INDUCTION PROGRAMME FOR RESEARCH SCHOLARS
CENTRE FOR RESEARCH, ANNA UNIVERSITY

DATA COLLECTION

29.03.22
Building blocks of Research

• Observation of phenomena
• Identifying a problem
• Developing hypotheses
• Developing research design
• Collecting data
• Analyzing data
• Interpreting results
• **Observing a phenomena**

  Broader problem - sudden drop in the sales, increase in the employee turnover, decrease in the number of customer; drop in production, drop in quality of product
Identifying a problem

seeking in-depth information regarding the facts being observed; information gathered through formal questionnaires, interviews or informal or causal talk with the concerned people; production data, quality data; developing investigative question
• **Constructing a theory**
  – integrate all the information in a logical manner so as to conceptualize and test the **factors** responsible for problem
  – Identify the critical variables contributing to the problems
  – Association or relationship among the variables
  – establish research objective
• Developing hypotheses
  – framing of testable hypotheses; hypotheses testing are called deductive research
  – Sometimes, the hypotheses which are not originally formulated get generated through the process of induction
  – After the collection of data an insight may occur based on which new hypotheses can be formulated
  – Hypotheses testing through deductive research and hypotheses generation through induction are both common

• Developing research design
  – Different designs- budget, feasibility; quantitative, qualitative
• **Collecting data**
  – Scientific method of collection; primary data, secondary data; sampling

• **Analyzing data**
  – Statistically analyze and validate hypothesis

• **Interpreting results**
  – Deduction, reporting
Data

• Data is a set of values of subjects with respect to qualitative or quantitative variables.

• Data are individual pieces of factual information recorded and used for the purpose of analysis.
• Data is a set of numbers or series of numerical data.

• Field data is raw data that is collected in an uncontrolled environment. Experimental data is data that is generated within the context of a scientific investigation by observation and recording.
• **Quantitative** data deals with numbers and can be measured objectively – How much, How many, how-often
  – Ex. dimensions such as height, width, and length. Temperature and humidity, prices, area and volume.

• **Qualitative** data deals with characteristics and descriptors that can't be easily measured, but can be observed subjectively—such as smells, tastes, textures, attractiveness, and color. – How, Why
• **Primary data** - original data collected specially for a purpose; someone collected the data from the original source first hand
  – Experimental data
• **Secondary data** - from a source other than the researcher; government census reports, other governmental databases, and administrative data; journals, books thesis
• **Tertiary sources** are an interpretation of a secondary source; index, bibliographies, dictionaries, encyclopedias, handbooks, directories and internet search engines.
Data Collection

• **Data collection** is the process of **gathering** and measuring information on variables of interest in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.
Accurate and appropriate data collection

Improper data

• inability to answer research questions accurately
• inability to repeat and validate the study
• distorted findings resulting in wasted resources
• misleading other researchers to pursue fruitless avenues of investigation
• causing harm to human participants and animal subjects
Integrity of data collection

- Intentional
- Sampling error

Common issues in data collection

- Data collection without hypothesis, concept; no connectivity between hypothesis and objectives of the work
- Wrong selection of factors and levels
- Wrong selection of experimental design
- Not testing the data for reliability, validity
- Not analyzing and interpreting data immediately
Primary data collection

Quantitative data

• *Discrete* data is a count – number of students in a class, number of defective pieces in a lot

• *Continuous* data, on the other hand, could be divided and reduced to finer and finer levels – height of students in a class
## Sampling Plan

<table>
<thead>
<tr>
<th>Size</th>
<th>Accuracy</th>
</tr>
</thead>
</table>
Qualitative data

• Binary data
  – Yes/No, right/wrong, true/false, accept/reject, like/dislike

• Nominal data
  – “nominal” - “name” and nominal scales are kind of like “names” or labels.

A sub-type of nominal scale with only two categories (e.g. male/female) is called “dichotomous.”
• Ordinal
  – Ordinal scales are typically the measures of non-numeric concepts like satisfaction, happiness, discomfort

How do you feel today?
- 1 – Very Unhappy
- 2 – Unhappy
- 3 – OK
- 4 – Happy
- 5 – Very Happy

How satisfied are you with our service?
- 1 – Very Unsatisfied
- 2 – Somewhat Unsatisfied
- 3 – Neutral
- 4 – Somewhat Satisfied
- 5 – Very Satisfied
• **Interval**

• Interval scales are numeric scales in which we know both the order and the exact differences between the values.

• statistical analysis - *central tendency* can be measured by mode, median, or mean; standard deviation

• An **interval scale** is a **scale** (of measurement) created by units of equal size - *examples* of variables that use **interval scales** would be time, temperature (Celsius), temperature (Fahrenheit), etc.
**QUANTITATIVE DATA**

- Answers “WHAT” “WHERE” “HOW” “WHEN” and “WHO”
- Based on numbers
- Larger sample size
- Statistical analysis
- Objective
- Closed-ended questions
- To validate hypothesis

**QUALITATIVE DATA**

- Answers “WHY” (which is very important)
- Based on opinions and experiences
- Smaller sample size
- Interviews & observation
- Subjective
- Open-ended questions
- To generate hypothesis or develop ideas
Characteristics of data collection

- Validity
- Reliability
- Practicability
- Economy
- Convenience
- Interpretability
Qualitative data collection

• Direct Personal Investigation
• Indirect Oral Interview
• Mailed Questionnaire
• Schedules
• Local agencies
PREPARATION OF QUESTIONNAIRE

- Personally administered questionnaire
- Mail questionnaire

Steps involved:
- Deciding the information to be collected
- Formulate the questions needed to obtain the information
  - Open ended versus closed questions
    - OE- list five factors which made to choose a particular investment proposal
    - CE- nominal, ordinal or Likert or ratio scale
  - Dichotomous questions
  - Multiple choice questions (Nominal) – single answer
  - Checklist questions (multiple responses to a single question)
  - Ranking questions – like check list question with ranking
– Positively and negatively worded questions - questionnaire should include both positively and negatively worded questions

– Double-barreled questions –
  • Do you like the flavour and the taste of the soft drink?

– Ambiguous question
  • Rate the level of job satisfaction
– Memory related questions - If the questions require respondents to recall experiences from a distance past that are very hazy in their memory, then the answers to such question might have bias or wrong.

– Leading / Loaded questions - Questions should not be asked in such a way that the respondents are forced or directed to respond in a manner that he would not have, under normal situations where all possible alternatives are given.

“Don’t you think that salary is the main reason for software employees to quit the job”?
• **Bad questions** - Any question that prevent or disturbs the fundamental communication between the researcher; Ex. incomprehensible questions, unanswerable question, leading or loaded questions, double barreled question, controversial.
• Decide on the wordings of the questions and layout of the questionnaire
  – vocabulary should be simple, direct and familiar to all respondents
  – words used should not give raise to ambiguity or vagueness – ex. ‘often’, ‘usually’; “Mention your place of origin” - district, state or country?
  – Double barreled question should be avoided
  – The instructions provided to answer the question should not be confusing the respondent
  – questions asked should be applicable to all the respondents
– **Simple short questions** should be asked instead of long ones
– Questions should not be asked in such a manner that it will *elicit socially desirable response*.
– “Do you think that physically challenged people should be given more weightage in employment opportunities?”
  • Irrespective of the true feeling of the respondent a socially desirable answer would be provided.
• **Sequencing of questions**

  – The order can encourage or discourage the commitment and promote or hinder the development of researcher-respondent rapport.

  – The sequence of questions asked in the questionnaire should lead the respondents from questions of general nature to specific nature.

  – It should **start with relatively easy questions** which does not involve much thinking and should progress to difficult questions. This facilitates easy and smooth progress of the respondents through the various items in the questionnaire.
Care should be taken to see that the positively and negatively worded questions addressing the same issue or concept are not placed continuously. For eg.,

- I am satisfied with the working environment
- I am not satisfied with the working environment

- **sequenced questions** would introduce **bias** in the response which is frequently referred to as the ordering effects.

- Randomly placing the questions would reduce bias in the response; difficulty in categorizing, coding and analyzing the responses.
Layout of questionnaire

• A neat, properly aligned and attractive questionnaire with a good introduction, instructions and well sequenced questions and response alternatives will make things easier for the respondents to answer.
  
  – **Introduction section**, the researcher can disclose his identity and communicate the purpose of the research.
  
  – It would motivate the respondents to answer the questions by conveying the importance of the research work and by specifying the importance of contribution from the respondent.
  
  – The researcher should also ensure the confidentiality of the information provided. The introduction section should end with a courteous note, thanking the respondent for the time devoted to respond to the survey.
• The questions - organized in a logical manner and numbered sequentially under appropriate sections.

• **Proper instructions** - to complete the questions in an unambiguous manner.

• **Questionnaire design** - respondent spends only **minimum time and effort** in completing the same.
• Questions relating to the personal profile of the respondents viz., name, gender, age, education, income, marital status etc., can appear in the beginning or at the end of the questionnaire. The questions should provide a range of response options rather than seeking an exact figure. It may be arranged at the end of questionnaire also.

• Avoiding the question relating to the name of the respondent would be better as it ensures anonymity and enhances the probability of response.

• Identification of the questionnaire with a particular respondent can be made by assigning number instead of asking name.
• The open ended questions should be put at the end so the respondent may find it easy to comment on the various aspects.

• The questionnaire should end with an expression of sincere thanks to the respondent for spending their valuable time and effort.

• The researcher can also include a courteous note, reminding the respondents check that all the items have been completed properly.
• Pretesting the questionnaire and correcting the problem
  – 15 respondents are sufficient for a short and straightforward questionnaire, whereas 25 may be needed in case of a long and complex questionnaire
• The researcher pretesting – conducted at initial stages to build more structure in to the test; Fellow researchers can be involved
• Participant pretesting – tested in the field by involving the participants or participant surrogates. Surrogates are those individuals with characteristics and backgrounds similar to the desired participants.
• **Collaborative pretest** - conducted by the researcher where the researcher informs and alerts the participants of their involvement in the preliminary test of questionnaire. This makes the participants as the collaborators in the process of refinement of the questionnaire. A detailed probing of the parts of the question, including the words and phrases is carried out.

• **Noncollaborative pretest** - researcher does not inform the participant that the activity is a pretest. However the probing of the questionnaire is done.
• E-Survey
  – User friendly
  – Less time consumption
  – Without ambiguity
  – If questionnaire is long, provision for save for later submission
  – Fast access
• 1. Avoid technical terms unlikely to be familiar to your respondents
  
  Bad example:  Do you worry that you may have halitosis? Yes/No
  To improve:  Do you worry that you may have bad breath? Yes/No

• 2. Avoid combination questions
  
  Bad example:  Do you like fruits and vegetables? Yes/No
  To improve:  Do you like fruits? Yes/No
  Do you like vegetables? Yes/No
• 3. Avoid using leading questions that imply the response that is wanted

• Bad example: Do you agree with most people that capital punishment should be restored? Yes/No

To improve: Do you believe that for some crimes capital punishment should be restored, should not be restored or do you have no opinion? Yes/No/No opinion
• **Complex Language or Sentence Structure**

• If you were in a foreign country and did not speak the language but knew someone who can do for you?

• a. Let the person that spoke the language speak for you

• b. Try to learn the language on your own

• c. Not try to learn the language or let the person that you knew that spoke the language speak for you.
• **Response Options Are Not Exhaustive**
• What is your current age?
  ○ 10 to 20
  ○ 21 to 30
  ○ 31 to 40

• **Response Options Are Not Mutually Exclusive**
• What is your current age?
  10 or less
• 10 to 20
  ○ 20 to 30
  ○ 30 or greater
• **Asking for What Respondents Don’t Know**
• Are most of your teachers show personal problems on students?
• **Overly Personal or Threatening Questions**
• **Off -Topic Questions**
• Prepare questionnaire for collecting information on
  – Purchase behaviour of customers in a retail store
  – usage of newly introduced compression bandage on pain management compared to a known brand
  – Effectiveness of newly introduced pain balm in reducing headache problem compared to a known brand
<table>
<thead>
<tr>
<th>Element Selection</th>
<th>Representation Basis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Probability</td>
</tr>
<tr>
<td>Unrestricted</td>
<td>Simple random</td>
</tr>
<tr>
<td>Restricted</td>
<td>Complex random</td>
</tr>
<tr>
<td></td>
<td>Systematic</td>
</tr>
<tr>
<td></td>
<td>Stratified</td>
</tr>
<tr>
<td></td>
<td>Cluster</td>
</tr>
<tr>
<td></td>
<td>Double</td>
</tr>
</tbody>
</table>
• **Probability Sampling**
  Probability sampling is where each sampling unit in the defined target population has a known nonzero probability of being selected in the sample.

• **Unrestricted or Simple Random sampling**
  In the unrestricted probability sampling design every element in the population has a known, equal nonzero chance of being selected as a subject.
  
  10 employees \((n = 10)\) are to be selected from 30 employees \((N = 30)\) – selection by drawing sample from lot.
  
  Probability of selection = \(\frac{\text{Size of sample}}{\text{Size of population}}\)

  If the target population consists of a larger number of sampling units,
  
  – Table of random numbers or computer generated random numbers can be used
• Restricted or Complex Probability Sampling

• *Systematic random sampling*
  – if 100 samples have to be drawn from a defined target population of 1000, the skip interval would be 10\((1000/100)\); the researcher would randomly select a starting point and take every 10\(^{th}\) until the entire target population is proceeded thorough.

  • Easy but hidden patterns
• **Stratified random sampling**
  – separation of defined target population into different groups called strata and the selection of sample from each stratum
  – minimize the variability within each stratum and maximize the difference between strata
  – *proportionate stratified sampling*, each stratum is properly represented
  – *disproportionate stratified sampling*, the sample size selected from each stratum is independent of that stratum’s proportion of the total defined target population
  – Ex. Poll opinion
Cluster sampling

- Cluster sampling is a probability sampling method.
- Each cluster is assumed to be the representative of the heterogeneity of the target population. Groups of elements that would have heterogeneity among the members within each group are chosen for study in cluster sampling.
- Several groups with intragroup heterogeneity and intergroup homogeneity are found.
- A random sampling of the clusters or groups is done and information is gathered from each of the members in the randomly chosen clusters (single stage), probability random selection for members from randomly selected clusters (two stage).
- Cluster sampling offers more of heterogeneity within groups and more homogeneity among the groups.
- Ex. Purchase behavior of public in city for air conditioners.
• Double sampling
• This is also called sequential or multiphase sampling.
• Opted when further information is needed from a subset of group from which some information has already been collected for the same study.
• It is called as double sampling because initially a sample is used in the study to collect some preliminary information of interest and later a subsample of this primary sample is used to examine the matter in more detail.
Nonprobability Sampling

• If the researcher is less concerned about generalizability; purpose is just to obtain some preliminary information in a quick and inexpensive way

• when the population size is unknown
• Convenience sampling
• collection of information from members of population who are conveniently available to provide it
• Used during the exploratory phase of research, best way of getting some basic information quickly and efficiently
• assumptions - target population is homogeneous
• **Purposive sampling**
  – **Judgment sampling**
    – Participants are selected according to an experienced individual’s belief that they will meet the requirements of the study.
    – The researcher selects sample members who conform to some criterion.
    – It is appropriate in the early stages of an exploratory study.
    – This is used when a limited number or category of people have the information that are being sought.
    – Assumption is that the researcher’s belief that the opinions of a group of perceived experts on the topic of interest are representative of the entire target population.
• Quota sampling
  – Selection of prospective participants according to pre-specified quotas - the demographic characteristics (gender, age, education, income, occupation etc..,) specific attitudes (satisfied, neutral, dissatisfied) or specific behaviours (regular, occasional, rare user of product).
  – The purpose of quota sampling is to provide an assurance that pre-specified subgroups of the defined target population are represented
• Snowball Sampling
  – A set of respondents are chosen who help the researcher to identify additional respondents to be included in the study.
  – This method of sampling is also called as referral sampling
Selection of sampling plan

- Research objectives
  - generalize the findings of the research study - probability sampling method
- Scope of the research
  - scope of the research project is local, regional, national or international
- Availability of resources
- Time frame
- Advanced knowledge of the target population
  - complete lists of the entire population elements - probability sampling method
- Degree of accuracy
- Perceived statistical analysis needs
  - Probability sampling