**CUHK STAT1013: Data Science Toolbox** Spring 2023

# Statistics in Python I

Overview - Statistics behind Data Science

Lecturer: Ben Dai

Statistics is a subject or tool for learning data into information. Before introducing fancy statistical methods to deal with massive data, we begin with some important probabilistic concepts and statistical frameworks to describe data.

## 1 Population

We will use the probabilistic term *experiment* in a general way to refer to some process, natural phenomena and procedure that produces a random outcome. For example,

- Outcome of flipping a coin
- Birth weight of a newborn
- Effect of Covid vaccine

**Definition 1.1** (Population)**.** A population $\Omega$ is a set of units or events which is of interest for an experiment.

An *experiment* usually can be represented as a random variable. A *random variable* is a function mapping the population to the real numbers, and its *distribution* is introduced to describe a relationship between events and the corresponding probabilities of those events.

**Types of random variables.** A random variable either has discrete or continuous range. Therefore, we have two types of random variables - *discrete* and *continuous*. Let's see the following examples.

- Flip a coin
  - Population: $\{\text{Head}\}$, $\{\text{Tail}\}$
  - Random Variable: $X \in \{0,1\}$ maps $\{\text{Head}\}$ to 0 and $\{\text{Tail}\}$ to 1
- Birth weight of a newborn
  - Population: $\{\text{all newborns}\}$
  - Random Variable: $X \in \mathbb{R}$ maps a newborn to a real number
- For *discrete random variables*, the distribution is quantified by a probability mass function:

$$p_X(x) := \mathbb{P}(X = x), \quad \text{for all } x \in \mathscr{X}.$$

- For *continuous random variables*, the distribution is quantified by a probability density function:

$$f_X(x) := F_X'(x), \quad \text{where } F_X(x) := \mathbb{P}(X \leq x).$$

**Examples of random variables.** Following are some examples of random variables:

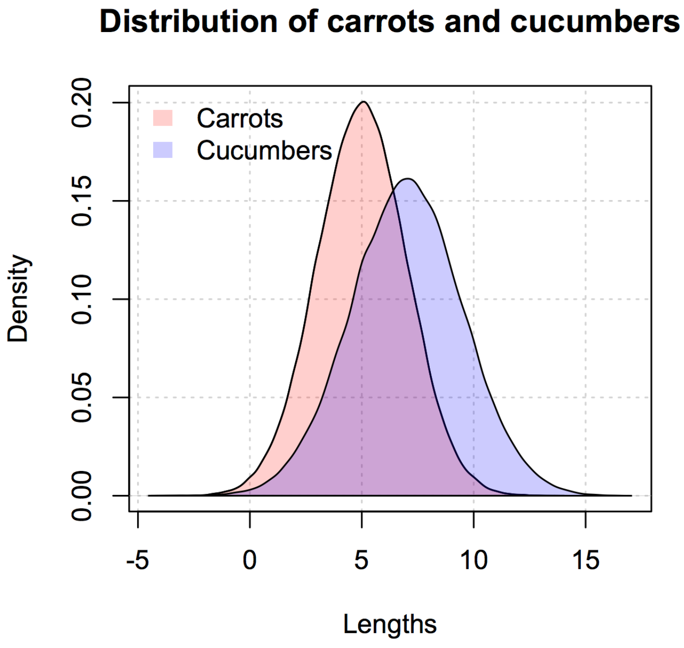
- Bernoulli Random Variable.
  - $X \sim \text{Bernoulli}(p)$, where $p$ represents the success probability.
  - Its probability mass function is defined as:

  $$\mathbb{P}(X=0) = 1-p, \quad \mathbb{P}(X=1) = p.$$

  - Flipping a coin with $\{\text{head}\}$ denoted as 1, and $\{\text{tail}\}$ denoted as 0, is a Bernoulli random variable
- Normal distribution
  - $X \sim N(\mu, \sigma^2)$, where $\mu$ and $\sigma$ represent the mean and standard deviation of the normal distribution.
  - Its probability density function is defined as:

  $$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

  - Length of a vegetable. Randomly select a carrot and measure its length following a normal distribution.



*Remark* 1.2. For most real applications, it is almost impossible to know the distribution of the random variable. The primary goal of statistics is to: using data/sample to learn/infer about the distribution or quantity of interest (e.g., expectation and variance) at the population level. For example, we want to know the effective percentage of a Covid vaccine.

# 2 Sampling

To infer the distribution of a random variable at the population level, we need sampling data from the population. We begin with a simple example.

**Example 2.1** (Polling)**.** *Suppose we want to know what percentage of HK adults own a laptop. Since it's impossible to ask all adult, we instead take a poll of 50 students from STAT1013.*

- *What is the population?* $\rightarrow$ *All HK adult.*
- *What is the sample?* $\rightarrow$ *The 50 students from STAT1013.*
- *Is this a good sample?* $\rightarrow$ *No!*

Independent and identically distributed (i.i.d.) sampling is a "good" way of obtaining a sample that is representative of the population.

**Definition 2.2** (Random sampling)**.** Given a random variable $X$, the collection of random variables $(X_1, \cdots, X_n)$ is said to be a random sample w.r.t. $X$ of size $n$ if they are independent and identically distributed (i.i.d.), i.e.,

1. $X_1, \cdots, X_n$ are independent random variables, and
2. they have the same distribution with $X$.

Again, in statistics we typically do not know $f_X(x)$ and want to use the random sample to infer something about $f_X(x)$.

**Example 2.3** (Descriptive Statistics)**.** *The first example of using the sample estimate to infer the population is descriptive statistics. Specifically, we may want to know the expectation (or population mean) of $X$ at the population level, i.e.,*

$$\mathbb{E}(X) = \int x f_X(x) dx.$$

*Since we have no idea about the population distribution of $X$, the population mean is, of course, an unknown constant. Yet, given a random sample $X_1, \cdots, X_n$, we can estimate it by the sample mean:*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

*More examples are included in the following table.*

| | *Population* | *Sample* |
|---|---|---|
| *Expectation* | $\mu = \mathbb{E}(X)$ | $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ |
| *Conditional Expectation* | $\mathbb{E}(X\|Y=0)$ | $\frac{1}{n}\sum_{i\in i:y_i=0} x_i$ |
| *Variance* | $\sigma^2 = \mathbb{E}(X-\mu)^2$ | $\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2$ |
| *Distribution* | $F_X(x) = \mathbb{P}(X \leq x)$ | $F_n(x) = \#(x_i \leq x)/n$ |

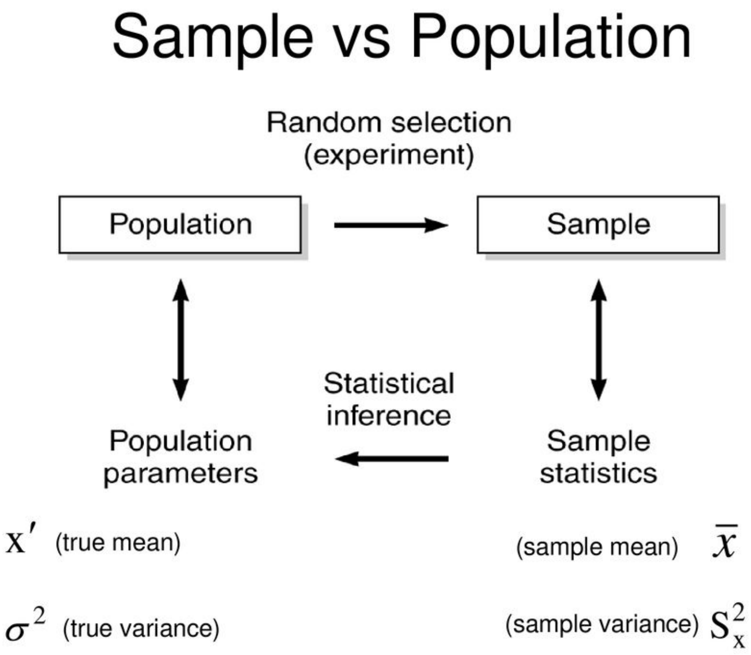


Figure 1: The relation between population and sample. [Source]

But how good the estimation is? Can we trust the *sample estimate*? What's the confidence?