Mathematical Statistics and Data Analysis

Lecture 1: Basic definitions, notations and ideas

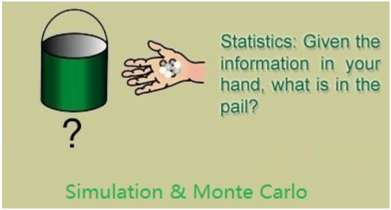
Lyu Ni

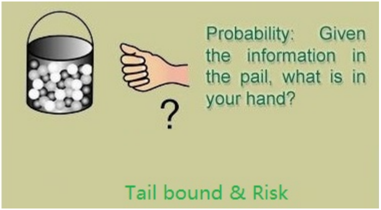
DaSE@ECNU
(lni@dase.ecnu.edu.cn)

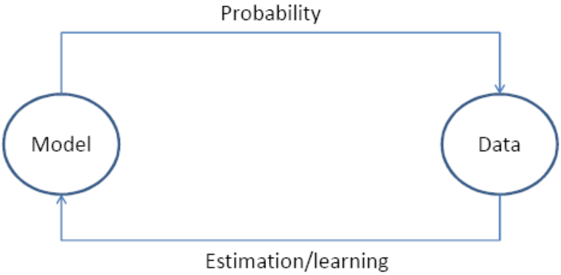
September 2, 2019

Outline

Introduction







Population vs Sample

- Population is a distribution.
- Sample space: the set of all possible outcomes. The sample space is denoted by Ω, and an element of Ω is denoted by ω.
  - Number of people in a queue, $\Omega = \{0, 1, 2, \cdots\}$
  - The time until the event is of interest, $\Omega = \{t | t \geq 0\}$.
  - Measure error, $\Omega = (-\infty, \infty)$.
- One-dimension, two-dimension and multi-dimension.
- Low dimension and high dimension.

Population vs Sample

- The model of measure error:

$$X = \mu + \varepsilon,$$

  where X is measured, μ is a quantity of interest and ε is error.
  - **Parameter**: μ is a sure but unknown quantity;
  - The assumption of ε:
    - The most frequent assumption: $\varepsilon \sim N(0, \sigma^2)$. Thus,

$$X \sim N(\mu, \sigma^2),$$

      where μ and σ^2 are two unknown parameters.
    - Suppose we know the population variance. Assume that $\varepsilon \sim N(0, \sigma_0^2)$, where σ_0^2 is a known constant.
    - Assume that ε has a symmetric distribution with zero.

Population vs Sample

- The model of measure error:

$$X = \mu + \varepsilon,$$

  where X is measured, μ is a quantity of interest and ε is error.
  - **Parameter**: μ is a sure but unknown quantity;
  - The assumption of ε:
    - The most frequent assumption: $\varepsilon \sim N(0, \sigma^2)$. Thus,

$$X \sim N(\mu, \sigma^2),$$

      where μ and σ^2 are two unknown parameters.
    - Suppose we know the population variance. Assume that $\varepsilon \sim N(0, \sigma_0^2)$, where σ_0^2 is a known constant.
    - Assume that ε has a symmetric distribution with zero.

Population vs Sample

- We extract n individuals from a population, denoted as $x_1, x_2, \cdots, x_n$.
- $x_1, x_2, \cdots, x_n$: **sample**.
- n: **sample size**.
- Simple Random Sampling:
  - Randomization;
  - Independence;
- $x_1, x_2, \cdots, x_n$ are **i**ndependently and **i**dentically **d**istributed (i.i.d) variables from a common distribution, F.
- The joint **c**umulative **d**istribution **f**unction (c.d.f) is

$$F(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F(x_i)$$

Population vs Sample

Example

- There is a batch of N products. We would like to investigate product defect rate p.
- We randomly pick up n products $(x_1, x_2, \cdots, x_n)$ to verify whether each product is qualified or not.
- The qualified product is denoted as 0; otherwise, it is denoted as 1.
- The population is a Bernoulli distribution, that is

$$P(x_i = 1) = p, P(x_i = 0) = 1 - p, i = 1, 2, \cdots, n$$

- Sample with replacement: $x_1, x_2, \cdots, x_n$ are i.i.d.
- Sample without replacement.

Population vs Sample

Example

- There is a batch of N products. We would like to investigate product defect rate p.
- We randomly pick up n products $(x_1, x_2, \cdots, x_n)$ to verify whether each product is qualified or not.
- The qualified product is denoted as 0; otherwise, it is denoted as 1.
- The population is a Bernoulli distribution, that is

$$P(x_i = 1) = p, P(x_i = 0) = 1 - p, i = 1, 2, \cdots, n$$

- Sample with replacement: $x_1, x_2, \cdots, x_n$ are i.i.d.
- Sample without replacement.