Chapter 1: Intro and overview

Research

- Systematic method consisting of enunciating the problem, formulating a hypothesis, collecting the facts or data, analysing the facts and reaching certain conclusions either in the form of solutions towards the concerned problem or in certain generalisations for some theoretical formulation.

Steps in conducting research

Example:
- choose a CS subject you're interested in
- think of a problem or issue you see in that area
- refine your interest to a possible project that involves one or more ways of solving that problem
- outline the steps you'd take to do the project work and test your ideas
- What is your hypothetical conclusion?
- how would you evaluate the quality of your solution?

Types of research

The classification of the types of a study on the basis of these perspectives is not mutually exclusive: that is, a research study classified from the viewpoint of 'application' can also be classified from the perspectives of 'objectives' and 'enquiry mode' employed.

The need for research

- To gain familiarity with a phenomenon or to achieve new insights into it.
- To accurately portray the characteristics of a particular individual, situation or a group.
- To determine the frequency with which something occurs or with which it is associated with something else.
- To test a hypothesis of a causal relationship between variables.

Theoretical Framework:

The theory provides a point of focus for attacking the unknown in a specific area.
If a relationship is found between two or more variables a theory should be formulated to explain why the relationship exists.
Theories are purposely created and formulated, never discovered.
Useful in organizing information gathered from literature review

Research Objectives:
-To investigate
-To study
-To compare
-To analyze

Chapter 2 : Planning and designing research study

Literature Review is a documentation of a comprehensive review of the published and unpublished work from secondary sources of data in the areas of specific interest to the researcher.

It is an extensive survey of all available past studies relevant to the field of investigation. It gives us knowledge about what others have found out in the related field of study and how they have done so.

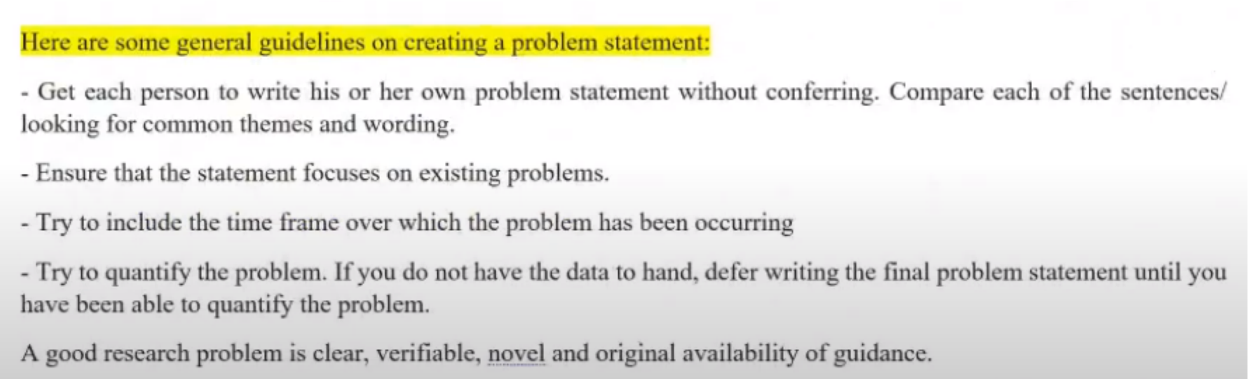
Variable:

•A variable is anything that can take on differing or varying values (attributes).

•For example; Age, Production units, Absenteeism, Sex, Motivation, Income, Height, Weight etc.

•Note: The values can differ at various times for the same object or person (or) at the same time for different objects or persons.

Problem statement



Here are some general guidelines on creating a problem statement:

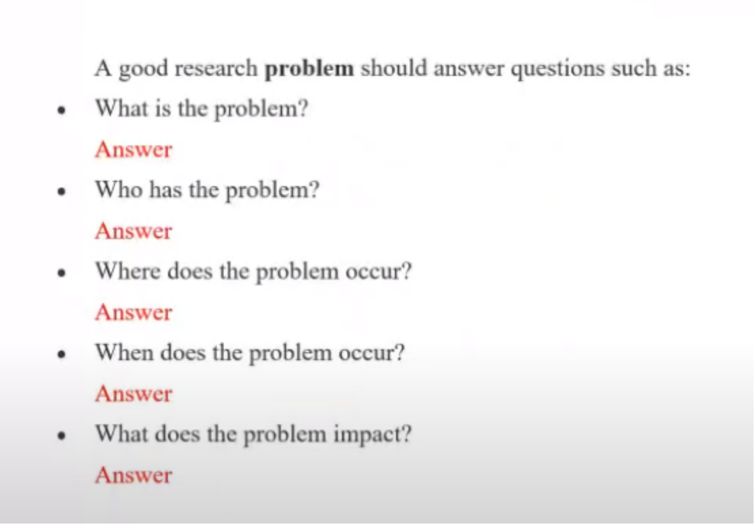
- Get each person to write his or her own problem statement without conferring. Compare each of the sentences/ looking for common themes and wording.

- Ensure that the statement focuses on existing problems.

- Try to include the time frame over which the problem has been occurring

- Try to quantify the problem. If you do not have the data to hand, defer writing the final problem statement until you have been able to quantify the problem.

A good research problem is clear, verifiable, novel and original availability of guidance.



A good research **problem** should answer questions such as:

- What is the problem?

  Answer

- Who has the problem?

  Answer

- Where does the problem occur?

  Answer

- When does the problem occur?

  Answer

- What does the problem impact?

  Answer

Chapter 3 General Types of Research Designs and Approaches

Quantitative Research

- Quantitative research is ‘explaining phenomena by collecting numerical data that are analyzed using mathematically based methods (in particular statistics)’.
-
- Quantitative research is based on the measurement of quantity or amount. It is applicable to phenomena that can be expressed in terms of quantity.
- It is an extensive survey of all available past studies relevant to the field of investigation. It gives us knowledge about what others have found out in the related field of study and how they have done so.

- Eg of quantitative research

- This could be questions like ““I think the lecture class is interesting’. We can develop a questionnaire that asks pupils to rate a number of statements as either ‘agree strongly’, ‘agree’, ‘disagree’ or ‘disagree strongly’, and give the answers a number (e.g. 1 for ‘agree strongly, 4 for strongly disagree).
- Now, we have quantitative data on pupil attitudes for lecture class.

Qualitative Research

- Qualitative research is concerned with qualitative phenomena, i.e., phenomena relating to or involving quality or kind.
- For instance, when we are interested in investigating the reasons for human behaviour (i.e., why people think or do certain things), we quite often talk of ‘Motivation Research’, an important type of qualitative research.
- This type of research aims at discovering the underlying motives and desires, using in depth interviews for the purpose.
-
- Qualitative research is especially important in the behavioural sciences where the aim is to discover the underlying motives of human behaviour.
- Through such research we can analyze the various factors which motivate people to behave in a particular manner or which make people like or dislike a particular thing.

Experimental study involves the researcher introducing the intervention that is assumed to be the cause of change and waiting until it has produced or has been given sufficient time to produce the change.

Non-experimental study consists of the researcher observing a phenomenon and attempting to establish what caused it. In this instance, the researcher starts from the effect or outcome and attempts to determine the causation.

In summary, non-experimental study starts from the effects to trace the cause.

After-only experimental design Example

- The researcher knows that a population is being exposed to an intervention and wishes to study its impact on the population.
- In this design, information on baseline (pre-test or before observation) is usually constructed on the basis of respondents' recall of the situation before the intervention or from information available in existing records.
- The change in the dependent variable is measured by the difference between the 'before' baseline and 'after' data sets.

Before-and-After experimental design Example

- It overcomes the problem of retrospectively constructing the 'before' observation by establishing it before the intervention is introduced to the study population.
- Then, when the program has been completely implemented, the 'after' observation is carried out to ascertain the impact attributable to the intervention.

Control group design

- In a study utilizing the control group design, the researcher selects two population groups instead of one: a control group and an experimental group.
- These groups are expected to be comparable as possible in every respect except for the intervention (the cause responsible for bring the change)
- The experimental group is exposed to the intervention.
- When it is assumed that the intervention has had an impact, an 'after' observation is made on both groups.
- Any difference in the 'before' and 'after' observations between the groups regarding the dependent variables is attributed to the intervention.

Chapter 4 Data Collection and Measurements strategies

Methods of Data Collection ***obj in midterm**

(New)

Primary data: Original source (collected by ourselves)

- Observation(direct and indirect, participants or non-participants), survey, interview and focus group

(Old/Existing)

Secondary data: Retrieve from existing data set (databases)

- Article, journal, internet, online resources, library (Document review and record)

Observation – direct and indirect (recording)

Four of the most popular data collection methods:

Direct observation

Experiments

Surveys – interviews, questionnaire

Direct observation

- Participant observation is when you, as a researcher, participate in the activities of the group being observed in the same manner as its members, with or without their knowing that they are being observed. *(e.g Rosenhan Experiment)*
- Non-participant observation, is when you, as a researcher, do not get involved in the activities of the group but remain a passive observer, watching and listening to its activities and drawing conclusions from this.
- Hawthorne effect: a change in the behaviour of persons or groups is attributed to their being observed

Direct: Non-participant Observation

- Non-participant Observation involves observing participants without actively participating (how would you behave when you know someone is observing?)
- Hawthorne effect: a change in the behaviour of persons or groups is attributed to their being observed
- Factors that may bias the results of observational studies can be broadly categorized as: selection bias resulting from the way study subjects are recruited or from differing rates of study participation depending on the subjects' cultural background, age, or socioeconomic status, information bias, measurement error, confounders, and further factors.
    - Halo effects: The **halo effect** is a cognitive bias in which an observer's overall impression of a person influences the observer's feelings and thoughts about that person's character.

Indirect Observation

- Recording Observation:
  - **Narrative Recording** – Records description of the interaction in their own words. (qualitative research)
  - **Using Scales** – develop a scale to rate the interaction or phenomenon. (one, two, three-directional: positive, neutral, negative)
  - **Categorical recording** – use categories to classify observation, e.g. passive/active (two categories)
  - **Recording on electronic devices** – recording using videotape or other electronic devices and then analyzed.

Survey: Interview

Methods of Data Collection – **Surveys**

- A survey solicits information from people; e.g. pre-election polls; marketing surveys.
- The *Response Rate* (i.e. the proportion of all people selected who complete the survey) is a key survey parameter.
- Surveys may be administered in a variety of ways, e.g.
  - Personal Interview,
  - Telephone Interview, and
  - Self-Administered Questionnaire.

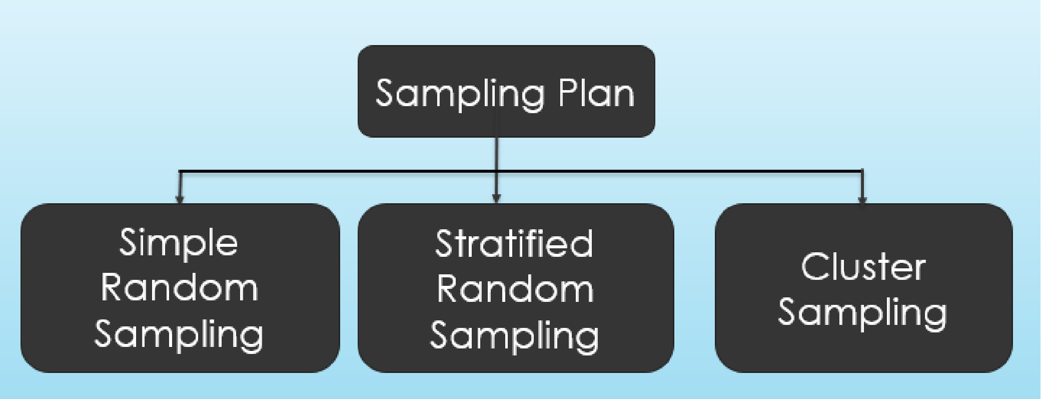
Data Sampling

- Selecting a subset of a whole population
- Statistical inference permits us to draw conclusions about a population based on a sample.
- Sampling (i.e. selecting a subset of a whole population) is often done for reasons of cost and practicality:
- it’s less expensive to sample 1,000 television viewers than 100 million TV viewers
- performing a crash test on every automobile produced is impractical

Sampled population and the target population should be similar to each other

Sampling Plan

- A sampling plan is just a method or procedure for specifying how a sample will be taken from a population.



Simple random sampling (SRS)

- A simple random sample is a sample selected in such a way that every possible sample of the same size is equally likely to be chosen.
- E.g. Drawing three names from a hat containing all the names of the students in the class is an example of a simple random sample: any group of three names is as equally likely as picking any other group of three names.

Stratified random Sampling

- A stratified random sample is obtained by separating the population into mutually exclusive sets, or strata, and then drawing simple random samples from each stratum.
- We can acquire information about the total population, make inferences within a stratum or make comparisons across strata.
- The accuracy of your estimate largely depends on the extent of variability or heterogeneity of the study population with respect to the characteristics that have a strong correlation with what you are trying to ascertain. It follows, therefore, that if the heterogeneity in the population can be reduced by some means for a given sample size you can achieve greater accuracy in your estimate. Stratified random sampling is based upon this logic. (more from chap 12 pg 203, Ranjit Kumar)
- Stratified sampling is one, in which the population is divided into homogeneous segments, and then the sample is randomly taken from the segments.

Cluster sampling

- Cluster sampling refers to a sampling method wherein the members of the population are selected at random, from naturally occurring groups called 'clusters'.
- A cluster sample is a simple random sample of groups or clusters of elements (vs. a simple random sample of individual objects).
- This method is useful when it is difficult or costly to develop a complete list of the population members or when the population elements are widely dispersed geographically.
- Cluster sampling may increase sampling error due to similarities among cluster members.

Data sampling - Sample Size

- Saturation occurs when adding more participants to the study does not result in additional perspectives or information

Sampling Errors

- Sampling error refers to differences between the sample and the population that exist only because of the observations that happened to be selected for the sample.
- Another way to look at this is: the differences in results for different samples (of the same size) is due to sampling error.
- E.g. Two samples of size 10 of 1,000 households. If we happened to get the highest income level data points in our first sample and all the lowest income levels in the second, this is a consequence of sampling error.
- Increasing the sample size will reduce this type of error.

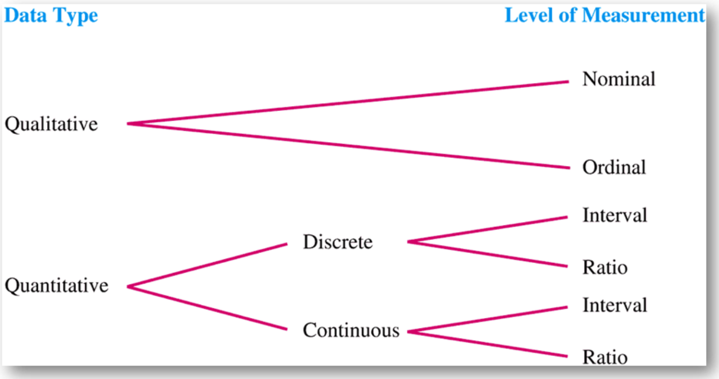
Non-Sampling Errors

- Non-sampling errors are more serious and are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly.
- Increasing the sample size will **NOT** reduce this type of error.

Non-sampling Error: Nonresponse Error

- Possible reasons- sensitive questions (unwilling to answer), questionnaire too long, bz respondents (wrong target)
- As mentioned earlier, the Response Rate (i.e. the proportion of all people selected who complete the survey) is a key survey parameter and helps in the understanding of the validity of the survey and sources of nonresponse error.

Summary of possible data types and levels of measurement



(Quantitative) Discrete vs Continuous Data:

- Nominal and ordinal scales are for discrete data
- data in which each item is a separate, whole unit
- E.g. number of individuals or species, members of particular groups (species, type of car, round or square, heavy or light, etc.)
- Interval and ratio scales are for continuous data
- data for points along a scale that, at least theoretically, could be subdivided
- E.g. temperature, weight, lengths, units of time, etc

Nominal level of measurement

- classifies data into names, labels or categories in which no order or ranking can be imposed.
- E.g. the number of courses offered in each of the different colleges.
- Tesco, Giants, Aeon
- Red,Yellow,Green
- Yes,no

Ordinal level of measurement

- classifies data into categories that can be ordered or ranked, BUT precise differences between the ranks do not exist.
- Generally, it does not make sense to do calculations with data at the ordinal level (we do not know the differences between them, equal? 30% more/less?)
- The first, second & third winners; Grade A,B,C..

*A zero point is an arbitrary point on a scale and does not indicate the absence of a quality or characteristic.

Interval level of measurement **(zero has meaning)**

- ranks data, precise differences between units(distance between units) of measure exist, but there is no meaningful zero. If a zero exists, it is an **arbitrary** point.
- E.g. IQ scores, it makes sense to talk about someone having an IQ 20 points higher than another person, but an IQ of zero has no meaning.

Ratio level of measurement **(zero has no meaning)**

- has all the characteristics of the interval level, but a true zero exists.
- Also, true ratios exist when the same variable is measured on two different members of the population.
- E.g. weight of an individual. It makes sense to say that a 150 lb adult weighs twice as much as a 75 lb. child.

- 0=**absolute**
- Because of the existence of true zero value, the ratio scale doesn't have negative values.

Precision and Accuracy in Measurement

- Greater precision does not always mean greater accuracy and accuracy does not guarantee precision.
- Digital watches can be precise to the hundredth of a second, but not necessarily accurate.

Measurement Scale- Attitudinal Scale

Likert scale

- Likert scale is a type of psychometric response scale in which responders specify their level of agreement to a statement.
- An example of 5 point Likert scale is: (1) Strongly disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly agree.

Thurstone Scale

- The Thurstone scale measures a respondent's attitude by using a series of "agree-disagree" statements of various weights. These statements help determine not only how a respondent feels, but how strongly they feel that way.

Guttman Scale

- The Guttman or "cumulative" scale measures how much of a positive or negative attitude a person has towards a particular topic.
- In this example here, Guttman scale is used to measure the acceptance of immigrants

Number of children: interval, discrete, quantitative.

Weight in kilogram: ratio, continuous, quantitative.
Hair color: nominal, qualitative
Economic status (low, medium, high): ordinal, qualitative
Pg 38

- ´Eye Color (blue, brown, green, hazel) **- qualitative (nominal)**
- ´Rating scale (poor, good, excellent) **- qualitative (ordinal)**
- ´ACT score **- quantitative, discrete (ratio)**
- ´Salary **- quantitative, continuous (ratio)**
- ´Age - **quantitative, discrete(ratio)**
- ´Ranking of high school football teams in Missouri **- qualitative (ordinal)**
- ´Nationality - qualitative (nominal)
- ´Temperature **- quantitative, continuous (interval)**
- ´Zip code - **qualitative (nominal)**

Pg 39
Classify each of the following variables as nominal, ordinal, interval or ratio and as discrete or continuous:

a) birth order - **quantitative, ordinal**
b) number of teeth - **qualitative, discrete (ratio)**
c) tail length -**qualitative , continuous (ratio)**
d) area of inhibited growth - **qualitative , continuous (ratio)**
e) species - **quantitative, nominal**
f) major (i.e. biology, sociology etc.) - **quantitative, nominal**
g) military rank - **quantitative, ordinal**

Chapter 5: Data Analysis and interpretation

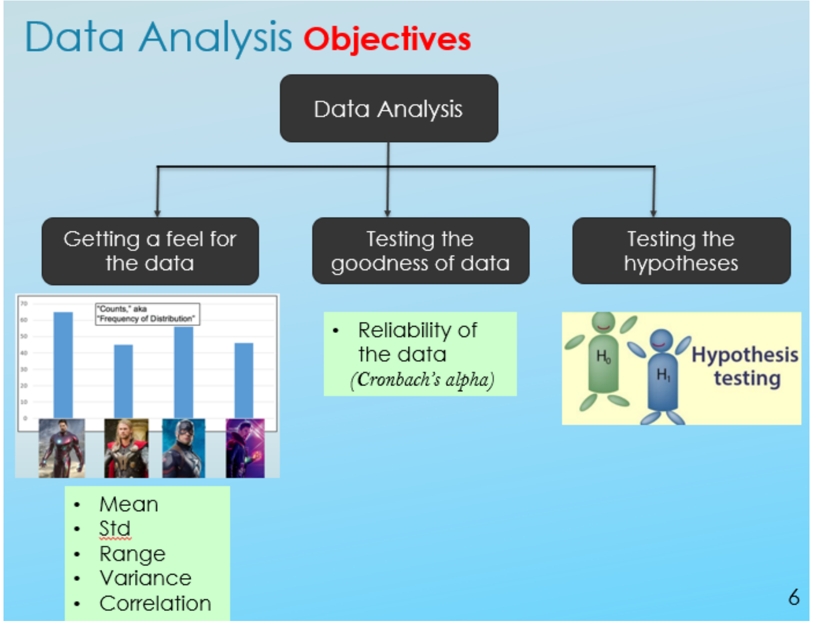
- **Data preprocessing** involves transforming raw data into an understandable format.
- **Data analysis and data interpretation** is an important and exciting step in the process of research. It is the process of assigning meaning to the collected information and determining the conclusions, significance and implications of the findings.

Data preprocessing

Data preprocessing in this sample spreadsheet involved the identification of unique values for ID column, the date format for birthday column, attribute dependencies, misspelling for country column, misfielded and missing values for city column and also the invalid values for gender column.

Data Preprocessing: Missing Value

- In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.
- There are various options to deal with missing values such as deleting the observations or variable, imputation with mean / median / mode.



Data Analysis **Objectives**

Data Analysis

Getting a feel for the data

Testing the goodness of data

Testing the hypotheses

- Mean
- Std
- Range
- Variance
- Correlation

- Reliability of the data *(Cronbach's alpha)*

Hypothesis testing

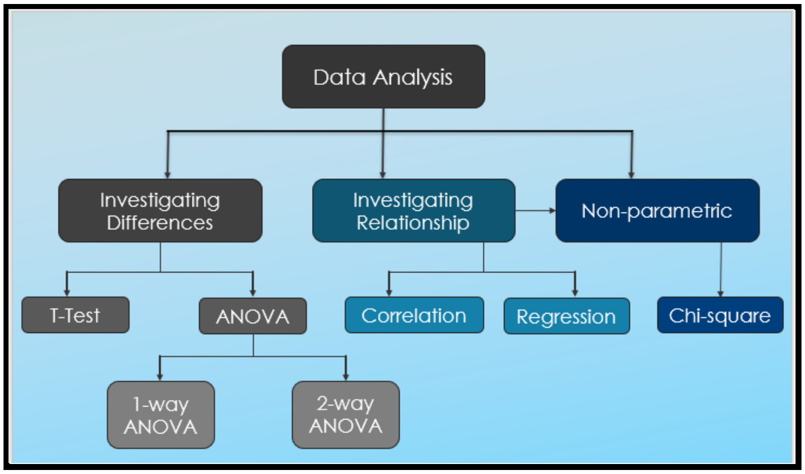
6

Cronbach's alpha (**midterm)

- Cronbach's alpha is one of the commonly used measures to assess the reliability, or internal consistency of a set of test items.

- We use Cronbach's alpha tests to see if multiple-question Likert scale surveys are reliable. These questions measure latent variables i.e hidden or unobservable variables like: a person's conscientiousness or openness.

One of the midterm favourite slides (mcq)



- Parametric tests assume underlying statistical distributions in the data.
- Whereas nonparametric tests do not rely on any distribution. They can thus be applied even if parametric conditions of validity are not met.
- We use **T-Test** and **ANOVA** to investigate the differences.
- We use **correlation** and **regression** to investigate relationships.
- And Chi-square is a non-parametric test. **Chi-square** is performed to test the relationships between categorical/nominal variables.

ANOVA test

- An ANOVA test is a way to find out if survey or experiment results are significant. In other words, ANOVA test helps you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis.
- One-way analysis of variance (ANOVA) tests how much the mean values of a numerical variable differ among the categories of a categorical variable.
- In one-way ANOVA we compare the means of two or more samples.
- T- Test à 2 samples