American University of Beirut

# STAT 210: Elementary Statistics for Sciences

2022–2023 Fall

Siamak Taati

Chapter 1
Introduction

# Practical information

## Lectures

(L2) Mondays and Wednesdays 12:00–13:00, Nicely 211

Instructor: Siamak Taati

Email: st71@aub.edu.lb

Office: Bliss Hall 312B

## Associated recitations

(E3) Thursdays 12:30–13:30, Bliss 205

Instructor: Alice Ashkar

(E4) Thursdays 17:00–18:00, Nicely 416

Instructor: Israa Fakih

(E6) Thursdays 17:00–18:00, Bliss 205

Instructor: Fatima Fneish

# Practical information

## Textbook

Introductory Statistics (global edition) by Prem S. Mann

## Moodle

Make sure to regularly check the course Moodle site for supplementary materials and announcements.

## Assessment

50%
- Midterm 1 (Thursday, October 6, 6pm; Nicely 500)
- Midterm 2 (Friday, November 4, 6pm; Nicely 500)

40%
- Final

10%
- Active participation/quizzes

# What is Statistics?

Statistics is the science/art of extracting reliable information from empirical data.

## Two branches of statistics

| | |
|---|---|
| Theoretical | Rationale, mathematical derivations and justifications |
| Applied | Applying the derived methods to real-world problems |

## Two types of statistical study

| | |
|---|---|
| Descriptive | Describing and summarizing data using tables, graphs and summary statistics |
| Inferential | Using statistical evidence (data) and probability models to draw conclusions |

# Population vs. sample

## Scenario

We would like to study the effect of sleep on the performance of university students.

We cannot possibly study *all* the university students at the same time (due to cost, time limitation, inconvenience, . . . ).

Instead, we randomly pick 100 students and record their sleeping habits and their study performance during an academic year, and use that as statistical evidence.

Here,

- Population refers to the set of all the university students
- Sample refers to the randomly picked students

# Data sets

| | gender | year | sleep | GPA |
|---|---|---|---|---|
| 1 | F | 2 | 8.16 | 2.93 |
| 2 | M | 4 | 7.60 | 3.04 |
| 3 | F | 4 | 8.54 | 3.28 |
| 4 | F | 1 | 7.44 | 3.21 |
| 5 | M | 2 | 7.93 | 3.38 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 100 | M | 2 | 7.10 | 3.49 |

- Each row corresponds to one of the picked students.
- Each column corresponds a variable.

# Data sets

| | gender | year | sleep | GPA |
|---|---|---|---|---|
| 1 | F | 2 | 8.16 | 2.93 |
| 2 | M | 4 | 7.60 | 3.04 |
| 3 | F | 4 | 8.54 | 3.28 |
| 4 | F | 1 | 7.44 | 3.21 |
| 5 | M | 2 | 7.93 | 3.38 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 100 | M | 2 | 7.10 | 3.49 |

- Each row corresponds to one of the picked students.
- Each column corresponds a variable.

## Sample vs. census data

This is an example of a sample data set.
If the data set contained data about *every* member of the population, then we would call it a census data set.

# Data sets

| | gender | year | sleep | GPA |
|---|---|---|---|---|
| 1 | F | 2 | 8.16 | 2.93 |
| 2 | M | 4 | 7.60 | 3.04 |
| 3 | F | 4 | 8.54 | 3.28 |
| 4 | F | 1 | 7.44 | 3.21 |
| 5 | M | 2 | 7.93 | 3.38 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 100 | M | 2 | 7.10 | 3.49 |

- Each row corresponds to one of the picked students.
- Each column corresponds a variable.

## Types of variables

- numerical (quantitative)
  - discrete [e.g., year of study]
  - continuous [e.g., average sleep hours per day]
- categorical (qualitative) [e.g., gender]

# Data sets

| | gender | year | sleep | GPA |
|---|---|---|---|---|
| 1 | F | 2 | 8.16 | 2.93 |
| 2 | M | 4 | 7.60 | 3.04 |
| 3 | F | 4 | 8.54 | 3.28 |
| 4 | F | 1 | 7.44 | 3.21 |
| 5 | M | 2 | 7.93 | 3.38 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 100 | M | 2 | 7.10 | 3.49 |

- Each row corresponds to one of the picked students.
- Each column corresponds a variable.

## Types of variables

- numerical (quantitative)
  - discrete [e.g., year of study]
  - continuous [e.g., average sleep hours per day]
- categorical (qualitative) [e.g., gender]

# Data sets

| | gender | year | sleep | GPA |
|---|---|---|---|---|
| 1 | F | 2 | 8.16 | 2.93 |
| 2 | M | 4 | 7.60 | 3.04 |
| 3 | F | 4 | 8.54 | 3.28 |
| 4 | F | 1 | 7.44 | 3.21 |
| 5 | M | 2 | 7.93 | 3.38 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 100 | M | 2 | 7.10 | 3.49 |

- Each row corresponds to one of the picked students.
- Each column corresponds a variable.

## Types of variables

- numerical (quantitative)
  - discrete [e.g., year of study]
  - continuous [e.g., average sleep hours per day]
- categorical (qualitative) [e.g., gender]

# Data sets

| | gender | year | sleep | GPA |
|---|---|---|---|---|
| 1 | F | 2 | 8.16 | 2.93 |
| 2 | M | 4 | 7.60 | 3.04 |
| 3 | F | 4 | 8.54 | 3.28 |
| 4 | F | 1 | 7.44 | 3.21 |
| 5 | M | 2 | 7.93 | 3.38 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 100 | M | 2 | 7.10 | 3.49 |

- Each row corresponds to one of the picked students.
- Each column corresponds a variable.

## Types of variables

- numerical (quantitative)
  - discrete [e.g., year of study]
  - continuous [e.g., average sleep hours per day]
- categorical (qualitative) [e.g., gender]

# Samples

In general, we would like the sample to be a good representative of the population.

## Types of samples

- Random sample: [make on a lottery]
  - with replacement [return each ticket before drawing the next]
  - without replacement [do not return the tickets]
- Non-random sample:
  - Convenience sample [e.g., pick the students in two sections of STAT 210]
  - Judgment sample [based on prior judgment of typical elements]

## Remark

*A random sample has a much higher chance of being a good representative of the population, hence it is always preferable to a non-random sample.*

# Errors and biases

Statistical studies are prone to errors and biases. The nature of these errors and biases could however be different.

## Types of error/bias

1. Sampling errors [Due to chance]
   The difference between the result from the sample, and a hypothetical result if we had access to data from the entire population.

## Remark

*Sampling errors are inevitable (due to chance). However, it is desirable to be able to measure how sampling errors affect the reliability of the results.*

# Errors and biases

Statistical studies are prone to errors and biases. The nature of these errors and biases could however be different.

## Types of error/bias

2. Systemic errors/biases [Due to human error]
   - Selection bias [sampling procedure not representative]
   - Non-response bias [some sampled students do not respond]
   - Response bias [some sampled students do not provide accurate responses]
   - Voluntary response bias [the data is based on responses from volunteers rather than a random sample]

## Remark

*Systemic errors/biases can be minimized by systematic sampling and following well-chosen protocols in collecting and handling data.*

# Errors and biases

Detect the sources of error!

## Example 1

A newspaper conducts a poll about the favorite candidate in the next presidential election, but only includes its readers.

# Errors and biases

Detect the sources of error!

## Example 1

A newspaper conducts a poll about the favorite candidate in the next presidential election, but only includes its readers.

$\longrightarrow$ Selection bias (non-random sample, convenience)

# Errors and biases

Detect the sources of error!

## Example 2

An NGO conducts a survey about a social issue by stopping random passersby on the street and asking them to fill in a 10-page questionnaire.

# Errors and biases

Detect the sources of error!

## Example 2

An NGO conducts a survey about a social issue by stopping random passersby on the street and asking them to fill in a 10-page questionnaire.

$\longrightarrow$ Non-response bias or voluntary response bias

# Errors and biases

Detect the sources of error!

## Example 3

A group of students conduct a study about the disadvantages of online classes over in-person classes by asking 3 randomly selected students to fill in a questionnaire.

# Errors and biases

Detect the sources of error!

## Example 3

A group of students conduct a study about the disadvantages of online classes over in-person classes by asking 3 randomly selected students to fill in a questionnaire.

$\longrightarrow$ Chance of sampling error will be high,
because the sample is very small.

# Random sampling

## Sampling techniques

- Simple: Make a lottery ticket for each student and put them in a box. Draw 100 tickets from the box.

# Random sampling

## Sampling techniques

- Simple: Make a lottery ticket for each student and put them in a box. Draw 100 tickets from the box.
- Systemic: Divide the population of all students into 100 groups of size roughly $n$. Pick a number $K$ between 1 and $n$ at random. Select the $K$-th student from each group.

# Random sampling

## Sampling techniques

- Simple: Make a lottery ticket for each student and put them in a box. Draw 100 tickets from the box.
- Systemic: Divide the population of all students into 100 groups of size roughly $n$. Pick a number $K$ between 1 and $n$ at random. Select the $K$-th student from each group.
- Stratified: Divide the population into subpopulations, called *stratas*. Draw a sample from each strata, and put these samples together.

# Random sampling

## Sampling techniques

- Simple: Make a lottery ticket for each student and put them in a box. Draw 100 tickets from the box.
- Systemic: Divide the population of all students into 100 groups of size roughly $n$. Pick a number $K$ between 1 and $n$ at random. Select the $K$-th student from each group.
- Stratified: Divide the population into subpopulations, called *stratas*. Draw a sample from each strata, and put these samples together.
- Cluster: Divide the population into (geographical) subpopulations, called *clusters*. Select a random sample of the clusters. Draw a sample from each selected cluster, and put these samples together.

# Random sampling

## Sampling techniques

- Simple: Make a lottery ticket for each student and put them in a box. Draw 100 tickets from the box.
- Systemic: Divide the population of all students into 100 groups of size roughly $n$. Pick a number $K$ between 1 and $n$ at random. Select the $K$-th student from each group.
- Stratified: Divide the population into subpopulations, called *stratas*. Draw a sample from each strata, and put these samples together.
- Cluster: Divide the population into (geographical) subpopulations, called *clusters*. Select a random sample of the clusters. Draw a sample from each selected cluster, and put these samples together.

## Remark

*In practice, these procedures are almost always realized using a computer.*

# Observational vs. experimental studies

## Scenario

We would like to study the possible relationship between salt and high blood pressure.

## Two types of study

1. Observational

# Observational vs. experimental studies

## Scenario

We would like to study the possible relationship between salt and high blood pressure.

## Two types of study

1. Observational
    a. We take a random sample of participants for the study.

# Observational vs. experimental studies

## Scenario

We would like to study the possible relationship between salt and high blood pressure.

## Two types of study

1. Observational
    a. We take a random sample of participants for the study.
    b. We collect data on their salt in-take and blood pressure over a period of time.

# Observational vs. experimental studies

## Scenario

We would like to study the possible relationship between salt and high blood pressure.

## Two types of study

1. Observational
    a. We take a random sample of participants for the study.
    b. We collect data on their salt in-take and blood pressure over a period of time.
    c. We study the collected data to derive conclusions.

# Observational vs. experimental studies

## Scenario

We would like to study the possible relationship between salt and high blood pressure.

## Two types of study

2. Randomized experiment [a.k.a. controlled or designed experiment]

# Observational vs. experimental studies

## Scenario

We would like to study the possible relationship between salt and high blood pressure.

## Two types of study

2. Randomized experiment [a.k.a. controlled or designed experiment]
    a. We take a random sample of participants for the study.

# Observational vs. experimental studies

## Scenario

We would like to study the possible relationship between salt and high blood pressure.

## Two types of study

2. Randomized experiment [a.k.a. controlled or designed experiment]
    a. We take a random sample of participants for the study.
    b. We randomly divide them into two groups and perform an experiment:

# Observational vs. experimental studies

## Scenario

We would like to study the possible relationship between salt and high blood pressure.

## Two types of study

2. Randomized experiment [a.k.a. controlled or designed experiment]
    a. We take a random sample of participants for the study.
    b. We randomly divide them into two groups and perform an experiment:
        i. The first group (the treatment group) receives a high-salt diet over the period of study.

# Observational vs. experimental studies

## Scenario

We would like to study the possible relationship between salt and high blood pressure.

## Two types of study

2. Randomized experiment [a.k.a. controlled or designed experiment]
    a. We take a random sample of participants for the study.
    b. We randomly divide them into two groups and perform an experiment:
        i. The first group (the treatment group) receives a high-salt diet over the period of study.
        ii. The second group (the control group) receives a average-salt diet over the period of study.

# Observational vs. experimental studies

## Scenario

We would like to study the possible relationship between salt and high blood pressure.

## Two types of study

2. Randomized experiment [a.k.a. controlled or designed experiment]
    a. We take a random sample of participants for the study.
    b. We randomly divide them into two groups and perform an experiment: [randomization]
        i. The first group (the treatment group) receives a high-salt diet over the period of study.
        ii. The second group (the control group) receives a average-salt diet over the period of study.

# Observational vs. experimental studies

## Scenario

We would like to study the possible relationship between salt and high blood pressure.

## Two types of study

2. Randomized experiment [a.k.a. controlled or designed experiment]
    a. We take a random sample of participants for the study.
    b. We randomly divide them into two groups and perform an experiment: [randomization]
        i. The first group (the treatment group) receives a high-salt diet over the period of study.
        ii. The second group (the control group) receives a average-salt diet over the period of study.
    c. We record the blood pressure of both groups over the period of study.

# Observational vs. experimental studies

## Scenario

We would like to study the possible relationship between salt and high blood pressure.

## Two types of study

2. Randomized experiment [a.k.a. controlled or designed experiment]
    a. We take a random sample of participants for the study.
    b. We randomly divide them into two groups and perform an experiment: [randomization]
        i. The first group (the treatment group) receives a high-salt diet over the period of study.
        ii. The second group (the control group) receives a average-salt diet over the period of study.
    c. We record the blood pressure of both groups over the period of study.
    d. We study the collected data to derive conclusions.

# Observational vs. experimental studies

Suppose the study of the gathered data shows an association between high salt in-take and high blood pressure.

## Question

*Can we conclude that high salt in-take causes high blood pressure?*

# Observational vs. experimental studies

Suppose the study of the gathered data shows an association between high salt in-take and high blood pressure.

## Question

*Can we conclude that high salt in-take causes high blood pressure?*

## Answer

- No (from an observational study)

  [What if there are other factors we have overlooked, which are the hidden cause for both?]

- Possibly yes (from an experimental study)

# Observational vs. experimental studies

Suppose the study of the gathered data shows an association between high salt in-take and high blood pressure.

## Question

*Can we conclude that high salt in-take causes high blood pressure?*

## Answer

- No (from an observational study)

  [What if there are other factors we have overlooked, which are the hidden cause for both?]

- Possibly yes (from an experimental study)

## Remark

*Association does not imply causation!*

# Observational vs. experimental studies

Suppose the study of the gathered data shows an association between high salt in-take and high blood pressure.

## Question

*Can we conclude that high salt in-take causes high blood pressure?*

## Answer

- No (from an observational study)

  [What if there are other factors we have overlooked, which are the hidden cause for both?]

- Possibly yes (from an experimental study)

## Remark

*Association does not imply causation!*

*An observational study can be used to detect association, but not causation (because there can be hidden factors).*

*Causation can only be inferred from a randomized experiment.*

# Placebo and blinding in randomized experiments

## Scenario

We would like to study the effect of a new type of pain-killer medicine in reducing pain in patients with sever migraine headaches.

We use a randomized experiment.

# Placebo and blinding in randomized experiments

## Placebo effect

A fake/dummy treatment which resembles an actual treatment (a placebo treatment) can sometimes lead to significant (though small) improvements in patients.

# Placebo and blinding in randomized experiments

## Placebo effect

A fake/dummy treatment which resembles an actual treatment (a placebo treatment) can sometimes lead to significant (though small) improvements in patients.

## Remark

*If not controlled, the placebo effect can lead to a bias in the result of designed experiments.*

# Placebo and blinding in randomized experiments

## Placebo effect

A fake/dummy treatment which resembles an actual treatment (a placebo treatment) can sometimes lead to significant (though small) improvements in patients.

## Remark

*If not controlled, the placebo effect can lead to a bias in the result of designed experiments.*

## Types of placebo control in experiments

# Placebo and blinding in randomized experiments

## Placebo effect

A fake/dummy treatment which resembles an actual treatment (a placebo treatment) can sometimes lead to significant (though small) improvements in patients.

## Remark

*If not controlled, the placebo effect can lead to a bias in the result of designed experiments.*

## Types of placebo control in experiments

- Single-blind experiment: The control group receive a placebo treatment. [They are blinded!]

# Placebo and blinding in randomized experiments

## Placebo effect

A fake/dummy treatment which resembles an actual treatment (a placebo treatment) can sometimes lead to significant (though small) improvements in patients.

## Remark

*If not controlled, the placebo effect can lead to a bias in the result of designed experiments.*

## Types of placebo control in experiments

- Single-blind experiment: The control group receive a placebo treatment. [They are blinded!]
- Double-blind experiment: In addition to the participants, the experimenters who interact with the participants are also blinded.

# Placebo and blinding in randomized experiments

## Placebo effect

A fake/dummy treatment which resembles an actual treatment (a placebo treatment) can sometimes lead to significant (though small) improvements in patients.

## Remark

*If not controlled, the placebo effect can lead to a bias in the result of designed experiments.*

## Types of placebo control in experiments

- Single-blind experiment: The control group receive a placebo treatment. [They are blinded!]
- Double-blind experiment: In addition to the participants, the experimenters who interact with the participants are also blinded.
- Triple-blind experiment: The participants, the experimenters, and the researchers analyzing the data are all blinded.