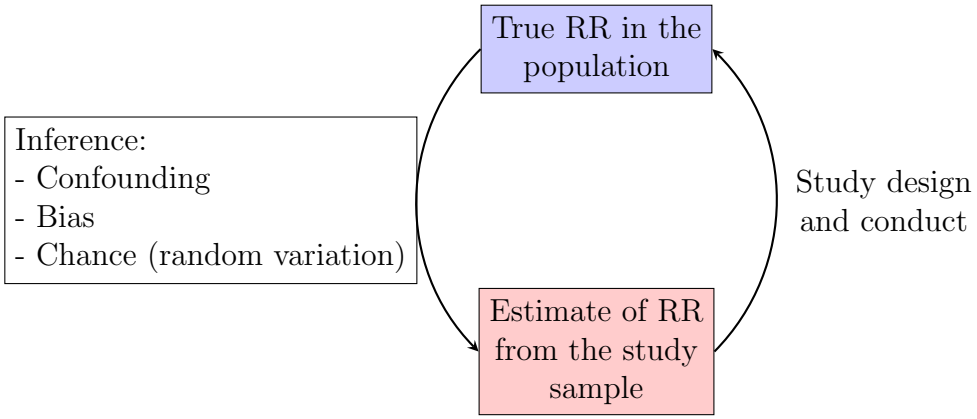
- **Purpose of Analytic Studies**: To test hypotheses (quantify population). For example, do government subsidy programs impact the profitability of fisheries? Does IT investment impact productivity in industry? Does a Mediterranean diet impact life expectancy?
- **Function of Replication**: To allow us to separate out true effects from chance effects.
- **Function of Control**: Provides context for evaluating the effect of interest.
- **Descriptive Studies**: The characteristics of people with a disease (person; place; time); lifestyle patterns in a population; attitudes to health care.
- **Well-defined and Not Well-defined Population**:
  - Well-defined: The collection of words in poems by W. B. Yeats. All patients diagnosed with colorectal cancer in New Zealand in 2015.
  - Not well-defined: The population of New Zealand. Right now? Past? Future?(Time). Target population for a particular cancer treatment. Which type/stage of cancer? Existing or future patients(time)? Over a certain age? On other medications?
- **Traits of Random Sampling**: Known-chance, Equal-chance. If doesn't have those traits, define the sample has representative or not.
- **Sampling Frame Definition**: List of all eligible sampling units from which the sample will be drawn. For example, to draw 200 out of 10,000 employees to form a sample, the roster of 10,000 employees, is the sampling frame. May be sourced from census, company data base or other secondary data. Completeness may be an issue when sourcing Sampling Frame. Sometimes need to use/combine multiple sample frames. Non-probability sampling techniques don't require a sampling frame.
- **Sources of Error for Sample Mean**:

$$\text{Sample mean} = \underbrace{\text{Population (true) mean}}_{} + \underbrace{\text{Error}}_{\underbrace{\text{Systematic error}}_{}\ \ \underbrace{\text{Random error}}_{}}$$

  - **Random Error**: Due to natural variability. Increasing the sample size will reduce the random fluctuations in the sample mean. Statistical methods allow us to quantify the influence of random error on our estimate.
  - **Systematic Error in a Descriptive Study (Bias)**: Due to aspects of the design or conduct of the study which systematically distort the results. Occurs if a sample is not representative of the population (Selection bias). Occurs if the information collected from the sample members is incorrect (Information bias). Cannot be reduced by increasing the sample size.
- **Probability Sampling**: We want our sampling frame to match the population of interest and provide a way to draw a sample. Probability sampling is important because it helps to justify the statistical models. For a finite population of size N draw a sample of size n such that each possible sample has the same probability of being selected.

- **Key Characteristic of Probability Sampling**: The key characteristic is that we know the probability of being selected for everyone in the sample frame.
- **Simplest Form of Probability Sampling**: Simple random sampling.
- **Types of Probability Sampling**: Simple random sampling, stratified random sampling, cluster sampling.
- **Traits of Simple Random Sampling**: Same chance of selection (e.g., Lotto).
- **Advantages of Stratified Sampling**: More precise estimate than for the same sample size from a simple random sample. Can take different sized samples from different strata (a device for reducing overall variability). Useful if you are interested in the results for each stratum and some of the strata are small. Example: colon cancer treatment, samples of colon cancer patients, stratified by ethnicity.
- **Types of Stratified Sampling**: Proportionate stratified sample, disproportionate stratified sample (equal number from each stratum).
- **Cluster Sampling**: The population may be composed of similar and naturally occurring groups. Dividing the population into a group/cluster (then selecting a sample from each cluster).
- **Types of Cluster Sampling**: One-stage, two-stage. Pros and cons for two stage cluster sampling: reduce cost & time, less precise.
- **Experimental Studies**: The researcher manipulates the conditions (intervenes in a natural process) and records the results. The aim is to control all other factors to isolate the effects of the intervention. Best way to study causation. Why randomisation? Randomisation can be used to ensure that effects of unmeasured factors are equalised across the intervention and control groups. Why NOT experimental studies? Ethical problems.
- **Observational Studies**: The investigator does not intervene, simply observes a naturally occurring process, and collects information. The idea is to get as close as possible to the information that would have been obtained if the experimental study could have been done. Cons: We can't know the confounding factors.
- **Case Control Study**: Outcome trace back to reason.
- **Traits of Randomised Controlled Trial (RCT)**:
  - Is considered the "gold standard" analytic study.
  - **Randomisation** - or random allocation, is used to create two comparable groups, one who will have the placebo treatment and the other the experimental treatment. At the end of follow-up any difference between the groups can be attributed to the difference in treatment.
  - **Control group** - is used to isolate the effects of the intervention.
- **What is blinding?**: Blinding refers to not knowing whether the participant is in the intervention or the control group. Several people may be blinded to the allocation including the participants, the people caring for patients, the people measuring outcomes, the lead researcher.

- **Pros and cons for RCT**:
  - Advantages: Experiment - the best way to test an hypothesis. If the trial is well conducted, differences in outcome can be attributed to the intervention.
  - Disadvantages: May not be ethical or feasible.
- **Example cohort study**: British doctors and smoking. Aim: to investigate the relationship between smoking and lung cancer.
- **Pros and cons for cohort study**:
  - Pros: Clear chronological order from reason to outcome. Can evaluate the relationship between multiple results and factor(s).
  - Cons: Large time consumption. Bias affect. Small sample size.
- **Characteristics of a case-control study**: Generally carried out to test hypotheses. Participants are chosen on the basis of their outcome status: a group with the outcome (cases) and a group without (controls). Information is collected from people with and without outcome about exposures that occurred in the past (retrospective). i.e. in general before disease was diagnosed.
- **Pros and cons for Case-control study**:
  - Advantages: Relatively quick. Smaller than cohort studies, particularly for rare outcomes. Can examine the effects of multiple exposures.
  - Disadvantages: Events have already occurred so the potential for bias is higher. It is very hard (if not impossible) to remove all the effects of confounding.
- **Sources of error in analytic studies**

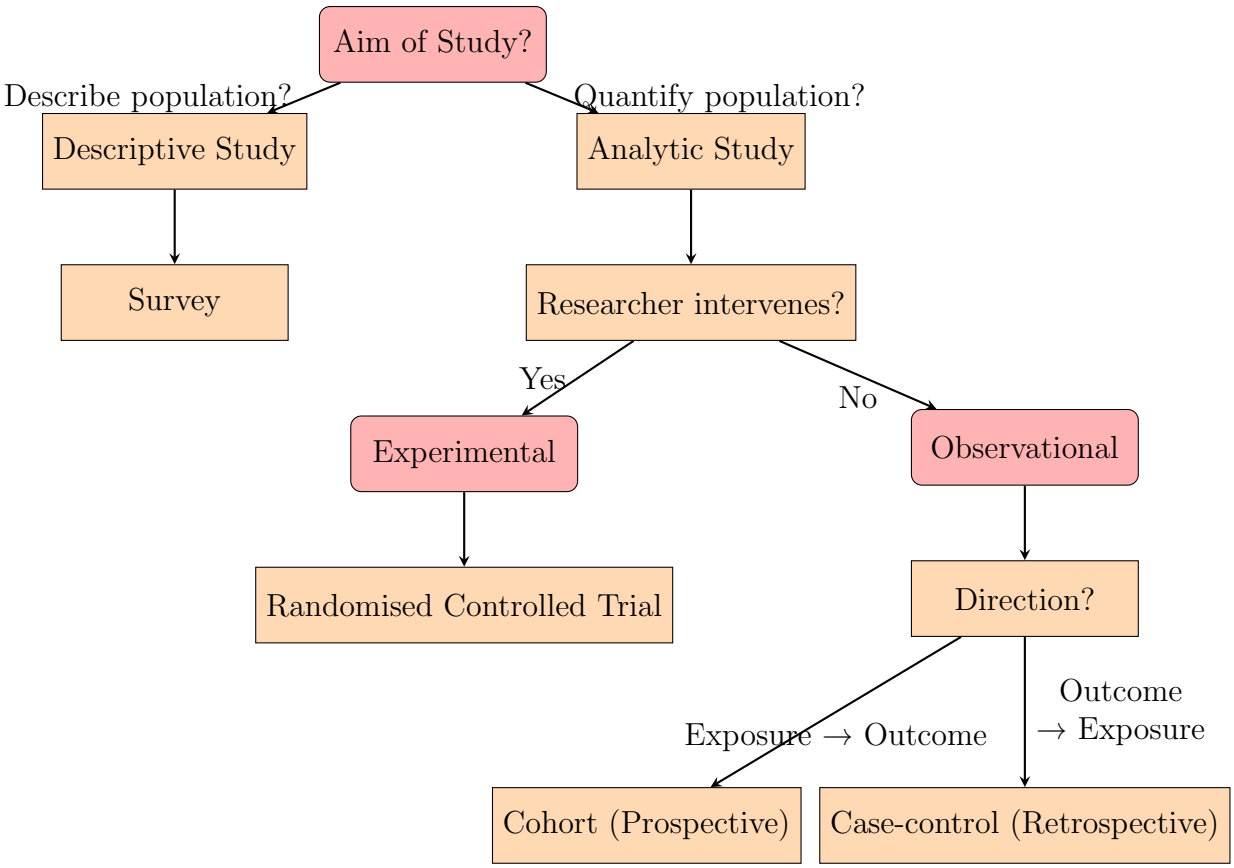


- **What is Confounding?**: Confounding is a distortion of the association between exposure and outcome caused by the presence of a third factor. A confounder is a variable which causes this distortion.
- **A variable must be both ( ) to become a confounder**:
  - associated with the exposure (independent of outcome);
  - and associated with the outcome (independent of exposure).
- **Bias in an analytic study**:

  - **Selection bias**: arising from the way participants are selected for inclusion in the study. In an analytic study, selection bias occurs if the selection processes cause a systematic difference between the groups of participants selected for the study. Prospective analytic studies rarely obtain participants through random sampling from a population. The issue of representativeness must be considered, but for analytic studies we consider it a generalisability issue rather than bias.
  - **Information bias**: arising from the way study information is obtained, interpreted and recorded. In an analytic study, information bias is a particular problem if there are systematic differences in the information obtained from groups under comparison in the study. Information bias may be introduced by the observer, the study individual (respondent), instruments used to collect the data (e.g., badly-designed questionnaire), or missing measurements (e.g., from loss to follow-up in a prospective study).

- **RCT, Cohort study, Case-control study**:
  - **Randomised controlled trial**: Analytic, experimental, prospective.
  - **Cohort study**: Analytic, observational, usually prospective.
  - **Case-control study**: Analytic, observational, retrospective.

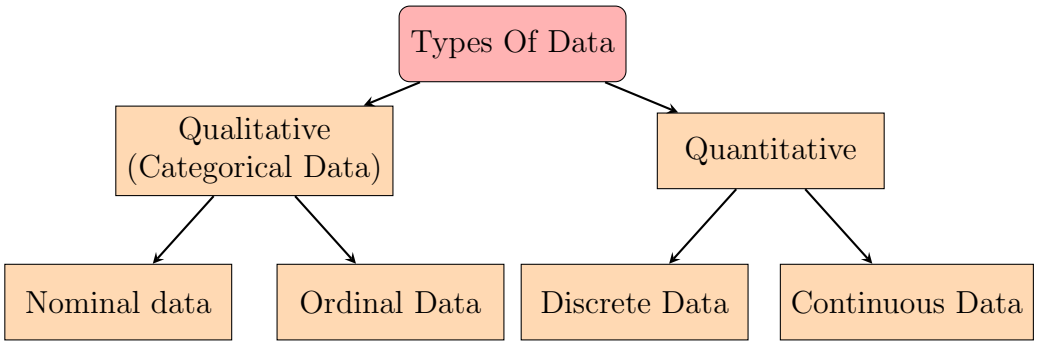
- **Summary for the classification**:



  - **Discrete**: A type of variable that can only take on specific values. These values are typically whole numbers or counts and cannot be subdivided further. For example, the number of children in a family is a discrete variable because it can only be a whole number (e.g., 1, 2, 3, etc.).

  - **Categorical**: Represent data that falls into specific categories or groups. The categories in nominal variables do not have any inherent order or ranking. Examples of nominal variables include gender (e.g., male, female), eye colour (e.g., blue, brown, green), or types of fruit (e.g., apple, banana, orange).
  - **Continuous**: Measurements that can take on any value within a specific range. They can be subdivided infinitely, and there are no gaps or interruptions in the possible values. Examples of continuous variables include height, weight, temperature, and time. These variables are often represented by real numbers and can include decimal values.
  - **Ordinal**: Similar to categorical variables, but they have an inherent order or ranking associated with their categories. The order represents the relative magnitude or importance of the categories, but the actual differences between the categories may not be uniform or measurable. Examples of ordinal variables include educational attainment (e.g., high school, bachelor's, master's, Ph.D.), socioeconomic status (e.g., low, medium, high), or survey ratings (e.g., strongly agree, agree, neutral, disagree, strongly disagree).
  - **If a data set is Categorical, must it also be Nominal?**: No. All nominal data is categorical data, but not all categorical data is nominal data. Nominal data refers specifically to categorical data without any order or hierarchy.

- **Types of data**



- **How to identify whether a study uses probability sampling?**: To find sampling frame.
- **Note for Stratified sampling**: Stratified sampling involves dividing the population into distinct subgroups (strata) based on certain characteristics.
- **Why Non-response can cause bias in surveys?**: because non-respondents tend to(maybe) behave differently compared to people who respond.