# Lecture 1:

5 modules:

1) Estimating parameters/distributions of data (5-6 lecs.)

2) Hypothesis testing (5-6 lecs)

Midterm

3) supervised learning (4-5 lecs)

4) unsupervised learning (4 lecs)

5) Introduction to statistical learning Thory. (4 lecs)

---

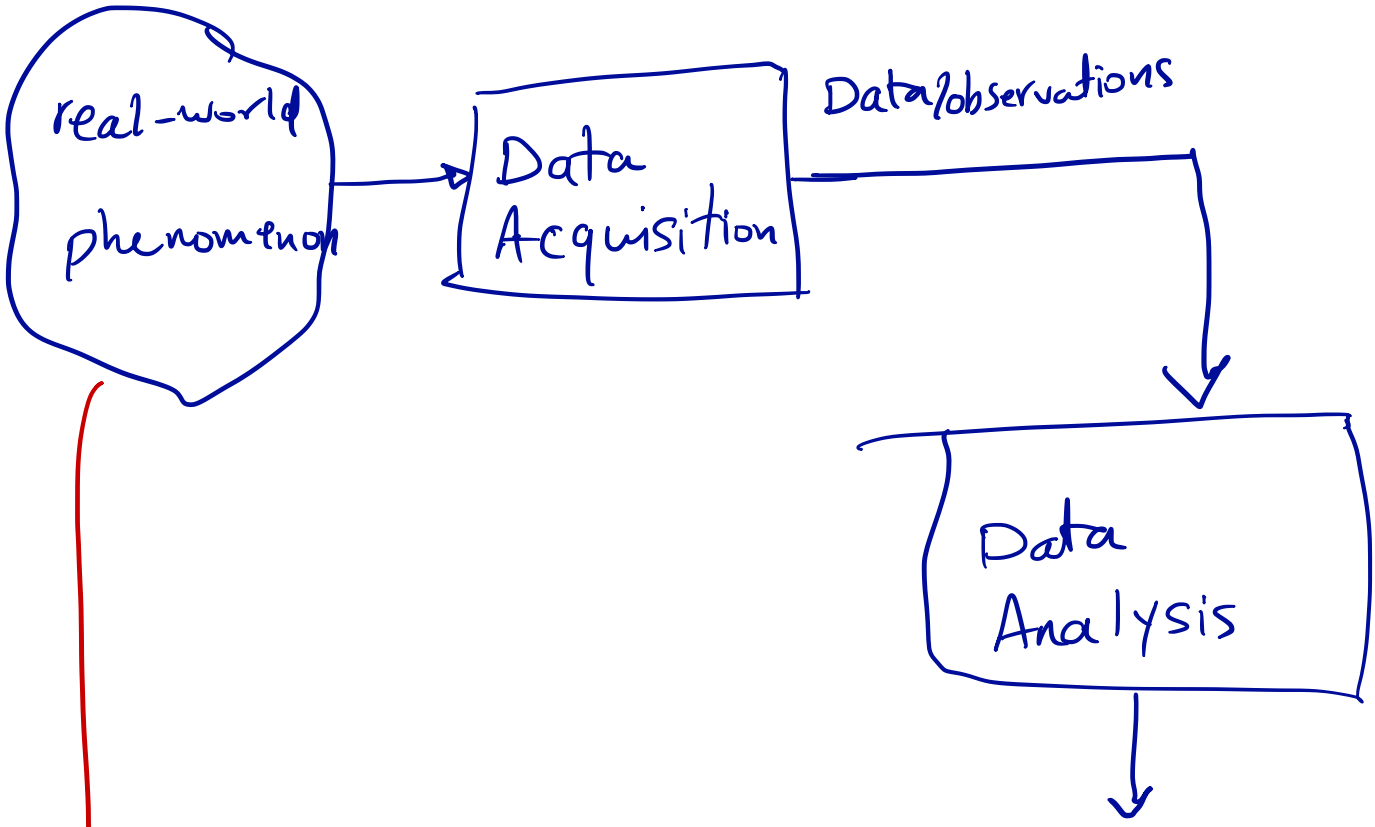
Grading:

Exercises: 40% → 7-8 HWs

Exams (Midterm, Final): 60%.

Textbooks → See Canvas.

# Lecture 2:




Examples of real-world phenomena:

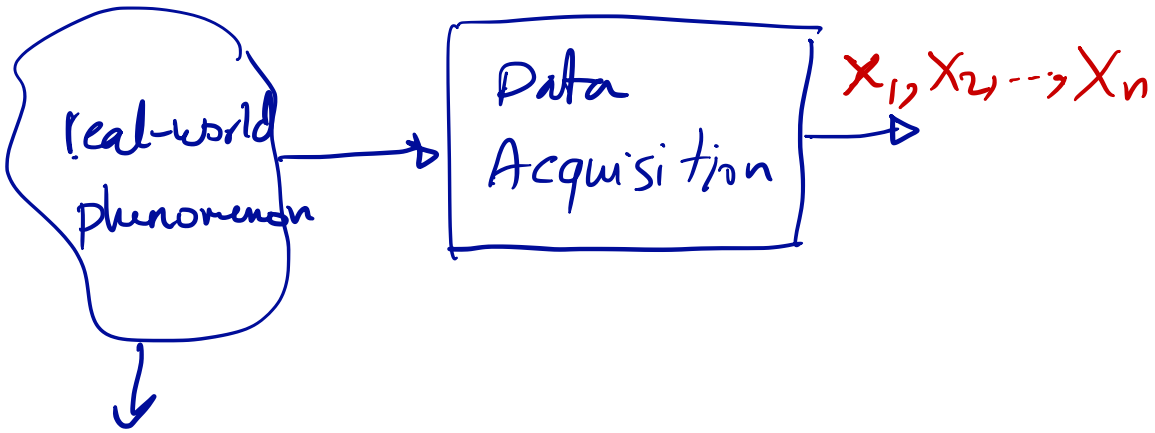
- finding the average hight of women in the US
- precentage of people who'd vote for a certain candidate in the state of Pennsylavania

- likelihood of a person to develop breast cancer within the next 5 years
- average speed of molecules in a specific gas
- likelihood of a user to click on a certain ad in her face book page

Data Science involves developing methods of acquisition / storage / analysis of data to effectively extract useful knowledge / information / insight / patterns.

# Estimating Basic Statistics of Data:

Consider the task of estimation the average hight of women across the US.




N = population size

$$\left( \text{task} = \text{find } \mu = \frac{y_1 + y_2 + \dots + y_N}{N} \right)$$

Typically, N is a very large number (e.g. N = tens of millions)

and the naive method that goes over the whole population is often inefficient or infeasible.

So the basic idea of data aquisition is to sample from the real-world phenomenon.

---

Random sampling: We will choose uniformly at random a subset of the population.

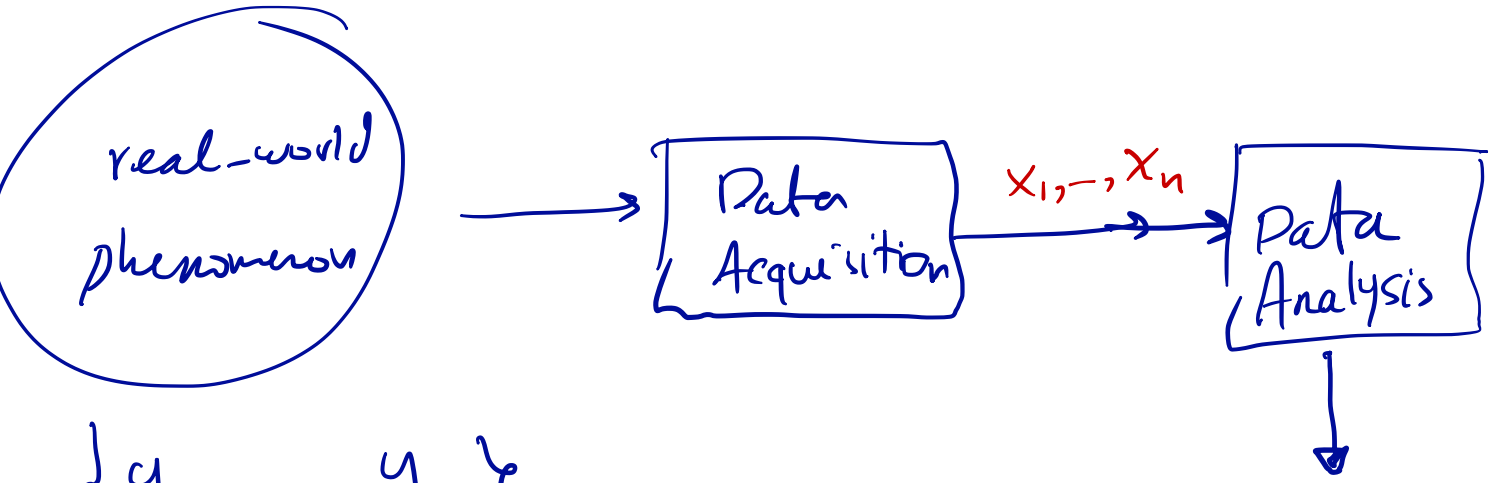
Let $X_i$ be the data corresponding to an individual chosen uniformly at random in the population.

$$X_i = \begin{cases} y_1 & \text{w.p. } \frac{1}{N} \\ y_2 & \text{w.p. } \frac{1}{N} \\ \vdots & \\ y_N & \text{w.p. } \frac{1}{N} \end{cases}$$

furthermore $X_i$'s are independently chosen.

Data: $X_1, X_2, \dots, X_n$

Typically $n$ is much smaller than $N$.



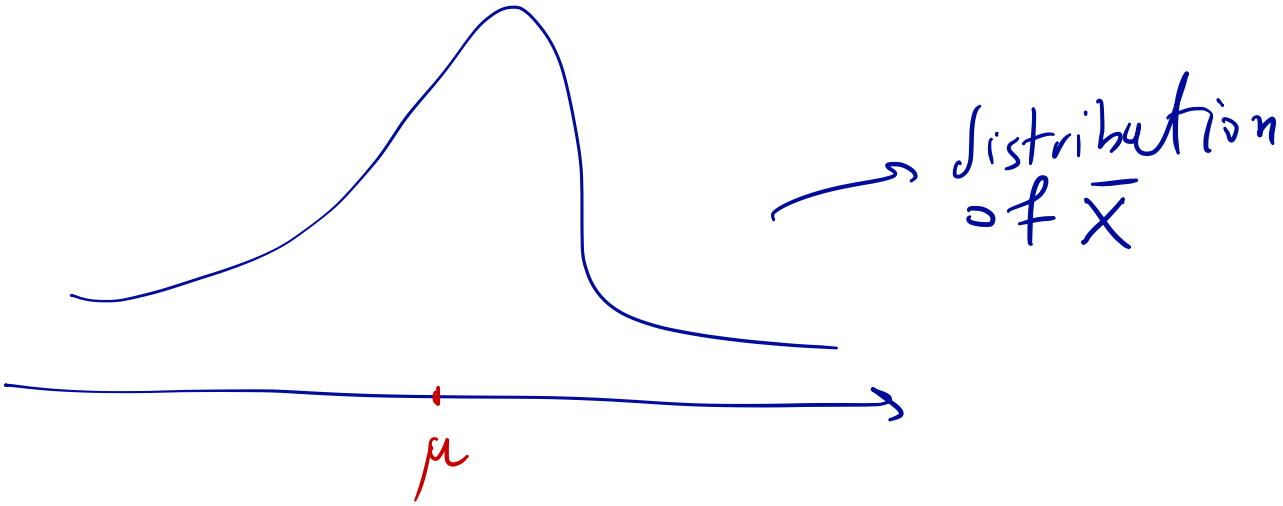

$\{y_1, \ldots, y_N\}$

question : Estimate the average of $y_i$'s (i.e. $\mu$) by only using the data $x_1, \ldots, x_n$.

simplest answer:

$$\bar{X} = \frac{x_1 + \cdots + x_n}{n}$$ (sample mean)

Note that $\bar{X}$ is a random object/variable.




Let's first compute the expected value of $\bar{X}$ and how it is related to $\mu$.

$$E[\bar{X}] = E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right]$$

$$= \frac{E[X_1] + \cdots + E[X_n]}{n}$$

Since $X_i$'s have exactly the same distribution, then

$$E[X_1] = \cdots = E[X_n]$$

---

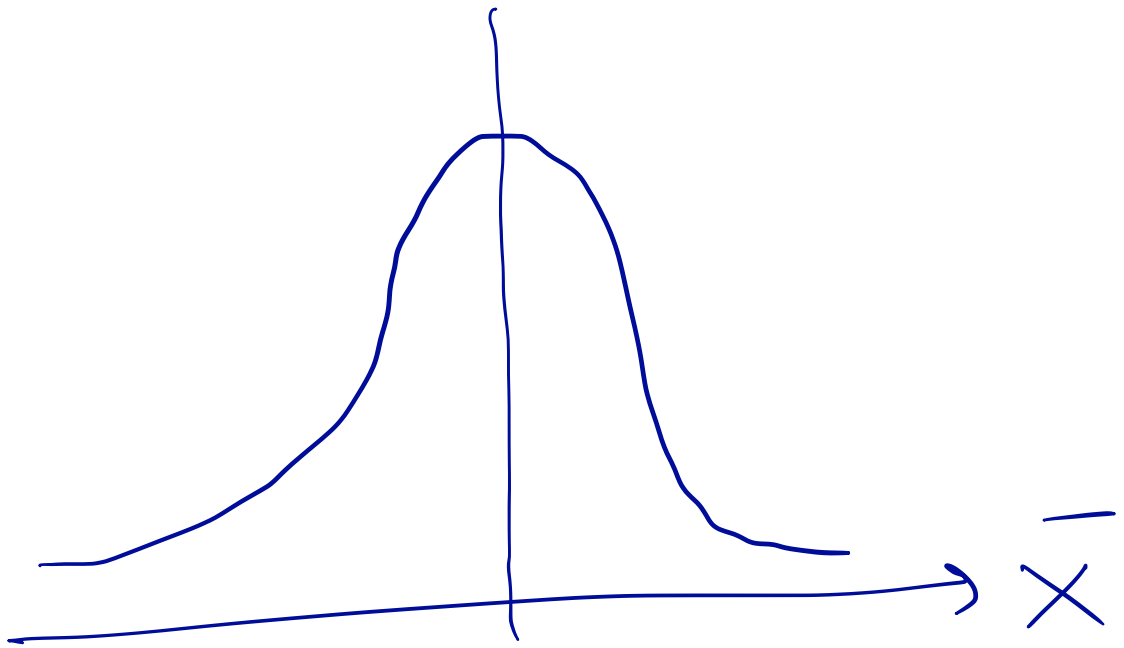
$$X_i = \begin{cases} y_1 & \frac{1}{N} \\ y_2 & \frac{1}{N} \\ \vdots & \\ y_N & \frac{1}{N} \end{cases}$$

$$\Rightarrow E[\bar{X}] = \frac{E[X_1] + E[X_1] + \cdots + E[X_1]}{n}$$

$$= E[X_1]$$

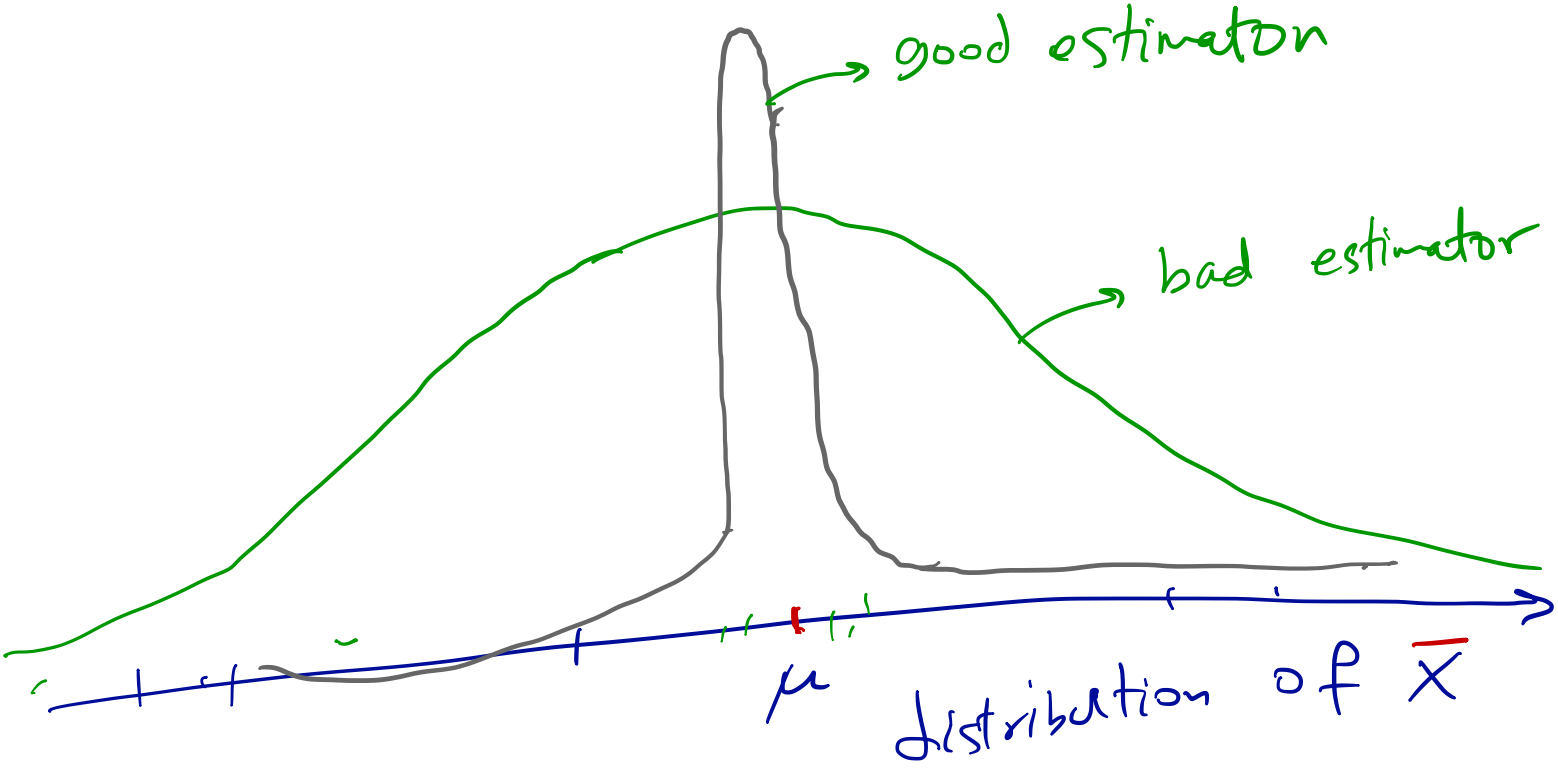
$$= \frac{1}{N} \cdot y_1 + \frac{1}{N} y_2 + \cdots + \frac{1}{N} y_N$$

$$= \frac{y_1 + \cdots + y_N}{N} = \mu$$




$$E[\bar{X}] = \mu$$

So, $\bar{X}$ is on average a very good estimator because it's expectation is $\mu$. The next question is how concentrated is $\bar{X}$ around its average $\mu$.




In order to see how concentrated $\bar{X}$ is around $\mu$, we need to compute ...

$$E\left[(\bar{X}-\mu)^2\right].$$