Data Collection and Analysis for Primary Research

Mayura Arandara

Points to be discussed

- Scientific Decision-making?
- What is data,
- Types of data and examples
- Data collection methods
- Terms used in data collection and analysis
- Sampling techniques
- Statistics in data analysis
- Basic data analysis methods

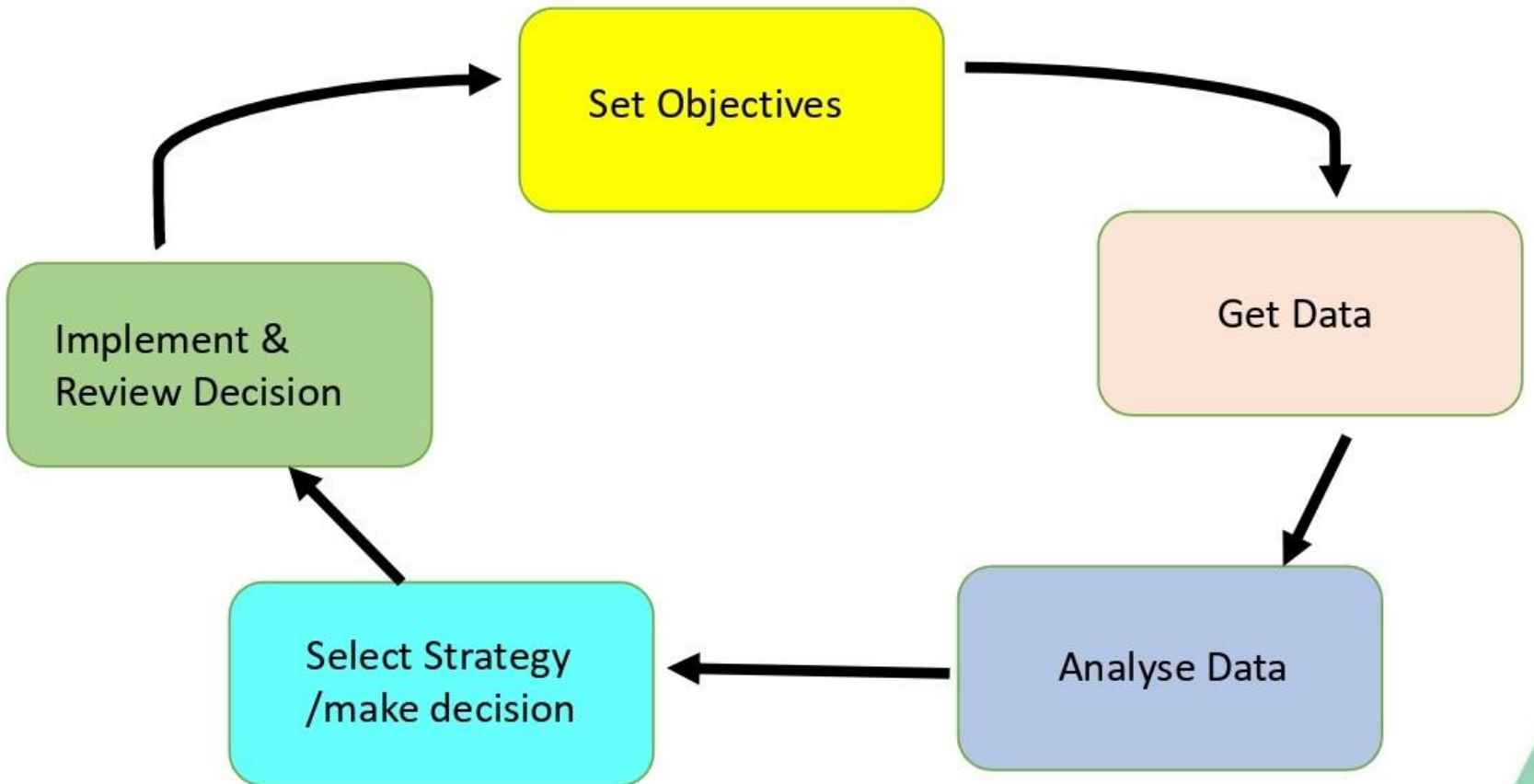
References

- Folz D.H, Survey Research for Public Administration
- Endagamge D. M., Researcher Development Program, Sri Lanka Institute of Development Administration
- Scientific Social Surveys and Research, 4th edition, NJ
- Siddisena K.P., Researcher Development Program, Sri Lanka Institute of Development Administration

Scientific Decision Making

- Scientific decision-making : the scientific process of making decisions based on evidence and data by adopting a systematic approach, instead of decisions based on intuition or trial and-error
- We need Evidence…….. Systematic approach…..Data
- Scientific decisions help to reduce the risk and form an analytical process which leads to
    - improve management efforts through already gathered information
    - predict the future of organization
    - correct errors based on the results from the past

Scientific Decision-Making Model



What is data

Data:

- The literally meaning of data is factual information about observation or experiment.
- Data are what is actually recorded by the researcher.
- Data are known facts or factual information.
- Data are gathered body of facts.
- Data refer to any observations or information collected during the research process.

Datum: is the singular of Data; is used to refer to any single observation or any single piece of information-e.g. age, any subject's sex.

- Generally speaking, we are cornered with many observations or information.

Information:

- Information is the knowledge derived from data/experience/study
- Data is raw material to derive information (Pauline Young, 1966)

(Ref: Scientific social surveys and research, 4thedition, New Jersey, Prentice Hall Inc.

Types of data

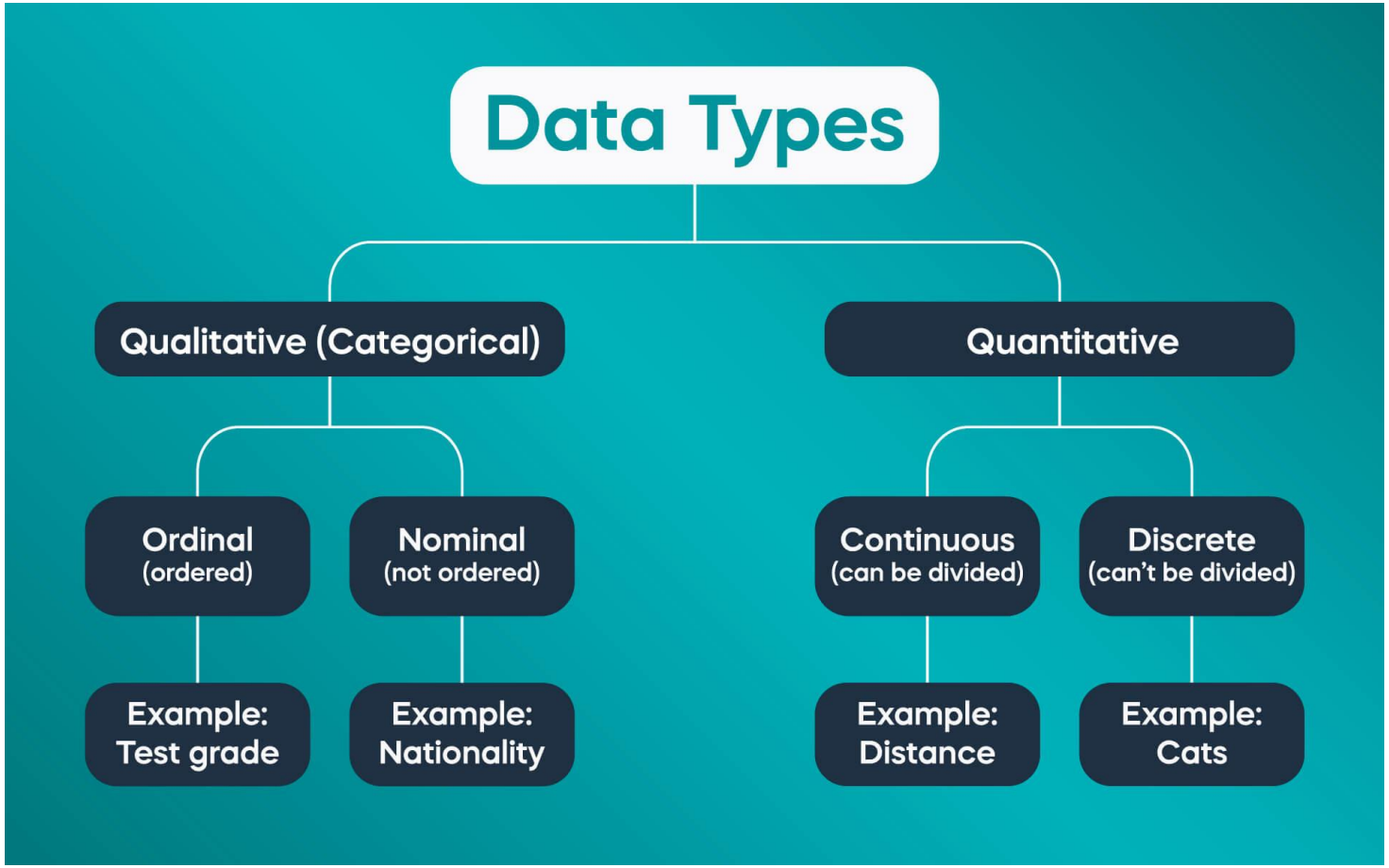
Level

- Primary data
- Secondary data

Arrangement

- Cross sectional data means that we have data from many units, at one point in time
- Time series data means that we have data from one unit, over many points in time
- Panel data (or time series cross section) means that we have data from many units, over many points in time

Types of data ctd.



Types of data ctd..

Observations	Average SAT Scores	Organization Form of Schools
School 1	540	Hierarchical
School 2	519	Flat
School 3	507	Hybrid
School 4	499	Hierarchical
....	...	...
School 50	557	Flat

Cross section

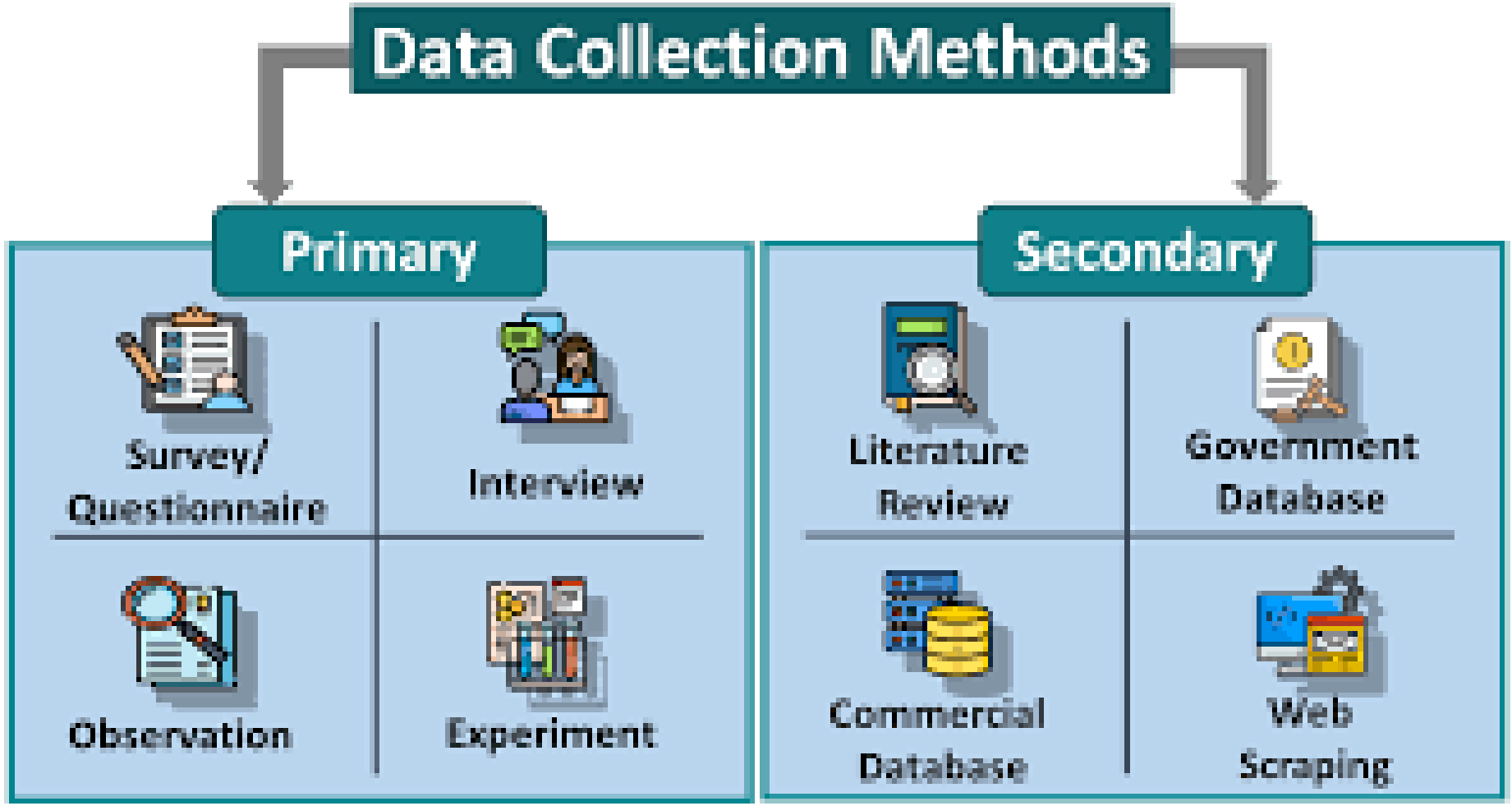
date date Date	charter_foreign bigint Integer
2008-01-01T00:00:00.000Z	57463
2008-02-01T00:00:00.000Z	56651
2008-03-01T00:00:00.000Z	95878
2008-04-01T00:00:00.000Z	100198
2008-05-01T00:00:00.000Z	130720
2008-06-01T00:00:00.000Z	196656
2008-07-01T00:00:00.000Z	195873
2008-08-01T00:00:00.000Z	152848
2008-09-01T00:00:00.000Z	88267
2008-10-01T00:00:00.000Z	76723
2008-11-01T00:00:00.000Z	26671
2008-12-01T00:00:00.000Z	30963
2009-01-01T00:00:00.000Z	35174

Time series

person	year	x	y
A	2018	3,5	85
A	2019	3,2	83
A	2020	3,8	88
B	2018	1,2	79
B	2019	1,5	83
B	2020	2,3	88
C	2018	5,6	75
C	2019	6	72
C	2020	5,8	78

Panel

Data collection methods



Survey questionnaire

- A questionnaire is a list of carefully structured questions, chosen after considerable testing, with a view to elicit reliable responses from a chosen sample. The aim is to find out what a selected group of participants do, think or feel.

- There are two types of questionnaire.
  1. Postal/Mail Questionnaire
  2. Questionnaire to be filled by enumerator/interviewer

❖ Advantages of a postal questionnaire

1. Low cost
2. Easy to manage

❖ Disadvantages of a postal questionnaire

1. Low response rate
2. Doubtful response/not targeted response
3. Biased response
4. Less probing facilities
5. Less accuracy

Questionnaire to be filled by enumerator

- There are two forms of questionnaire
  1. Household Questionnaire (H Schedule)
  2. Individual Questionnaire

-Advantages of having HH Questionnaire

a. Obtaining HH information

b. Identification of the targeted respondents

2. Individual Questionnaire

a. Questions related to **Background information**

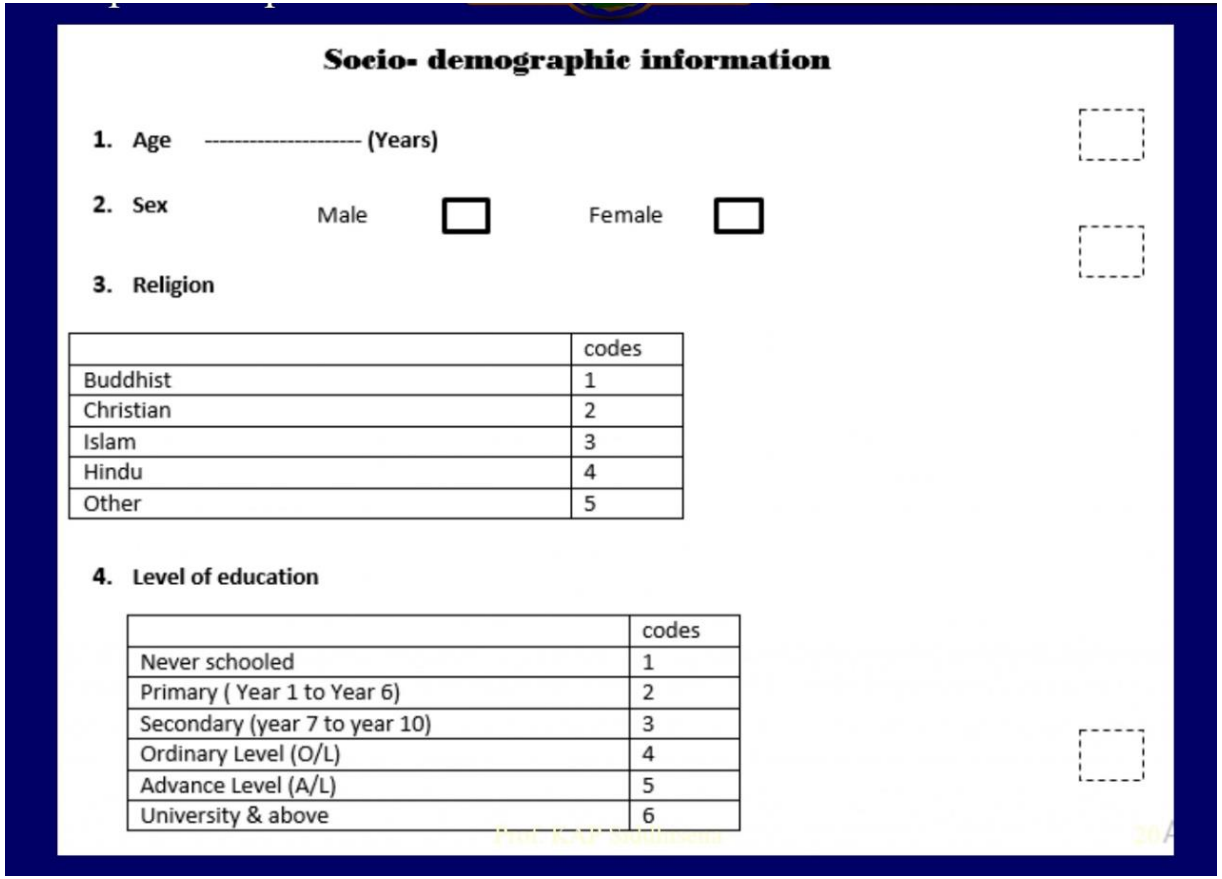
b. Questions related to **subject matter**

Two Types of Questions (Question design)

- Questionnaire design is concerned with the type of questions, their working, the reliability and validity of the responses. Because designing questions is a crucial element of many of the data collection method.

- There are two types of questions
    1. Structured Questions (Close-ended Questions)
    2. Unstructured Questions (Open-ended questions)

A sample questionnaire



Socio- demographic information

1. **Age** ---------- **(Years)**

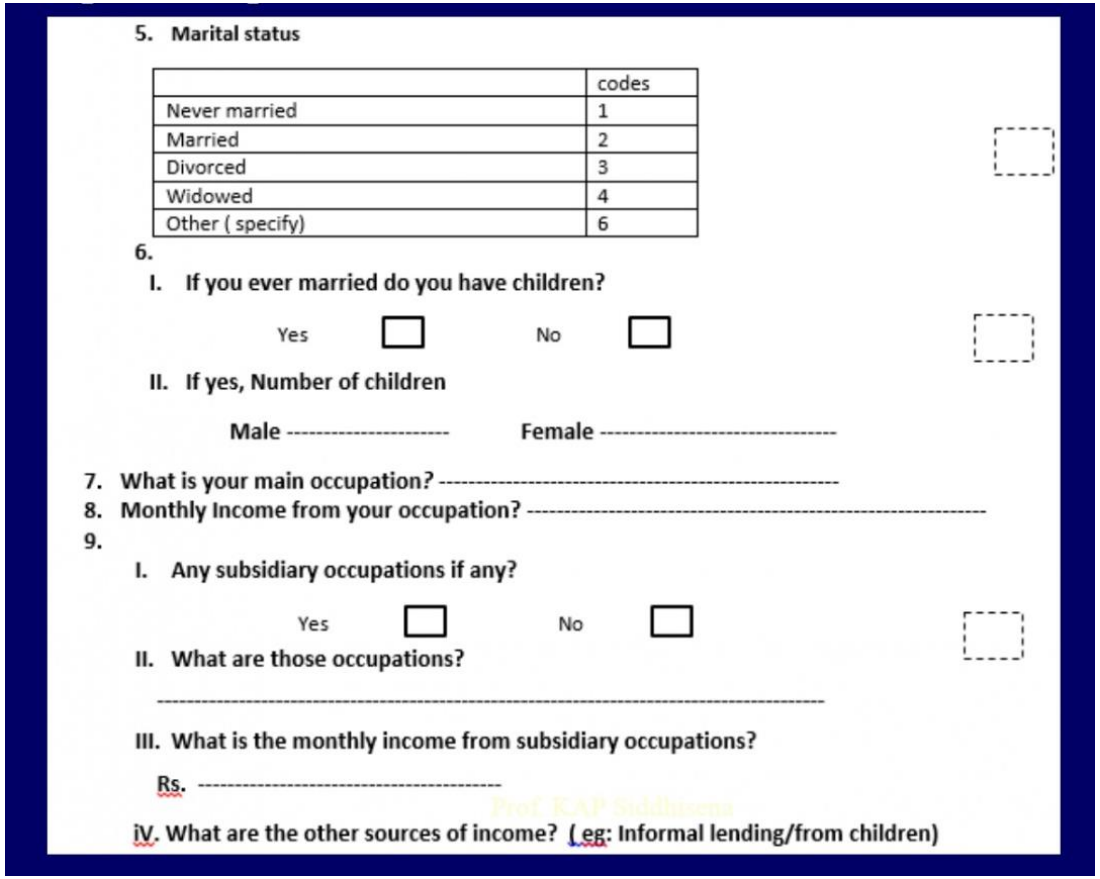
2. **Sex** Male ☐ Female ☐

3. **Religion**

	codes
Buddhist	1
Christian	2
Islam	3
Hindu	4
Other	5

4. **Level of education**

	codes
Never schooled	1
Primary (Year 1 to Year 6)	2
Secondary (year 7 to year 10)	3
Ordinary Level (O/L)	4
Advance Level (A/L)	5
University & above	6



5. **Marital status**

	codes
Never married	1
Married	2
Divorced	3
Widowed	4
Other (specify)	6

6.

I. **If you ever married do you have children?**

Yes ☐ No ☐

II. **If yes, Number of children**

Male ---------- **Female** ----------

7. **What is your main occupation?** ----------
8. **Monthly Income from your occupation?** ----------
9.

I. **Any subsidiary occupations if any?**

Yes ☐ No ☐

II. **What are those occupations?**

III. **What is the monthly income from subsidiary occupations?**

Rs. ----------

iV. **What are the other sources of income? (eg: Informal lending/from children)**

Interview

Interview Method

- The term interview is constructed out of two words ‘inter + view’. View means to look at to have a perception of and ‘inter’ means between two objects or points.
- Therefore the interview refers to the act of perceiving as conducted between two separate points or two separate persons.

- There are several types of interview:

- 1. Face to face interview
- one to one interview
- one to many
- many to one
- many to many interview

- 2. voice to voice interview
- 3. Screen to Screen interview
- **Another classification:**
  1. Structured or standardized interview
  2. Unstructured or non-standardized interview
  3. Semi-structured interview
  4. In-depth interview

Observation

- This method entails sending observer to record what actually happens whilst it is happening (e.g. Traffic census).
- This is a self reporting and self administrative method of data collection.
- Observation can take place in a laboratory setting or in a natural setting. A natural setting is a "research environment that would have existed had researchers never studied it (vogt, 1993).

- Non-Participant Observation

Non-participant observation is to observe and record what people do in terms of their actions and their behaviour without the researcher being involved.

- The observer is separate from the activities taking place and the subject of the research may or may not be aware that they are being observed.
- Number of ways to do non-participant observation: looking or observing, using video or a still camera.

❖ **Participant Observation**

❖ *Is a method of collecting data where the researcher is fully involved with the participants and the phenomena being researched.*

❖ *An important method adopted in anthropological research- heavily used in relation to ethnographic studies.*

Experimental study

- An experimental study is a standard method of data collection that involves the manipulation of the samples by applying some form of treatment prior to data collection

Measurement	Treated (Scholarship)	Control (No scholarship)
Attendance	182 days (202)	151 days (202)
Test score average	78.3	54.7
Extra curriculum activities	6	2

Other methods

FOCUS GROUPS

- What are focus groups?
  - Focus groups are group discussions that gathers together people from similar backgrounds and experiences to discuss a specific topic of interest to the researcher. The discussion is guided by a group leader (called a moderator) who asks questions and tries to help the group to have a natural and free conversation with each other.
  - Focus groups are aimed at encouraging participants to talk with each other, rather than answer questions back to the moderator.
  - It also focused because the participants usually share a common characteristic. This may be age, sex, educational background, religion, or something directly related to the topic.
  - The group interaction of focus groups is important because it gives us some understanding of how the people are thinking about the topic.

Other methods

Case studies

- A particular individual or a community or an object or an event or a specific action, which is called a case, is an element in the data collection.
- Case study is a full description of the above selected case; eg. women who had sexual harassments in the middle-east.
- A small number of cases however cannot be easily generalized to a larger population.
- Drawing boundaries as to what include in a case study and what to exclude remains an issue; the major guide line here is to remember what is it a case of

Semiotics/Semeotics

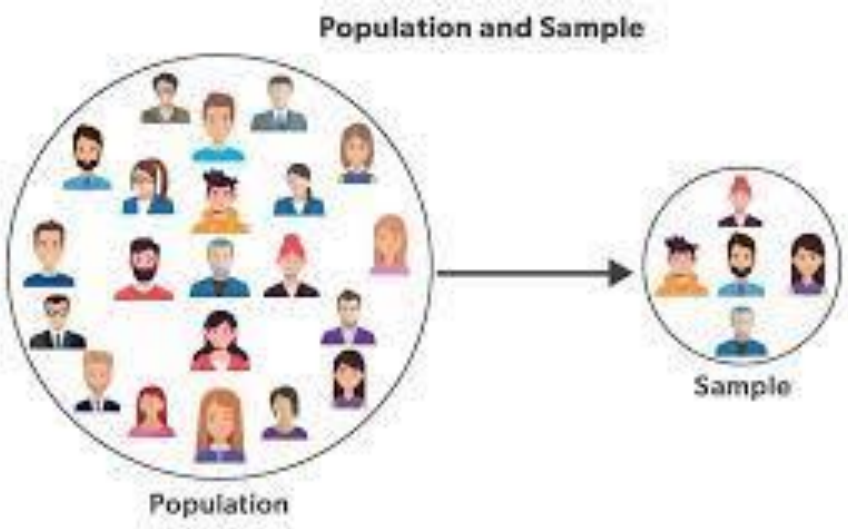
- Semiotics/Semeotics are the study of signs and symbols and their use or interpretation. In other words, Semiotics can include signs, logos, gestures and other linguistic and nonlinguistic communication methods. As a word, semiotics derives from the Greek sēmeiōtikós, which describes the action of interpreting signs. Eg. **pedestrian** signs, traffic signs, toilets identifications etc.
- e.g. pedestrian crossing



- Anthropologists, sociologists and ethnograpgers are using semiotics for their anthropological, historical, archiological studies.

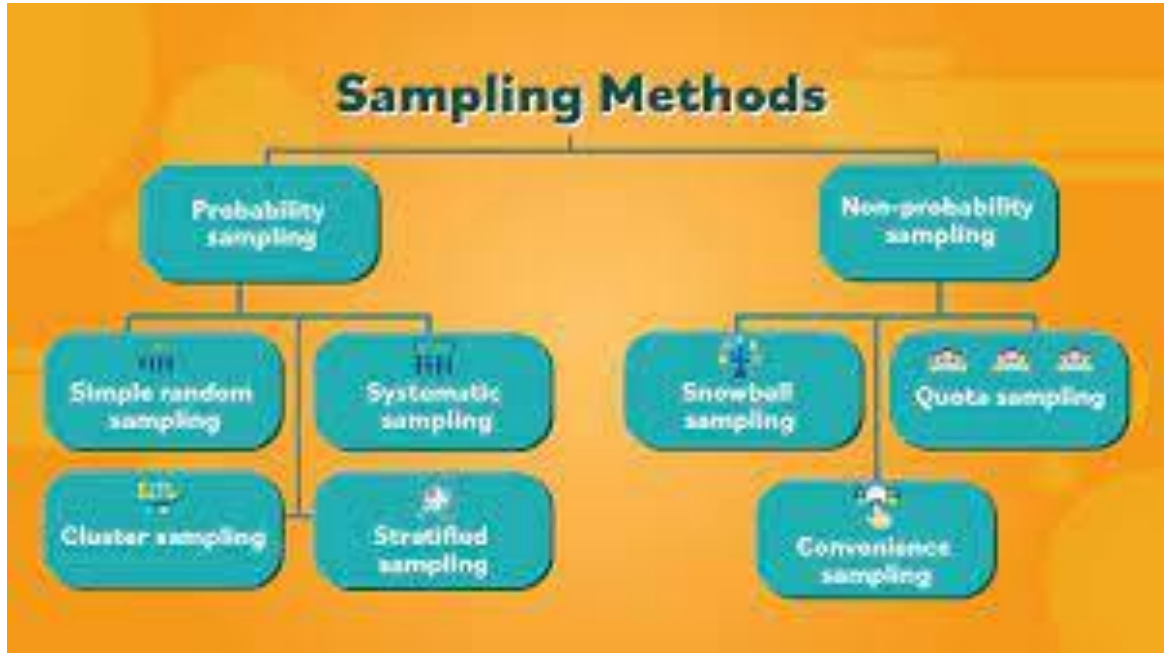
Terms used in data collection and analysis

- A population is the entire group that you want to draw conclusions about (that interest an investigator). N can be very large or even infinite
- A sample is an observed subset of population values with sample size given by n (the specific group that you will collect data from)



Terms used in data collection and analysis

- Simple random sampling. With simple random sampling, every element in the population has an equal chance of being selected as part of the sample. ...
- Systematic sampling
- Stratified sampling
- Cluster sampling
- Convenience sampling
- Quota sampling
- Convenience sampling
- Snowball or referral sampling



Determining the sample size

- An important step in planning a statistical study.
- Part of designing a high-quality study.
- A study must be of adequate size, relatives to the goal of the study.
- An under-sized study is a waste of resources for not having the capability to produce useful result.
- An over-sized study uses more resources than necessary.
- Different study designs need different methods of sample size calculation and one formula cannot be used in all designs

Determining the sample size

CALCULATING SAMPLE SIZE

$$\text{SAMPLE SIZE} = \frac{\frac{z^2 \times p\,(1-p)}{e^2}}{1 + \left(\frac{z^2 \times p\,(1-p)}{e^2\,N}\right)}$$

$$n_0 = \frac{z^2 \cdot p \cdot (1 - p)}{e^2}$$

WHERE,
N = POPULATION SIZE
p = PERCENTAGE/PROPORTION PICKING A CHOICE
e = SAMPLING ERROR (SE) / MARGIN OF ERROR (ME)
z = Z-SCORE (THE NUMBER OF STANDARD DEVIATIONS A GIVEN PROPORTION IS AWAY FROM THE MEAN)

Desired Confidence Level	Z-score
80%	1.28
85%	1.44
90%	1.65
95%	1.96
99%	2.58

MUST REMEMBER:

- SMALLER MARGIN OF ERROR MEANS LARGER SAMPLE SIZE
- HIGHER CONFIDENCE LEVEL MEANS LARGER SAMPLE SIZE

Determining the sample size

	Confidence level = 95%			Confidence level = 99%		
	Margin of error			Margin of error		
Population size	5%	2,5%	1%	5%	2,5%	1%
100	80	94	99	87	96	99
500	217	377	475	285	421	485
1.000	278	606	906	399	727	943
10.000	370	1.332	4.899	622	2.098	6.239
100.000	383	1.513	8.762	659	2.585	14.227
500.000	384	1.532	9.423	663	2.640	16.055
1.000.000	384	1.534	9.512	663	2.647	16.317



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993

Terms used in data collection and analysis

- **Unit of Analysis (UoA): (Study Unit) -** The major entity that is being analyzed in a study. It can be defined as individuals (most common), groups, social organizations and social artifacts.

- **Unit of Observation (UoO):** Individual, who is observed to collect data to make conclusions regarding the Unit of Analysis. Most of the times these two are coincide.

- **Variable -** is the characteristic of the individual to be measured or observed.



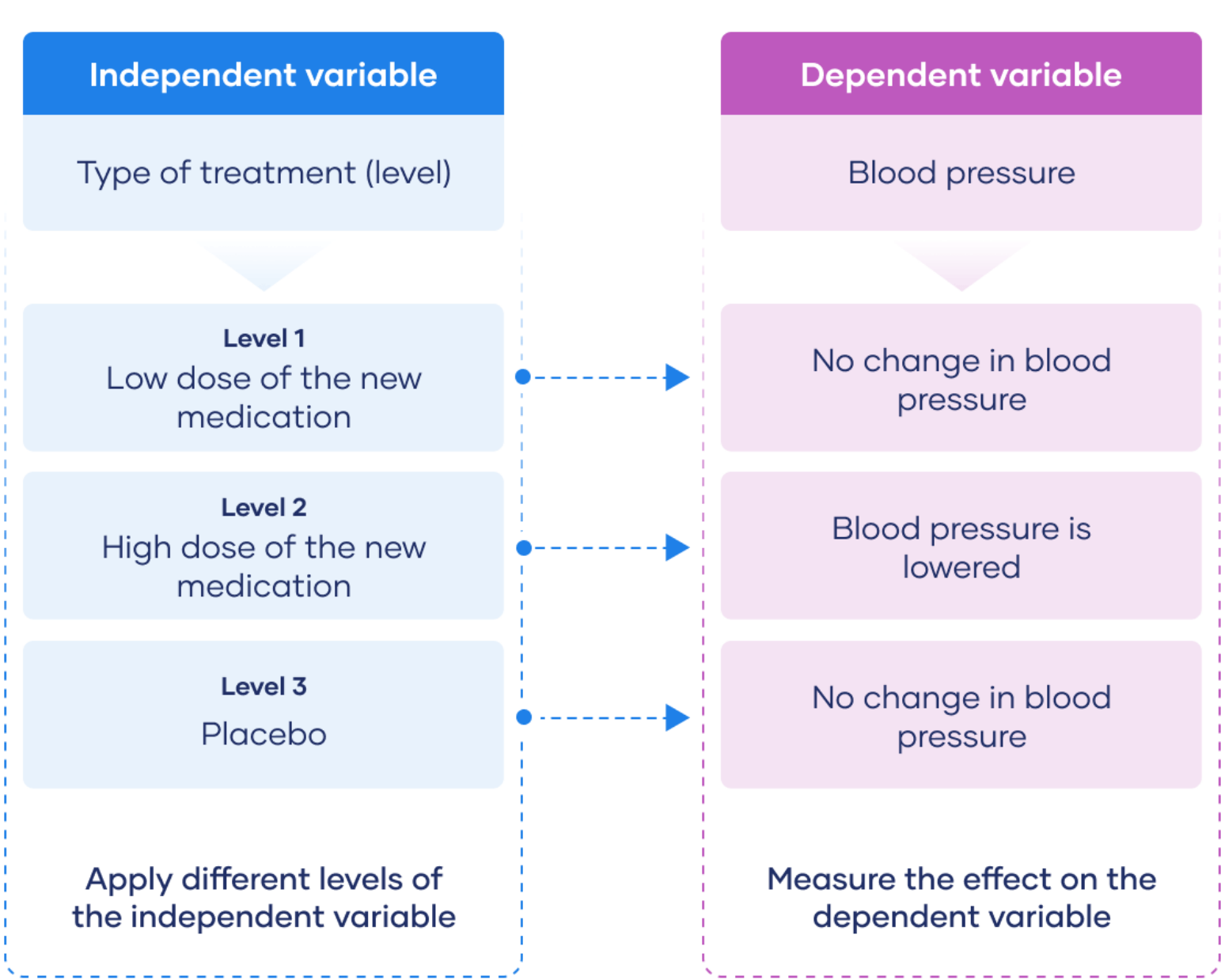
Terms used in data collection and analysis

Parameter and Statistic

- A parameter is a specific characteristic of a population
- A statistic is a specific characteristic of a sample

Independent and dependent and controlled variables

- The independent variable is the one that is changed by the scientist.
- To insure a fair test, a good experiment has only ONE independent variable


Independent variable
Type of treatment (level)
Level 1
Low dose of the new medication
Level 2
High dose of the new medication
Level 3
Placebo
Apply different levels of the independent variable
Dependent variable
Blood pressure
No change in blood pressure
Blood pressure is lowered
No change in blood pressure
Measure the effect on the dependent variable

Terms used in data collection and analysis

Mediator

- A mediator variable explains the relationship between two other variables

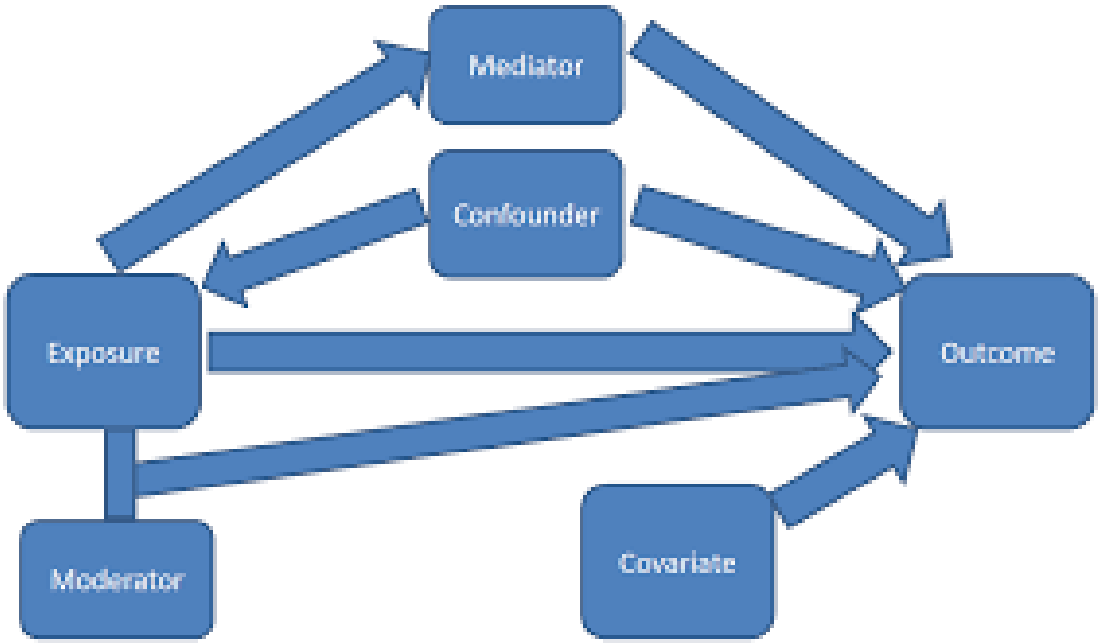
Moderator

- A moderator variable affects the direction or strength of the relationship between two other variables


Mediator
SHOWS CONNECTION
Independent variable
SLEEP QUALITY
Dependent variable
WORK
Moderator
CHANGING STRENGTH
AND DIRECTION

A

B


Mediator
Confounder
Exposure
Outcome
Moderator
Covariate

Terms used in data collection and analysis

Reliability

- Reliability refers to how consistently a method measures something. If the same result can be consistently achieved by using the same methods under the same circumstances, the measurement is considered reliable
- You measure the temperature of a liquid sample several times under identical conditions. The thermometer displays the same temperature every time, so the results are reliable

Tests to check for reliability

- Cronbach Alpha --- internal consistency (across items)
- Pearson correlation coefficient ---test–retest reliability (across time)

Terms used in data collection and analysis

Validity

- Validity refers to how accurately a method measures what it is intended to measure. If research has high validity, that means it produces results that correspond to real properties, characteristics, and variations in the physical or social world
- High reliability is one indicator that a measurement is valid. If a method is not reliable, it probably isn’t valid
- Validity is a judgment based on various types of evidence
- Ex: Validity can be estimated by comparing research results to other relevant data or theories

Terms used in data collection and analysis

Data cleaning

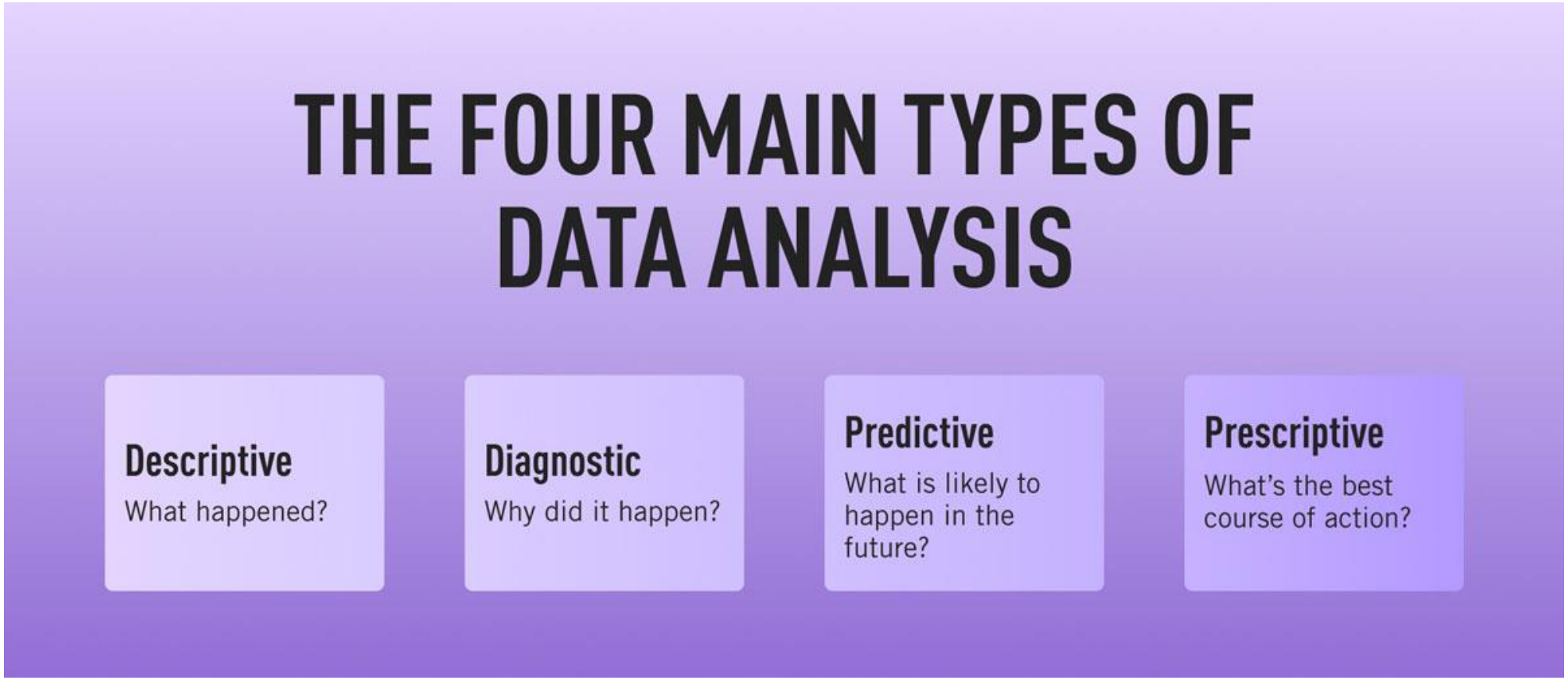
- Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset

Data editing

- Data editing is the application of checks to detect missing, invalid or inconsistent entries or to point to data records that are potentially in error

Data Analysis

- Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data



Data Analysis

Quantitative Data Analysis Methods



Descriptive Analysis

The first level of analysis, this helps researchers find absolute numbers to summarize individual variables and find patterns.

A few examples are...

- **Mean:** numerical average
- **Median:** midpoint
- **Mode:** most common value
- **Percentage:** ratio as a fraction of 100
- **Frequency:** number of occurrences
- **Range:** highest and lowest values



Inferential Analysis

These complex analyses show the relationships between multiple variables to generalize results and make predictions.

A few examples are...

- **Correlation:** describes the relationship between 2 variables
- **Regression:** shows or predicts the relationship between 2 variables
- **Analysis of variance:** tests the extent to which 2+ groups differ

Data analysis

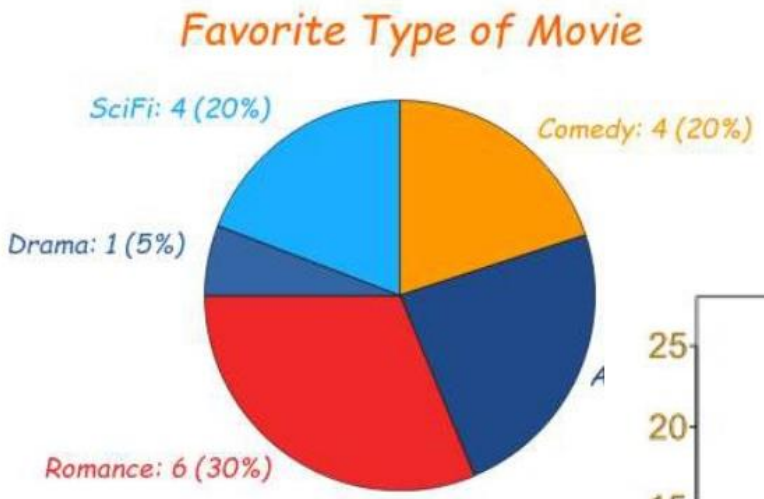
Qualitative Approach

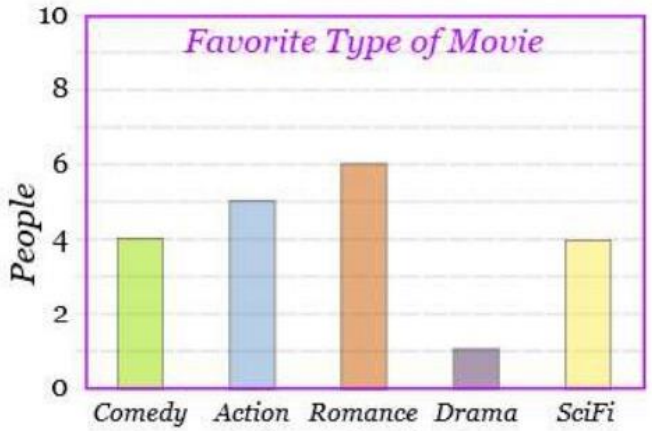
- Qualitative Data Analysis: The qualitative data analysis method derives data via words, symbols, pictures, and observations.
  - This method doesn't use statistics. The most common qualitative methods include:
    - Content Analysis, for analyzing behavioral and verbal data.
    - Thematic Analysis, entails searching across a data set to identify, analyze, and report repeated patterns.
    - Narrative Analysis, for working with data culled from interviews, diaries, surveys.
    - Grounded Theory, for developing causal explanations of a given event by studying and extrapolating from one or more past cases.

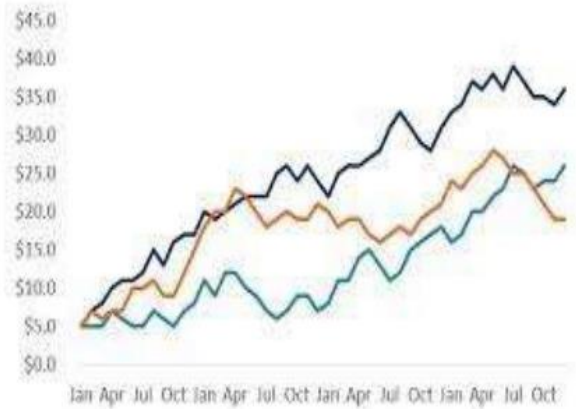


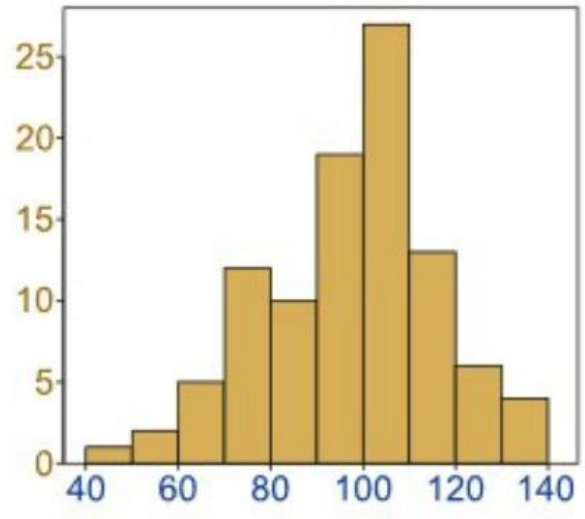
Descriptive statistics

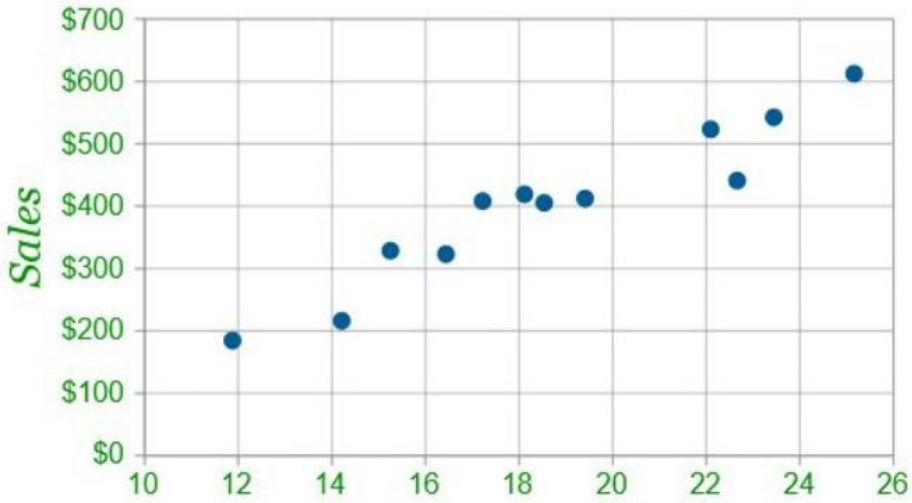
- Graphical methods











Descriptive statistics

Numerical Presentation Tables

Month	Revenue
JAN	$28,361
FEB	$14,744
MAR	$19,407
APR	$15,891
MAY	$21,277
JUN	$21,530
JUL	$17,990
AUG	$21,838
SEP	$20,174
OCT	$20,025
NOV	$48,055
DEC	$24,318

	Baseball	Basketball	Football	Total
Male	13	15	20	48
Female	23	16	13	52
Total	36	31	33	100

	Baseball	Basketball	Football	Total
Male	36.11	48.39	60.61	36.11
Female	63.89	51.61	39.39	63.89
Total	100	100	100	100

	Baseball	Basketball	Football	Total
Male	27.08	31.25	41.67	100.00
Female	44.23	30.77	25.00	100.00
Total	36	31	33	100

Descriptive statistics

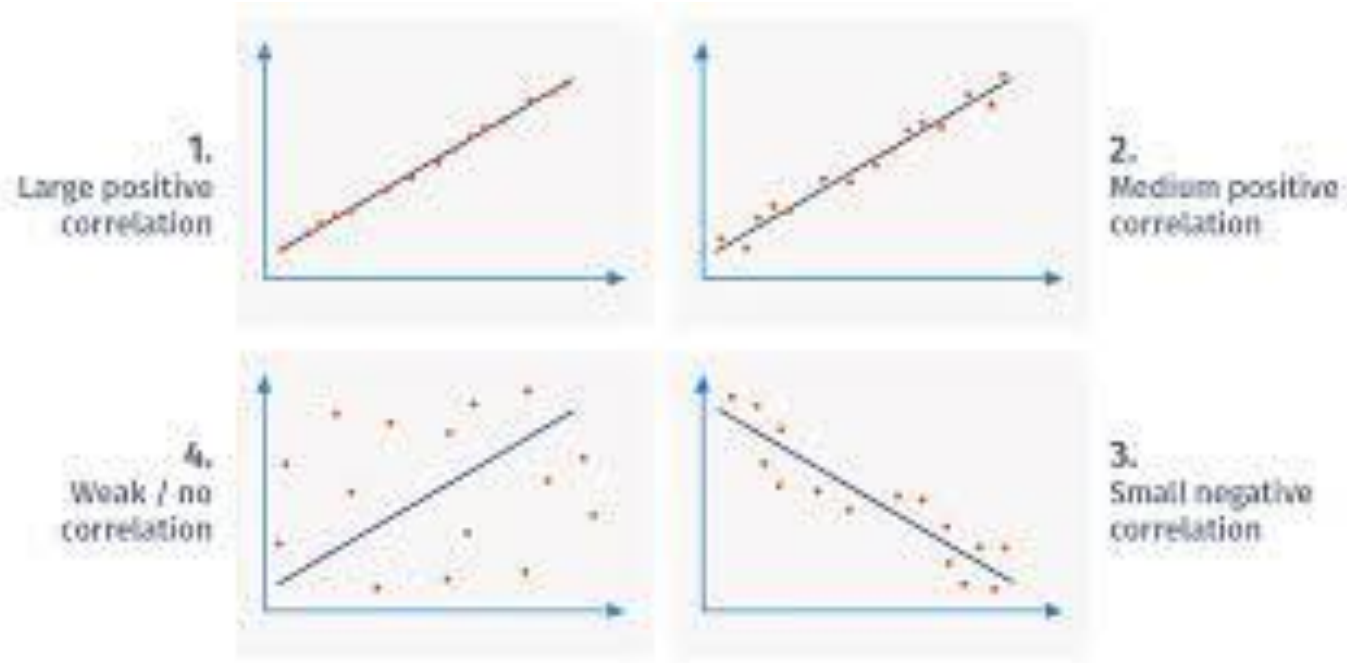
Method used in academic writing

Descriptive Statistics

Variable	Obs	Mean	Std.Dev.	Min	Max
price	74	6165.257	2949.496	3291	15906
mpg	74	21.297	5.786	12	41
rep78	69	3.406	.99	1	5
headroom	74	2.993	.846	.846	5
trunk	74	13.757	4.277	5	23
weight	74	3019.459	777.194	1760	4840
length	74	187.932	22.266	142	
turn	74	39.649	4.399	31	
displacement	74	197.297	91.837	79	
gear_ratio	74	3.015	.456	2.19	
foreign	74	.297	.46	0	1

Inferential Analysis

- Correlation analysis in research is a statistical method used to measure the strength of the linear relationship between two variables and compute their association



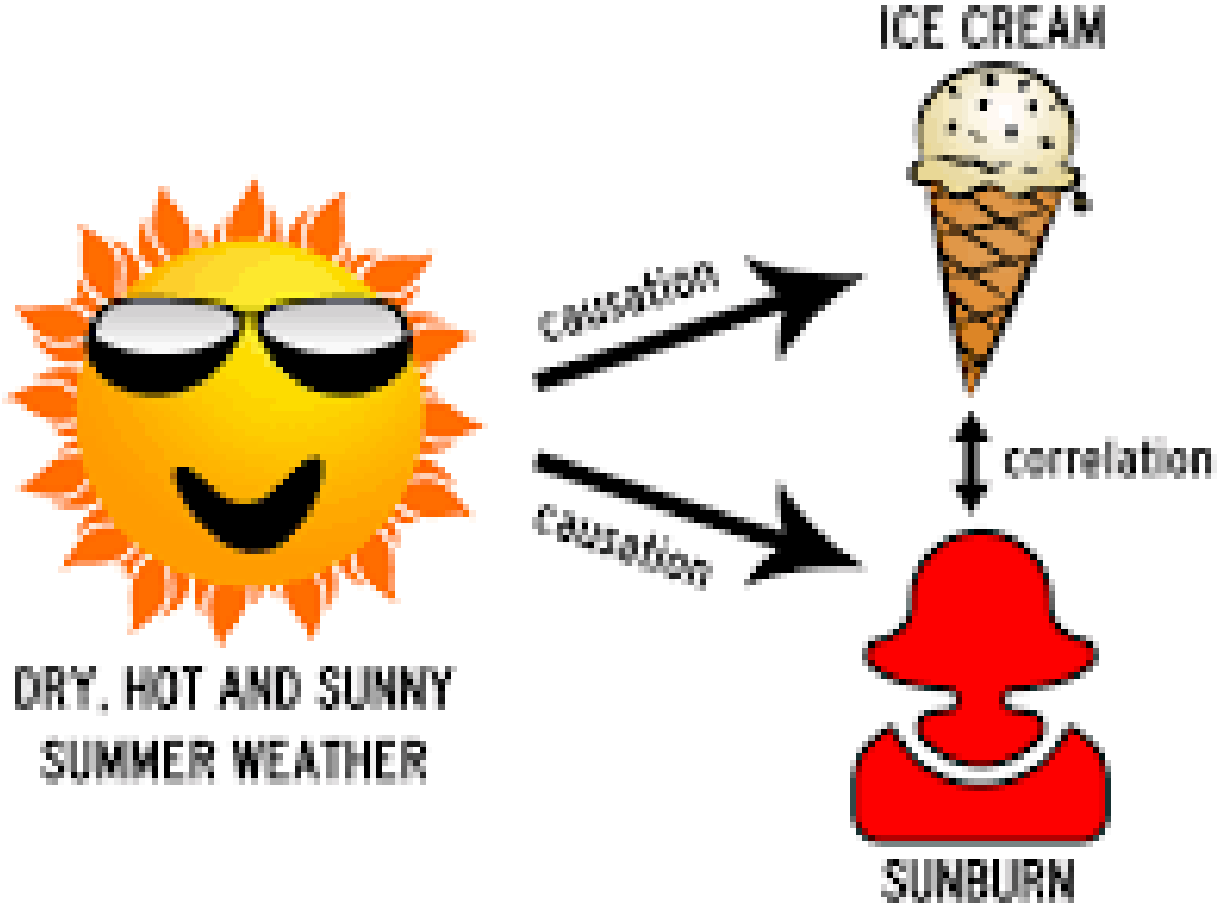
Inferential analysis



Doing regression analysis in MS-Excel



Correlation vs Causation



Advanced Identification Strategies

- Randomized control Trial
- Randomized encouragement design
- Regression discontinuity design
- Instrumental variable approach
- Different in differences
- Fixed effect estimation
- Propensity score matching
- Causal Mediation Analysis

Thank you