HIROSHIMA UNIVERSITY

Fundamental Data Science (30104001)

Lecture 13 — Methods for data collection

Jorge N. Tendeiro

Hiroshima University

Today

- Sample survey, sample selection, random sampling methods.
- Experimental study, observational study.
- Examples of using randomness.

Sample survey

Sample survey

Suppose you want to know the ratings of a TV show among all Japanese households.

There are several methods we could use:

- Survey using measuring instruments (people meter or online meter).

  > *Set measuring instruments in target households to measure ratings.*

- Survey by questionnaire.

  > *Enter viewing information of target households to measure ratings.*

The big question is:

> *Why survey only some target households instead of surveying all households?*

Sample survey

Problems with surveying all households:

- It is costly and time consuming.

  Example:
  It is *expensive* and *time consuming* to install measuring instruments.
  It takes *too much time* to collect data through questionnaires.

- Compliance rates can be low.

  Example:
  People *refuse* to install measuring instruments.
  People *refuse* to answer questionnaires.

Sample survey is frequently done to limit the number of households to survey (avoid surveying all households).

Population and sample

Population:

> *The entire set of objects under investigation.*

Example.

- Survey of program viewers among households.
  Population: all households.
- Survey of academic achievements of national primary school students.
  Population: all primary school students in a country.

Sample:

> *Subset of the population.*

Complete *vs* sample survey

Complete survey:

> *Survey the entire population of interest.*

Example.

National population census.

Sample survey:

> *Survey a part of the population of interest.*

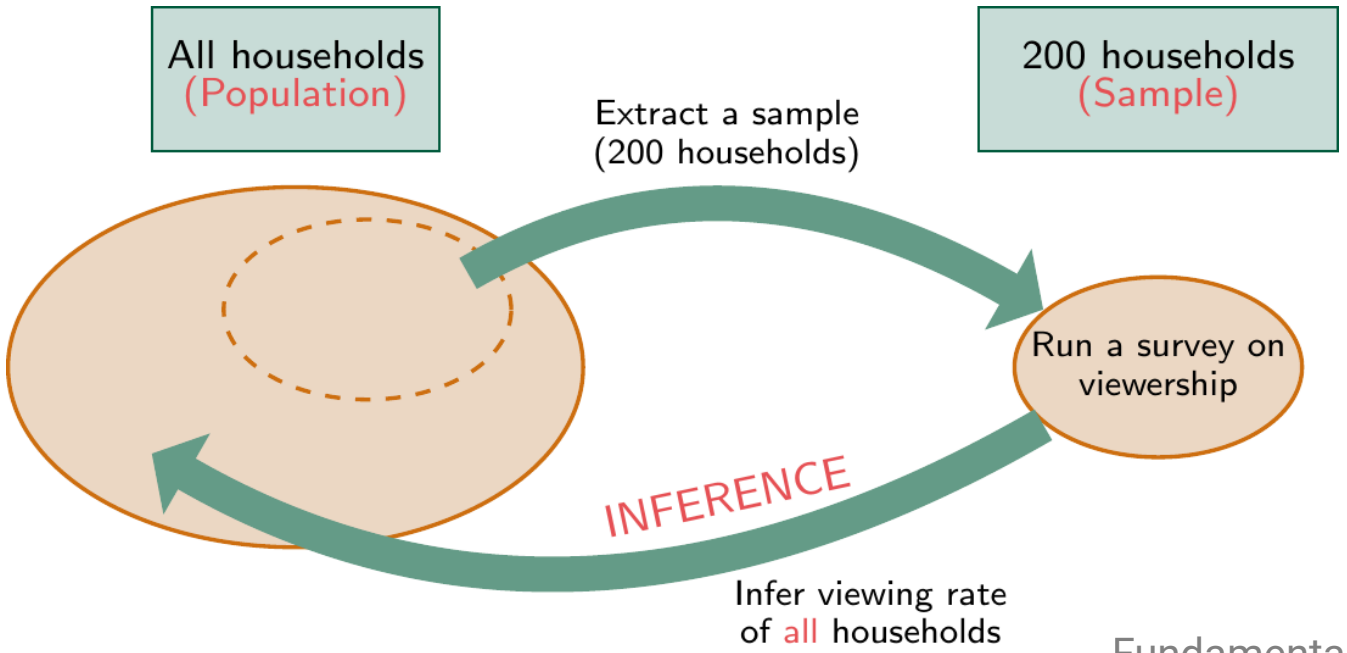
Sample surveys are used when it is cost-, time- and physical-wise difficult to perform a complete survey.

Example.

Poll.

Sample survey — Example

Suppose you want to know the ratings of a TV show among all Japanese households.



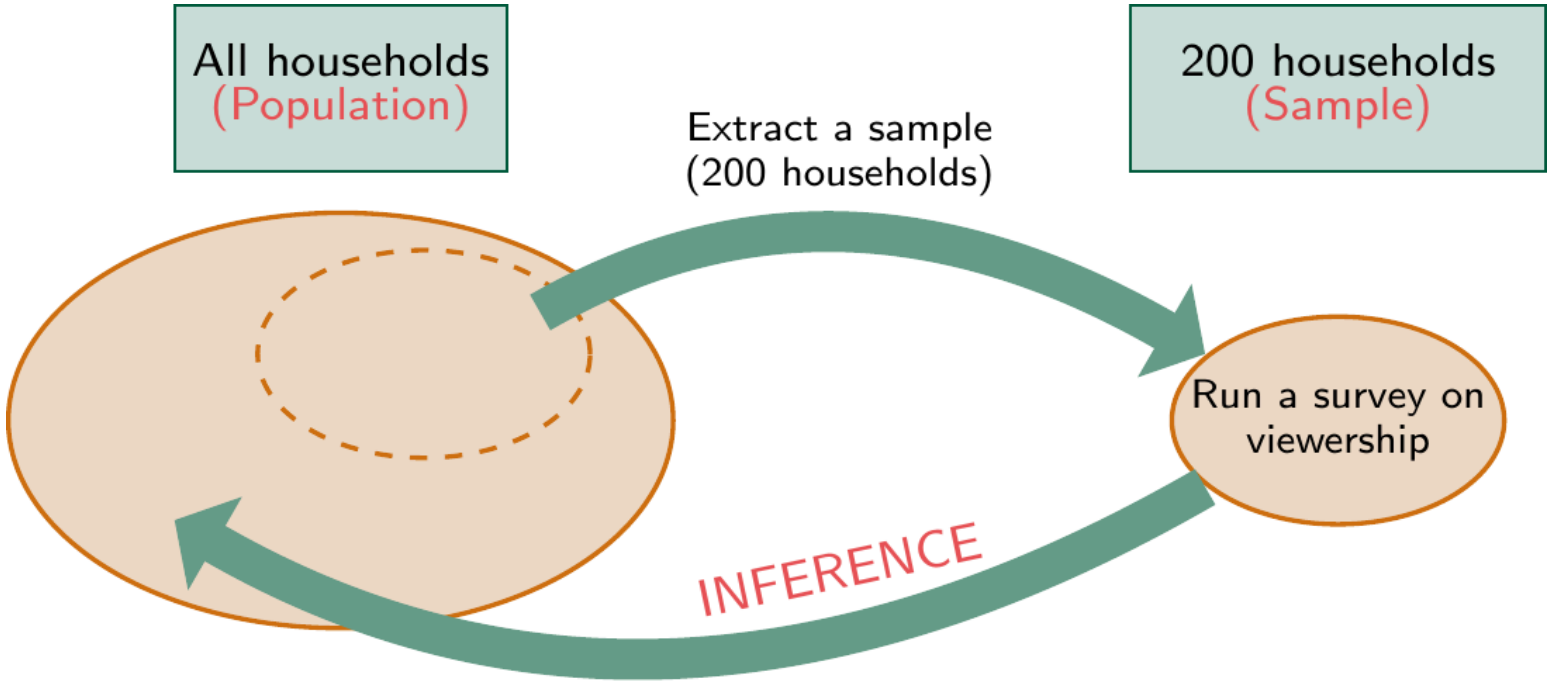
Surveys — Summary

	Complete survey	Sample survey
Subject of a study	Population	Sample (representative part of the population)
Cost	High	Low

Both complete survey and sample survey intend to study the target population.

Sample selection

Sample selection



Notes on sample selection

Two examples of things to avoid when sampling.

Example:

> *Conduct TV ratings survey only among households in Hiroshima, knowing that the TV show was shot in Hiroshima.*

TV ratings may become higher than the actual viewership of the population (i.e., all households in Japan).
This is known as *undercoverage bias* (when some population members are inadequately represented in the sample), which induces sample bias.

Example:

> *Exclude households from where the leading actor is from.*

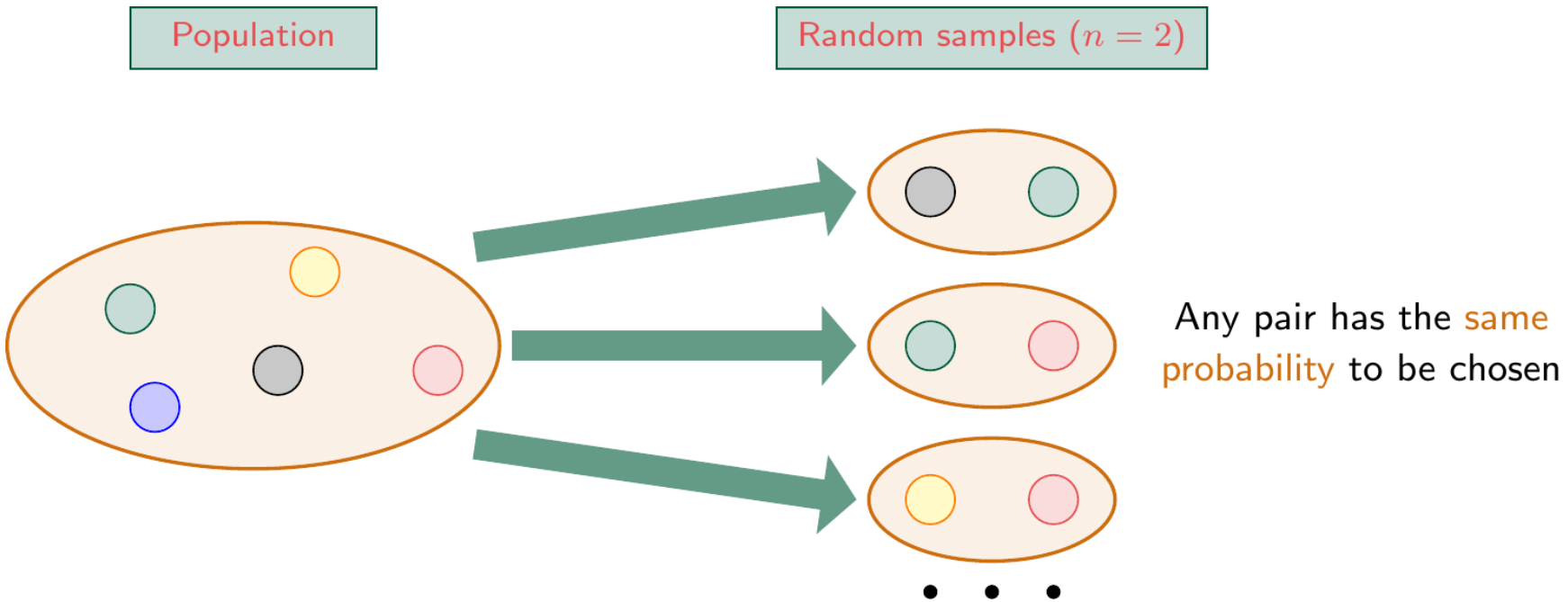
Including subjectivity into the sampling procedure.
This is known as *exclusion bias* (when some population members are intentionally excluded from the sample), which induces sample bias.

Random sampling

Random sampling:

> *Select a sample by using a lottery or dice system to eliminate subjectivity and select randomly.*

(Strict definition: All objects in the population may be selected with equal probability.)



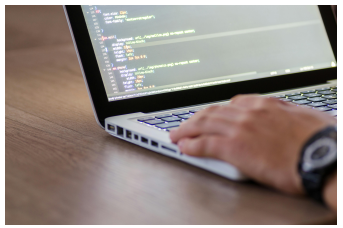
Random sampling

Random sampling can be performed by using various strategies to select a sample:

Dice

Lottery

Lottery machine



Generate pseudo random numbers via a computer

Reasons for random sampling:

- To (theoretically) eliminate bias and subjectivity.
  The goal is that the characteristics of the sample are as close as possible to those of the population.
- Random sampling is a major premise to perform inferential statistics.

Question

For each of the four examples below:

1. What is the population?
2. Is random sampling being used?

Example 1

> *In order to survey 1,000 people from the general consumers nationwide, people were categorized by sex $\times$ age groups. As we collected data, we assigned data into the above-described categorized groups until the number reached 1,000 people.*

Example 2

> *Because Product A is developed for two-generation families living in a suburb area, we identified households who matched these conditions. Only when such a household kindly agreed to participate on a study, we collected data to search for the needs of Product A.*

Question

For each of the four examples below:

1. What is the population?
2. Is random sampling being used?

Example 3

> *To study product feedback of a service from 300 users, we identified users of the service within a company. Then, we asked them to recruit their friends who also used the same service to join the study. We repeatedly asked participants to recruit their friends to collect data.*

Example 4

> *To broadcast about an approval rating (public opinion) of Tokyo governor among Tokyo citizens during a news program tomorrow, we allocated staff members at 3 main Tokyo stations from the morning till the evening on the day before the news program to interview passersby. In case someone refused to answer their questions, they interviewed the next person. In this way, they have collected answers from more than 1000 people.*

Ref: Nikkei search company, random sampling (https://www.nikkei-r.co.jp/glossary/id=1663)

Question — ANSWER

For each of the four examples below:

1. What is the population? $\longrightarrow$ In red.
2. Is random sampling being used? $\longrightarrow$ **NO.** The reasons are in blue.

Example 1

> *In order to survey 1,000 people from the general consumers nationwide, people were categorized by sex $\times$ age groups. As we collected data, we assigned data into the above-described categorized groups until the number reached 1,000 people.*

Example 2

> *Because Product A is developed for two-generation families living in a suburb area, we identified households who matched these conditions. Only when such a household kindly agreed to participate on a study, we collected data to search for the needs of Product A.*

Question — ANSWER

For each of the four examples below:

1. What is the population? ⟶ In red.
2. Is random sampling being used? ⟶ **NO.** The reasons are in blue.

Example 3

> *To study product feedback of a service from 300 users, we identified users of the service within a company. Then, we asked them to recruit their friends who also used the same service to join the study. We repeatedly asked participants to recruit their friends to collect data.*

Example 4

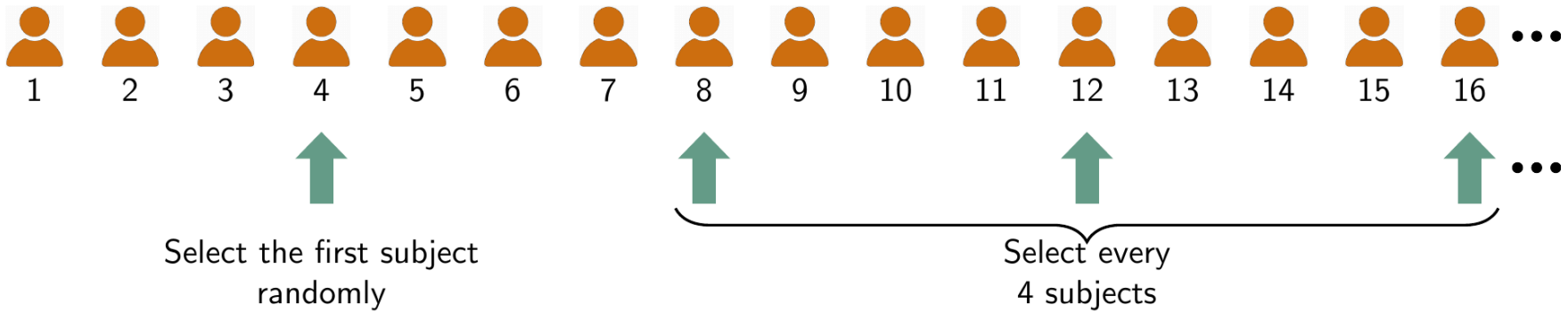
> *To broadcast about an approval rating (public opinion) of Tokyo governor among Tokyo citizens during a news program tomorrow, we allocated staff members at 3 main Tokyo stations from the morning till the evening on the day before the news program to interview passersby. In case someone refused to answer their questions, they interviewed the next person. In this way, they have collected answers from more than 1000 people.*

Methods for random sampling

Systematic sampling

Systematic sampling:

> *Sequentially numbering all subjects of the population and randomly choosing the first, but after the second one onward, subsequently sample with a certain interval.*



Example: Investigation from the Basic Resident Register or business office list.

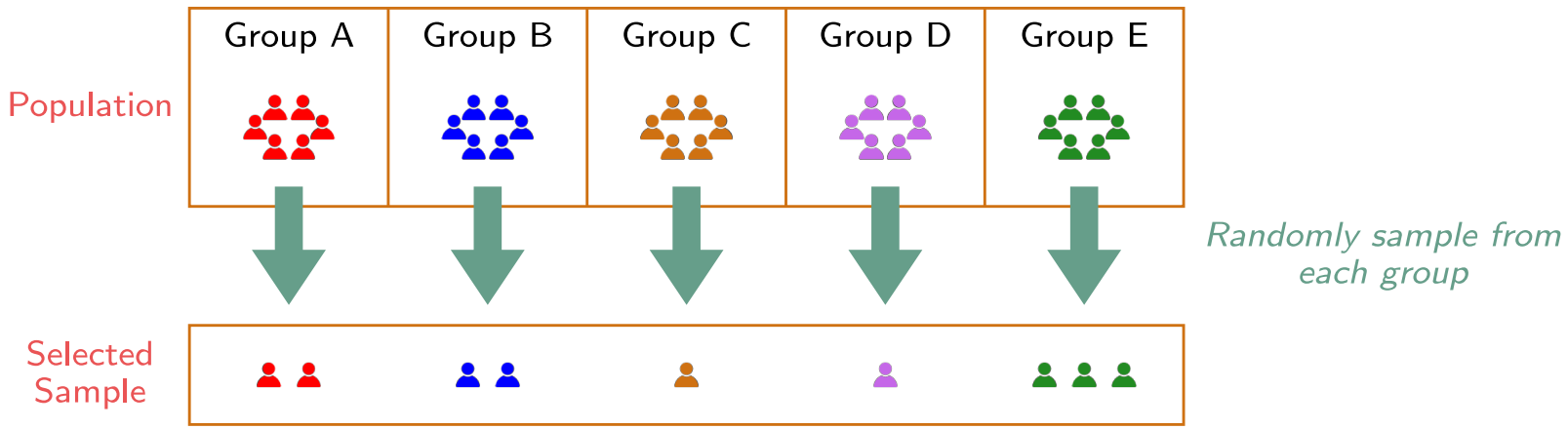
> *Basic Resident Register: summarize resident register per household.*

With systematic sampling, it is less likely that people from the same household are selected, and therefore it is unlikely to have bias.

Stratified sampling

Stratified sampling:

> *Divide the population into several groups and randomize samples from each group.*

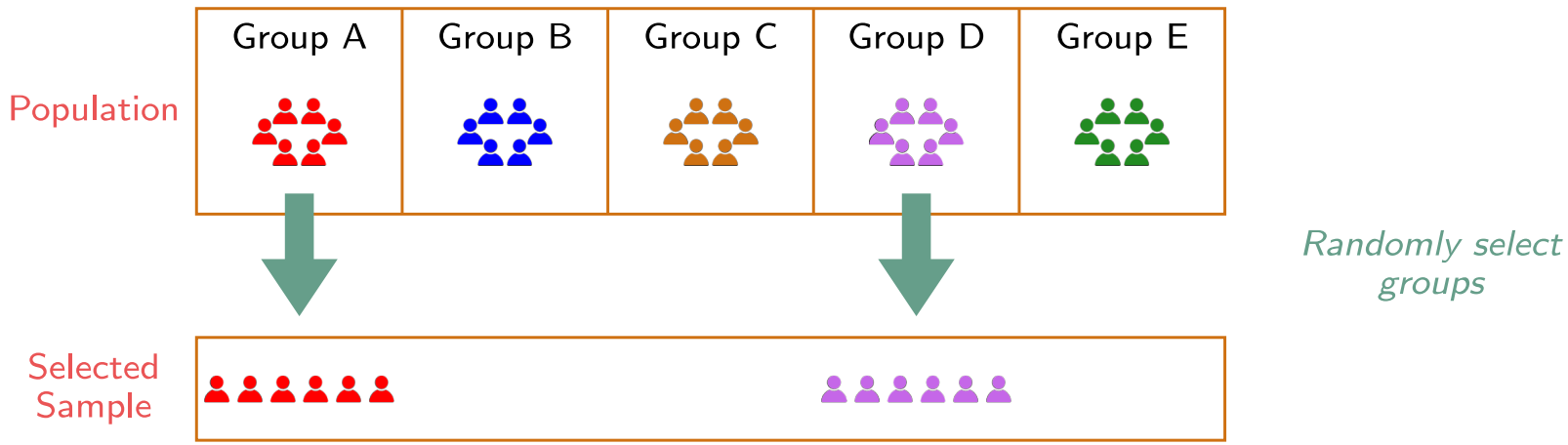


Example: American political opinion polls (grouped by religion, etc.).
Such polling should reflect the overall public opinion.
Random sampling must take into account the relative group sizes in the population.

Cluster sampling

Cluster sampling:

> *Divide the population into several groups.*
> *Then, select groups randomly and perform a complete survey for the selected groups.*



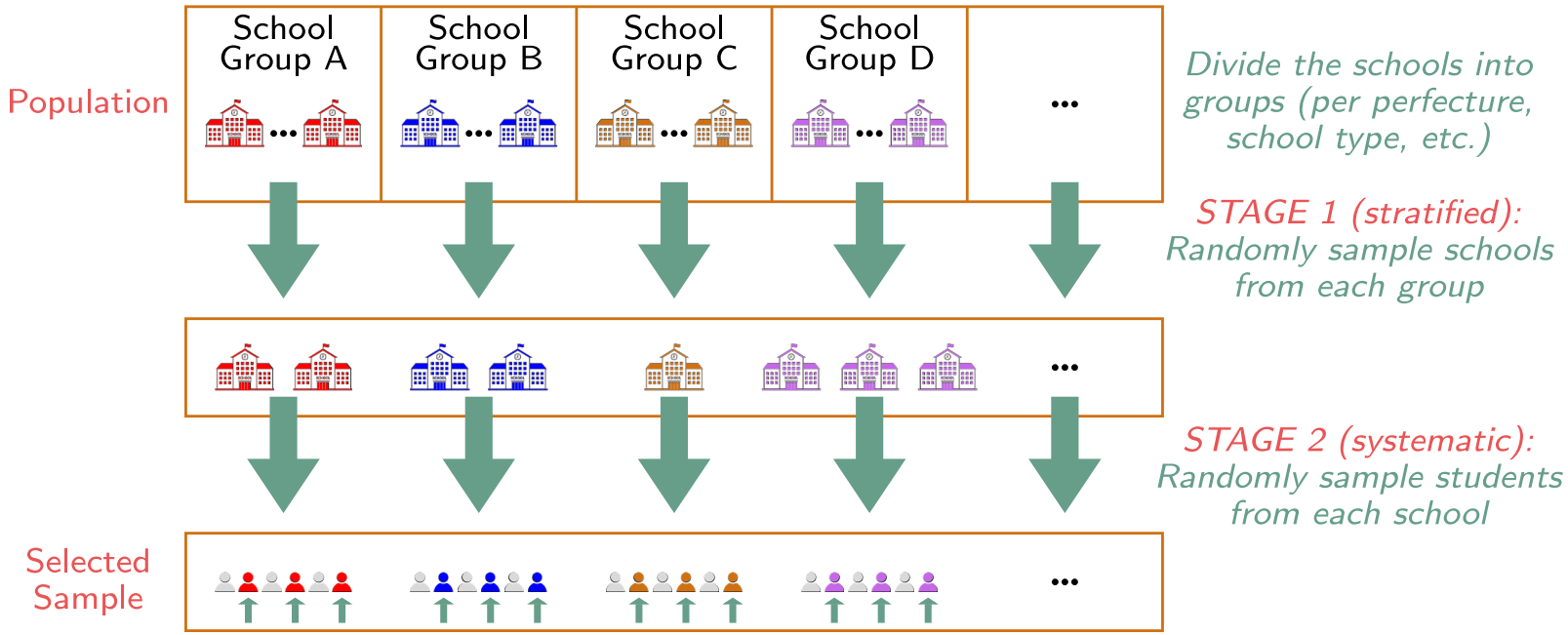
Example: National Livelihood Survey (randomly sample survey districts from about 900 thousand districts).
This sampling scheme allows to cut costs since the target districts are limited.

Two-stage sampling

Two-stage sampling:

> *Do sampling in two stages (stratified + systematic sampling).*

Example: Annual Report of School Health Statistics Research (development survey).



Sampling — Example

A poll conducted by NHK.
An example of a survey where investigators visit 3,600 survey target people in Japan.

1. *Sampling survey locations.*
   1.1. Divide the entire Japan into 18 blocks.

   1.2. In each block, order cities by their sizes and composition ratio of employees by industry. Then, systematic sampling is conducted from 300 sites in Japan, in proportion to the size of the population of each block.

2. *Sample study target.*
   From the Basic Resident Register of the municipality of the survey site, sample 12 study target people in an equidistant manner for each site (systematic sampling).

Reference: https://www.nhk.or.jp/bunken/yoron/nhk/process/sampling.html

Experimental and observational studies

Experimental and observational studies

Researchers are often interested in studying how a phenomenon influences outcomes.

Examples:

> *Do you lose weight by eating diet food A?*

- Phenomenon = *eating diet food A*
- Outcome = *lose weight*

> *If you smoke, do lung cancer incidence rates increase?*

- Phenomenon = *smoking*
- Outcome = *cancer incident rates increase*

There are two types of studies that we can conduct in order to study such effects:

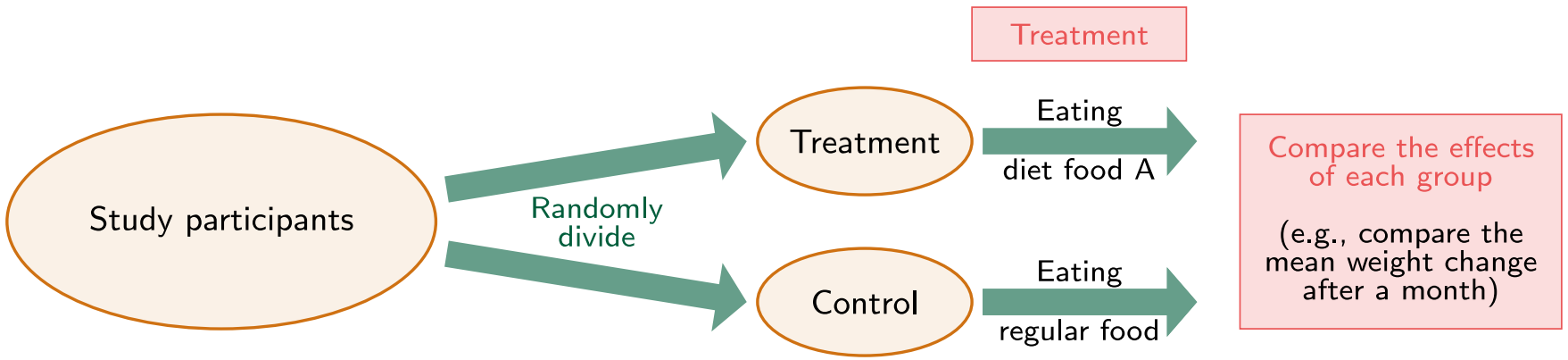
- Experimental studies.
- Observational studies.

Experimental study

Experimental study:

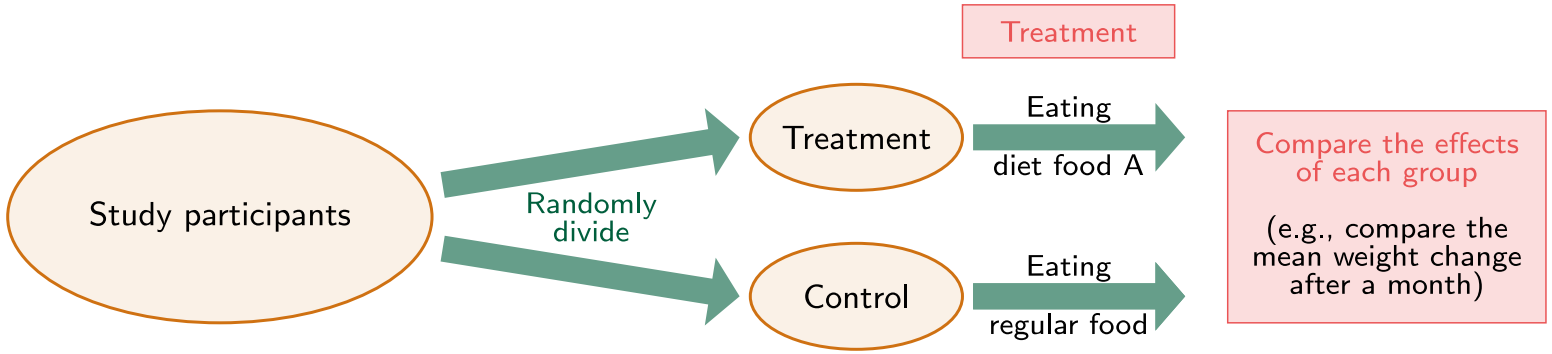
> *To study the effect of an intervention (e.g., a new diet food), subjects are randomly assigned to either a treatment group (eat the diet food) or a control group (eat the usual food). This way, groups are (about) equal at baseline, and differences after the experiment may be ascribed to the treatment.*

Example: *Do you lose weight by eating diet food A?*



- Treatment group: Group to assign a treatment.
- Control group: Group to serve as a basis of comparison.

Experimental study — Some notes



- Consider a placebo effect:

  > *Effect due to believing in the treatment effect, while belonging to the control group.*

  For example, some patients with a disease can actually improve by taking a medication without an effect. It is important to study an effect by considering that there may be a placebo effect.

- Beware of ethical issues.

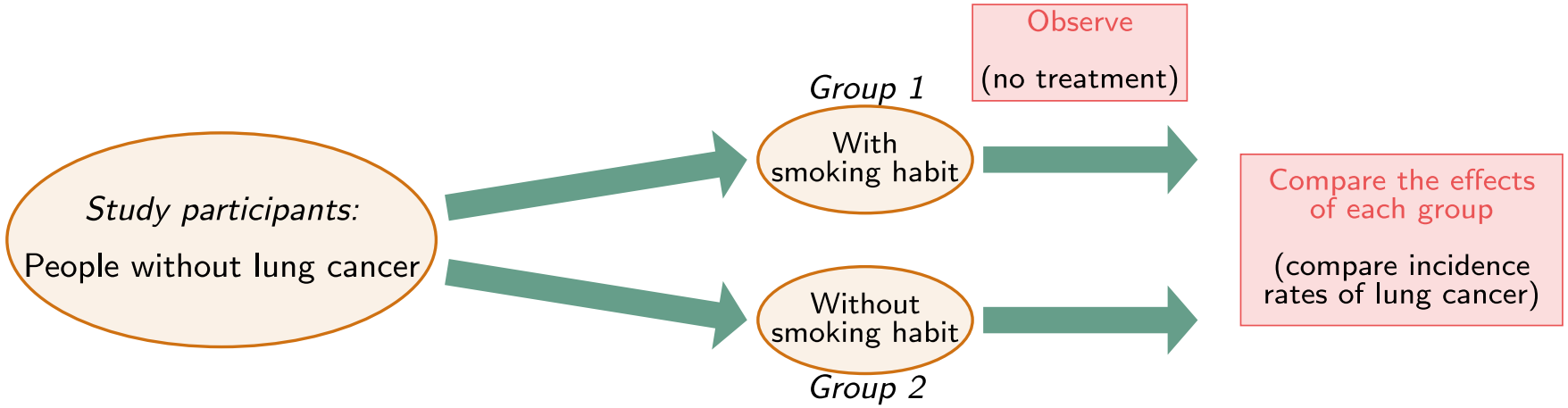
  For example, forcing the treatment group to smoke in order to examine the incidence rates of lung cancer is completely unacceptable.

Observational study

Observational study:

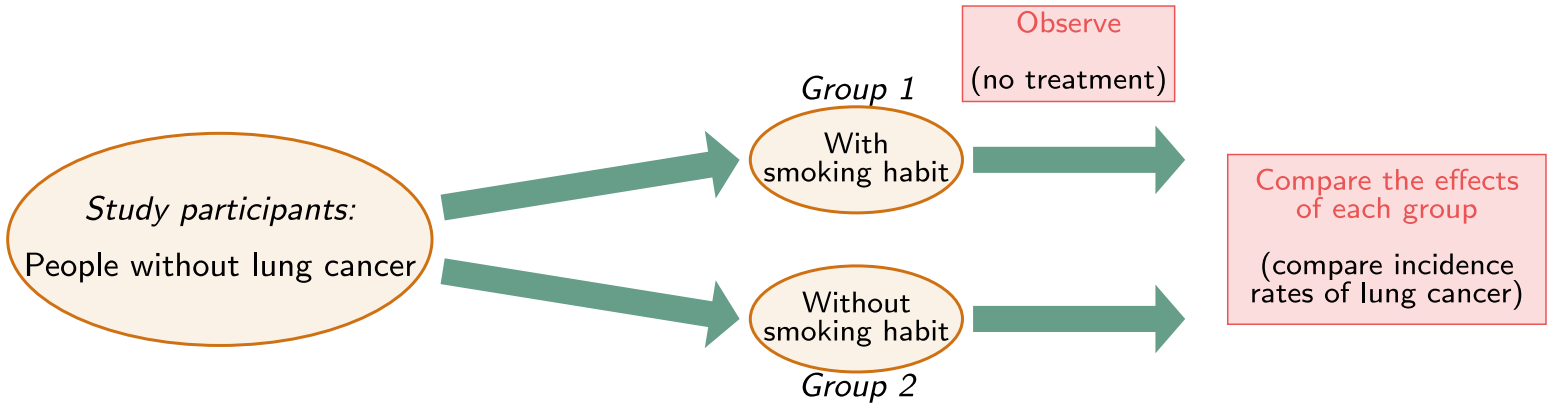
> *No treatment is administered to study objects.*
> *Instead, we study an effect of a phenomenon by collecting data from an observation.*

Example: *If you smoke, do lung cancer incidence rates increase?*



Observational studies are typically done when experimental studies are unfeasible.

Observational study — Some notes



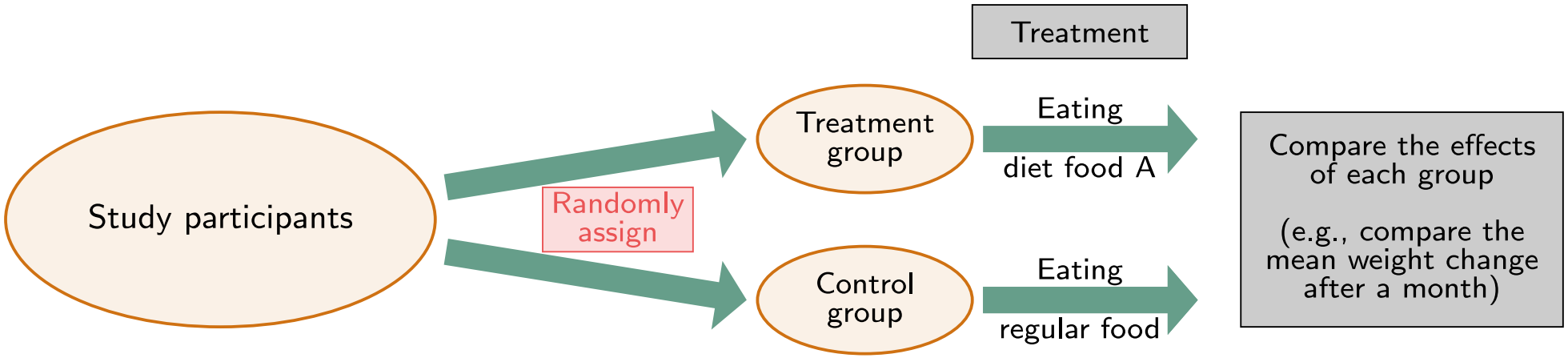
- Because participants were not randomly allocated, Groups 1 and 2 may differ on many variables other than 'smoking habit'.
  Hence, the 'smoking habit' may be confounded with other, uncontrolled, effects.
- We should therefore always consider other possible causes which could influence an outcome.
  Example: Consider drinking habit, gender, etc.

Examples of using randomness

Randomized controlled trial (RCT)

Randomized controlled trial:

> *An experiment where people are randomly assigned to a treatment group or a control group.*

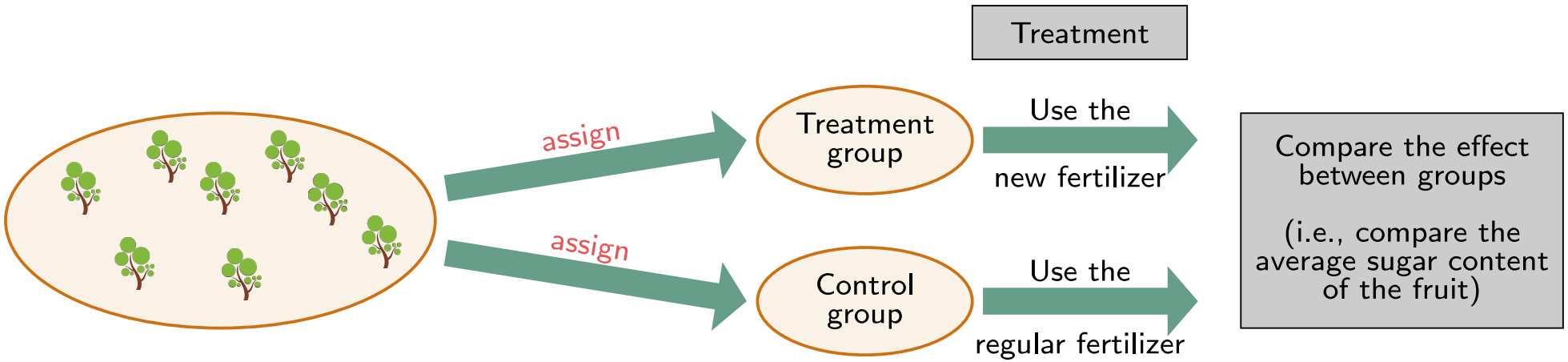


Q: Why do we assign people randomly?

A: To control for possible influences due to other factors (e.g., exercise habits, etc.).
This way, the treatment and control groups become 'equal'.
It is then possible to compare the treatment effect.

Rationale of RCT

Example: *Does feeding a new fertilizer to a sapling increase the sugar content of the fruit?*



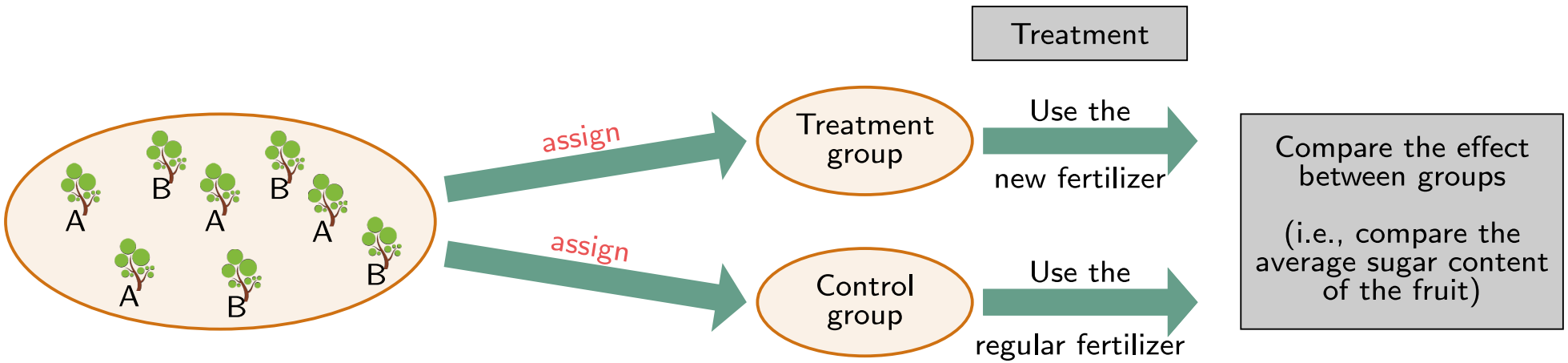
The question here is:

> *Does random assignment to each group actually matter to learn about the treatment effect?*

Rationale of RCT

Consider the following scenario.
Unbeknownst to us:

- There are actually two types of sapling genotypes (Type A, Type B).
  For simplicity, assume that both genotypes occur equally likely.



Rationale of RCT

Unbeknownst to us:

- The sugar content of the fruit varies with genotype:

> *Type A has on average 4g of sugar more than Type B, per 100g of fruit.*

- The treatment effect is the same across both genotypes (it increases the sugar content by 1g). (Assume we did not know this, but this were the reality.)

Fertilizer \ Genotype	Type A	Type B
New	16	12
Regular	15	11

Note. In grams sugar/100g.

Rationale of RCT — Random assignment

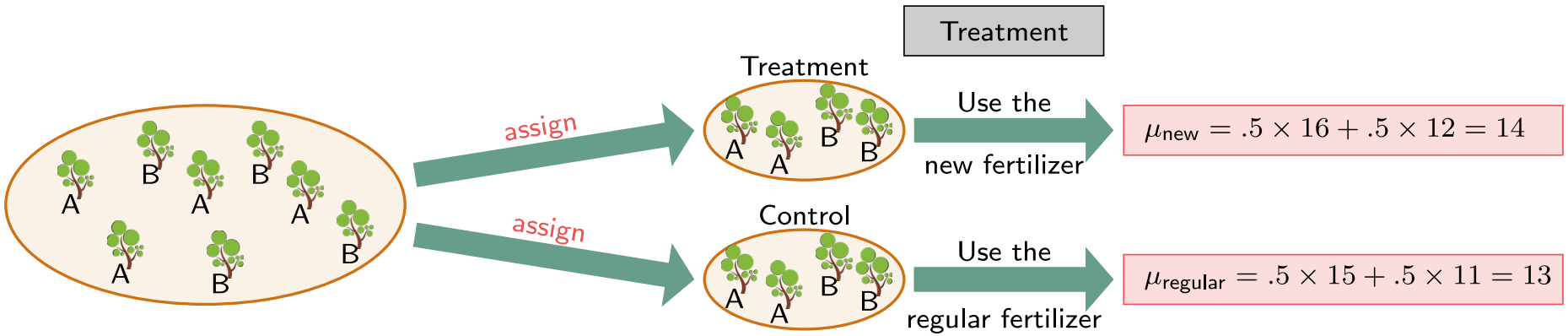
If saplings are randomly assigned to each group then we except that, on average,

$$\mu_{\text{new}} = 50\%(16) + 50\%(12) = 14$$
$$\mu_{\text{regular}} = 50\%(15) + 50\%(11) = 13$$

Fertilizer \ Genotype	Type A	Type B
New	16	12
Regular	15	11

Note. In grams sugar/100g.

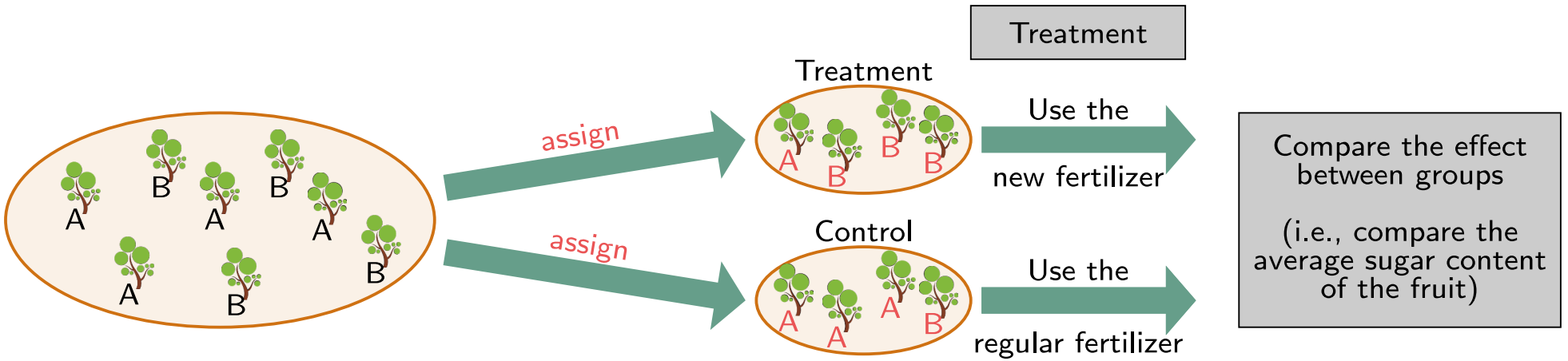
We conclude that the new treatment works:
It increases the sugar amount on average by 1g of sugar per 100 grams.



Rationale of RCT — Non- random assignment

Suppose that, unfortunately, assignment was poorly performed:

- 75% of saplings in the treatment group are of Type B.
- 25% of saplings in the treatment group are of Type A.



Rationale of RCT — Non- random assignment

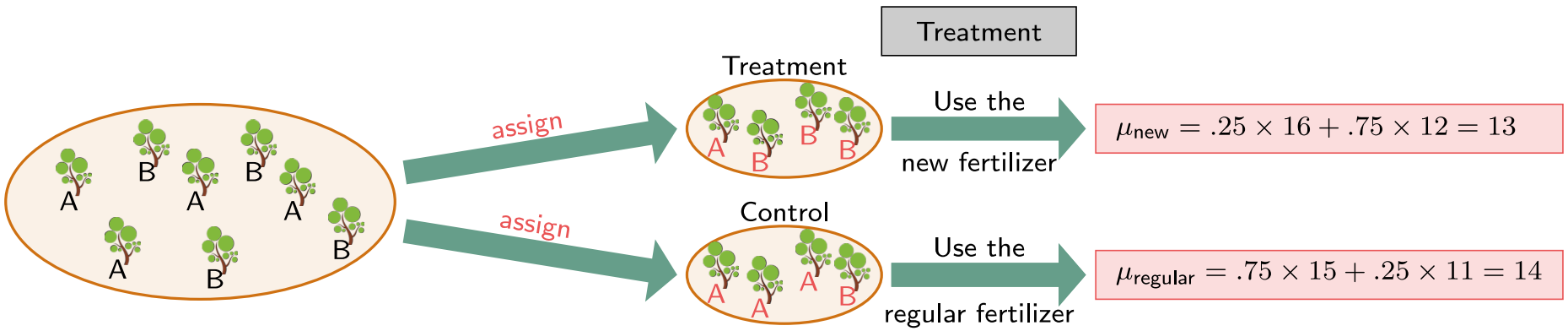
Now,

$$\mu_{\text{new}} = 25\%(16) + 75\%(12) = 13$$
$$\mu_{\text{regular}} = 75\%(15) + 25\%(11) = 14$$

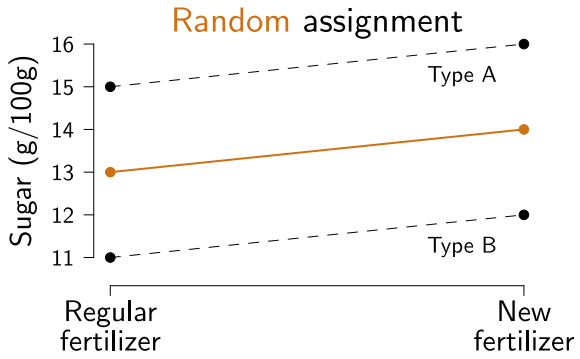
Fertilizer \ Genotype	Type A	Type B
New	16	12
Regular	15	11

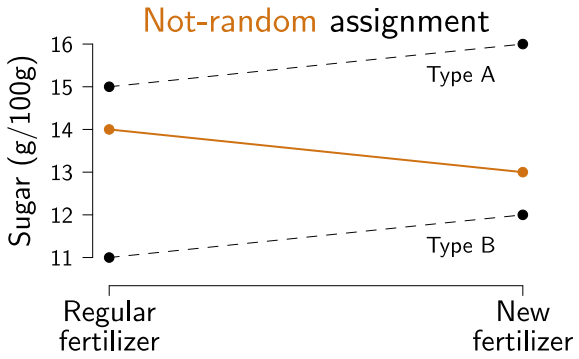
Note. In grams sugar/100g.

We conclude that the new treatment does not work:
It decreases the sugar amount on average by 1g of sugar per 100 grams.



Rationale of RCT — Summary





- Random assignment controls for the effect of unmeasured factors (like genotype in our example).
- Although the treatment works for both genotypes (1g of sugar on average), non-random assignment actually reversed the treatment effect. This is known as Simpson's paradox.

The reversal of the treatment effect occurred because:

- The proportion of genotypes in each group were very different from the true proportions (due to poor assignment).
- Type B has lower sugar average than Type A.
- The treatment group was mostly composed on Type B saplings (75%).

A/B testing

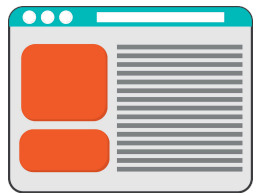
A/B testing:

> *A method to compare two plans, A and B, in marketing or business context.*

Suppose we want to compare the effectiveness of two webpage designs:

Classical design.

A new design.



As the outcome variable we will use click rate.

Our question is:

> *Which of the two webpage designs leads to a higher click rate?*

A/B testing

Here is an example on how to use of A/B testing to compare the click rates:

1. Randomly assign each website visitor to a group ('classical' or 'new'), say, for a month.

2. Show the classical webpage design to subjects in the 'classical' group and the new design to subjects in the 'new' group.

3. Record the click rate of each group.

Webpage design	Click rate
Classical	1.1%
New	1.6%

4. Choose the new webpage design since it has a higher clicking rate.

Summary

We learned about

- Sample survey, sample selection, random sampling methods.

	Complete survey	Sample survey
Subject of a study	Population	Sample (representative part of the population)
Cost	High	Low

- Experimental study, observational study.
- Examples of using randomness:
  - Randomized control trial.
  - A/B testing.