13

Note on sample survey

Dr. Biplab Dhak

Scientific sample surveys are cost-efficient and reliable ways to collect population-level information such as social, economic, demographic and health data. The objective is to achieve compatible, consistent and the best quality survey results. The sampling should be guided by a number of general principles. The key principles of sampling are as follows.

Sampling frame

A probability sample can only be drawn from an existing sampling frame which is a complete list of statistical units covering the target population. Since the construction of a new sampling frame is likely to be too expensive, any surveys should use a pre-existing sampling frame which is officially recognized. This is possible for most of the countries where there has been a population census in recent years. Census frames are generally the best available sampling frame in terms of coverage and reliability of an organization. However, an evaluation of the quality of the frame should be checked while using. The selection of sample frames however depends on the objective of the study or sampling unit or target population. For example, if the target population is primary school students, the sampling frame needs to be prepared based on data with school administration. For a multi-stage survey, a sampling frame should exist for each stage of selection. The sampling unit for the first stage of selection is called the Primary Sampling Unit (PSU); the sampling unit for the second stage of selection is called the Secondary Sampling Unit (SSU), and so on.

Evaluation of the sampling frame

No matter what kind of sampling frame will be used, it is always necessary to check the quality of the frame before selecting the sample. Following are things that need to be checked when using a conventional sampling frame:

• Coverage

- Distribution
- Identification and coding
- Measure of size
- Consistency

There are several easy but useful ways to check the quality of a sampling frame. For example, for a census frame, check the total population of the sampling frame and the population distribution among urban and rural areas and among different regions/administrative units obtained from the frame with that from the census report. Any important differences may indicate that there may be coverage problems. If the frame provides information on population and households for each Enumeration Areas (EA), then the average number of household members can be calculated, and a check for extreme values can help to find incorrect measures of size of the PSUs. If information on population by sex is available for each EA, then a sex ratio can be calculated for each EA, and a check for extreme values can help to identify non-residential EAs. If the EAs are associated with an identification (ID) code, then check the ID codes to identify miscoded or misplaced EAs. A sampling frame with full coverage and of good quality is the first element for a survey.

Stratification

Stratification is the process by which the survey population is divided into subgroups or strata that are as homogeneous as possible using certain criteria. Explicit stratification is the actual sorting and separating of the units into specified strata. Within each stratum, the sample is designed and selected independently. The principal objective of stratification is to reduce sampling errors. In a stratified sample, the sampling errors depend on the population variance existing within the strata but not between the strata. Another major reason for stratification is that, where marked differences exist between subgroups of the population (e.g., urban vs. rural areas), stratification allows for a flexible sample design that can be different for each subgroup. Stratification should be introduced only at the first stage of sampling. At the household selection stage, systematic sampling should be used for convenience.

Household listing and pre-selection of household

It is considered to be good practice that households be pre-selected prior to the start of fieldwork rather than in the field which may lead to the bias selection. In order to prevent bias, no changes

or replacements should not be allowed in the field. The list is usually obtained from a household listing operation conducted before the main survey. The household listing operation may be combined with the main survey to form a single field operation. Combining the household listing and survey data collection in one field operation is less expensive; however, it provides incentive to leave households off the household list to reduce workload, thus reducing the representativeness of the survey results. Close supervision is needed during the field work to prevent this problem. Separate listing and data collection operations are thus required for this reason.

Sampling errors and non-sampling errors

The estimates from a sample survey are affected by two types of errors: sampling errors and non-sampling errors. Sampling errors are the representative errors due to sampling of a small number of eligible units from the target population instead of including every eligible unit in the survey. Sampling errors are related to the sample size and the variability among the sampling units. Sampling errors can be statistically evaluated after the survey. Non-sampling errors result from problems during data collection and data processing, such as failure to locate and interview the correct household, misunderstanding of the questions on the part of either the interviewer or the respondent, and data entry errors. Non-sampling errors are related to the capacity of the implementing organization, and experience shows that (1) non-sampling errors are always the most important source of error in a survey, and (2) it is difficult to evaluate the magnitude of non-sampling errors once a survey is complete. Theoretically, with the same survey methodology and under the same survey conditions the larger the sample size, the better the survey precision. However, this relationship does not always hold true in practice, because non-sampling errors tend to increase with survey scale and sample size.

CONCEPT OF STANDARD ERROR

The standard deviation of sampling distribution of a statistic is known as its standard error (S.E) and is considered the key to sampling theory. The utility of the concept of standard error in statistical induction arises on account of the following reasons:

1. The standard error helps in testing whether the difference between observed and expected type frequencies could arise due to chance. The criterion usually adopted is that if a difference is less than 3 times the S.E., the difference is supposed to exist as a matter of chance and if the difference

is equal to or more than 3 times the S.E chance fails to account for it, and we conclude the difference as a significant difference. This criterion is based on the fact that at $X \pm 3$ (S.E.) the normal curve covers an area of 99.73 percent. Sometimes the criterion of 2 S.E. is also used in place of 3 S.E.

2. The standard error gives an idea about the reliability and precision of a sample. The smaller the S.E The greater the uniformity of sampling distribution and hence, greater is the reliability of sample. Conversely, the greater the S.E., the greater the difference between observed and expected frequencies. In such a situation the unreliability of the sample is greater. The size of S.E., depends upon the sample size to a great extent and it varies inversely with the size of the sample.
3. The standard error enables us to specify the limits within which the parameters of the population are expected to lie with a specified degree of confidence. Such an interval is usually known as confidence interval.

Sampling distribution: We are often concerned with sampling distribution in sampling analysis. If we take certain number of samples and for each sample compute various statistical measures such as mean, standard deviation, etc., then we can find that each sample may give its own value for the statistic under consideration. All such values of a particular statistic, say mean, together with their relative frequencies will constitute the sampling distribution of the particular statistic, say mean. Accordingly, we can have sampling distribution of mean, or the sampling distribution of standard deviation or the sampling distribution of any other statistical measure. It may be noted that each item
in a sampling distribution is a particular statistic of a sample. The sampling distribution tends quite closer to the normal distribution if the number of samples is large. The significance of sampling distribution follows from the fact that the mean of a sampling distribution is the same as the mean of the universe. Thus, the mean of the sampling distribution can be taken as the mean of the universe.

IMPORTANT SAMPLING DISTRIBUTIONS

Some important sampling distributions, which are commonly used, are: (1) sampling distribution of mean; (2) sampling distribution of proportion; (3) student's 't' distribution; (4) F distribution; and (5) Chi-square distribution.

SAMPLE SIZE AND ITS DETERMINATION

In sampling analysis the most important part is: What should be the size of the sample. If the sample size ('n') is too small, it may not serve to achieve the objectives and if it is too large, we may incur huge cost and waste resources. As a general rule, one can say that the sample must be of an optimum size. Size of the sample should be determined by a researcher keeping in view the following points:

(i) Nature of universe: Universe may be either homogeneous or heterogeneous in nature. If the items of the universe are homogenous, a small sample can serve the purpose. But if the items are heterogeneous, a large sample would be required. Technically, this can be termed as the dispersion factor.

(ii) Number of classes proposed: If many class-groups (groups and sub-groups) are to be formed, a large sample would be required because a small sample might not be able to give a reasonable number of items in each class-group.

(iii) Nature of study: If items are to be intensively and continuously studied, the sample should be small. For a general survey the size of the sample should be large, but a small sample is considered appropriate in technical surveys.

(iv) Type of sampling: Sampling technique plays an important part in determining the size of the sample. A small random sample is apt to be much superior to a larger but badly selected sample.

(v) Standard of accuracy and acceptable confidence level: If the standard of accuracy or the level of precision is to be kept high, we shall require a relatively larger sample.

(vi) Availability of finance: In practice, size of the sample depends upon the amount of money available for the study purposes. This factor should be kept in view while determining the size of sample.

Cochran's Sample Size Formula

The Cochran formula is widely used to calculate an ideal sample size given a desired level of precision, desired confidence level, and the estimated proportion of the attribute present in the population.

The Cochran formula is:

$N_o = Z^2 pq/e^2$

Where:

e is the desired level of precision (i.e. the margin of error),
p is the (estimated) proportion of the population which has the attribute in question,
q is 1 – p.
The z-value is found in a Z table.

Example

Suppose we are doing a study on the inhabitants of a large town, and want to find out how many households take breakfast in the mornings. Suppose we have prior knowledge that half of the families take breakfast. So p = 0.5. Now let's say we want 95% confidence, and at least 5 percent—plus or minus—precision. A 95 % confidence level gives us Z values of 1.96, from the normal tables, so we get

((1.96)2 (0.5) (0.5)) / (0.05)2 = 385.

So a random sample of 385 households in our target population should be enough to give us the confidence levels we need.

Modification for the Cochran Formula for Sample Size Calculation In Smaller Populations

If the population we're studying is small, we can modify the sample size we calculated in the above formula by using this equation:

$n = n_0 / 1 + (n_0 - 1) / N$

Here n0 is Cochran's sample size recommendation, N is the population size, and n is the new adjusted sample size. In our earlier example, if there were just 1000 households in the target population, we would calculate

$385 / (1 + (384 / 1000)) = 278$

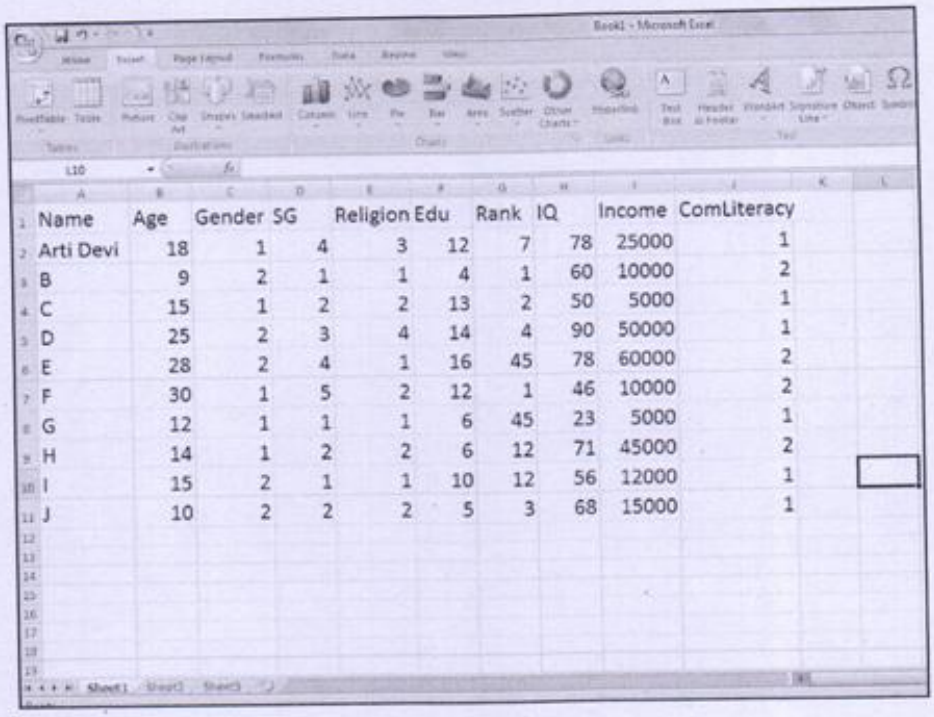
So for this smaller population, all we need are 278 households in our sample; a substantially smaller sample size.

Theorization and Representation of Social Data

Dr. Biplab Dhak

Socio-Economic Status of Students in Patna

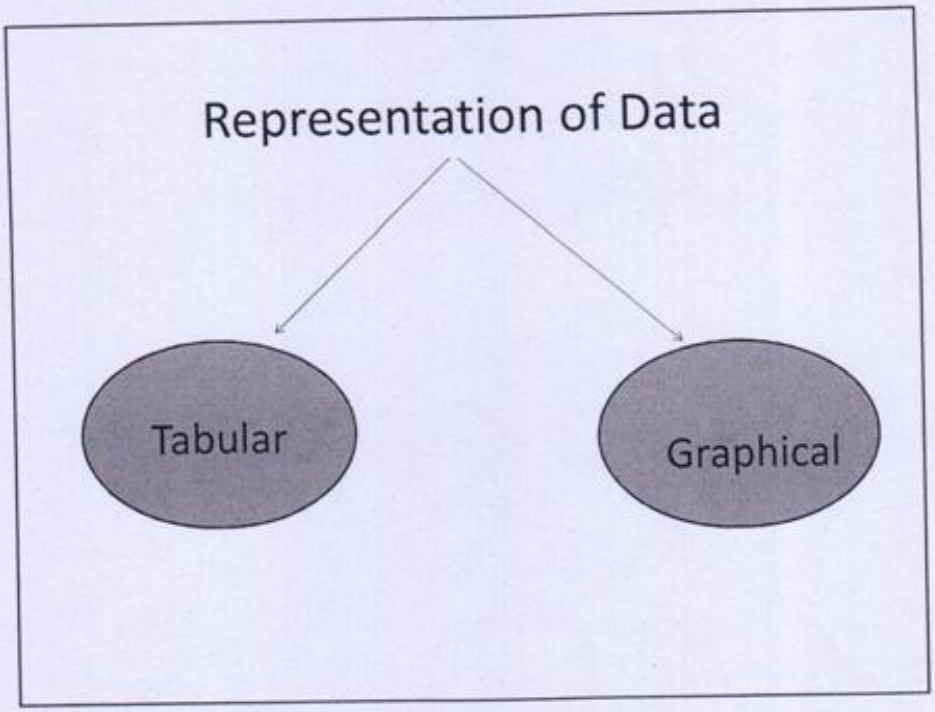
Sl No	Questions	Response
1	Name	Arti Devi
2	Age	18
3	Gender	1
	Male-1; Female-2	
4	Social Group	4
	SC-1; ST-2; OBC-3; General-4, Others-5	
5	Religions	3
	Hindu-1, Muslim-2, Chiristian-3, Sikh-4, Jain-5, Others (Specify).	
6	Years of Education	12
7	Rank in the class	7
8	IQ Score	78
9	Household Income (monthly)	25,000



Name	Age	Gender	SG	Religion	Edu	Rank	IQ	Income	ComLiteracy
Arti Devi	18	1	4	3	12	7	78	25000	1
B	9	2	1	1	4	1	60	10000	2
C	15	1	2	2	13	2	50	5000	1
D	25	2	3	4	14	4	90	50000	1
E	28	2	4	1	16	45	78	60000	2
F	30	1	5	2	12	1	46	10000	2
G	12	1	1	1	6	45	23	5000	1
H	14	1	2	2	6	12	71	45000	2
I	15	2	1	1	10	12	56	12000	1
J	10	2	2	2	5	3	68	15000	1

Type of Data

- Nominal
- Ordinal
- Interval
- Ratio



Tabulation

- Rate
- Ratio
- Frequency Distribution
- Mean, Median, Mode etc.

Graphical

- **Column**
- **Bar**
- **Line**
- **Pie**
- **Area**
- **Scatter**

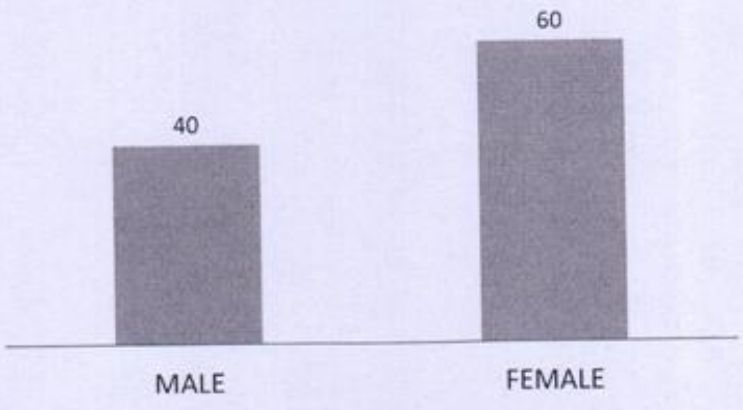
Percentage Distribution of Studies

Gender	%	N
Male	50.0	5
Female	50.0	5
Social Group		
SC	30.0	3
ST	30.0	3
OBC	10.0	1
General	20.0	2
Others	10.0	1
Total		10

Rate of Computer Literacy by Sex

MALE	40
FEMALE	60

Figure 1: Rate of Computer Literacy by Sex



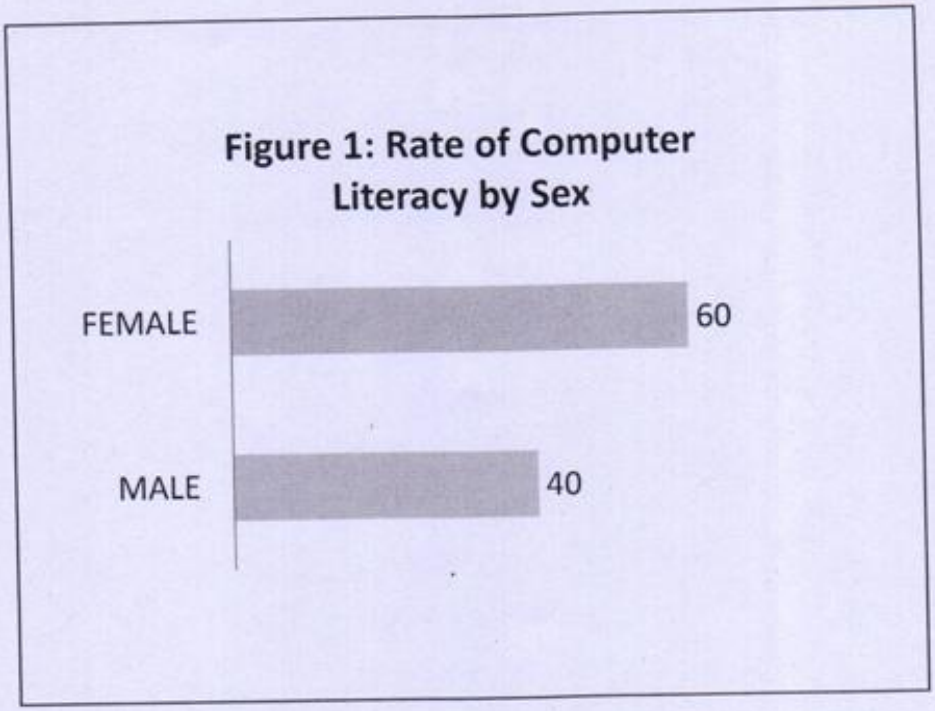
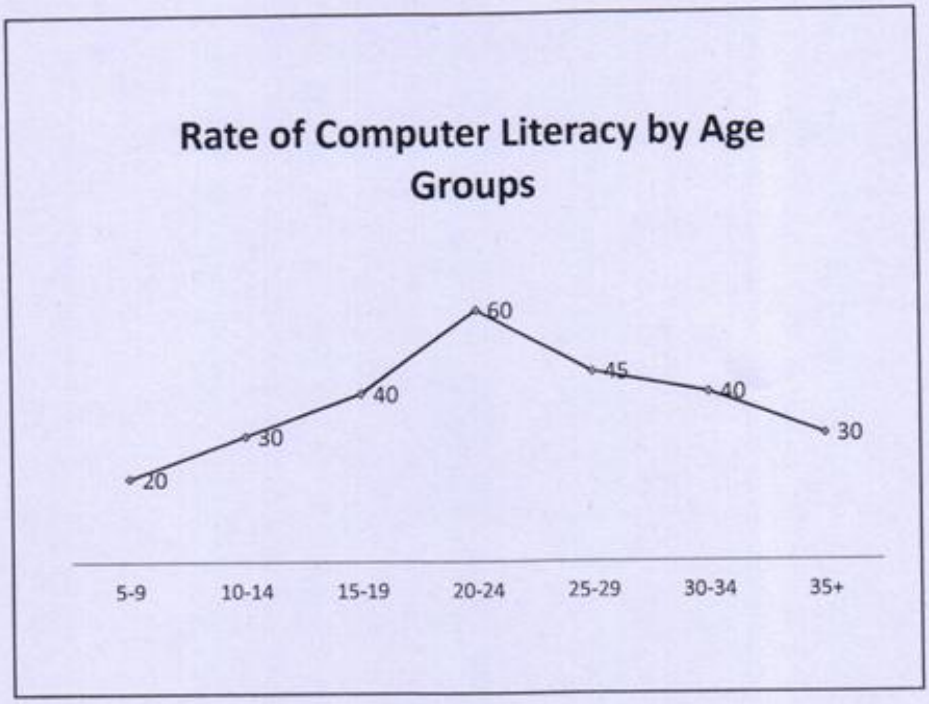
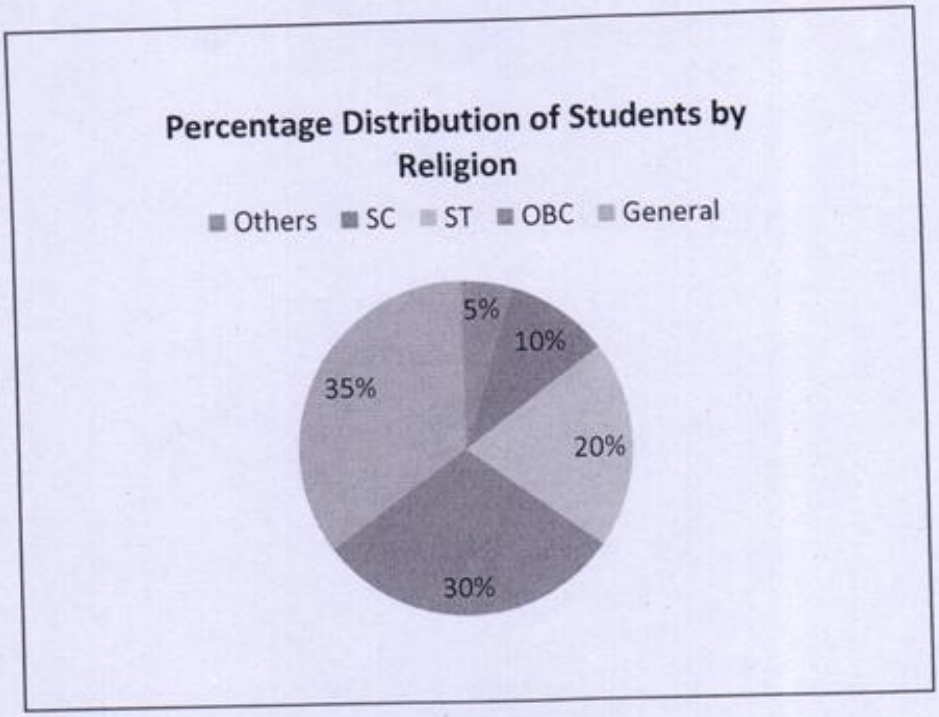
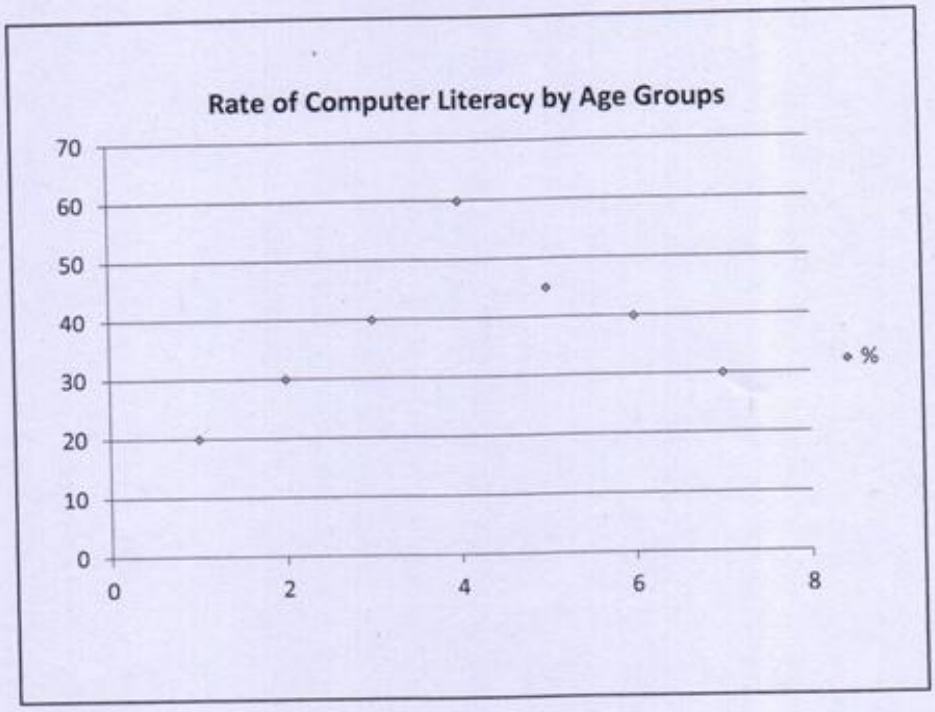


Figure 1: Rate of Computer Literacy by Sex
FEMALE
60
MALE
40


Rate of Computer Literacy by Age Groups
20
30
40
60
45
40
30
5-9
10-14
15-19
20-24
25-29
30-34
35+


Percentage Distribution of Students by Religion
Others
SC
ST
OBC
General
5%
10%
20%
30%
35%


Rate of Computer Literacy by Age Groups
70
60
50
40
30
20
10
0
0
2
4
6
8
%

What is qualitative research?

"Development of concepts which help us to understand social phenomena in natural (rather than experimental) settings, giving due emphasis to the meanings , experiences and views of the participants."

Dimensions of qualitative methods

Understanding context

- How economic, political, social, cultural, environmental and organizational factors influence health

Understanding people

- How people make sense of their experiences of health and disease

Understanding interaction

- How the various actors involved in different public health activities interact each other

Common qualitative study designs

Study design	Description
Ethnography	Portrait of people- study of the story and culture of a group usually to develop cultural awareness & sensitivity
Phenomenology	Study of individual's lived experiences of events- e.g. the experience of AIDS care
Grounded theory	Going beyond adding to the existing body of knowledge-developing a new theory about a phenomenon-theory grounded on data
Participatory action research	Individuals & groups researching their own personal beings, socio-cultural settings and experiences
Case study	In-depth investigation of a single or small number of units at a point (over a period) in time. E.g. Evaluation of s service

What is qualitative data?

- Data that are not easily reduced to numbers
- Data that are related to concepts, opinions, values and behaviours of people in social context
- Transcripts of individual interviews and focus groups, field notes from observation of certain activities, copies of documents, audio/video recordings...

Types of Qualitative Data

Structured text, (writings, stories, survey comments, news articles, books etc)

Unstructured text (transcription, interviews, focus groups, conversation)

Audio recordings, music Video recordings (graphics, art, pictures, visuals)

Qualitative data collection methods

- Observation
- Interview
- Focus Group Discussion

Preparing transcript

- Transcribe word by word (verbatim) Consider non-verbal expressions
- Try to do the transcribing yourself
- Be patient-Time consuming

What is Qualitative Data Analysis?

- Qualitative Data Analysis (QDA) is the range of processes and procedures whereby we move from the qualitative data that have been collected into some form of explanation, understanding or interpretation of the people and situations we are investigating.

- QDA is usually based on an interpretative philosophy. The idea is to examine the meaningful and symbolic content of qualitative data

Approaches in analysis

- Deductive approach – Using your research questions to group the data and then look for similarities and differences – Used when time and resources are limited – Used when qualitative research is a smaller component of a larger quantitative study

- Inductive approach – Used when qualitative research is a major design of the inquiry – Using emergent framework to group the data and then look for relationships

Terms used in Qualitative data analysis

- Theory: A set of interrelated concepts, definitions and propositions that presents a systematic view of events or situations by specifying relations among variables
- Themes: idea categories that emerge from grouping of lower-level data points
- Characteristic: a single item or event in a text, similar to an individual response to a variable or indicator in a quantitative research. It is the smallest unit of analysis
- Coding: the process of attaching labels to lines of text so that the researcher can group and compare similar or related pieces of information
- Coding sorts: compilation of similarly coded blocks of text from different sources in to a single file or report Indexing: process that generates a word list comprising all the substantive words and their location within the texts entered in to a program

The Process of Qualitative data analysis

Step 1: Organize the data

Step 2: Identify framework

Step 3: Sort data in to framework

Step 4: Use the framework for descriptive analysis

Step 5: Second order analysis

Types of qualitative analysis

- Content analysis
- Narrative analysis
- Discourse analysis
- Framework analysis
- Grounded theory

Content analysis

- Content analysis is the procedure for the categorization of verbal or behavioural data for the purpose of classification, summarization and tabulation

- The content can be analyzed on two levels – Descriptive: What is the data? – Interpretative: what was meant by the data?

Narrative analysis

- Narratives are transcribed experiences
- Every interview/observation has narrative aspect-the researcher has to sort-out and reflect up on them, enhance them, and present them in a revised shape to the reader
- The core activity in narrative analysis is to reformulate stories presented by people in different contexts and based on their different experiences

Discourse analysis

- A method of analyzing a naturally occurring talk (spoken interaction) and all types of written texts
- Focus on ordinary people method of producing and making sense of everyday social life: How language is used in everyday situations? – Sometimes people express themselves in a simple and straightforward way –

What is a theory?

< A. A logically interrelated set of propositions about empirical reality. These propositions are comprised of:

1. Definitions: Sentences introducing terms that refer to the basic concepts of the theory
2. Functional relationships: Sentences that relate the basic concepts to each other. Within these we have – a. Assumptions or axioms – b. Deductions or hypotheses
3. Operational definitions: Sentences that relate some theoretical statement to a set of possible observations

What do theories do?

– 1. Help us classify things: entities, processes, and causal relationships
– 2. Help us understand how and why already observed regularities occur
– 3 . Help us predict as yet unobserved relationships
– 4. Guide research in useful directions
– 5. Serve as a basis for action. "There is nothing so practical as a good theory."

Table 1: Rate of morbidity and Hospitalization

	Morbidity	Hospitalization
Place of Residence		
Rural	9.5	2.3
Urban	10	3.4
Gender		
Male	10.5	3.1
Female	8.6	2.0
MPCE		
1st	9.0	1.5
2nd	8.9	2.0
3rd	9.5	2.8
4th	10.8	4.2
Type of latrine		
Service	6.9	1.9
Pit	9.3	3.3
Septic/flush	9.4	3.3
No Latrine	9.0	2.0
All	9.6	2.6

Figure 1: Duration of Ailment (Number of days)

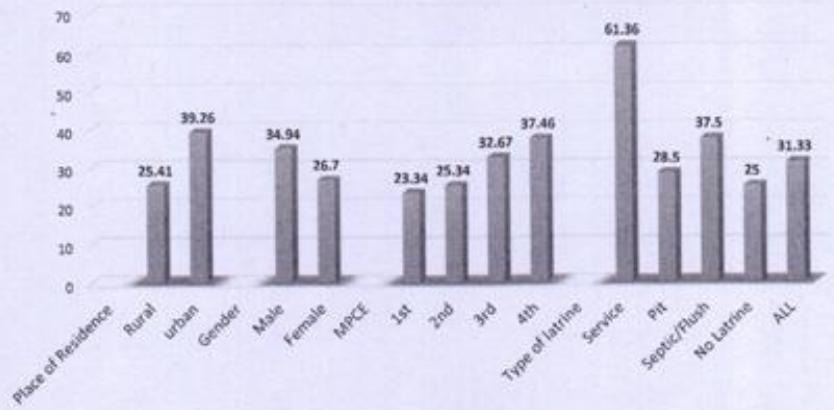


Table 2: Results of logistic regression of Morbidity and hospitalization

	Morbidity	Hospitalization
Place of Residence		
Rural	1	1
urban	1.01	1.07$_a$
Gender		
Male	1	1
Female	0.79$_a$	0.65$_a$
MPCE		
1st	1	1
2nd	1.23$_a$	1.46$_a$
3rd	1.24$_a$	1.97$_a$
4th	1.26$_a$	1.94$_a$
Type of latrine		
Service	1	1
Pit	0.70$_a$	0.49$_a$
Septic/Flush	0.41$_{1a}$	0.26$_a$
No Latrine	0.48$_a$	0.36$_a$
Constant	3.50	5.46

Note: 'a' denotes p<0.01; Religions, social group are taken into controlled.

Figure 2: Didn't seek treatment (%)

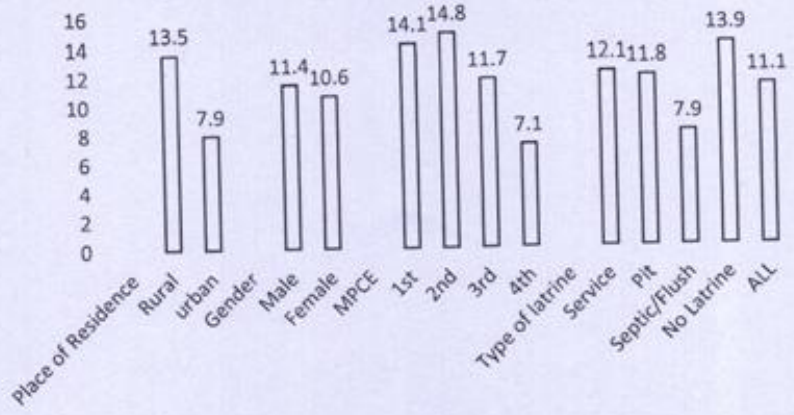


Table 3: Results of logistic regression of not seeking health care

	Odds Ratio
Place of Residence	
Rural	1
urban	1.40[a]
Gender	
Male	1
Female	1.07
MPCE	
1st	1
2nd	1.54[a]
3rd	1.65[a]
4th	1.55[a]
Type of latrine	
Service	1
Pit	0.41[b]
Septic/Flush	0.43[a]
No Latrine	0.34
Constant	-0.363

Note: 'a' denotes p<0.01; 'b' denotes p<0.05; Religion, social group are taken into controlled.

Table 5: Distribution of diseases by MPCE quartile classes

Diseases	1st	2nd	3rd	4th	All
Fever with loss of consciousness	2.6	3.7	3.8	3	3.3
Fever with rash	1.6	3.6	2.9	2.1	2.6
Fever due to diphtheria, whooping cough	6.3	8.3	5.9	6.5	6.7
All other fevers	40.7	35	36.7	40.2	38.2
Diarrhoeas	**10.1**	**8.2**	**7.8**	**5.4**	**7.4**
Acute upper respiratory infections	19.1	22.2	22.6	22.9	22.1
Cough with sputum	6.3	4.3	4.9	4.8	5
Bronchial asthma				1.2	
Gastric and peptic ulcers	2.5	1.1	1.7	1.3	1.6
Skin infection	1.9	3.5	2.2	2.5	2.5
Illness in the new-born	1.5	1.2	1.4	1.3	1.3
others	7.4	8.9	10.1	8.8	9.3
All	100	100	100	100	100

Table 4: Distribution of diseases by type of latrine

Diseases	Service	Pit	Septic/Flush	No latrine	All
Fever with loss of consciousness	12.1	3.3	3.2	3.2	3.3
Fever with rash	6.1	3.2	1.9	2.9	2.6
Fever due to diphtheria, whooping cough	6.1	4.8	7.3	7	6.7
All other fevers	30.3	42	37	37.1	38.2
Diarrhoeas	15.2	6.3	6.8	8.5	7.4
Acute upper respiratory infections	15.2	18.6	24.5	21.1	22.1
Cough with sputum	3.0	5.2	4.7	5.1	5.0
Bronchial asthma	3.0	-	1	-	-
Gastric and peptic ulcers	-	1.4	1.6	1.6	1.6
Skin infection	3.0	3	2.2	2.7	2.5
Illness in the new-born	3.0	1.1	1.2	1.6	1.3
others	3.0	11.1	8.6	9.2	9.3

Table 6: Causes of Death of children aged 1-59 months in India, WHO 2016

Cause of death	% Distribution
HIV/AIDS	0.9
Diarrhoeal diseases	22.2
Measles	5.5
Meningitis/encephalitis	4.2
Malaria	1.4
Acute lower respiratory infections	28.3
Prematurity	4.9
Birth asphyxia and birth trauma	1.2
Other communicable, perinatal and nutritional conditions	10
Congenital anomalies	6
Other no communicable diseases	7.7
Injuries	7.8
All	100.1

What does one mean by textual data?

"any text which constitutes a relevant and necessary source material for answering the questions one is interested in"

-open responses to questionnaires, -newspaper editorials,-commentaries, -articles, -different kinds of reports, e.g. company annual reports, -newspaper reports, etc., -journal articles, -advertisements, -public speeches, -conversations, -interviews,-letters

Content analysis

- Words
- Themes (is simple sentence)
- Character (person)
- Concept
- Items
- Paragraph

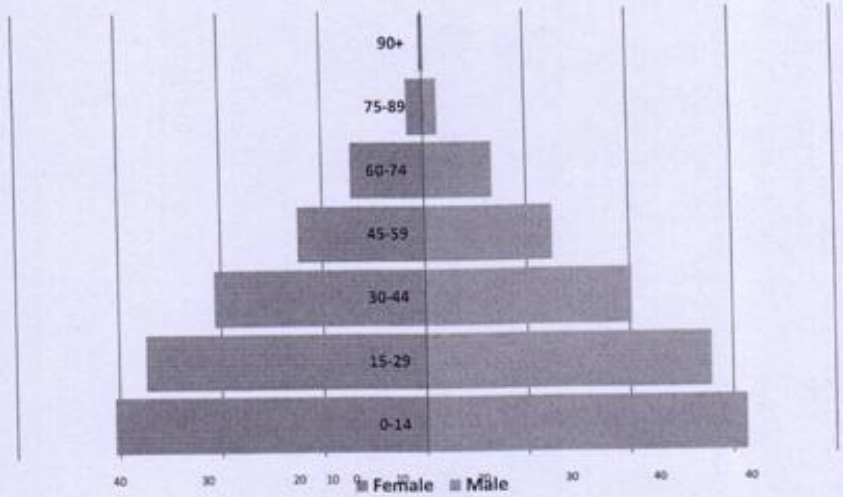
Table 1: Population in India by age groups and sex

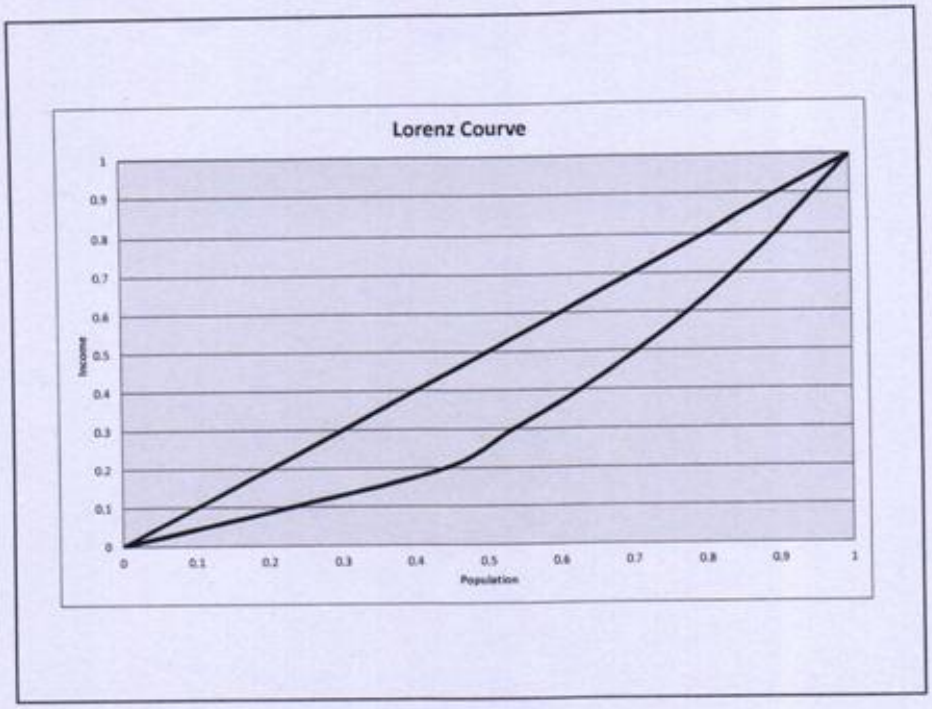
Age group	Male	Female
0-14	194301418	178049191
15-29	172859894	160406254
30-44	125100705	121022747
45-59	77422219	73081403
60-74	41292037	42025263
75-89	8536868	9297393
90+	1236309	1448844
All	620749450	585331095

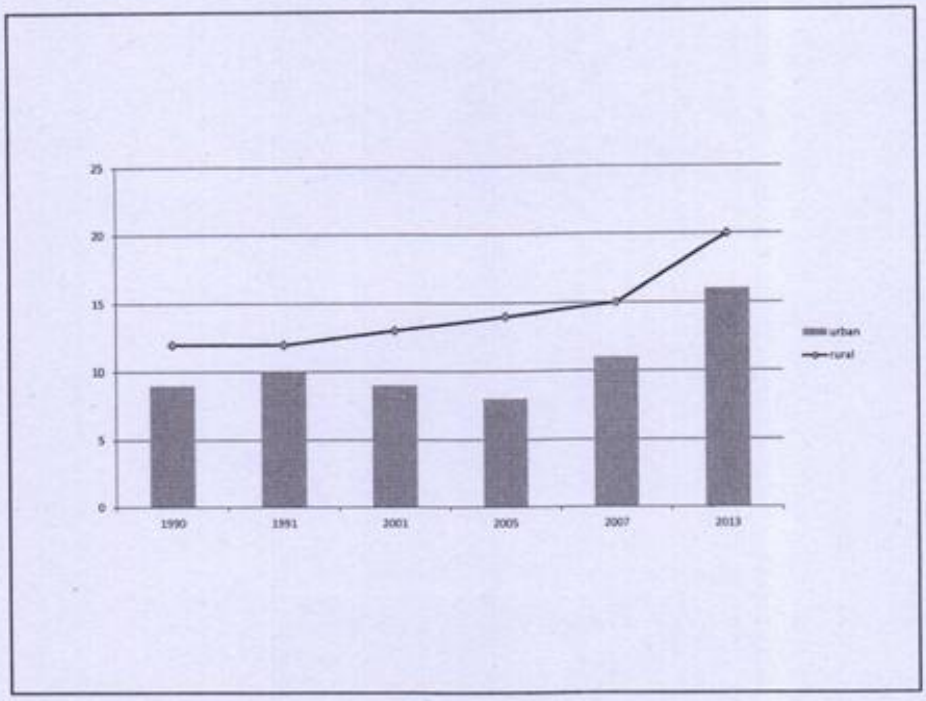
Table 2: Percentage distribution of population in India by age group and sex

Age Group	Male	Female
0-14	31.3	30.4
15-29	27.8	27.4
30-44	20.2	20.7
45-59	12.5	12.5
60-74	6.7	7.2
75-89	1.4	1.6
90+	0.2	0.2

Figure: Age Sex Pyramid of Population in India




Lorenz Courve
Income
Population
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


25
20
15
10
5
0
1990
1991
2001
2005
2007
2013
urban
rural