Imperial College London

Lecture 01: Statistical Models

Statistical Modelling I

Dr. Riccardo Passeggeri

Outline

1. Background and Scope

2. Statistical Models

3. Using Models

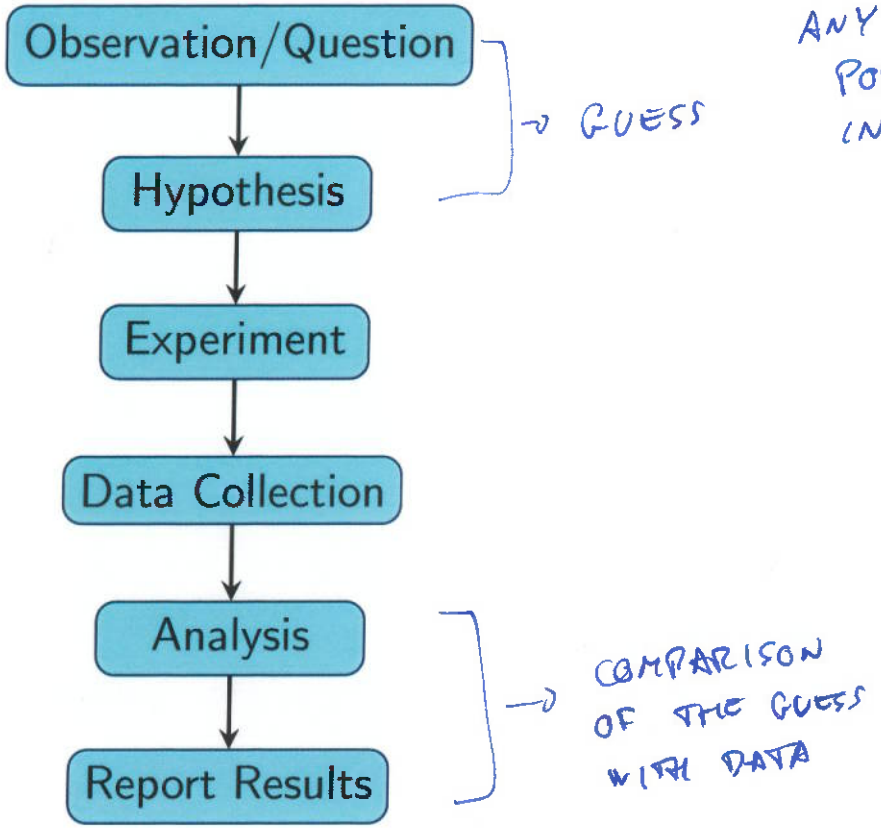
Background and Scope

Science

How scientists do science?

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

The Scientific Method and Statistics



ANY SCIENTIFIC QUESTION NEED TO BE POSED AS A STATISTICAL QUESTION IN ORDER TO BE TESTED

Statistical modelling

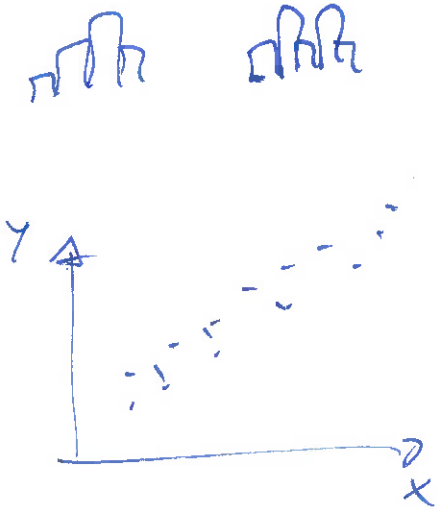
Where does the randomness come from?

- measurament 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

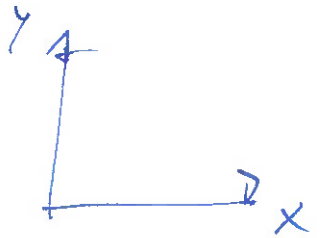
Some Conventions

Key: observed data y is realization of random Y.

$y = (y_1, \dots, y_n)$ $\quad$ $Y = (Y_1, \dots, Y_n)$

- Random variables: X, Y, Z
- Data: x, y, z
- Parameters: θ, α, β
- Estimators: $\hat{\theta}, \hat{\alpha}, \hat{\beta}$

- Outcome: Y, y (DEPENDENT)
- Covariate: X, x (INDEPENDENT)



$Y = \beta_0 + \beta_1 X + \varepsilon$

$(\Omega, \mathcal{F}, \mathbb{P})$

SAMPLE SPACE — σ-ALGEBRA — PROB. MEASURE

$Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$

$Y : \omega \to Y(\omega)$

Parametric Models

$N(\theta, 1)$, $\theta \in \mathbb{R}$

Let P_θ be a probability distribution with parameter θ. For example,

SINCE OBSERVED DATA ARE THE OUTCOME OF AN UNDERLYING R.V. Y, THEN A STATISTICAL MODEL IS THE SPECIFICATION OF THE DISTRIBUTION OF Y UP TO A PARAMETER $\theta \in (H)$

Definition

A **statistical model** is a collection of probability distribtions $\{P_\theta : \theta \in \Theta\}$ over a sample space. The set Θ 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. $P_{\theta_1} = P_{\theta_2} \Rightarrow \theta_1 = \theta_2$

$N(\theta^2, 1)$, $\theta \in \mathbb{R}$

Independent and Identically Distributed (iid)

Important special case:

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

Example: Guess the Weight

Observed data

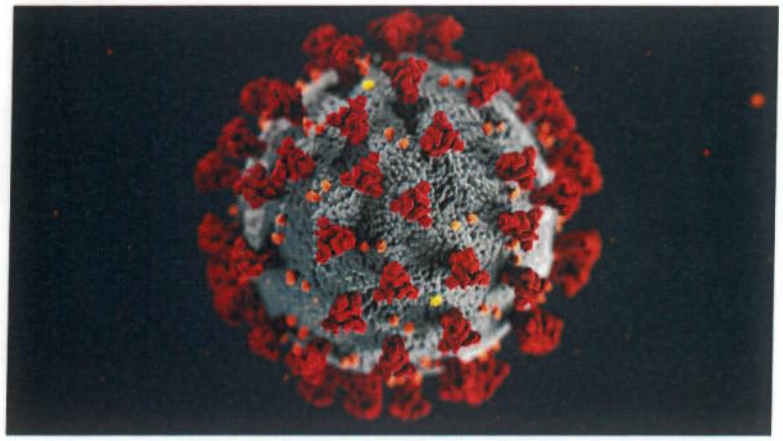
- People guess the weight of an ox
- n guesses $y_1, \ldots, y_n$
- The true weight is 543.4 kg

Statistical model

$N(543.4, \sigma^2)$, $\sigma \in [0, \infty)$

Example: Clinical Trial

Observed data



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

Statistical model

$$x_i = \begin{cases} 0 & \text{NO DRUG} \\ 1 & \text{DRUG} \end{cases}$$

$$Y_i \mid x_i = 0 \sim ? \; N(\mu_0, 1)$$

$$Y_i \mid x_i = 1 \sim ? \; N(\mu_1, 1)$$

SCIENTIFIC Q: DOES THE DRUG HAVE AN EFFECT?

STATISTICAL Q: DOES μ_0 AND μ_1 DIFFER "SIGNIFICANTLY"?

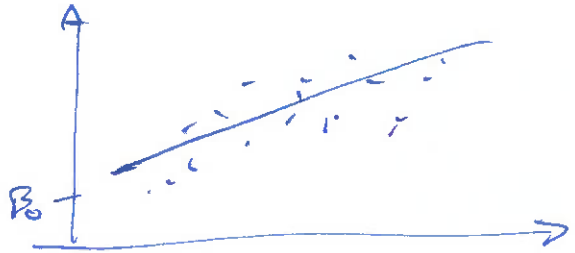
Example: Do Taller People Have Higher Incomes?

Observed data

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

Statistical model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



STAT Q: IS β_1 SIGNIFICANTLY DIFFERENT FROM 0?

Using Models

Remember: Statistical Answers to Scientific Questions

Quantify distributions

OX WEIGHT GUESS

Compare distributions

CLINICAL TRIAL

Predict observations

TALLER PEOPLE HIGHER INCOME

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?