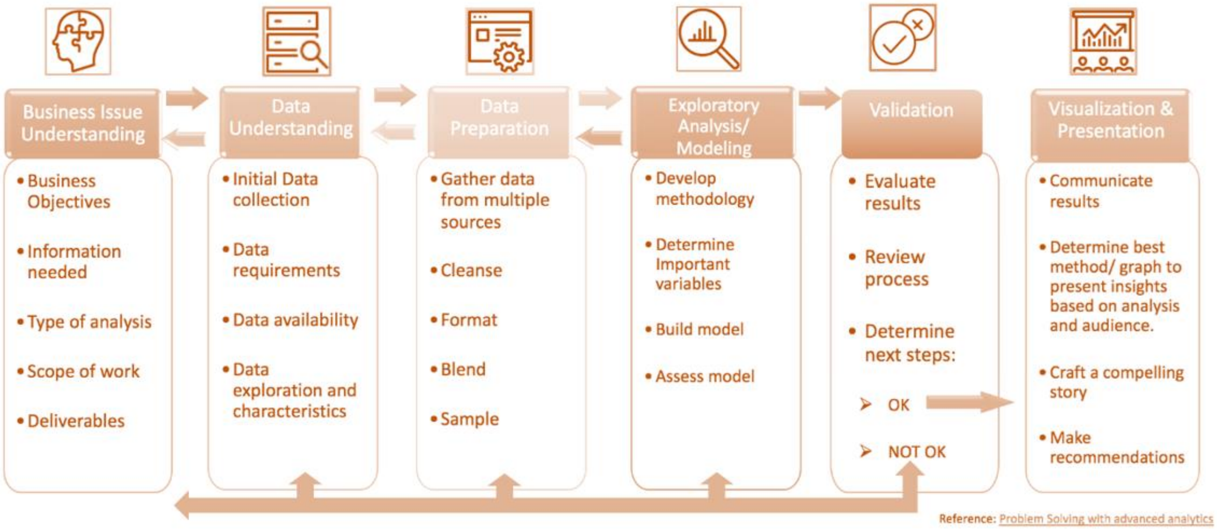
A Brief Introduction to Data Generation/Collection Processes

Contents

1. Overview

1.1. Data Analysis Project Life Cycle



Business Issue Understanding
- Business Objectives
- Information needed
- Type of analysis
- Scope of work
- Deliverables

Data Understanding
- Initial Data collection
- Data requirements
- Data availability
- Data exploration and characteristics

Data Preparation
- Gather data from multiple sources
- Cleanse
- Format
- Blend
- Sample

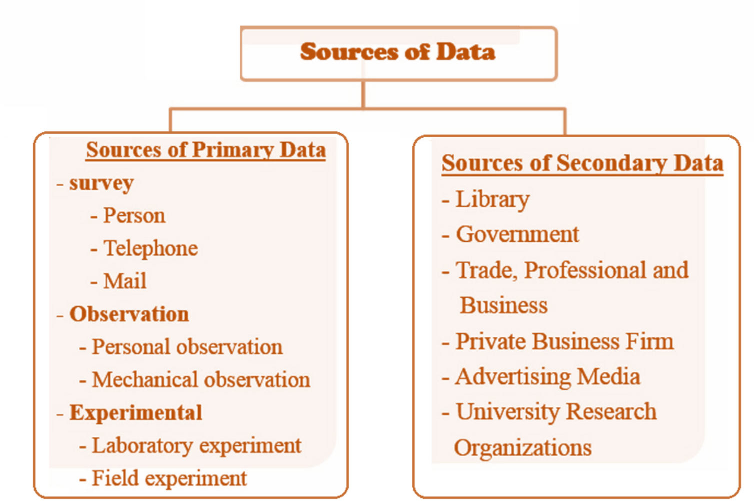
Exploratory Analysis/ Modeling
- Develop methodology
- Determine Important variables
- Build model
- Assess model

Validation
- Evaluate results
- Review process
- Determine next steps:
  - OK
  - NOT OK

Visualization & Presentation
- Communicate results
- Determine best method/ graph to present insights based on analysis and audience.
- Craft a compelling story
- Make recommendations

Reference: Problem Solving with advanced analytics

1.2. Data Sources



Sources of Data

Sources of Primary Data
- **survey**
  - Person
  - Telephone
  - Mail
- **Observation**
  - Personal observation
  - Mechanical observation
- **Experimental**
  - Laboratory experiment
  - Field experiment

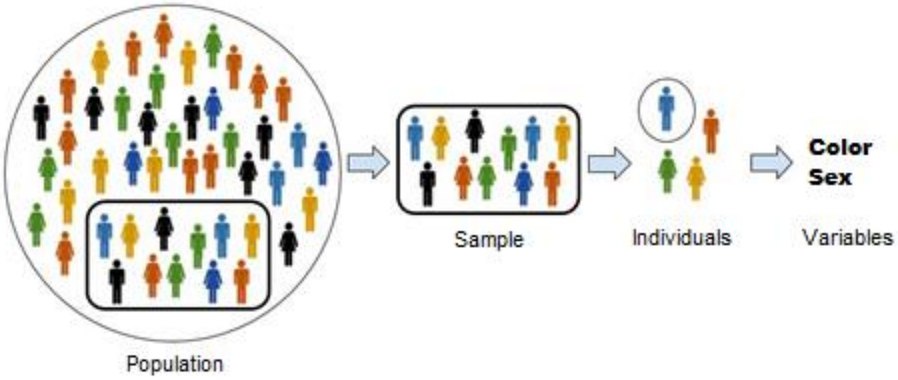
Sources of Secondary Data
- Library
- Government
- Trade, Professional and Business
- Private Business Firm
- Advertising Media
- University Research Organizations

2. Sampling Plans

A sampling plan is a statistical procedure that outlines the steps for collection a sunset of representatives of the population of study.

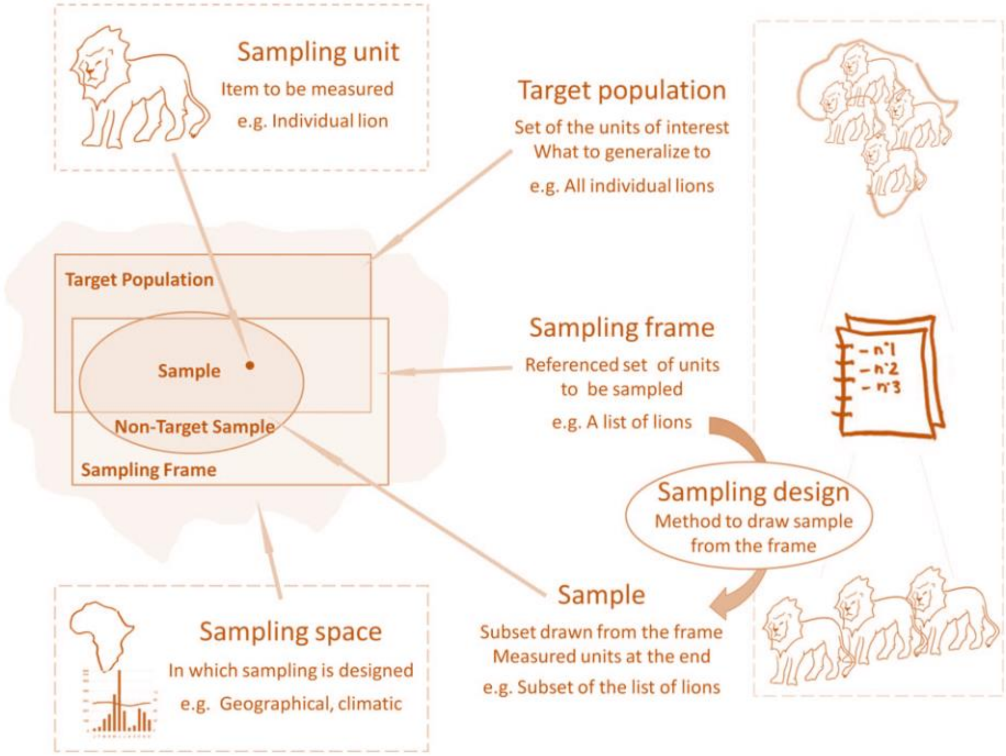
2.1. What is Sampling

Sampling is a process used to obtain a sample (subset) from the population of study.



2.2. Vocabulary of Sampling

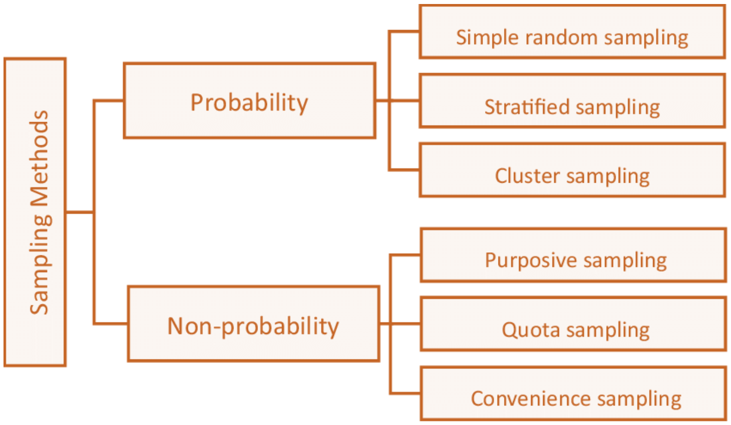
The next figure explains some of the important terms used in sampling.



2.3. Why Sampling?

- Less cost is studying a subset of population, in particular, the experiment is destructive.
- Less human error in recording
- Less time-consuming

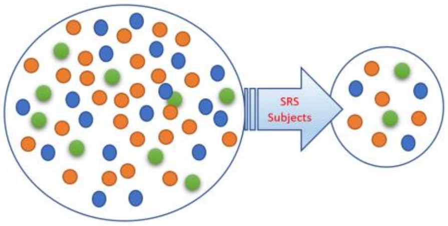
2.4. Methods Plans



3. Probability Sampling Plans

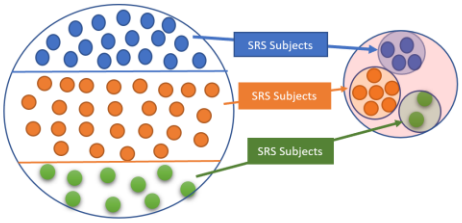
3.1. Simple Random Sample

Simple random sampling: A sample is chosen randomly from the population. Every object in the population has an equal chance of being chosen as part of the sample.



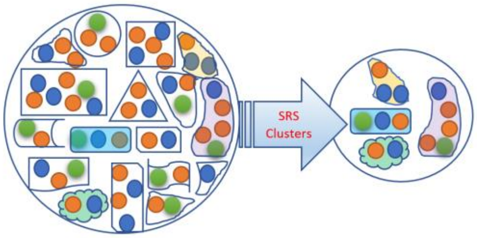
3.2. Stratified Sampling

Stratified Sampling takes SRS samples from individual strata and combined them to obtain a stratified random sample.



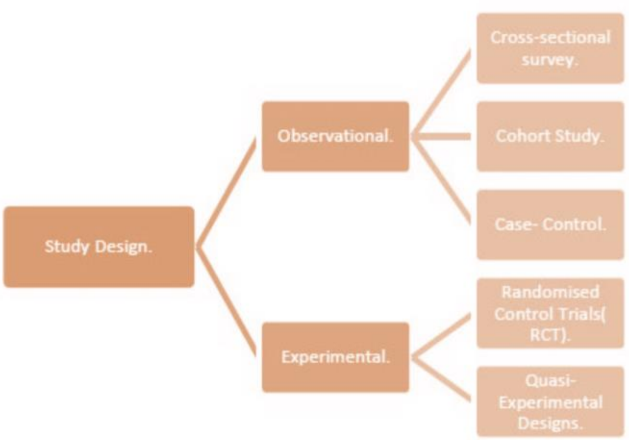
3.3. Cluster Sampling

Cluster sampling takes SRS based on clusters and combine the sampled clusters to obtain a random sample based on clusters. In this sampling plan, the sampling units are the clusters.



4. Study Designs

There are two major study designs.

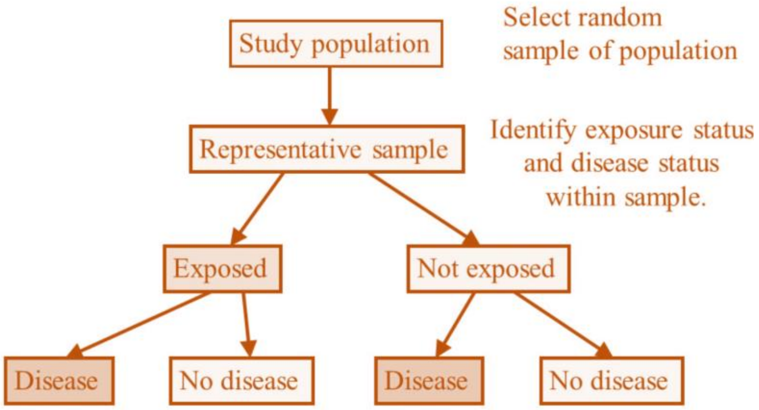


4.1. Observational Studies

There three major observational study designs.

4.1.1. Cross-sectional Study

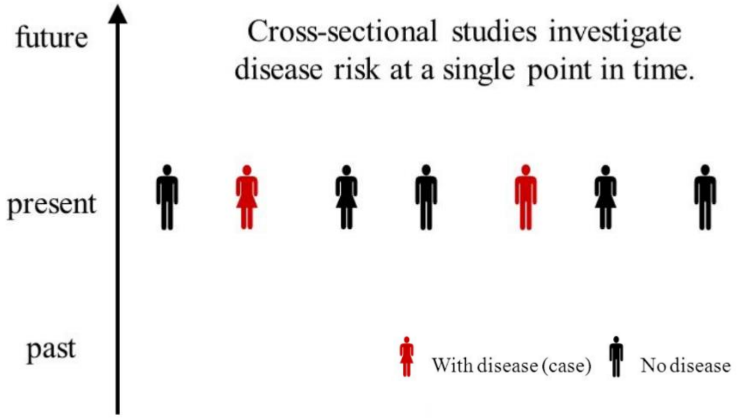
Cross-sectional study: involves data collection from a population, or a representative subset, at one specific point in time.


Study population
Select random sample of population
Representative sample
Identify exposure status and disease status within sample.
Exposed
Not exposed
Disease
No disease
Disease
No disease

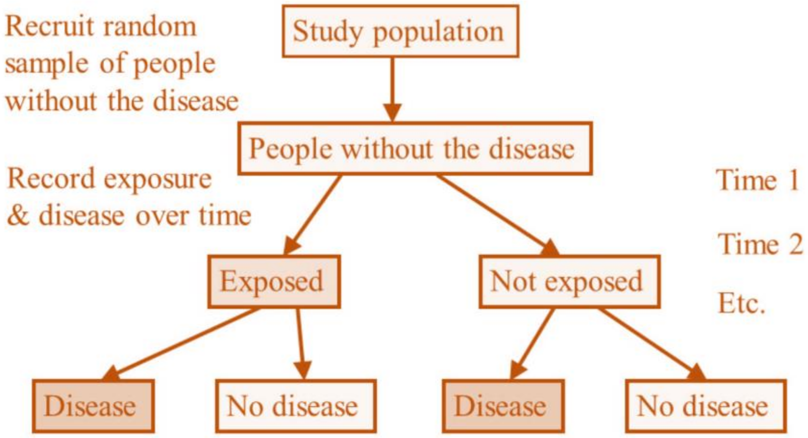
Cross-sectional studies are simple in design and are aimed at finding out the prevalence of a phenomenon, problem, attitude or issue by taking a snapshot or cross-section of the population. This obtains an overall picture as it stands at the time of the study.

For example, a cross-sectional design would be used to assess demographic characteristics or community attitudes. These studies usually involve one contact with the study population and are relatively cheap to undertake.

Pre-test/post-test studies measure the change in a situation, phenomenon, problem, or attitude. Such studies are often used to measure the efficacy of a program. These studies can be considered as a variation of the cross-sectional design as they involve two sets of cross-sectional data collection on the same population to determine if a change has occurred.


future
Cross-sectional studies investigate disease risk at a single point in time.
present
past
With disease (case)
No disease

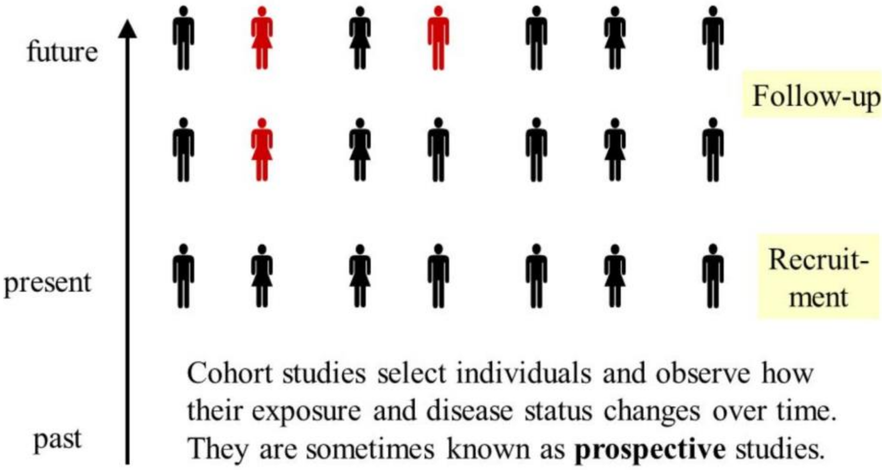
4.1.2. Prospective Study (Cohort Study, longitudinal Study)


Recruit random sample of people without the disease
Study population
People without the disease
Record exposure & disease over time
Time 1
Time 2
Etc.
Exposed
Not exposed
Disease
No disease
Disease
No disease

Prospective studies seek to estimate the likelihood of an event or problem in the future. Thus, these studies attempt to predict what the outcome of an event is to be. General science experiments are often classified as prospective studies because the experimenter must wait until the experiment runs its course to examine the effects.

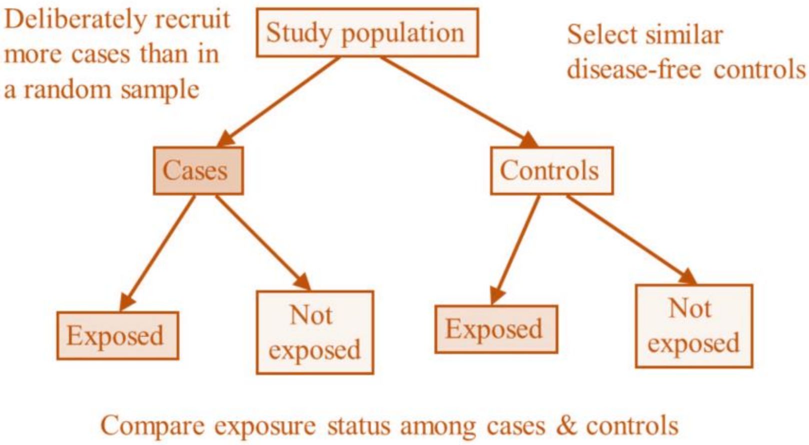
Randomized controlled trials are always prospective studies and often involve following a "cohort" of individuals to determine the relationship between various variables.

Longitudinal studies follow study subjects over a long period of time with repeated data collection throughout. Some longitudinal studies last several months, while others can last decades. Most are observational studies that seek to identify a correlation among various factors. Thus, longitudinal studies do not manipulate variables and are not often able to detect causal relationships.

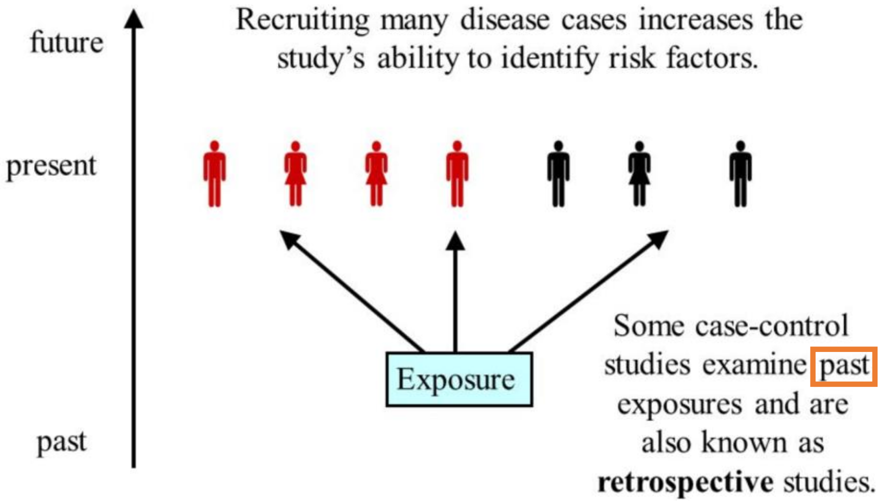


Cohort studies select individuals and observe how their exposure and disease status changes over time. They are sometimes known as **prospective** studies.

4.1.3. Retrospective Study (Case-control Study)


Deliberately recruit more cases than in a random sample
Study population
Select similar disease-free controls
Cases
Controls
Exposed
Not exposed
Exposed
Not exposed
Compare exposure status among cases & controls

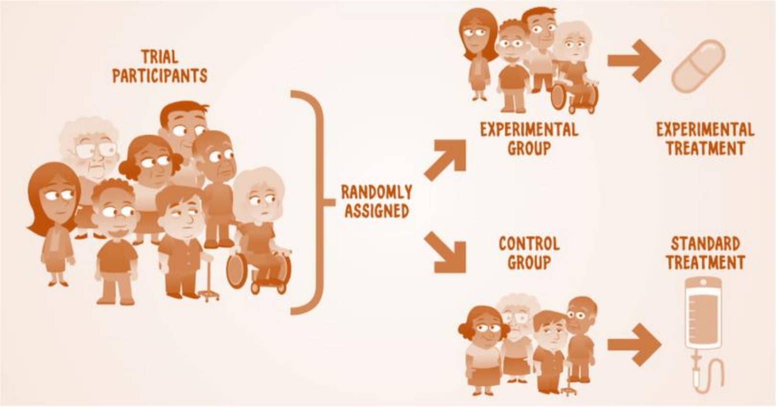
Retrospective studies investigate a phenomenon or issue that has occurred in the past. Such studies most often involve secondary data collection, based upon data available from previous studies or databases. For example, a retrospective study would be needed to examine the relationship between levels of unemployment and street crime in NYC over the past 100 years.


future
present
past
Recruiting many disease cases increases the study's ability to identify risk factors.
Exposure
Some case-control studies examine past exposures and are also known as **retrospective** studies.

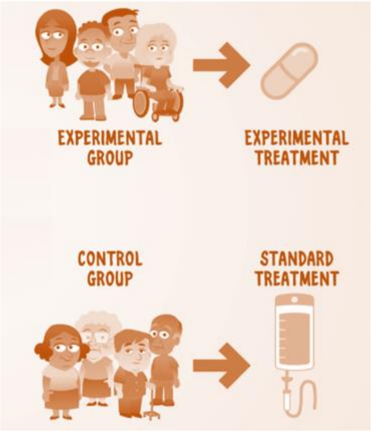
4.2. Experimental Studies

We list three experimental study designs here. The randomized clinical trial is the most commonly used in clinical study.

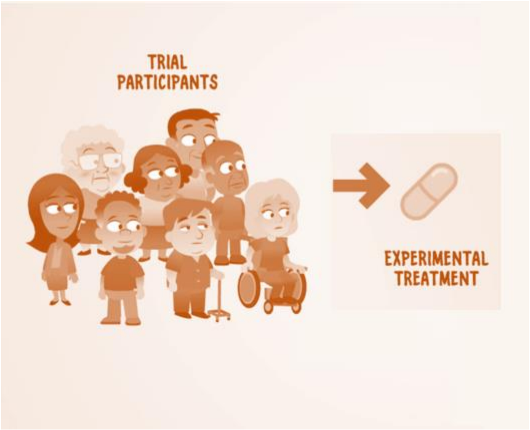
4.2.1. Randomized Clinical Trials (CRT)



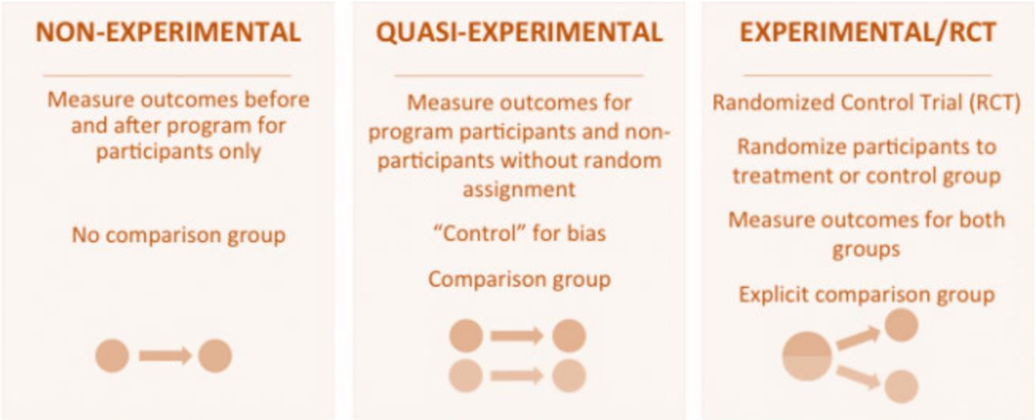
4.2.2. Quasi-experimental Design (no-random assignment!)



4.2.3. Non-experiment design (no controls, i.e., no-comparison)



4.2.4. Summary



NON-EXPERIMENTAL

Measure outcomes before and after program for participants only

No comparison group

QUASI-EXPERIMENTAL

Measure outcomes for program participants and non-participants without random assignment

"Control" for bias

Comparison group

EXPERIMENTAL/RCT

Randomized Control Trial (RCT)

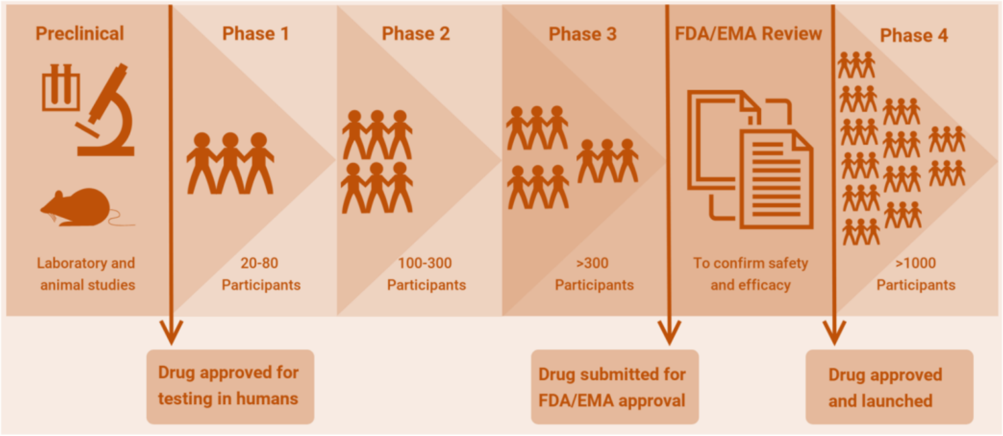
Randomize participants to treatment or control group

Measure outcomes for both groups

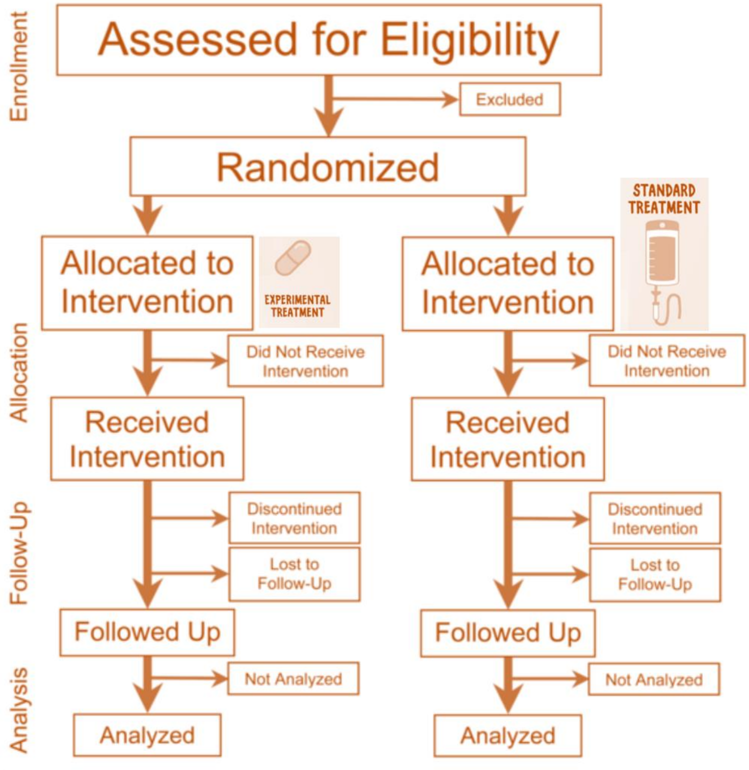
Explicit comparison group

5. Randomized Clinical Trials (RCT)

5.1. Phases of RCT


Preclinical
Laboratory and animal studies
Phase 1
20-80 Participants
Phase 2
100-300 Participants
Phase 3
>300 Participants
FDA/EMA Review
To confirm safety and efficacy
Phase 4
>1000 Participants
Drug approved for testing in humans
Drug submitted for FDA/EMA approval
Drug approved and launched

5.2. Challenges in RCT



6. Data Set Creation

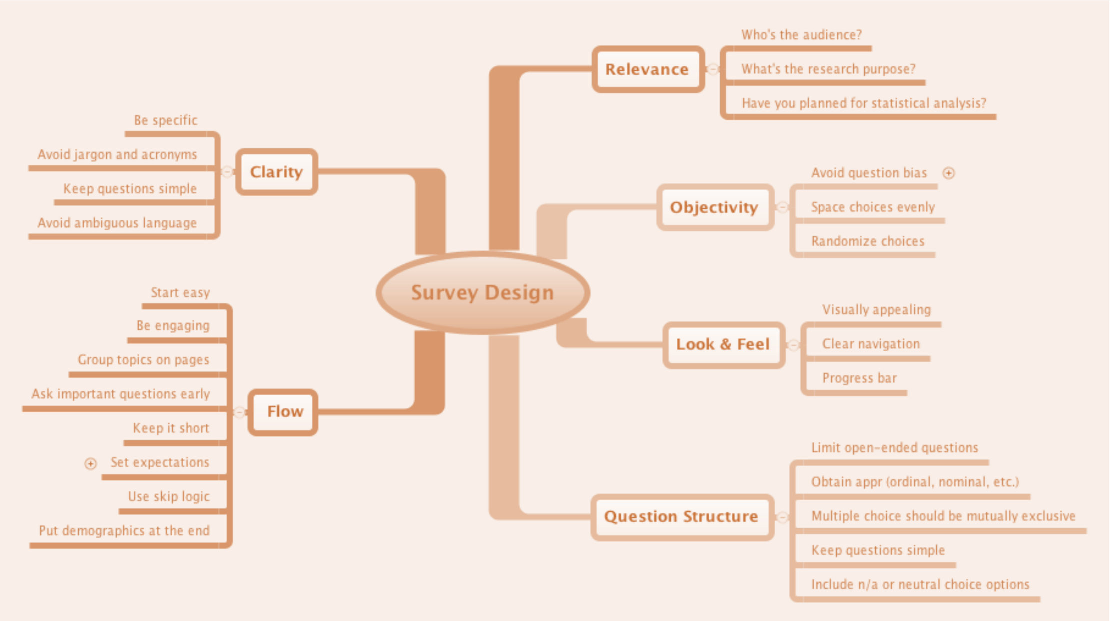
6.1. Data Set Layout

Once a sample of subjects (human subjects or other subjects such as animals, plants, companies, geographic regions, etc.) is selected, we can record the information from each subject.

The layout of a data set has the following form:

	Var_1	Var_2	Var_3	…..	Var_n	Var_(n+1)
Subj 01						
Subj 02						
Subj 03						
Subj 04						
…………						
Subj n						

6.2. Survey Questionnaire Design

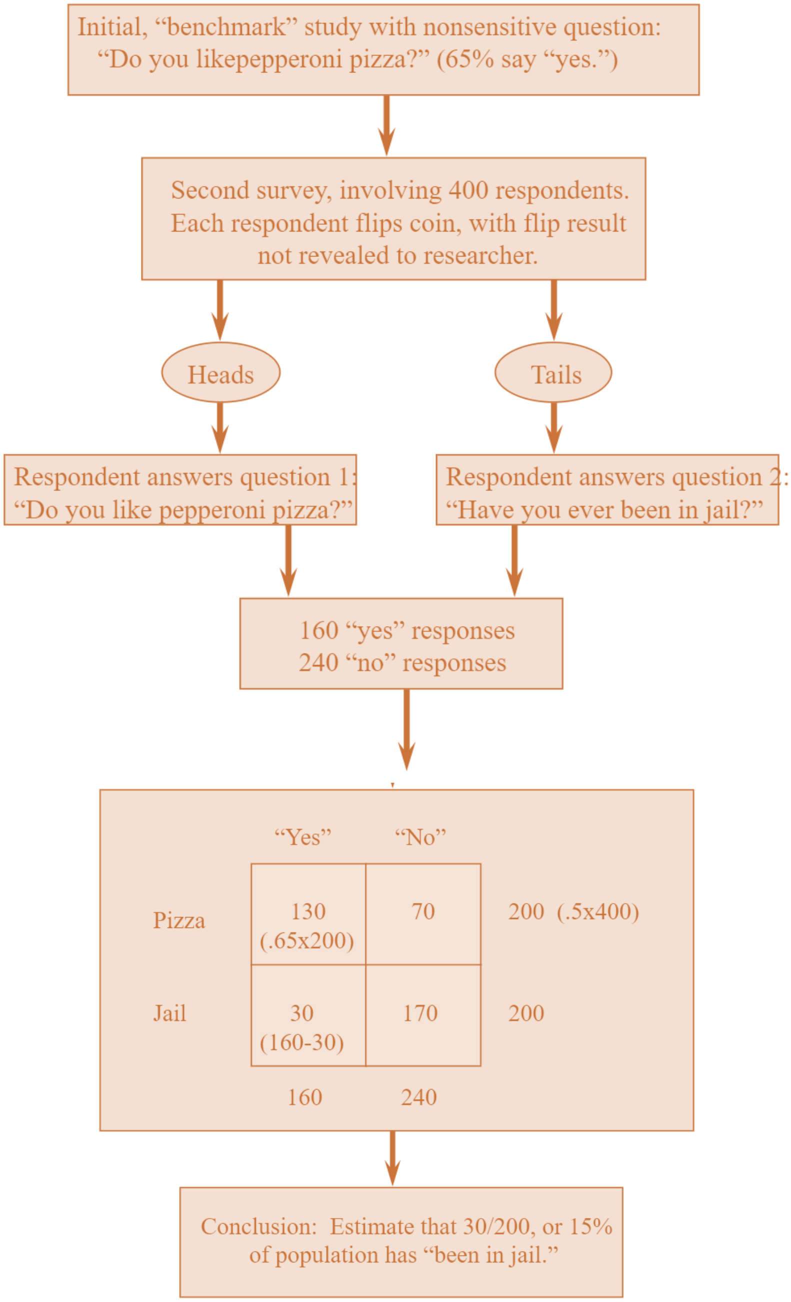


6.3. Handling Sensitive Survey Questions

- Deliberately phrase the question to make it less sensitive to obtain truthful information.

- Randomized response techniques (RRT) – There are different RRTs.

Example 1. ``*Have you ever been in jail?"*



Initial, "benchmark" study with nonsensitive question: "Do you likepepperoni pizza?" (65% say "yes.")

Second survey, involving 400 respondents. Each respondent flips coin, with flip result not revealed to researcher.

Respondent answers question 1: "Do you like pepperoni pizza?"

Respondent answers question 2: "Have you ever been in jail?"

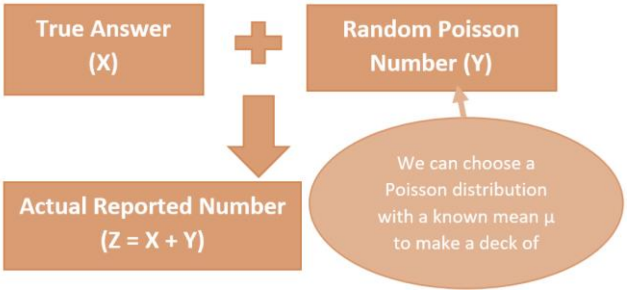
160 "yes" responses
240 "no" responses

	"Yes"	"No"	
Pizza	130 (.65x200)	70	200 (.5x400)
Jail	30 (160-30)	170	200
	160	240	

Conclusion: Estimate that 30/200, or 15% of population has "been in jail."

Example 2. Average number of illegal drugs used by high school students in the past three months.

Question: How many times did you use an illegal drug?


True Answer (X)
Random Poisson Number (Y)
Actual Reported Number (Z = X + Y)
We can choose a Poisson distribution with a known mean μ to make a deck of

Therefore,

$$E[X] = E[Z] - E[Y] \approx \overline{Z} - \mu$$