



# Lecture 01: Statistical Models

## Statistical Modelling I

Dr. Riccardo Passeggeri

# Outline

---

1. Background and Scope

2. Statistical Models

3. Using Models

# Background and Scope

# Science

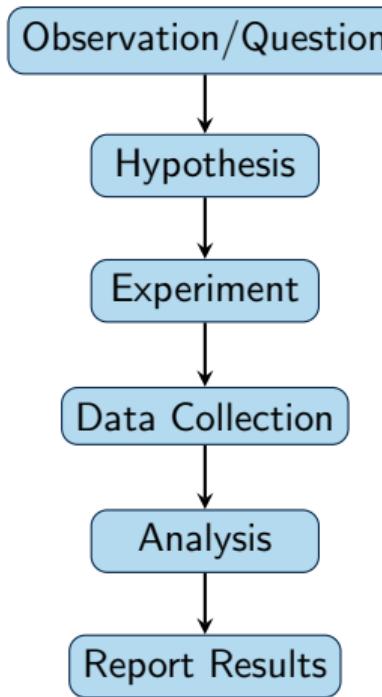
---

How scientists do science?

- ▶ Guess
- ▶ Compute the consequences of the guess
- ▶ Compare them with experiments

# The Scientific Method and Statistics

---



# Statistical modelling

---

Where does the randomness come from?

- ▶ measurement errors and technical noises (Higgs boson discovery announcement  
<https://www.youtube.com/watch?v=0CugLD9HF94> )
- ▶ impossibility of repeating the exact same experiment
- ▶ impossibility of repeating the experiment

# Statistical Answers to Scientific Questions (for this Module!)

---

Quantify distributions

Compare distributions

Predict observations

**Other common tasks:** clustering observations or variables.

# Module Outline

---

- ▶ **1st half:** deriving, evaluating and applying estimators, confidence intervals and hypothesis tests based on parametric models.
- ▶ **2nd half:** deriving, evaluating and applying estimators, confidence intervals and hypothesis tests based on the theory of linear models.

# Statistical Models

# Some Conventions

---

Key: observed data  $y$  is realization of random  $Y$ .

- ▶ Random variables:  $X, Y, Z$
- ▶ Data:  $x, y, z$
- ▶ Parameters:  $\theta, \alpha, \beta$
- ▶ Estimators:  $\hat{\theta}, \hat{\alpha}, \hat{\beta}$
- ▶ Outcome:  $Y, y$
- ▶ Covariate:  $X, x$

# Parametric Models

Let  $P_\theta$  be a probability distribution with parameter  $\theta$ . For example,

## Definition

A **statistical model** is a collection of probability distributions  $\{P_\theta : \theta \in \Theta\}$  over a sample space. The set  $\Theta$  of all possible parameter values is called the **parameter space**. In this module,  $\Theta \subseteq \mathbb{R}^p$ , so that we consider **parametric models**.

A statistical model is generally required to be such that distinct parameter values give rise to distinct distributions, i.e.

# Independent and Identically Distributed (iid)

---

Important special case:

- ▶  $y = (y_1, \dots, y_n) \in \mathbb{R}^n \Rightarrow Y = (Y_1, \dots, Y_n)$  is a random vector
- ▶ Then statistical model specifies the joint distribution of  $Y_1, \dots, Y_n$  up to  $\theta \in \Theta$
- ▶  $Y_1, \dots, Y_n$  is a **random sample** if  $Y_i$ 's are **iid**

## Example: Guess the Weight

---

Observed data



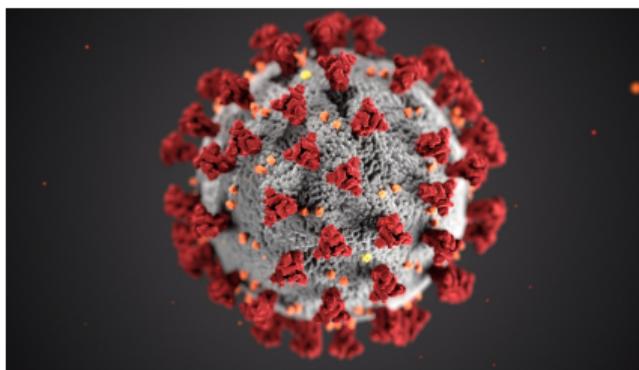
Statistical model

- ▶ People guess the weight of an ox
- ▶  $n$  guesses  $y_1, \dots, y_n$
- ▶ The true weight is 543.4 kg

## Example: Clinical Trial

---

Observed data



Statistical model

- ▶ Compare new and old treatment
  - ⇒ or vaccine
- ▶  $n$  treatment assignments  $x_1, \dots, x_n$
- ▶  $n$  times until recovery  $y_1, \dots, y_n$

# Example: Do Taller People Have Higher Incomes?

Observed data



Statistical model

- ▶ Compare income across heights
- ▶  $n$  heights  $x_1, \dots, x_n$
- ▶  $n$  incomes  $y_1, \dots, y_n$

Background and Scope  
oooooo

Statistical Models  
oooooooo

Using Models  
●oooo

# Using Models

# Remember: Statistical Answers to Scientific Questions

---

Quantify distributions

Compare distributions

Predict observations

# Assessing Statistical Models

---

*"All models are wrong, but some are useful."*

We want our parametric model to

- ▶ agree well with observed data
- ▶ be simple and interpretable

## Looking Ahead

---

How do we “know” which estimators, confidence intervals, and hypothesis tests to use?