs to make this effort little bit complete. Mainly thanks to Director and Principal Sir to motivate for making this lab manual.

C6P – Statistical Methods in Geography

1. Construction of data matrix with each row representing an aerial unit (districts / blocks / mouzas / towns) and corresponding columns of relevant attributes.

2. Based on the above, a frequency table, measures of central tendency and dispersion would be computed and interpreted.

3. Histograms and frequency curve would be prepared on the dataset.

4. From the data matrix a sample set (20%) would be drawn using, random, and systematic and stratified methods of sampling and locate the samples on a map with a short note on methods used.

5. Based on of the sample set and using two relevant attributes, a scatter diagram and regression line would be plotted and residual from regression would be mapped with a short interpretation.

1. Construction of data matrix with each row representing an aerial unit (districts / blocks / mouzas / towns) and corresponding columns of relevant attributes.

Data collected either from Primary or Secondary source need to be systematically presented as these are invariably in unsystematic or rudimentary form. Such raw data fail to reveal any meaningful information. The data should be rearranged and classified in a suitable manner to understand the trend and message of the collected information.

Discrete Frequency Distribution:

Here different observations are not written as in simple array. Here we count the number of times any observation appears which is known as frequency. The literary meaning of frequency is the number or occurrence of a particular event/ score in a set of samples. It shows the number of observations from the data set that fall into each of the classes.

Grouped Frequency Distribution:

The quantitative phenomena under study is termed as Variable. Variables are of two kinds: (i) continuous variable, and (ii) discrete variable. Those variables which can take all the possible values in a given specified range are termed as Continuous variable. For example, age (it can be measured in years, months, days, hours, minutes, seconds etc.) , weight (lbs), height (in cms), etc. On the other hand, those variables which cannot take all the possible values within the given specified range, are termed as discrete variables. For example, number of children, marks obtained in an examination (out of 200), etc.

There are three methods for describing the class limits for distribution

- Exclusive method
- Inclusive method
- True or actual class method

1. Exclusive method: In this method of class formation, the classes are so formed that the upper limit of one class also becomes the lower limit of the next class. Exclusive method of classification ensures continuity between two successive classes. In this classification, it is presumed that score equal to the upper limit of the class is exclusive, i.e., a score of 40 will be included in the class of 40 to 50 and not in a class of 30 to 40.

2. Inclusive method: In this method classification includes scores, which are equal to the upper limit of the class. Inclusive method is preferred when measurements are given in the whole numbers. Above example may be presented in the following form by using inclusive method of classification.

3. True or Actual class method: In inclusive method upper class limit is not equal to lower class limit of the next class. Therefore, there is no continuity between the classes. However, in many statistical measures continuous classes are required. To have continuous classes, it is assumed that an observation or score does not just represent a point on a continuous scale but an internal of unit length of which the given score is the middle point. Thus, mathematically, a score is internal when it extends from 0.5 units below to 0.5 units above the face value of the score on a continuum. These class limits are known as true or actual class limits.

<u>Types of Grouped Frequency Distributions</u>

There are various ways to arrange frequencies of a data array based on the requirement of the statistical analysis or the study. A few of them are discussed below.

i) Open End Frequency Distribution: Open end frequency distribution is one which has at least one of its ends open. Either the lower limit of the first class or upper limit of the last class or both are not specified.

ii) Relative frequency distribution: A relative frequency distribution is a distribution that indicates the proportion of the total number of cases observed at each score value or internal of score values.

iii) Cumulative frequency distribution: Sometimes investigator is interested to know the number of observations less than a particular value. This is possible by computing the cumulative frequency. A cumulative frequency corresponding to a class-interval is the sum of frequencies for that class and of all classes prior to that class.

iv) Cumulative relative frequency distribution: A cumulative relative frequency distribution is one in which the entry of any score of class-interval expresses that score's cumulative frequency as a proportion of the total number of cases.

You are given, below the district wise distribution of population in west Bengal as per 2011 census. Arrange the data in frequency distribution table. **[Univariate Frequency Distribution, Continuous Data, after two decimal place, Inclusive Method]**

Serial Number	District Name	Population (As per Census 2011)	Population Divided by 100000
01	Bankura	3596674	35.96
02	Burdwan	7717563	77.2
03	Birbhum	3502404	35.2
04	Coochbehar	2819086	28.2
05	Dakshin Dinajpur	1676276	16.76
06	Hoogly	5519145	55.2
07	Howrah	4850029	48.5
08	Jalpaiguri	3872846	38.7
09	Kalingpong	1846823	18.5
10	Kolkata	4496694	44.97
11	Maldah	3988845	39.9
12	Murshidabad	7103807	71.04
13	Nadia	5167600	51.68
14	North-24th Paragana	10009781	100.1
15	Paschim Medinipur	5913457	59.1
16	Purba Medinipur	5095875	50.96
17	Purulia	2930115	29.3
18	South-24th Paragana	8161961	81.62
19	Uttar Dinajpur	3007134	30.07

District Wise all Population divided with 100000, arrange the data in ascending order.

We have

16.76, 18.50, 28.20, 29.30, 30.07, 35.20, 35.96, 38.70, 39.90, 44.97, 48.50, 50.96, 51.68, 55.20, 59.10, 71.04, 77.20, 81.62, 100.01

Solution

Step - I

Number of class= 1+3.322 log N

Or, 1+3.322 log 19

=5.2

=5 (Approx.)

Step - II

Class interval = Highest value – Lowest value ÷ Number of class

= (100.1 – 16.76) ÷ 5

= 16.67 (Approx.)

= 18 (Approx.)

2. Based on the above, a frequency table, measures of central tendency and dispersion would be computed and interpreted.

TABULATION OF DATA

Components of a Statistical Table

The main components of a table are given below:

1. Table number: When there are more than one tables in a particular analysis, a table should be marked with a number for their reference and identification. The number should be written in the center at the top of the table.

2. Title of the table: Every table should have an appropriate title, which describes the content of the table. The title should be clear, brief, and self-explanatory. Title of the table should be placed either centrally on the top of the table or just below or after the table number.

3. Caption: Captions are brief and self-explanatory headings for columns. Captions may involve headings and sub-headings. The captions should be placed in the middle of the columns. For example, we can divide students of a class into males and females, rural and urban, high SES and Low SES etc.

4. Head note: This is written at the extreme right hand below the title and explains the unit of the measurements used in the body of the tables.

5. Footnote: This is a qualifying statement which is to be written below the table explaining certain points related to the data which have not been covered in title, caption, and stubs.

6. Source of data: The source from which data have been taken is to be mentioned at the end of the table. Reference of the source must be complete so that if the potential reader wants to consult the original source they may do so.

Class Interval	Tally Mark	Class Frequency	Class Limit		Class Boundary	Class Mark/Mid Value	Width	Frequency Density	Relative Frequency	Percentage Frequency
			Low	High						
16-33	𝍸	5	16	33	15.5 – 33.5	24.5	18	0.3	0.3	30%
34-51	𝍸 ///	8	34	51	33.5 – 51.5	42.5	18	0.4	0.4	40%
52-69	//	2	52	69	51.5 – 69.5	60.5	18	0.1	0.1	10%
70-87	///	3	70	87	69.5 – 87.5	78.5	18	0.2	0.2	20%
88-105	/	1	88	105	87.5 -105.5	96.5	18	0.05	0.05	5%

CONCEPT OF CENTRAL TENDENCY

Central tendency is defined as "the statistical measure that identifies a single value as representative of an entire distribution." It aims to provide an accurate description of the entire data. It is the single value that is most typical/representative of the collected data. The term "number crunching" is used to illustrate this aspect of data description. The mean, median and mode are the three commonly used measures of central tendency.

➢ ARITHMETIC MEAN

The arithmetic mean is commonly known as mean. It is a measure of central tendency because other figures of the data congregate around it. Arithmetic mean is obtained by dividing the sum of the values of all observations in the given data set by the number of observations in that set. It is the most commonly used statistical average in the disciplines such as commerce, management, economics, finance, production, etc. The arithmetic mean is also called as simple Arithmetic Mean.

Calculate arithmetic means from the following data

Class Boundary	15.5-33.5	33.5-51.5	51.5-69.5	69.5-87.5	87.5-100.5
Frequency	5	8	2	3	1

Step - I

Calculation of arithmetic mean ($\bar{x}$) by direct method (Equal Class Interval)

Class Boundary	Class Width	Mid Value (x)	Frequency (f)	fx
15.5-33.5	18	24.5	5	122.5
33.5-51.5	18	42.5	8	340
51.5-69.5	18	60.5	2	121
69.5-87.5	18	78.5	3	235.5
87.5-100.5	18	96.5	1	96.5
			Σ f = 19	**Σ fx = 915.5**

Step - II

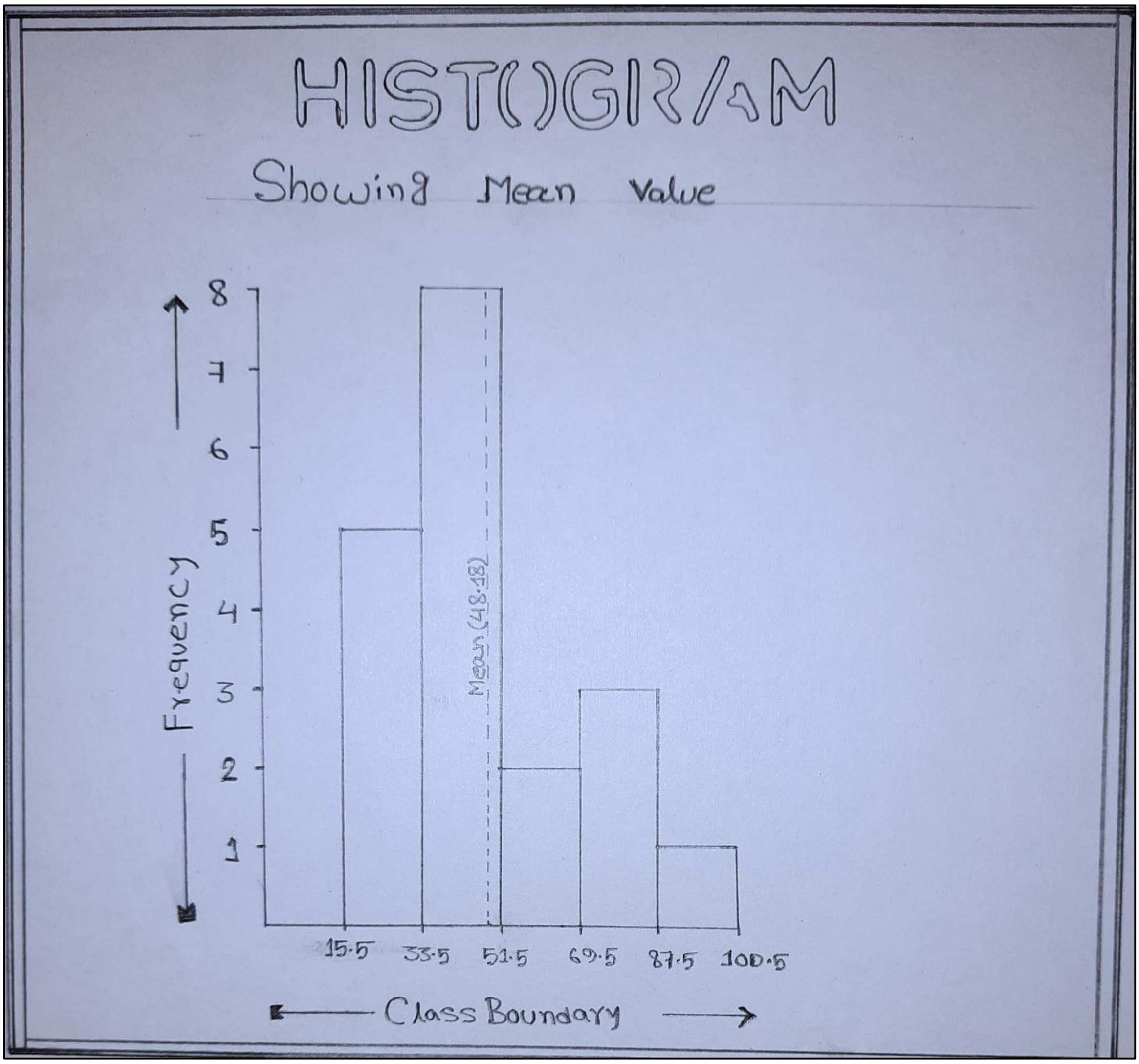
Applying the formula of arithmetic mean

Mean $(\overline{x}) = \Sigma fx \div \Sigma f \div N$

$= 915.5 \div 19$ Where, f = Frequency

$= 48.18$ x = Variable

Σf = Total frequency



➢ MEDIAN

The median is also a measure of central tendency. Unlike arithmetic mean, this median is based on the position of a given observation in a series arranged in an ascending or I descending order. Therefore, it is called a positional average. It has nothing to do with the magnitude of all the observations, as in the case of arithmetic mean. Simply, median refers to the middlemost value of the variable when they are arranged in order of magnitude. The position of the median in a series is such that an equal number of items 'lie on either side of it. Median of a given series is the value of the variable that divides the series-into two equal parts. It is the most central point of a series where half of the items lie above this value and the remaining half lie below this value. In the case of a frequency curve the median is that value of the variable which splits the area into two equal parts.

Merits and Limitations of Median

You have studied the meaning, methods of computation and properties of median. Now let us discuss the merits and limitations of median.

- **Merits**

1) For an open-ended distribution, such as income distribution, the median gives a more representative value.
2) Since median is not distorted by the extreme items, in some cases it is preferred over. mean as the latter is likely to be distorted by extreme values.
3) For dealing the qualitative phenomena, median is the most suitable average.
4) Since median minimises the total absolute deviations, median is preferred in the situations wherein the total geographical distance is to be minimised. For example, there is a conference of five top executives from five different cities of India lying. Almost in a straight line. The city located at a median distance would be a more proper place for the conference.
5) While taking a decision to buy a particular brand of tyre, when only one or two tyres are to be bought, the brand with greater median run will be preferred. Similarly, in buying a washing machine, the machine with greater median life will be preferred, rather than one with a greater mean life.

- **Limitations**

1) Median is not capable of algebraic treatment. That means we cannot have a combined median of two or more groups, unless all the items of the groups are known.
2) It is described, sometimes, as an insensitive measure as it is not based on all items of the series.
3) It is affected more by sampling fluctuations than the value of mean.
4) The computational formula of a median is in a way an interpolation under the assumption that the items in the median class are uniformly distributed, which is not very me.

5) The impression created by median in some cases may be illusory and deceptive because its value is determined strictly by the value of middle observation(s). For example, in lotteries the median value of the prize won by a ticket is always zero when all tickets are considered (more than 50% of the tickets will not get any prize). This median value of prize will, not help in analysing the prizes offered by lotteries as the matter of interest may be the first prize out of a number of prizes offered.

Calculate median from the following data

Class Boundary	15.5-33.5	33.5-51.5	51.5-69.5	69.5-87.5	87.5-100.5
Frequency	5	8	2	3	1

Step - I

Calculation of median (Equal Class Interval)

Class Boundary	**Class Width**	**Frequency (f/N)**	**Cumulative Frequency Less than type**	**Remarks**
15.5-33.5	18	5	5	
33.5-51.5	18	8 (fm)	13 (fc)	
51.5-69.5	18	2	15	
69.5-87.5	18	3	18	
87.5-100.5	18	1	19	
	Σ f=19			

Step - II

Identification of Median Class

(f ÷ N) ÷ 2

= 19÷2

= 9.5

Step - III

So, the Median class will be 33.5 - 51.5

Step - IV

Applying the formula of median

Median = $L_1 + (\frac{\frac{N}{2} - fc}{fm}) \times i$

$=33.5+\frac{\frac{19}{2}-5}{8}\times18$

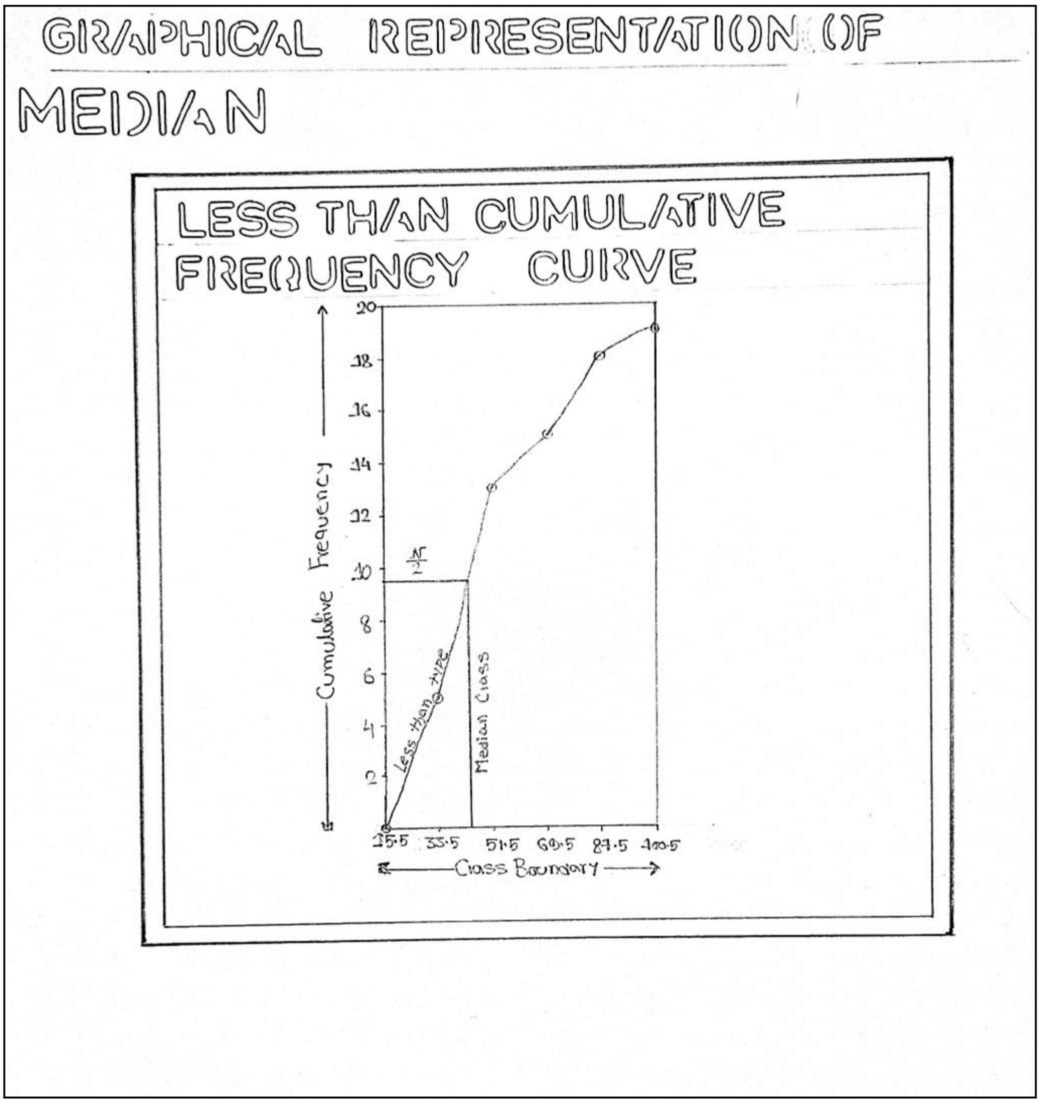
$=33.5+9.5-5/8\times18$ Where, L_1 = Lower Class Boundary of Median Class

$=33.5+4.5/8\times18$ fc = Cumulative frequency of pre median class

$=33.5+10.12$ fm = Frequency of median class

$=43.62$ f = Total frequency

I = Class width of the median class



➢ MODE

Mode is often considered to be that value of the variate which occurs most frequently. But it is not exactly true for every frequency distribution. Rather it is that value of the variate around which the other items tend to concentrate most heavily. It shows the centre of concentration of the frequency in and around a given value. If is not the centre of gravity like mean. It is a positional measure similar to median.

Merits

1) 1). In certain situations, mode is the only suitable average, e.g., modal size of garments, modal size of shoes, modal wages, modal balance of depositors in a bank, etc.
2) It is used to describe qualitative phenomena. For instance, if a printing press turns out five impressions which we rate very sharp, sharp, sharp, blurred and sharp, then the modal value is sharp.
3) For the preference of consumers' product, the modal preference is regarded. A restaurant owner who specialises in one dish may wish to know the modal preference of his potential clientele.
4) In the case of skewed distribution, mode is the indicator of the point of heaviest concentration.
5) It is very profitably used in market research.
6) Even if one or more classes are open-ended, mode can be used.

Limitations

1) Too often, there is no modal value. It is a useless measure, when there are more than one mode.
2) It is not capable of further algebraic treatment.
3) It is an ill-defined measure. Therefore, different formulas yield somewhat different answers.
4) It is not based on all the items of the data.
5) The value of the mode is affected significantly by the size of the class-intervals
6) Although a mode is the value of a variate that occurs most frequently, its frequency does not represent a majority of the total frequencies.

Calculate mode from the following data

Class Boundary	15.5-33.5	33.5-51.5	51.5-69.5	69.5-87.5	87.5-100.5
Frequency	5	8	2	3	1

Step - I

Calculation of mode ($\overline{x}$) by direct method (Equal Class Interval)

Class Boundary	Class Width	Frequency (f/N)	Remarks
15.5-33.5	18	5	
33.5-51.5	18	8	Mode

51.5-69.5	18	2	
69.5-87.5	18	3	
87.5-100.5	18	1	

Step - II

From the above analysis table, we have found that the mode is 33.5 to 51.5, as its frequencies occur the maximum times.

Step - III

Mode is estimated by the formula:

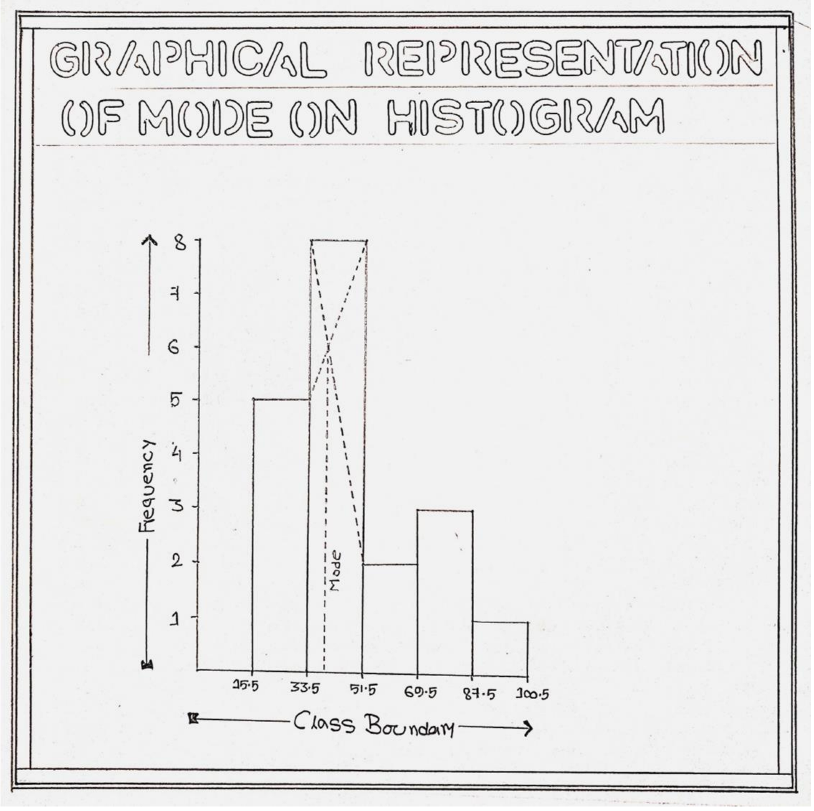
Mode= $L_1+ \Delta_1/ \Delta_1+ \Delta_2 \times i$

$=33.5+3/3+6\times18$ Where, L_1 = Lower Class Boundary of Modal Class

$=33.5+3/9\times18$ Δ_1 = Difference of frequency between modal and pre modal class

$=33.5+6$ Δ_2 = Difference of frequency between modal and post modal class

$=39.5$ i = interval of modal class (class width)



➢ MEASURES OF DISPERSION

According to Spiegel, the degree to which numerical data tend to spread about an average value is called the variation or dispersion of data. Actually, there are two basic kinds of a measure of dispersion (i) Absolute measures and (ii) Relative measures. The absolute measures of dispersion are used to measure the variability of a given data expressed in the same unit, while the relative measures are used to compare the variability of two or more sets of observations. Following are the different measures of dispersion: 1. Range 2. Quartile Deviation 3. Mean Deviation 4. Standard Deviation and Variance.

Standard Deviation Standard deviation (SD) is defined as the positive square root of variance.

Calculation table for standard deviation:

Class Boundary	**Mid Value**	**Frequency (f/N)**	**fx**	$\bar{x}$	**d(x- $\bar{x}$)**	**d^2**	**fd^2**
15.5-33.5	24.5	5	122.5	48.05	-235.5	-554.60	-2773
33.5-51.5	42.5	8	340		-5.55	-30.80	-246.4
51.5-69.5	60.5	2	121		12.45	155.00	310
69.5-87.5	78.5	3	235.5		30.45	927.20	2781.6
87.5-100.5	100.5	1	94		45.95	2111.40	2111.4
							Σ fd^2=2183.6

Mean ($\bar{x}$) = Σ fx ÷ ΣN

= 913 ÷ 19 Where, f=Frequency

= 48.05 x=Variable

Σ f = Total frequency

Standard Deviation= $\sqrt{\frac{\Sigma \mathbf{fd2}}{N}}$

$=\sqrt{\frac{2183.6}{19}}$

$=\sqrt{114.926}$

=10.720 (Approx)

Co-efficient of Variance = $\frac{\text{SD}}{\bar{x}}$×100

= $\frac{10.720}{48.05}$×100

= 0.223×100 = 22.3

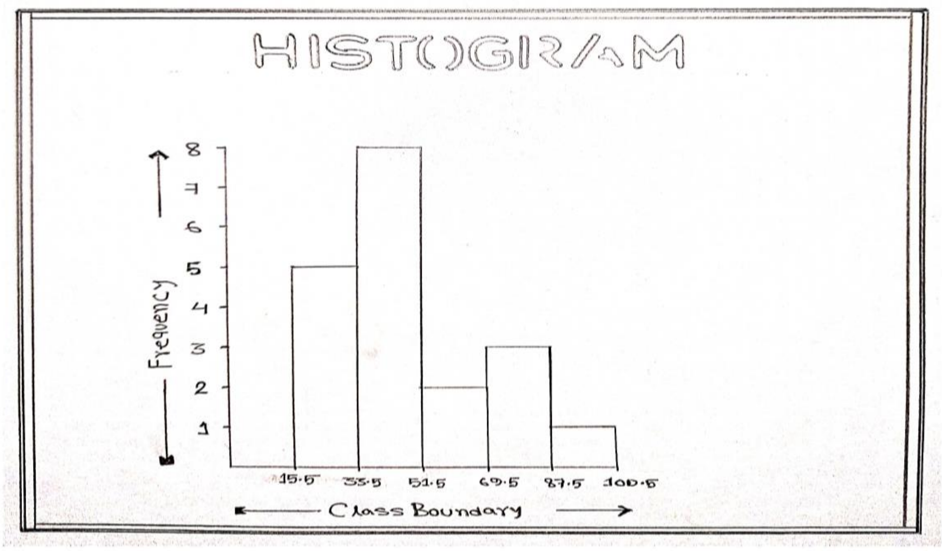
3. Histograms and frequency curve would be prepared on the dataset.

Histogram

Histogram is a rectangular diagram where the area of each rectangle is proportional to the frequency of the respective class. Remember that histogram is appropriate for continuous data arranged into class intervals. It is not used for discrete data.

Table for Histogram:

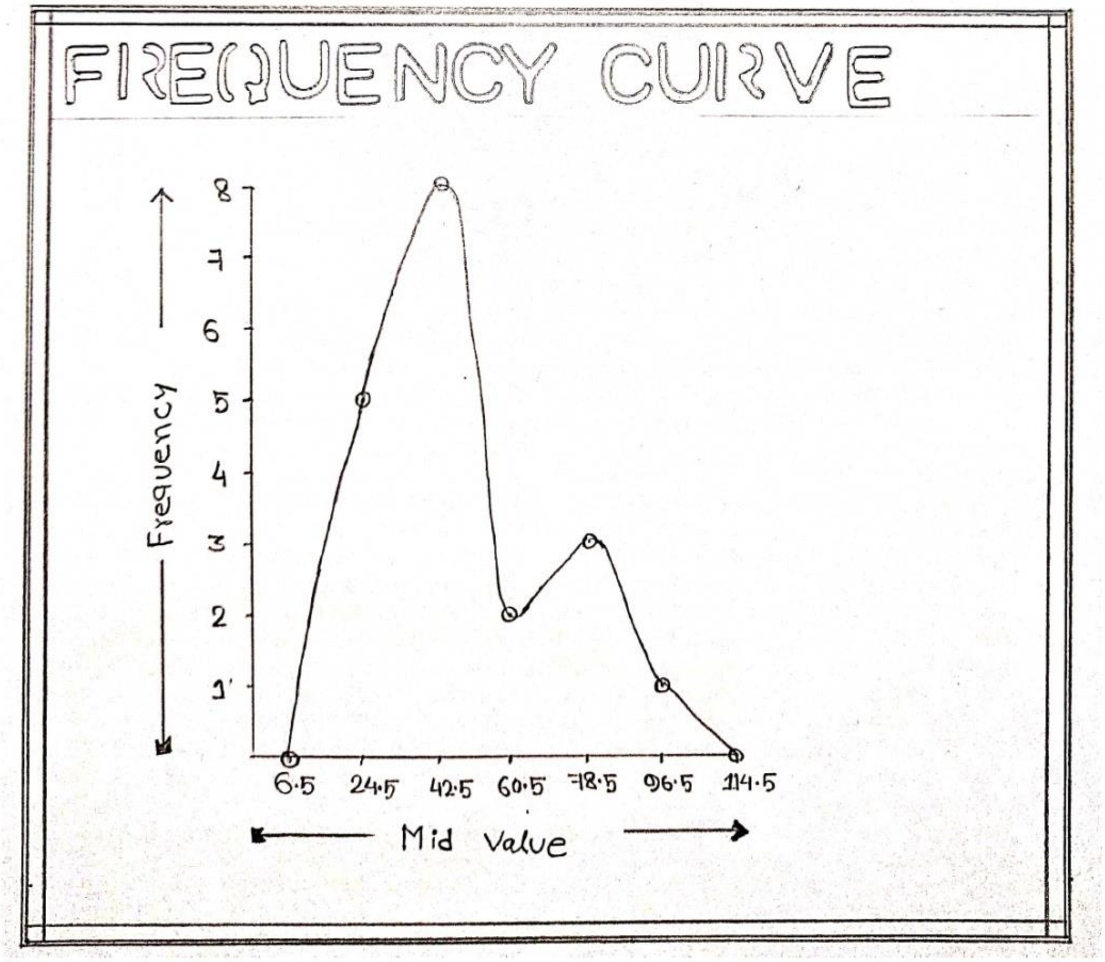
Class Boundary	15.5-33.5	33.5-51.5	51.5-69.5	69.5-87.5	87.5-100.5
Frequency	5	8	2	3	1



- ## Frequency Curve

A frequency curve is a smooth curve for which the total area is taken to be unity. It is a limiting form of a histogram or frequency polygon.

The frequency curve for a distribution can be obtained by drawing a smooth and free hand curve through the midpoints of the upper sides of the rectangles forming the histogram.


FREQUENCY CURVE
Frequency
8
7
6
5
4
3
2
1
6.5
24.5
42.5
60.5
78.5
96.5
114.5
Mid Value

4. From the data matrix a sample set (20%) would be drawn using, random, and systematic and stratified methods of sampling and locate the samples on a map with a short note on methods used.

When you conduct research about a group of people, it’s rarely possible to collect data from every person in that group. Instead, you select a sample. The sample is the group of individuals who will actually participate in the research.

To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. There are two types of sampling methods:

- **Probability sampling** involves random selection, allowing you to make strong statistical inferences about the whole group.
- **Non-probability sampling** involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

Population vs. sample

First, you need to understand the difference between a population and a sample, and identify the target population of your research.

- The **population** is the entire group that you want to draw conclusions about.
- The **sample** is the specific group of individuals that you will collect data from.

The population can be defined in terms of geographical location, age, income, and many other characteristics.

It can be very broad or quite narrow: maybe you want to make inferences about the whole adult population of your country; maybe your research focuses on customers of a certain company, patients with a specific health condition, or students in a single school.

It is important to carefully define your target population according to the purpose and practicalities of your project.

If the population is very large, demographically mixed, and geographically dispersed, it might be difficult to gain access to a representative sample.

Sampling frame

The sampling frame is the actual list of individuals that the sample will be drawn from. Ideally, it should include the entire target population (and nobody who is not part of that population).

Example

You are doing research on working conditions at Company X. Your population is all 1000 employees of the company. Your sampling frame is the company's HR database which lists the names and contact details of every employee.

Sample size

The number of individuals you should include in your sample depends on various factors, including the size and variability of the population and your research design. There are different sample size calculators and formulas depending on what you want to achieve with statistical analysis.

Probability Sampling Methods

Probability sampling means that every member of the population has a chance of being selected. It is mainly used in quantitative research. If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.

There are four main types of probability sample.

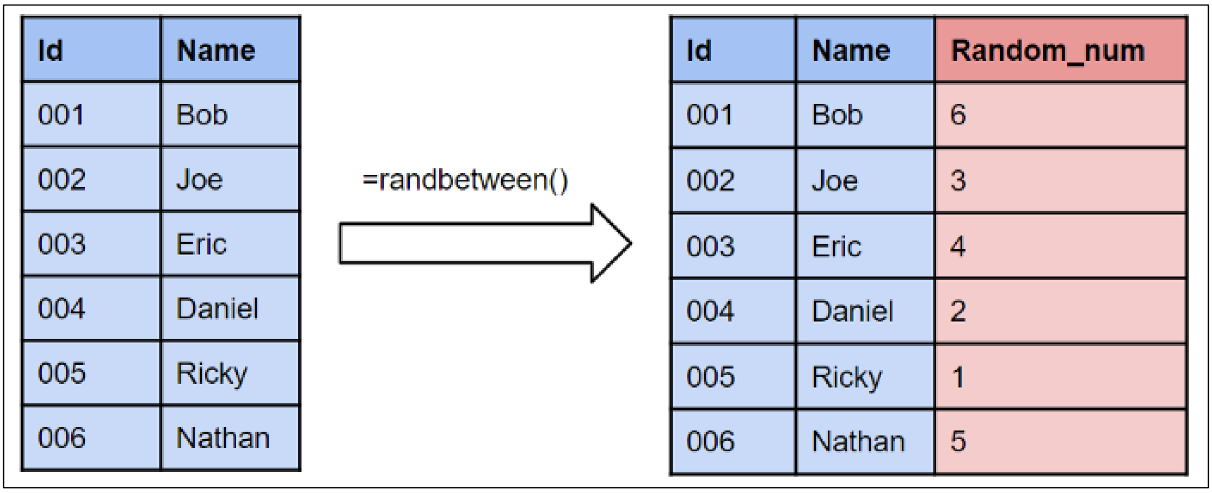
1. Simple Random Sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

Example

You want to select a simple random sample of 100 employees of Company X. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.



Id	Name
001	Bob
002	Joe
003	Eric
004	Daniel
005	Ricky
006	Nathan

=randbetween()

Id	Name	Random_num
001	Bob	6
002	Joe	3
003	Eric	4
004	Daniel	2
005	Ricky	1
006	Nathan	5

Pros and Cons:

- **Strong external validity:** Allows researchers to generalize results from the sample to the entire population being studied.
- **Relative speed and efficiency compared to the census:** A simple random sample allows researchers to learn about an entire population much faster and more efficiently than collecting data from every member of the population.
- **Expensive:** Contacting a large, randomly selected group of people requires lots of resources.
- **Time consuming:** Although this method is faster than conducting a census, gathering data from a large, random sample is often slow when compared to other methods.
- **Not always possible:** Researchers may wish to study a group for which there is no organized list (sampling frame) to randomly sample from.

Simple random sampling methods

Researchers follow these methods to select a simple random sample:

1. They prepare a list of all the population members initially, and then each member is marked with a specific number (for example, there are nth members, then they will be numbered from 1 to N).
2. From this population, researchers choose random samples using two ways: random number tables and random number generator software. Researchers prefer a random number generator software, as no human interference is necessary to generate samples.

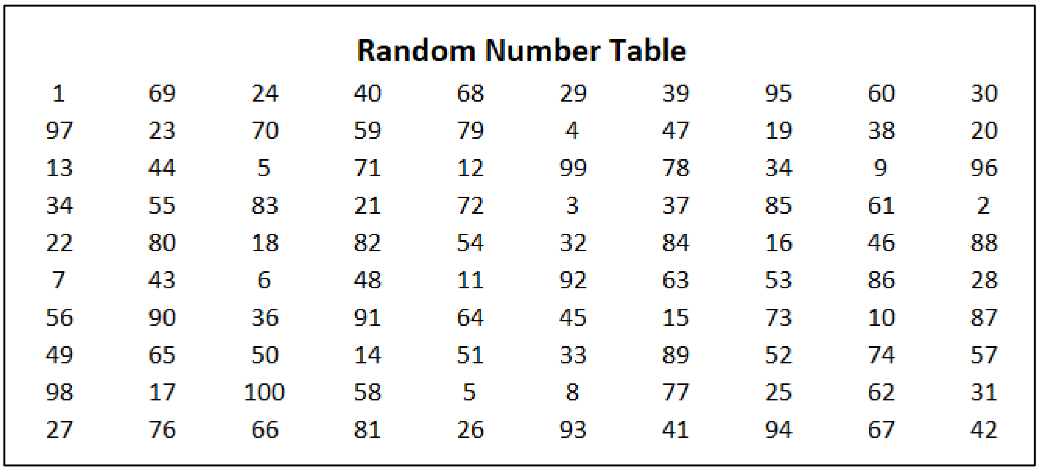
Two approaches aim to minimize any biases in the process of simple random sampling:

- **Method of lottery**

Using the lottery method is one of the oldest ways and is a mechanical example of random sampling. In this method, the researcher gives each member of the population a number. Researchers draw numbers from the box randomly to choose samples.

- **Use of random numbers**

The use of random numbers is an alternative method that also involves numbering the population. The use of a number table similar to the one below can help with this sampling technique.



Random Number Table

1	69	24	40	68	29	39	95	60	30
97	23	70	59	79	4	47	19	38	20
13	44	5	71	12	99	78	34	9	96
34	55	83	21	72	3	37	85	61	2
22	80	18	82	54	32	84	16	46	88
7	43	6	48	11	92	63	53	86	28
56	90	36	91	64	45	15	73	10	87
49	65	50	14	51	33	89	52	74	57
98	17	100	58	5	8	77	25	62	31
27	76	66	81	26	93	41	94	67	42

Simple random sampling formula

Consider a hospital has 1000 staff members, and they need to allocate a night shift to 100 members. All their names will be put in a bucket to be randomly selected. Since each person has an equal chance of being selected, and since we know the population size (N) and sample size (n), the calculation can be as follows:

P=1- N-1/N. N-2/N-1…. N-n/N-(n-1)

Cancelling = 1-N-n/N

= n/N

= 100/1000

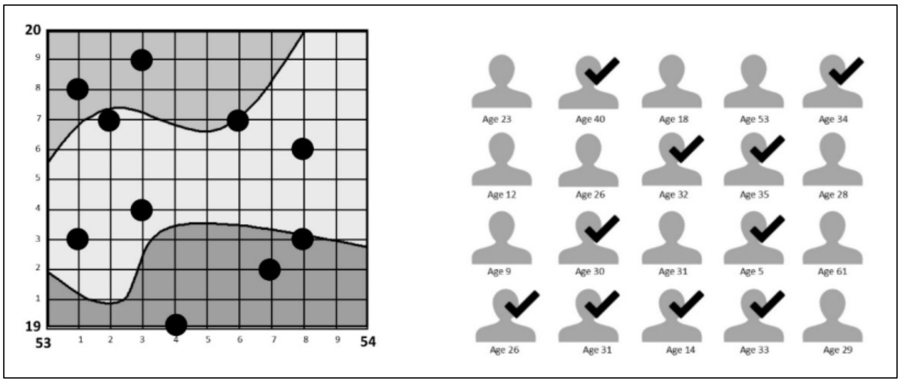
= 10%

Example of simple random sampling

Follow these steps to extract a simple random sample of 100 employees out of 500.

1. **Make a list** of all the employees working in the organization. (As mentioned above there are 500 employees in the organization, the record must contain 500 names).
2. **Assign a sequential number** to each employee (1,2,3…n). This is your sampling frame (the list from which you draw your simple random sample).
3. **Figure out what your sample size is going to be**. (In this case, the sample size is 100).
4. **Use a random number generator** to select the sample, using your sampling frame (population size) from Step 2 and your sample size from Step 3. For example, if your sample size is 100 and your population is 500, generate 100 random numbers between 1 and 500.

Graphical Presentation:



Following data set is female literacy rate (%) of 25 blocks of Purba Medinipur district, 2011.

1. Prepare a sample [n = 5 (20% of data set)] from the given data set through applying random sampling techniques.
2. Derive the sample arithmetic mean.

SI. No.	C.D.Block	Female Literacy Rate (%)	SI. No.	C.D.Block	Female Literacy Rate (%)
1	Tamluk	83.74	14	Pataspur-II	80.53
2	Sahid Matangini	80.89	15	Bhagabanpur-I	82.50
3	Paskura-I	78.07	16	Egra-I	78.72
4	Kolaghat	78.37	17	Egra-II	79.45
5	Moyna	80.24	18	Khejuri-I	84.36
6	Nandakumar	80.07	19	Khejuri-II	79.80
7	Chandipur	82.93	20	Bhagabanpur-II	86.29
8	Mahisadal	80.84	21	Ramnagar-I	81.72
9	Nandigram-I	80.71	22	Ramnagar-II	83.37
10	Nandigram-II	84.88	23	Contai-I	83.73
11	Sutahata	80.09	24	Deshapran	87.02

12	Haldia	81.97	25	Contai-III	84.75
13	Pataspur-I	79.90	-	-	-

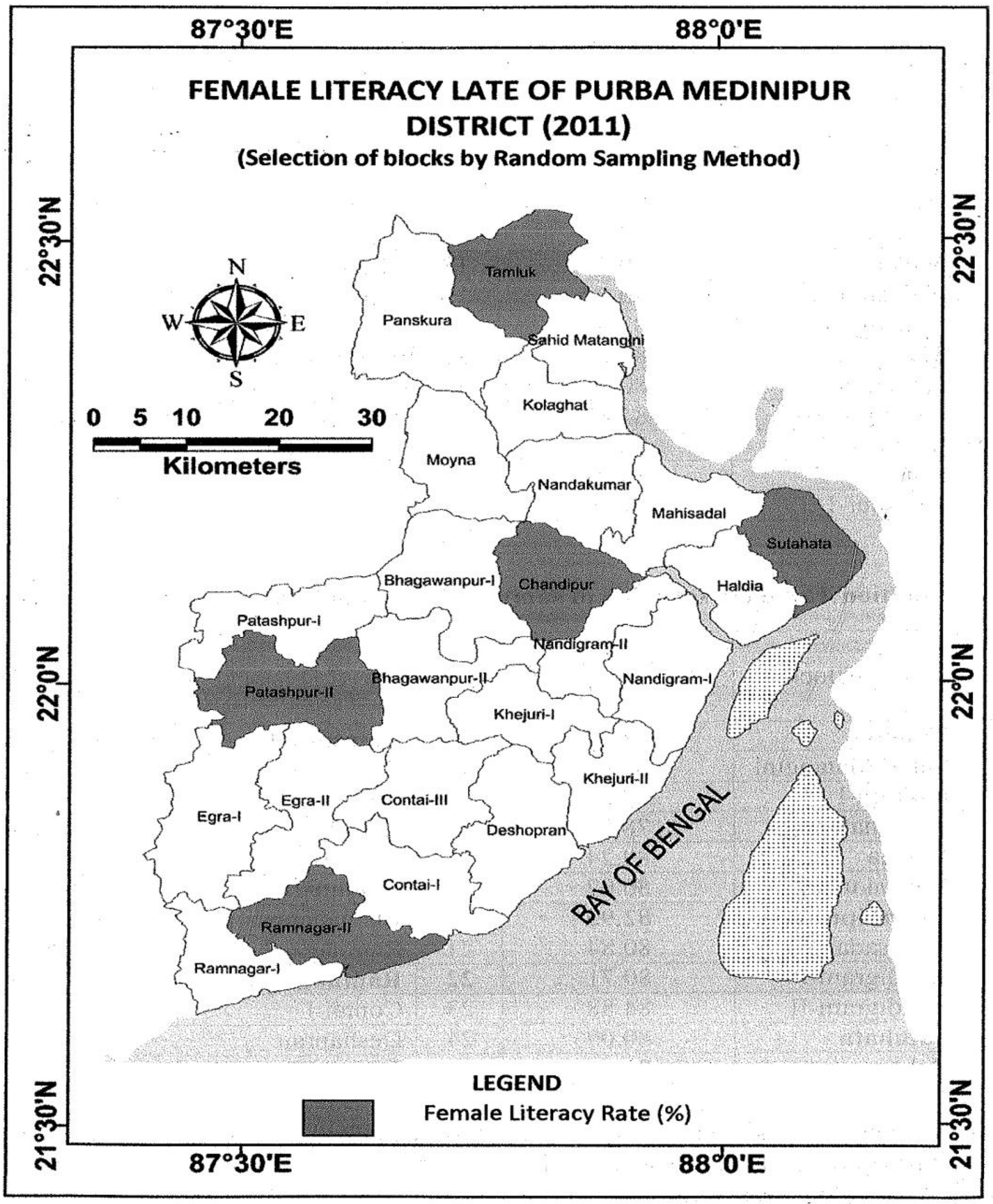
Computation Table for Selection of Samples from Population by Random Sampling

SI. No.	**C.D.Block**	**Female Literacy Rate (%)**	**SI. No.**	**C.D.Block**	**Female Literacy Rate (%)**
1	**Tamluk**	**83.74**	**14**	**Pataspur-II**	**80.53**
2	Sahid Matangini	80.89	15	Bhagabanpur-I	82.50
3	Paskura-I	78.07	16	Egra-I	78.72
4	Kolaghat	78.37	17	Egra-II	79.45
5	Moyna	80.24	18	Khejuri-I	84.36
6	Nandakumar	80.07	19	Khejuri-II	79.80
7	**Chandipur**	**82.93**	20	Bhagabanpur-II	86.29
8	Mahisadal	80.84	21	Ramnagar-I	81.72
9	Nandigram-I	80.71	**22**	**Ramnagar-II**	**83.37**
10	Nandigram-II	84.88	23	Contai-I	83.73
11	**Sutahata**	**80.09**	24	Deshapran	87.02
12	Haldia	81.97	25	Contai-III	84.75

1. Selections of blocks by random sampling are **Tamluk (83.74%); Chandipur (82.93%); Sutahata (80.09%); Pataspur-II (80.53%) & Ramnagar-II (83.37%)**.
2. Sample Mean $= \frac{\sum X}{n} = \frac{410.66\%}{5} = \mathbf{82.13}\%$, Population Mean $= \frac{\sum X}{n} = \frac{2044.94\%}{25} = \mathbf{81.80}\%$

<u>Interpretation</u>

Here Simple random sampling techniques are applied for selection of samples. When homogeneous data but no. of population is quite low are given then this method is applied. There is no significant difference between population mean & sample mean (82.13% - 81.80%) = 0.33%. So it is assumed that the selection of sample is very much correct.



FEMALE LITERACY LATE OF PURBA MEDINIPUR DISTRICT (2011)

(Selection of blocks by Random Sampling Method)

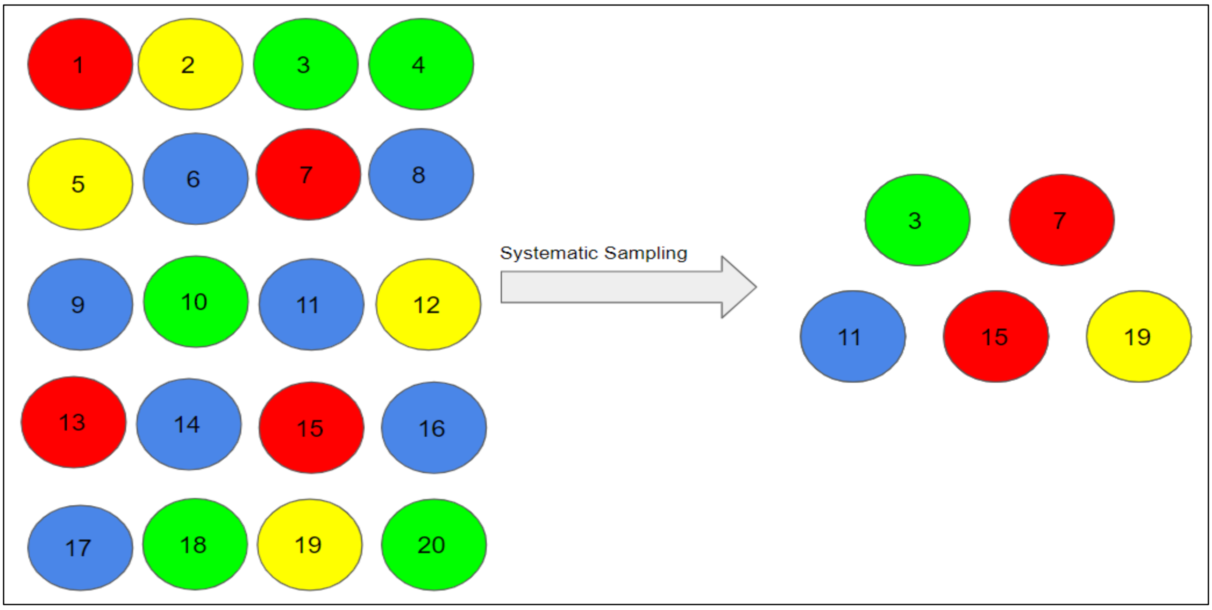
2. Systematic Sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Example

All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.



Pros and Cons:

- **External validity:** Allows generalization from the sample to the population being studied.
- **Relative speed:** Faster than contacting all members of the population or simple random sampling.
- **Limited feasibility:** This sampling method is not possible without a list of all members of the population.

Steps to form a sample using the systematic sampling technique

Here are the steps to form a systematic sample:

- **Step one:** Develop a defined structural audience to start working on the sampling aspect.
- **Step two:** As a researcher, figure out the ideal size of the sample, i.e., how many people from the entire population to choose to be a part of the sample.
- **Step three:** Once you decide the sample size, assign a number to every member of the sample.
- **Step four:** Define the interval of this sample. This will be the standard distance between the elements. For example, the sample interval should be 10, which is the result of the division of 5000 (N= size of the population) and 500 (n=size of the sample).

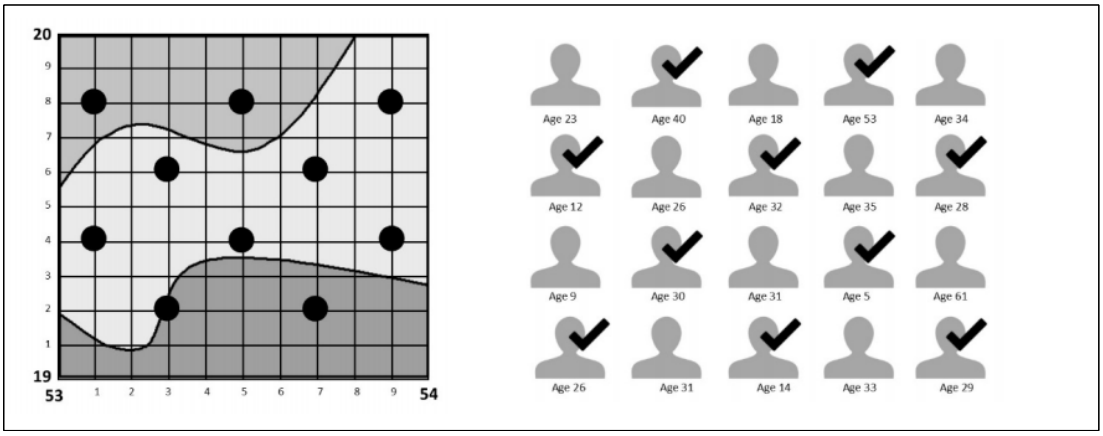
Systematic Sampling Formula for interval (i) = N/n = 5000/500 = 10

- **Step five:** Select the members who fit the criteria which in this case will be 1 in 10 individuals.
- **Step six:** Randomly choose the starting member (r) of the sample and add the interval to the random number to keep adding members in the sample. r, r+i, r+2i, etc. will be the elements of the sample.

How systematic sampling works

When you are sampling, ensure you represent the population fairly. Systematic sampling is a symmetrical process where the researcher chooses the samples after a specifically defined interval. Sampling like this leaves the researcher no room for bias regarding choosing the sample. To understand how systematic sampling exactly works, take the example of the gym class where the instructor asks the students to line up and asks every third person to step out of the line. Here, the instructor has no influence over choosing the samples and can accurately represent the class.

Graphical Presentation:



Following data set is female literacy rate (%) of 25 blocks of Purba Medinipur district, 2011.

1. Prepare a sample [n = 6 (24% of data set)] from the given data set through applying systematic sampling techniques without replacement after arranging all the data items in array and selecting every 5th item starting from first one .
2. Derive the sample arithmetic mean.

SI. No.	C.D.Block	Female Literacy Rate (%)	SI. No.	C.D.Block	Female Literacy Rate (%)
1	Tamluk	83.74	14	Pataspur-II	80.53
2	Sahid Matangini	80.89	15	Bhagabanpur-I	82.50
3	Paskura-I	78.07	16	Egra-I	78.72
4	Kolaghat	78.37	17	Egra-II	79.45
5	Moyna	80.24	18	Khejuri-I	84.36
6	Nandakumar	80.07	19	Khejuri-II	79.80
7	Chandipur	82.93	20	Bhagabanpur-II	86.29
8	Mahisadal	80.84	21	Ramnagar-I	81.72
9	Nandigram-I	80.71	22	Ramnagar-II	83.37
10	Nandigram-II	84.88	23	Contai-I	83.73
11	Sutahata	80.09	24	Deshapran	87.02
12	Haldia	81.97	25	Contai-III	84.75
13	Pataspur-I	79.90	-	-	-

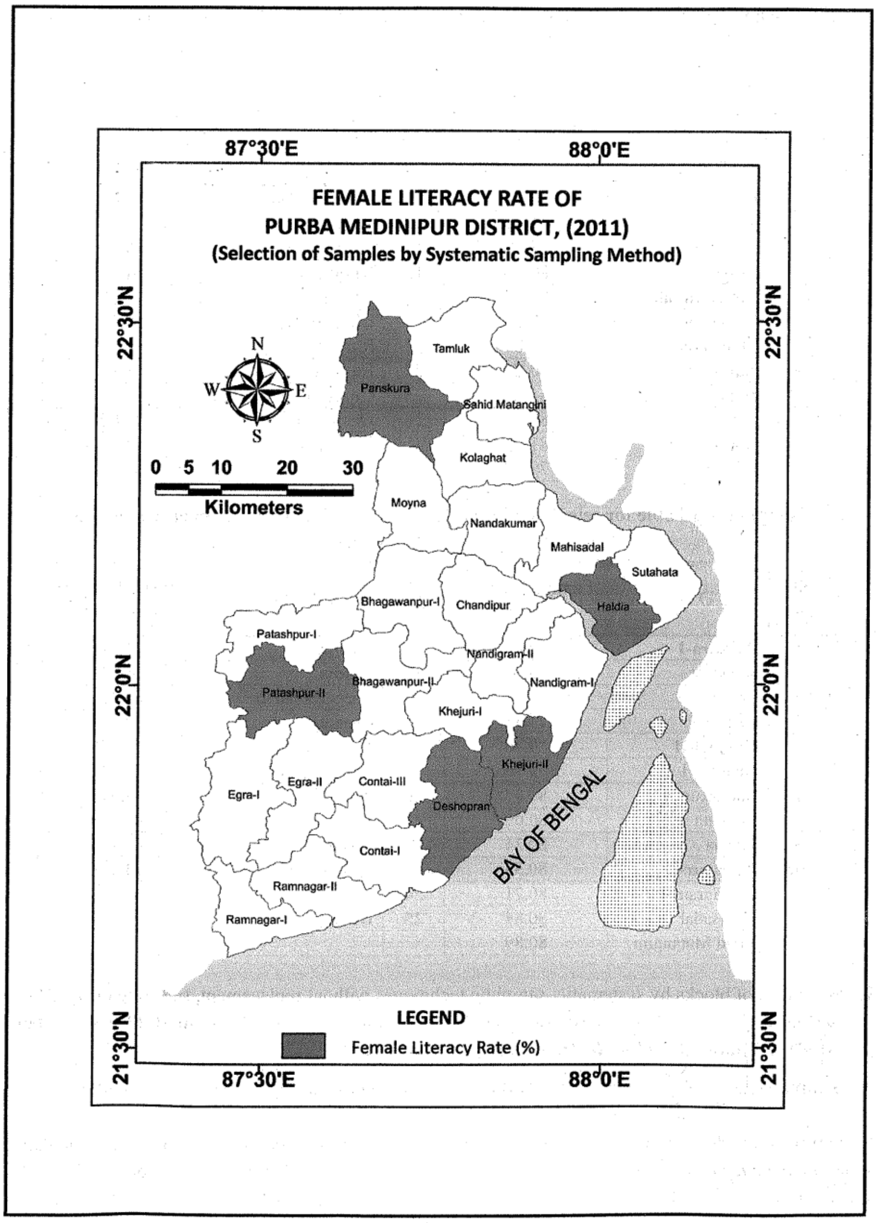
Computation Table for Selection of Samples from Population by Systematic Sampling (Ascending Order)

SI. No.	C.D.Block	Female Literacy Rate (%)	SI. No.	C.D.Block	Female Literacy Rate (%)
1	**Paskura-I**	**78.07**	14	Ramnagar-I	81.72
2	Kolaghat	78.37	**15**	**Haldia**	**81.97**
3	Egra-I	78.72	16	Bhagabanpur-I	82.50
4	Egra-II	79.45	17	Chandipur	82.93
5	**Khejuri-II**	**79.80**	18	Ramnagar-II	83.37
6	Pataspur-I	79.90	19	Contai-I	83.73
7	Nandakumar	80.07	**20**	**Tamluk**	**83.74**
8	Sutahata	80.09	21	Khejuri-I	84.36
9	Moyna	80.24	22	Contai-III	84.75
10	**Pataspur-II**	**80.53**	23	Nandigram-II	84.88
11	Nandigram-I	80.71	24	Bhagabanpur-II	86.29
12	Mahisadal	80.84	**25**	**Deshapran**	**87.02**
13	Sahid Matangini	80.89	-	-	-

1. Selections of blocks by systematic sampling techniques without replacement, selecting every 5th item starting from first one is **Paskura-I (78.07%), Khejuri-II (79.80), Pataspur-II (80.53%), Haldia (81.97%), Tamluk(83.74%) & Deshapran (87.02%)**.
2. Sample Mean $= \frac{\sum X}{n} = \frac{491.13\%}{6} = \mathbf{81.86}\%$, Population Mean $= \frac{\sum X}{n} = \frac{2044.94\%}{25} = \mathbf{81.80}\%$

<u>Interpretation</u>

Here Systematic sampling technique is applied for selection of samples from population. When homogeneous data, but no. of population is quite more than this method is applied. There is no significant difference between population mean & sample mean (81.86% - 81.80%) = 0.06 %. So it is assumed that the selection of sample is very much correct.



3. Stratified Sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

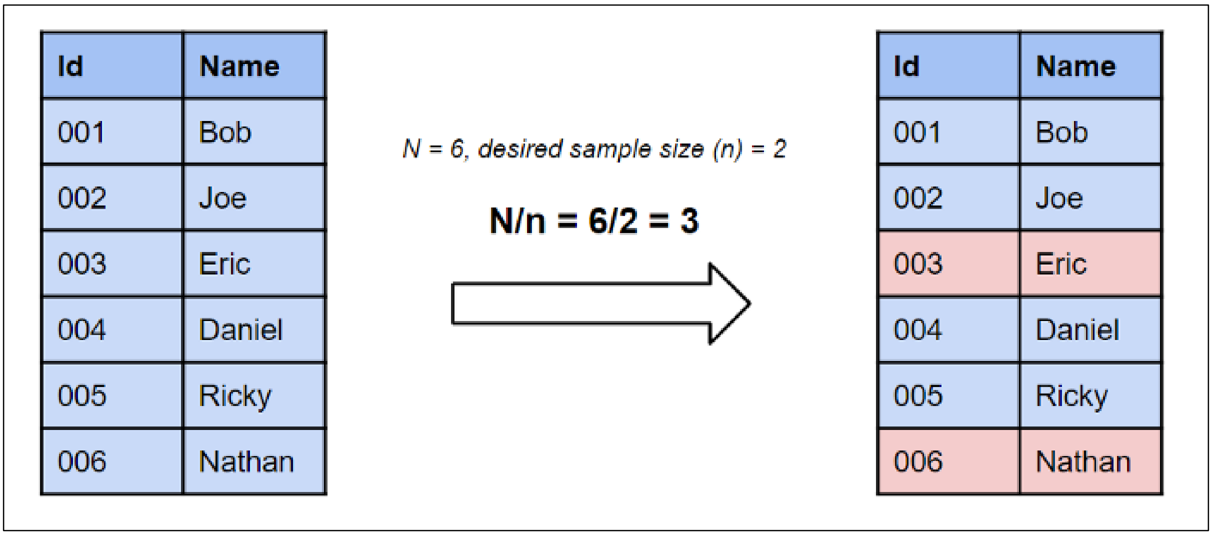
What should be the size of the sample chosen from each stratum?

The size of the sample you select will vary based on several factors:

- **Scale**
  In general, to analyse and draw meaningful conclusions, you need a large sample that can provide you with sufficient data from the total population.
- **Practicality**
  From a practical standpoint, if you have a larger population, you want to also have a sample size that does not require a lot of administration to collect and manage.
- **Accuracy**
  You want a sample size that is going to accurately represent the total population to make the findings as truthful as possible.

Example

The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.



Id	Name
001	Bob
002	Joe
003	Eric
004	Daniel
005	Ricky
006	Nathan

N = 6, desired sample size (n) = 2

N/n = 6/2 = 3

Id	Name
001	Bob
002	Joe
003	Eric
004	Daniel
005	Ricky
006	Nathan

Pros and Cons:

- **External validity:** Maintains the researcher's ability to generalize from the sample to the entire population being studied.
- **Representation:** By selecting important groups to sample within before beginning data collection, the researchers can ensure adequate representation of small and minority groups.

Calculation Methods:

- **Proportionate Stratified Sampling:**

In this approach, each stratum sample size is directly proportional to the population size of the entire population of strata. That means each stratum sample has the same sampling fraction.

Proportionate Stratified Random Sampling Formula: $n_h = (N_h / N) * n$

$\mathbf{n_h}$ = Sample size for hth stratum

$\mathbf{N_h}$ = Population size for hth stratum

N = Size of entire population

n = Size of entire sample

If you have 4 strata with 500, 1000, 1500, 2000 respective sizes and the research organization selects ½ as sampling fraction. A researcher has to then select 250, 500, 750, 1000 members from the respective stratum.

Stratum	A	B	C	D
Population Size	500	1000	1500	2000
Sampling Fraction	1/2	1/2	1/2	1/2
Final Sampling Size Results	250	500	750	1000

Irrespective of the sample size of the population, the sampling fraction will remain uniform across all the strata.

- **Disproportionate Stratified Sampling:**

Sampling fraction is the primary differentiating factor between the proportionate and disproportionate stratified random sampling. In disproportionate sampling, each stratum will have a different sampling fraction.

The success of this sampling method depends on the researcher’s precision at fraction allocation. If the allotted fractions aren’t accurate, the results may be biased due to the overrepresented or underrepresented strata.

Stratum	A	B	C	D
Population Size	500	1000	1500	2000
Sampling Fraction	1/2	1/3	1/4	1/5
Final Sampling Size Results	250	333	375	400

Following is a classic stratified random sampling example:

Let’s say, 100 (N_h) students of a school having 1000 (N) students were asked questions about their favourite subject. It’s a fact that the students of the 8th grade will have different subject preferences than the students of the 9th grade. For the survey to deliver precise results, the ideal manner is to divide each grade into various strata.

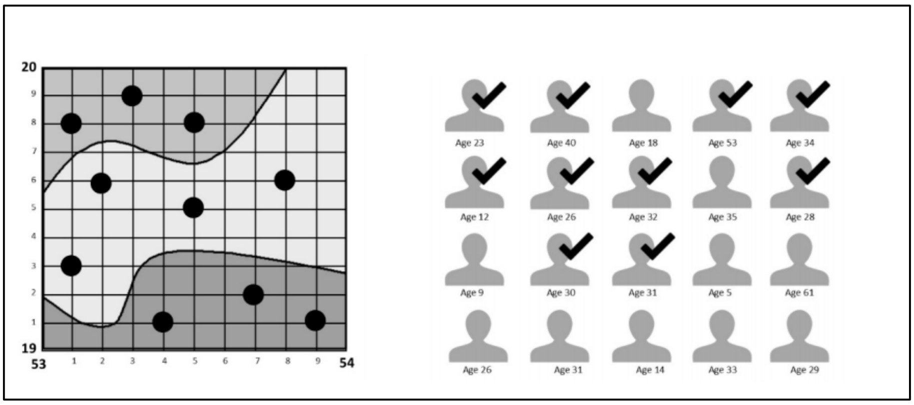
Here’s a table of the number of students in each grade:

Grade	**Number of students (n)**
5	150
6	250
7	300
8	200
9	100

Calculate the sample of each grade using the stratified random sampling formula:

Stratified Sample (n_5) = 100 / 1000 * 150 = 15
Stratified Sample (n_6) = 100 / 1000 * 250 = 25
Stratified Sample (n_7) = 100 / 1000 * 300 = 30
Stratified Sample (n_8) = 100 / 1000 * 200 = 20
Stratified Sample (n_9) = 100 / 1000 * 100 = 10

Graphical Presentation:



Following data set is female literacy rate (%) of 25 blocks of Purba Medinipur district, 2011.

1. Prepare a sample [n = 5 (20% of data set)] from the given data set through applying stratified sampling techniques.
2. Derive the sample arithmetic mean.

SI. No.	C.D.Block	Female Literacy Rate (%)	SI. No.	C.D.Block	Female Literacy Rate (%)
1	Tamluk	83.74	14	Pataspur-II	80.53
2	Sahid Matangini	80.89	15	Bhagabanpur-I	82.50
3	Paskura-I	78.07	16	Egra-I	78.72
4	Kolaghat	78.37	17	Egra-II	79.45
5	Moyna	80.24	18	Khejuri-I	84.36
6	Nandakumar	80.07	19	Khejuri-II	79.80
7	Chandipur	82.93	20	Bhagabanpur-II	86.29
8	Mahisadal	80.84	21	Ramnagar-I	81.72
9	Nandigram-I	80.71	22	Ramnagar-II	83.37
10	Nandigram-II	84.88	23	Contai-I	83.73
11	Sutahata	80.09	24	Deshapran	87.02
12	Haldia	81.97	25	Contai-III	84.75
13	Pataspur-I	79.90	-	-	-

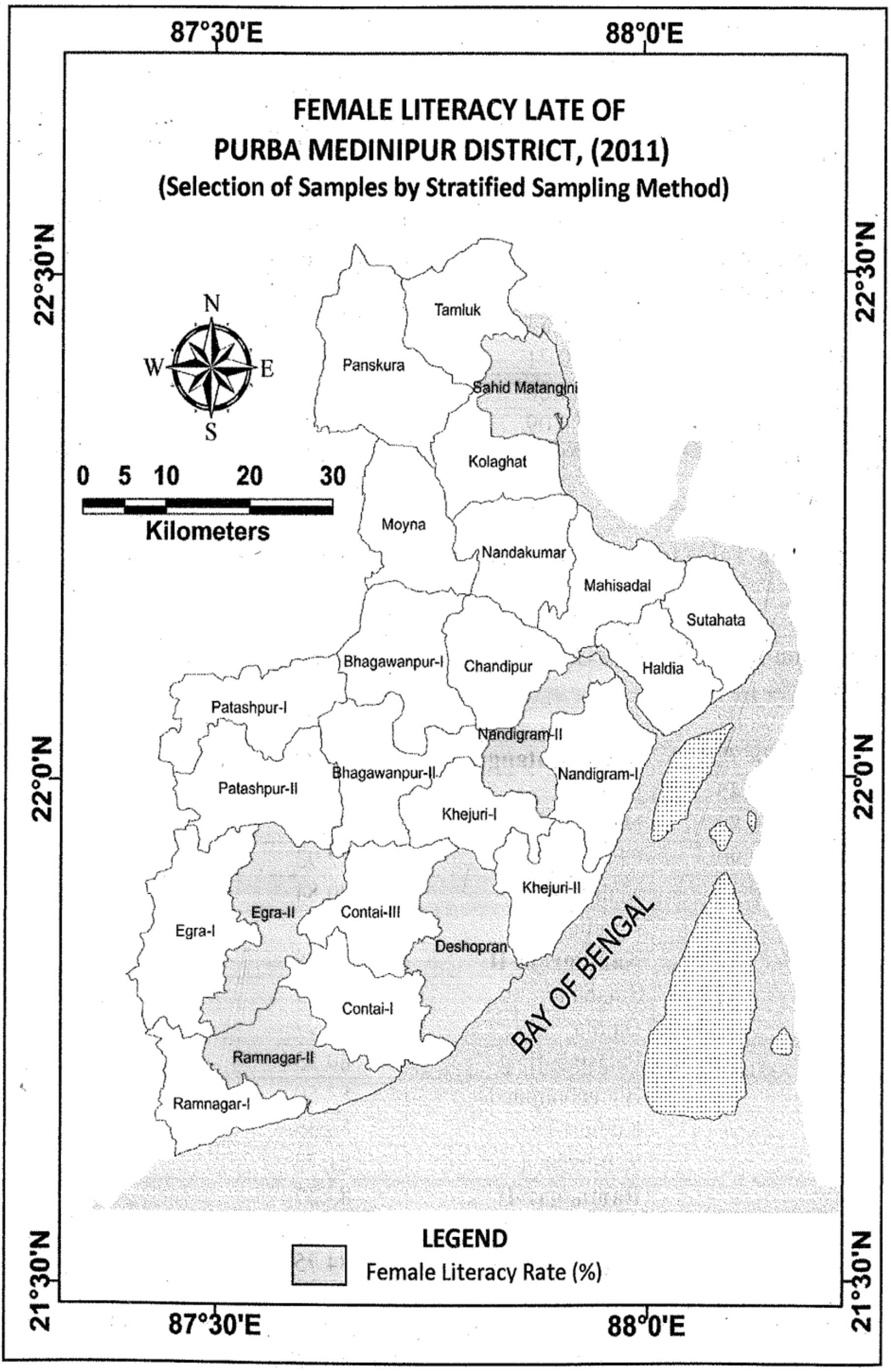
Computation Table for Selection of Samples from Population by Stratified Sampling

Female Literacy Rate (%)					
Low Education Rate (Below 80 %)		**Medium Education Rate (80% - 85 %)**		**Medium Education Rate (Above 85%)**	
Paskura-I	78.07	Tamluk	83.74	Bhagabanpur-II	86.29
Egra-I	78.72	**Sahid Matangini**	**80.89**	**Deshapran**	**87.02**
Egra-II	**79.45**	Moyna	80.24		
Khejuri-II	79.80	Nandakumar	80.07		
Pataspur-I	79.90	Chandipur	82.93		
Kolaghat	78.37	Mahisadal	80.84		
		Nandigram-I	80.71		
		Nandigram-II	**84.88**		
		Sutahata	80.09		
		Haldia	81.97		
		Pataspur-II	80.53		
		Bhagabanpur-I	82.50		
		Khejuri-I	84.36		
		Ramnagar-I	81.72		
		Ramnagar-II	**83.37**		
		Contai-I	83.73		
		Contai-III	84.75		

1. Selections of blocks from different stratification layer are **Egra-II (79.45%), Sahid Matangini (80.89%), Nandigram-II (84.88%), Ramnagar-II (83.37%) & Deshapran (87.02%)**.
2. Sample Mean $= \frac{\sum X}{n} = \frac{415.61\%}{5} = \mathbf{81.86}\%$, Population Mean $= \frac{\sum X}{n} = \frac{2044.94\%}{25} = \mathbf{81.80}\%$

Interpretation

Here Stratified sampling techniques are applied for selection of samples. When heterogeneous data set is given then to reduce heterogeneity, some stratification layering has been done, after that samples are collected from each layer, and then this method is applied. There is no significant difference between population mean & sample mean (83.12% - 81.80%) = 1.32 %. So it is assumed that the selection of sample is very much correct.



FEMALE LITERACY LATE OF PURBA MEDINIPUR DISTRICT, (2011)

(Selection of Samples by Stratified Sampling Method)

5. Based on of the sample set and using two relevant attributes, a scatter diagram and regression line would be plotted and residual from regression would be mapped with a short interpretation.

The following data shows rainfall and rainy days of a year:

Months	**J**	**F**	**M**	**A**	**M**	**J**	**Ju**	**A**	**S**	**O**	**N**	**D**
Rainy days	0.8	1.3	2.0	3.0	6.0	12.6	16.4	17.0	18.8	6.9	1.2	0.4
Rainfall (cm)	1.4	1.6	2.6	5.1	10.3	28.0	32.7	31.4	29.4	13.4	1.7	0.7

1. Draw a scatter diagram with regression line on the basis of above data.
2. What will be the rainfall (cm) if the rainy days are 5 & 15?
3. Interpret the nature of variables with the help of product moment correlation co-efficient method.
4. Calculate the probable error of the correlation coefficient to justify the nature of correlation.
5. Find out the standard error of estimate for the model.
6. Justify the relationship exists between the rainy days and rainfall with the application of test of significance.
7. Compute the percentage of variance in rainfall explained by rainy days.
8. Calculate the co-efficient of determination. (R^2)

Scatter diagram and regression line can be calculated by 3 methods.

a. Least square method
b. Regression co-efficient method
c. Correlation co-efficient method

Prepare a Residual map showing corresponding between rural area & rural population of Purba Medinipur district and interpret its spatial pattern of the following data.

Sl. No.	C.D.Block	Area (Sq. Km) (X)	Rural Population (in ´000) (Y)	Sl. No.	C.D.Block	Area (Sq. Km) (X)	Rural Population (in ´000) (Y)
1	Tamluk	133.86	207.064	14	Pataspur-II	191.74	175.056
2	Sahid Matangini	97.82	183.987	15	Bhagabanpur-I	174.24	222.677
3	Paskura-I	246.92	283.303	16	Egra-I	218.01	167.163
4	Kolaghat	147.91	239.646	17	Egra-II	184.71	178.763
5	Moyna	154.51	220.330	18	Khejuri-I	130.51	132.992
6	Nandakumar	165.70	262.998	19	Khejuri-II	137.46	139.463
7	Chandipur	137.58	176.704	20	Bhagabanpur-II	180.20	192.162
8	Mahisadal	146.48	199.613	21	Ramnagar-I	139.43	161.986
9	Nandigram-I	181.84	202.032	22	Ramnagar-II	163.27	156.054
10	Nandigram-II	105.74	117.945	23	Contai-I	155.27	170.894
11	Sutahata	79.54	118.629	24	Deshapran	184.55	170.938
12	Haldia	170.34	97.992	25	Contai-III	160.52	157.793
13	Pataspur-I	172.26	166.977				

Necessary Formula: Absolute Residuals (e_i) = (Y_i -$\hat{Y}$)

$$\hat{Y} = a + bX$$

$$b = \frac{\Sigma XY - \frac{\Sigma X.\Sigma Y}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}} \quad \text{.........} \ (1)$$

$$a = \overline{Y} - b\overline{X} \quad \text{....................} \ (2)$$

Y_i = Dependent Variable

$\hat{Y}$ = Predicted dependent variable

b = Regression Co-efficient

a = ‘Y’ intercept

$\overline{Y} = \frac{\Sigma Y}{n} \quad \overline{X} = \frac{\Sigma X}{n}$

n = No. of data pairs

Putting the value **n** = 25, **ΣX** = 3960.41, **ΣY** = 4503.16, **ΣX^2** = 658513.22, **ΣXY** = 730843.64 in equation... (1), then we gets,

$$\text{Regression Co-efficient (b)} = \frac{\Sigma XY - \frac{\Sigma X.\Sigma Y}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$$

$$b = \frac{730843.64 - \frac{3960.41 \text{ x } 4503.16}{25}}{658513.22 - \frac{(3960.41)^2}{25}}$$

$$b = \frac{730843.64 - 713374.40}{658513.22 - 627393.89}$$

$$\mathbf{b = 0.56}$$

Y intercept, $\mathbf{a} = \bar{Y} - b\bar{X}$

$a = 180.13 – 0.56 \text{ x } 158.42$

$a = 180.13 – 88.72$

$\mathbf{a = 91.41}$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{4503.16}{25} = 180.13$$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{3960.41}{25} = 158.42$$

So the regression equation is $\mathbf{\hat{Y} = a + bX}$ [$\mathbf{\hat{Y} = 91.41 + 0.56\ X}$]

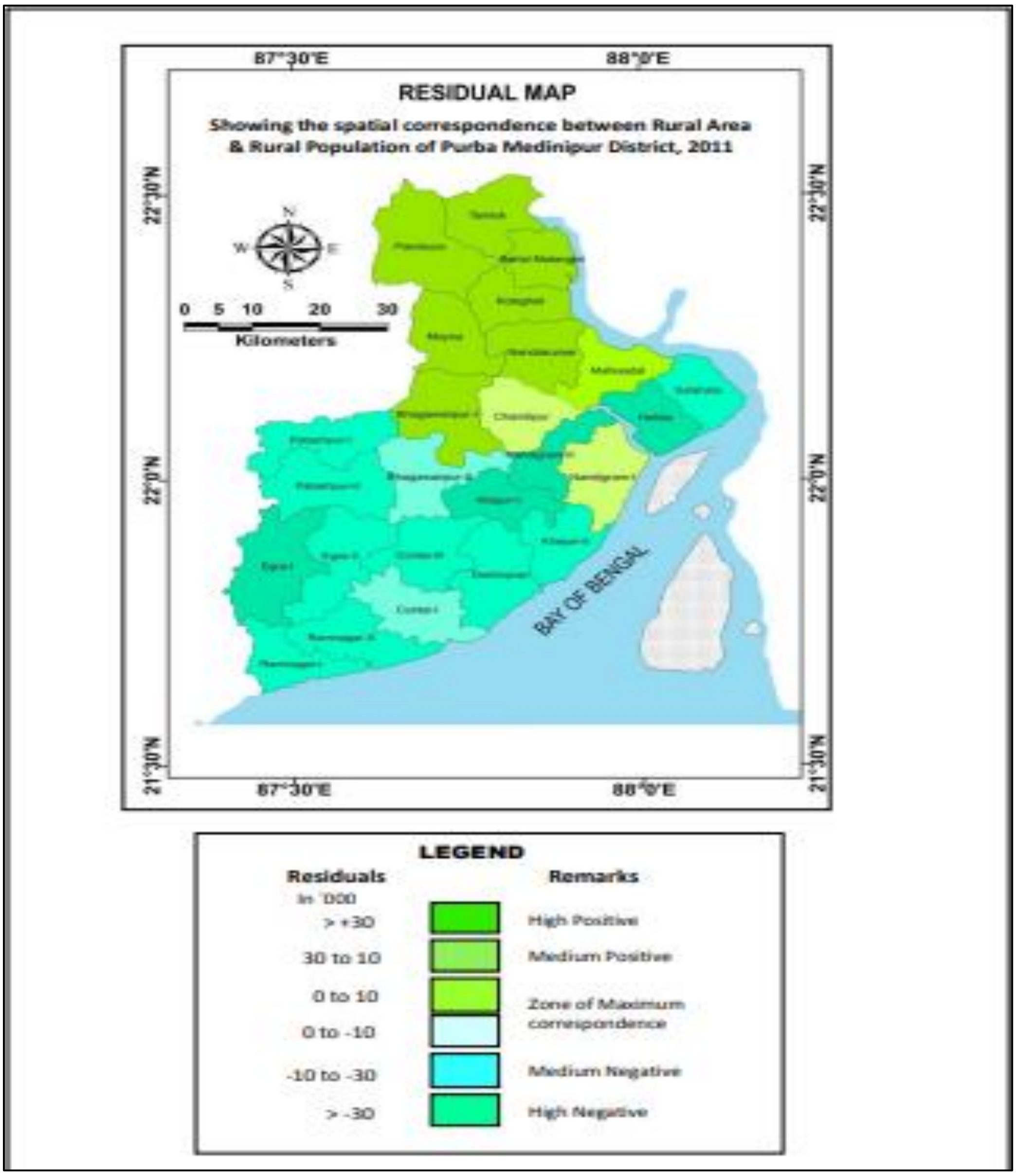
Predicted Rural Population of Tamluk ($\mathbf{\hat{Y}}$) = 91.41+ 0.56 X

= 91.41+ 0.56 x 133.86

= ***166.37**

Computation table of residuals of rural population & rural area

Sl. No.	C.D. Block	Area (Sq. Km) (X)	Rural Population (in ´000) (Y)	X^2	XY	$\hat{Y}$ Predicted Rural Population (in ´000)	Residuals ($Y_i - \hat{Y}$)
1	Tamluk	133.86	207.064	17918.50	27717.59	*166.37	40.694
2	Sahid Matangini	97.82	183.987	9568.75	17997.61	146.18	37.807
3	Paskura-I	246.92	283.303	60969.49	69953.18	229.69	53.613
4	Kolaghat	147.91	239.646	21877.37	35446.04	174.24	65.406
5	Moyna	154.51	220.330	23873.34	34043.19	177.94	42.39
6	Nandakumar	165.70	262.998	27456.49	43578.77	184.07	78.928
7	Chandipur	137.58	176.704	18928.26	24310.94	168.45	8.254
8	Mahisadal	146.48	199.613	21456.39	29239.31	173.44	26.173
9	Nandigram-I	181.84	202.032	33065.79	36737.50	193.24	8.792
10	Nandigram-II	105.74	117.945	11180.95	12471.50	150.62	-32.675
11	Sutahata	79.54	118.629	6326.61	9435.75	135.95	-17.321
12	Haldia	170.34	97.992	29015.72	16691.96	186.80	-88.808
13	Pataspur-I	172.26	166.977	29673.51	28763.46	187.88	-20.903
14	Pataspur-II	191.74	175.056	36764.23	33565.24	198.78	-23.724
15	Bhagabanpur-I	174.24	222.677	30359.58	38799.24	188.98	33.697
16	Egra-I	218.01	167.163	47528.36	36443.21	213.50	-46.337
17	Egra-II	184.71	178.763	34117.78	33019.31	194.85	-16.087
18	Khejuri-I	130.51	132.992	17032.86	17356.79	164.50	-31.508
19	Khejuri-II	137.46	139.463	18895.25	19170.58	168.39	-28.927
20	Bhagabanpur-II	180.20	192.162	32472.04	34627.59	192.322	-0.16
21	Ramnagar-I	139.43	161.986	19440.72	22585.71	169.49	-7.504
22	Ramnagar-II	163.27	156.054	26657.09	25478.94	182.84	-26.786
23	Contai-I	155.27	170.894	24108.77	26534.71	178.36	-7.466
24	Deshapran	184.55	170.938	34058.70	31546.61	194.76	-23.822
25	Contai-III	160.52	157.793	25766.67	25328.93	181.30	-23.507
		ΣX **3960.41**	**ΣY** **4503.16**	**ΣX^2** **658513.22**	**ΣXY** **730843.64**		



Interpretation:

Through Residual mapping Purba Medinipur as, a district show great disparity in the distribution of the total rural area (sq km.) with respect to the rural population, 2011. There are only 10 blocks out of 25 blocks of Purba Medinipur district shows as positive relation between rural area & rural population; remaining 15 blocks of the district are negative relation between rural area & rural population of the district. The residual map of spatial correspondence between rural area and total rural population of Purba Medinipur district reveals that there are five blocks Chandipur, Nandigram-I, Bhagabanpur-II, Ramnagar-I & Contai-I choropleths value 0 to + 10 & - 10 shows maximum correspondence. Fertile soil, favourable climate, developed irrigation system; developed transport system is the probable cause of this correspondence distribution. Remaining blocks shows either positive or negative distribution due to insuffiency of these variables.

DISCLAIMER

This self-learning material is based on different books, journals and web sources.