# Mathematical statistics

September $18^{th}$, 2018

Lecture 7: Introduction to parameter estimation

# Where are we?

| | |
|---|---|
| **Week 1** | Probability reviews |
| **Week 2** | Chapter 6: Statistics and Sampling Distributions |
| **Week 4** | **Chapter 7: Point Estimation** |
| **Week 7** | Chapter 8: Confidence Intervals |
| **Week 10** | Chapter 9: Test of Hypothesis |
| **Week 14** | Regression |

# Overview

7.1 Point estimate
- unbiased estimator
- mean squared error

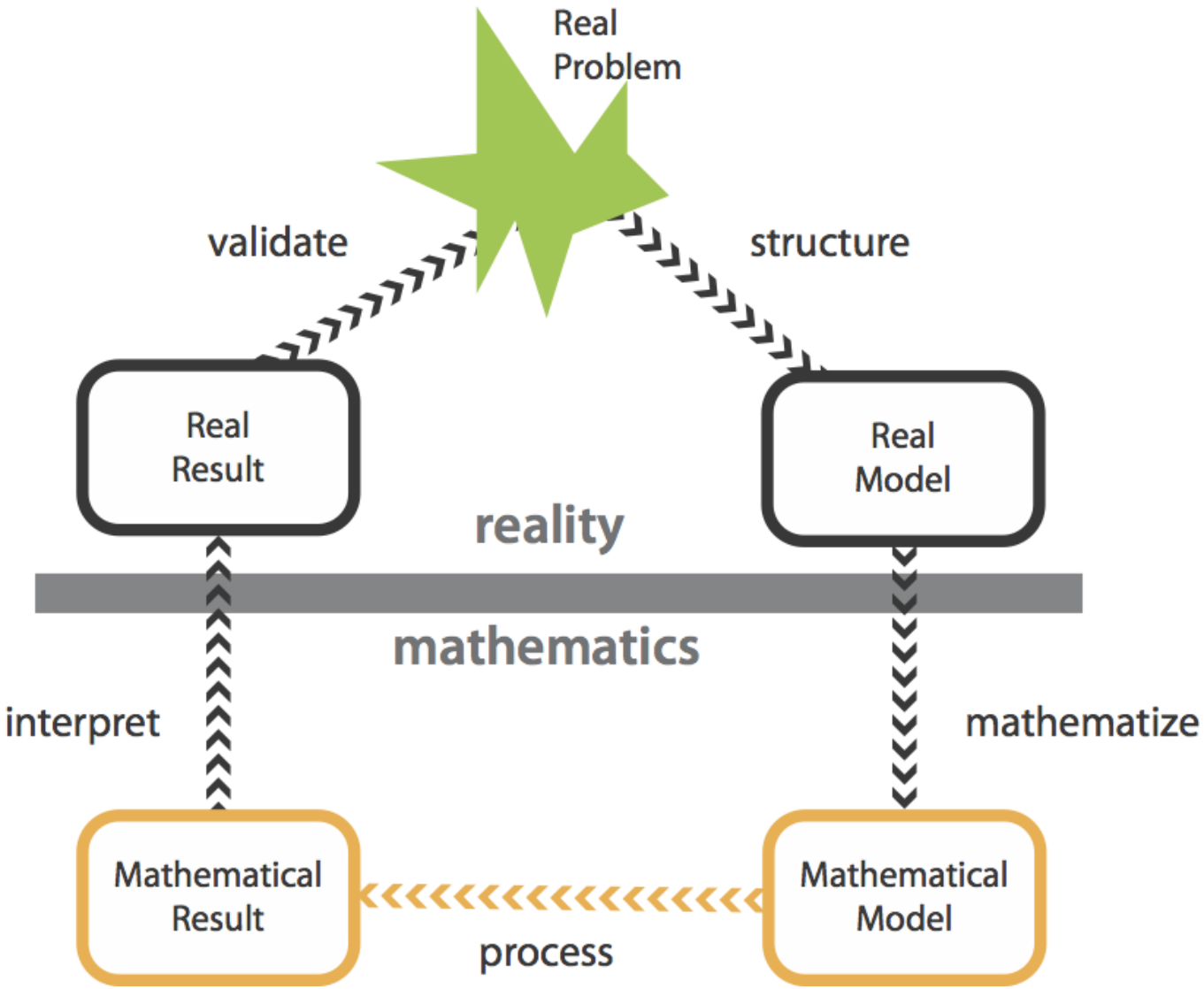
7.2 Methods of point estimation
- method of moments
- method of maximum likelihood.

7.3 Sufficient statistic

7.4 Information and Efficiency
- Large sample properties of the maximum likelihood estimator

- Bootstrap

# Mathematical modelling



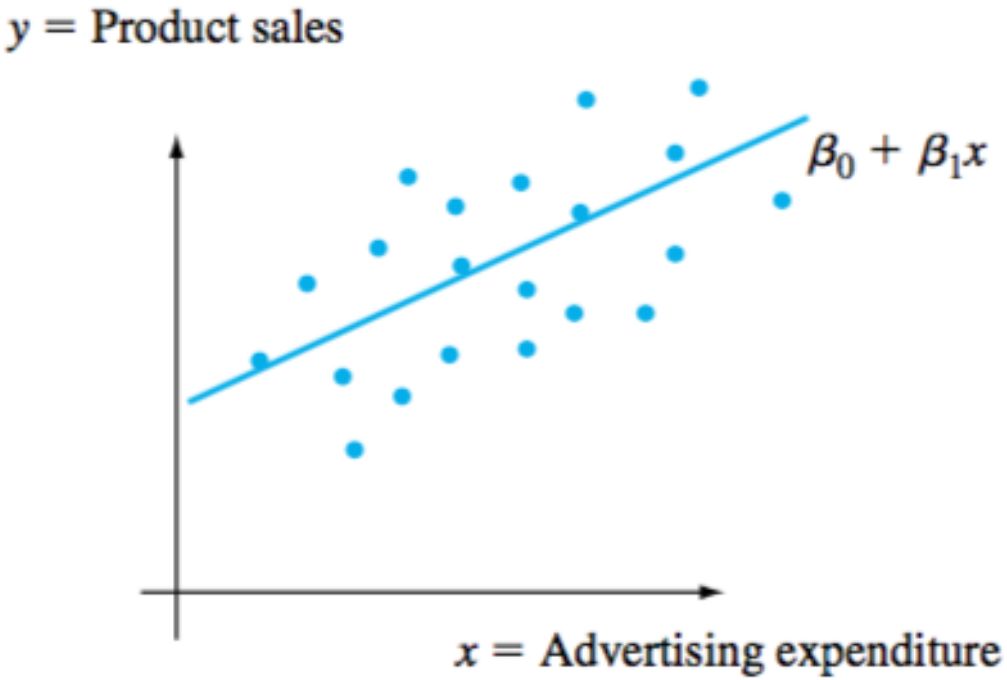
# Parameter estimation

- In a mathematical model, parameters are used to define a whole family of functions that relate the inputs and the outputs
- Example:

$$y = ax + b$$

  represents a family of linear functions parameterized by $(a, b)$
- Parameter estimation: from collected data, determine the values of the parameter

# Deterministic modelling vs. Stochastic modelling

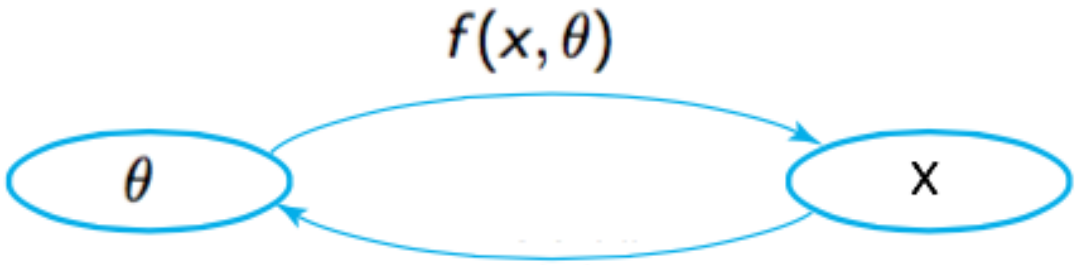



Mathematical model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

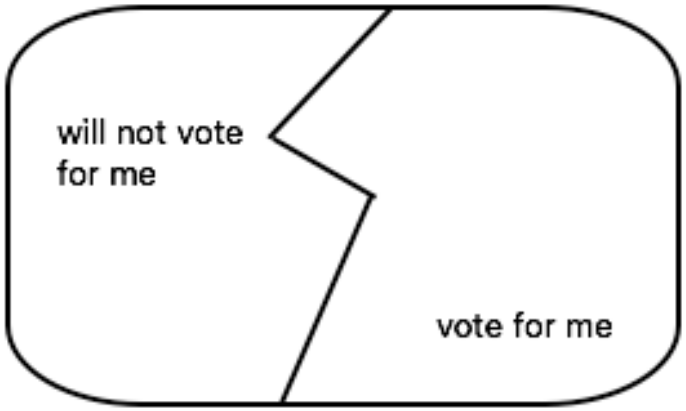
# Question of this chapter

- Given a random sample $X_1, \ldots, X_n$ from a distribution with pmf/pdf $f(x, \theta)$ parameterized by a parameter $\theta$
- Goal: Estimate $\theta$



# Example 1

- Setting: I'm running for president of the US
- I want to estimate how many people support me




- Denote
  - A: the total number of people who will vote for me
  - B: the total number of people who will not

$$p = \frac{A}{A + B}$$

is an unknown quantity that I'm interested in

# Step 1: Random sample

- Choose one random person.
- Record the response by a random variable $X$
  - Yes $\rightarrow X = 1$
  - No $\rightarrow X = 0$
- The pmf of $X$ is as follows

| x | 0 | 1 |
|---|---|---|
| p(x) | 1-p | p |

- Repeat 2000 times $\rightarrow$ a sample $X_1, X_2, \ldots, X_{2000}$
- Obtained data: $x_1 = 1, x_2 = 0, \ldots, x_{2000} = 1$
- Summary statistics: $n_{yes} = 1200$, $n_{no} = 800$
- Question: What is a good estimate of $p$?

# Step 2: Analysis

- A good estimate of $p$ is

$$\hat{p} = \frac{n_{yes}}{n} = \frac{1200}{2000} = 0.6$$

# Step 2: Analysis

- A good estimate of $p$ is

$$\hat{p} = \frac{n_{yes}}{n} = \frac{1200}{2000} = 0.6$$

- A more proper way to write $\hat{p}$

$$\hat{p} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \bar{X}$$

- The strong law of large number

$$\hat{p} = \bar{X} \approx E[X]$$

and

$$E[X] = p.1 + (1 - p).0 = p$$

# Step 2: Analysis

Central Limit Theorem: $(n > 40)$

$$P\left[-1.96 \leq \frac{\hat{p} - E[X]}{\sigma x/\sqrt{n}} \leq 1.96\right] = 95\%$$

or

$$P\left[p - 1.96\frac{1}{\sqrt{n}}\sqrt{p(1-p)} \leq \hat{p} \leq p - 1.96\frac{1}{\sqrt{n}}\sqrt{p(1-p)}\right] = 95\%$$

# Step 2: Analysis

- Simplified expression:

$$P\left[\hat{p} - 1.96\frac{\hat{p}(1-\hat{p})}{\sqrt{n}} \leq p \leq \hat{p} + 1.96\frac{\hat{p}(1-\hat{p})}{\sqrt{n}}\right] = 95\%$$

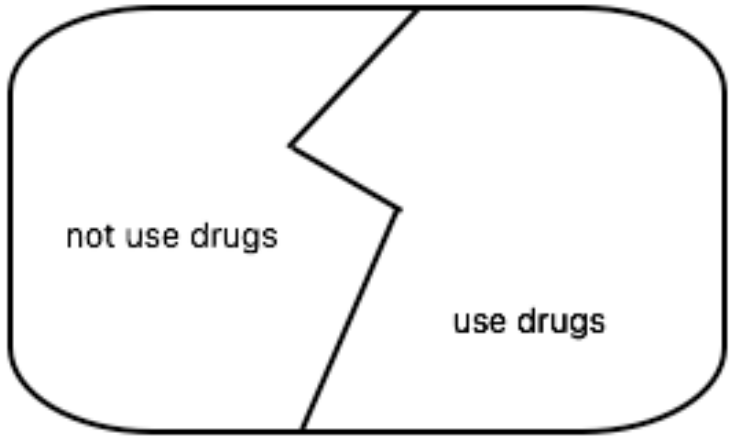
- Plug $\hat{p} = 0.6$ in, we can say

$$0.579 \leq p \leq 0.621$$

with 95% confidence

# Example 2: Private learning

- I want to estimate the number of UD students who have used drugs in the last 6 months




- Denote
  - A: the total number of people who have used drugs
  - B: the total number of people who have not

$$p = \frac{A}{A + B}$$

is an unknown quantity of interest

# Step 1: Random sample

- Choose one random person. Denote their unknown true response by $X$

| x | 0 | 1 |
|---|---|---|
| p(x) | 1-p | p |

- Give them a biased coin that turns head 55% of the time

| z | tail | head |
|---|---|---|
| p(z) | 0.45 | 0.55 |

- Ask the person to toss the coin, see the outcome and does not show it to anyone
  - tail $\rightarrow$ tell the truth
  - head $\rightarrow$ lie
- The outcome is recorded by a random variable $Y$

# Step 1: Random sample

- Repeat 2000 times $\rightarrow$ a sample $Y_1, Y_2, \ldots, Y_{2000}$
- Obtained data: $y_1 = 1, y_2 = 0, \ldots, y_{2000} = 1$
- Summary statistics: $n_{yes} = 1080$, $n_{no} = 920$
- Question: What is a good estimate of $p$?