통계분석

Statistical Analysis

Syllabus

- Time : 10:00~12:00, Wed
- Venue : Online (Google Meet)
- Homepage : https://class.ust.ac.kr

Grade: Letter Grades (A, B, C, D, F)

- Attendance : 20%
- Assignments : 40%
- Exams : Mid-term (20%) & Final (20%)

A+	95 ≤ pts ≤ 100	A0	90 ≤ pts < 95
B+	85 ≤ pts < 90	B0	80 ≤ pts < 85
C+	75 ≤ pts < 80	C0	70 ≤ pts < 75
D+	65 ≤ pts < 70	D0	60 ≤ pts < 65
F	pts < 60		

Academic Integrity

- Please DO NOT cheat at assignments and exams
  - Do not copy other students' solutions or any other materials on the web.
  - Solve assignments and exams on your own with your originality.
- If your cheating is found, then F(fail) will be given.

A+	95 ≤ pts ≤ 100	A0	90 ≤ pts < 95
B+	85 ≤ pts < 90	B0	80 ≤ pts < 85
C+	75 ≤ pts < 80	C0	70 ≤ pts < 75
D+	65 ≤ pts < 70	D0	60 ≤ pts < 65
F	pts < 60		

Course Schedule

Mar.	8	15	22	29		4
Apr.	5	12	19(?)	26		4
May.	3	10	17	24	31	5
Jun.	7	14	21	28		4

50 mins lecture	10 mins break
50 mins lecture	Q&A (10mins)

Probability and Statistics

Probability and Statistics

- Statistics (통계학)?

> The study of the collection, analysis, interpretation, presentation, and organization of **data**
>
> From https://wikipedia.org

> 수량적인 비교를 기초로 많은 사실(**데이터**)을 관찰하고 처리하는 방법을 연구하는 학문
>
> 제대로 시작하는 기초통계학, 노경섭

Data (데이터, 자료)

- 자료(Data) = “문제 해결을 위한 원재료로서 **처리되지 않은** 문자나 숫자 또는 일련의 사실이나 기록들의 **모임**”

  Collections of unprocessed text, numbers, a series of facts, or documents.

- 원소(element) = “자료를 이루는 기본 구성 단위” 혹은 최소 단위

  Basic single or minimal unit that constitutes data

- 모집단(population) = “관심의 대상이 되는 **모든** 원소들의 집합”

  A set of all items or events which is of interest for some experiment

Population (모집단)

- 모집단(population) = “관심의 대상이 되는 **모든** 원소들의 집합”

  A set of all items or events which is of interest for some experiment

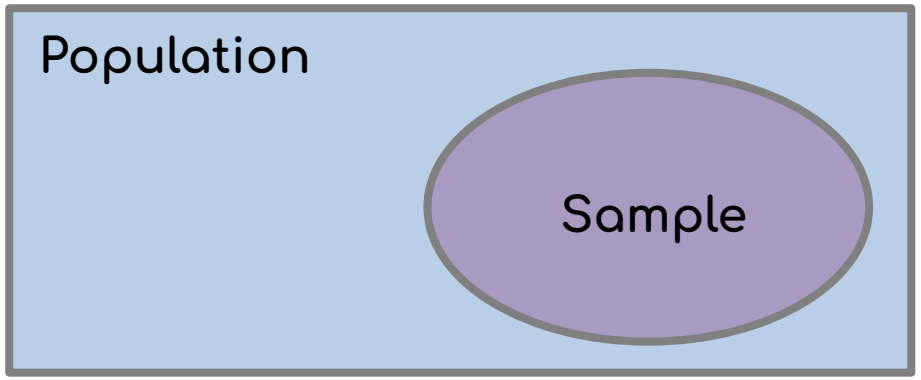
- Set of all stars within the Milky Way galaxy
- All individuals who received a B.S. in engineering last year
- All automobiles sold in Korea last year
- All burgers sold at McDonald’s in Daejeon

Population (모집단) and Sample (표본)

- It would be better to collect data or information from a population of our interest. But, due to limitations on time, money, and other resources, the **"population survey" (전수조사)** is not the ideal one.

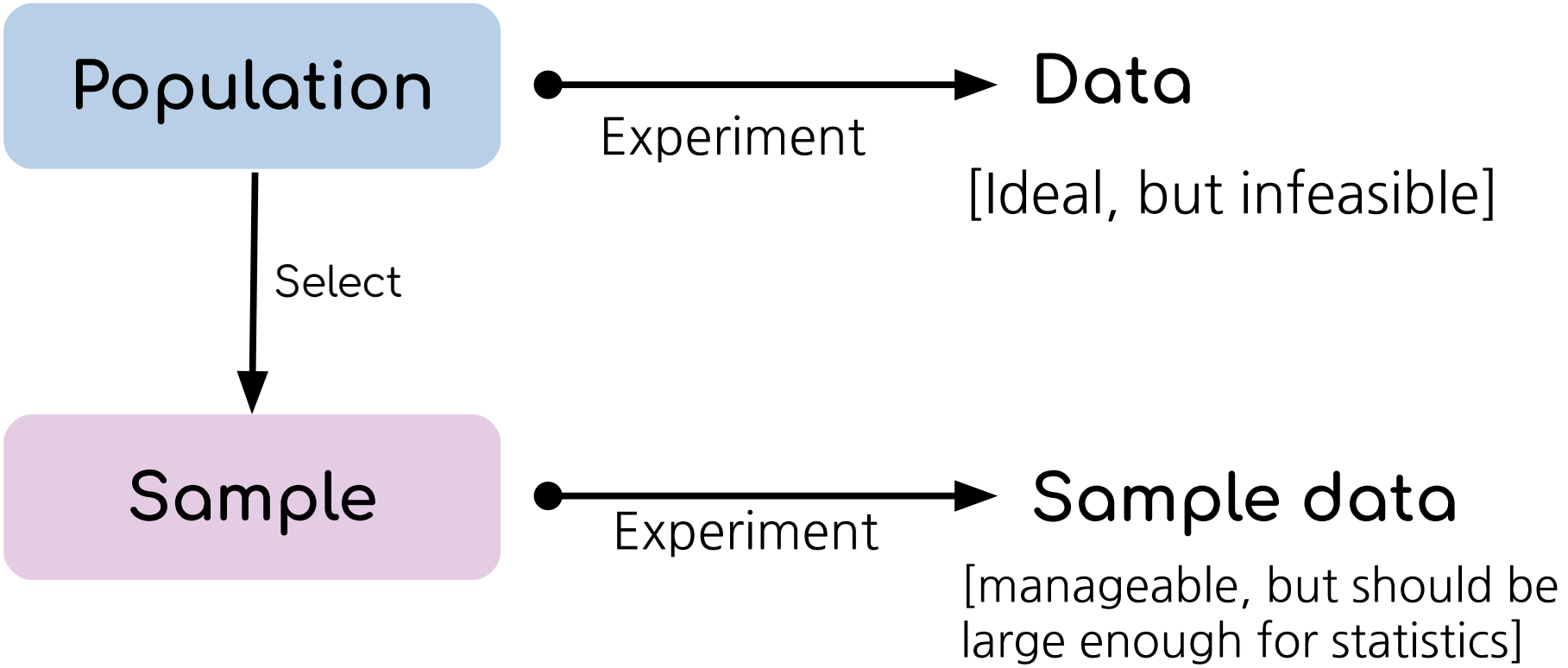
- Instead, some part (subset) of the population is collected, and related survey is performed on the subset. Such a subset is called a **sample (표본)**, and the survey on the sample is the **"sample survey" (표본조사).**

- **Sample =** A subset of the population, on which real experiments or surveys will be conducted.

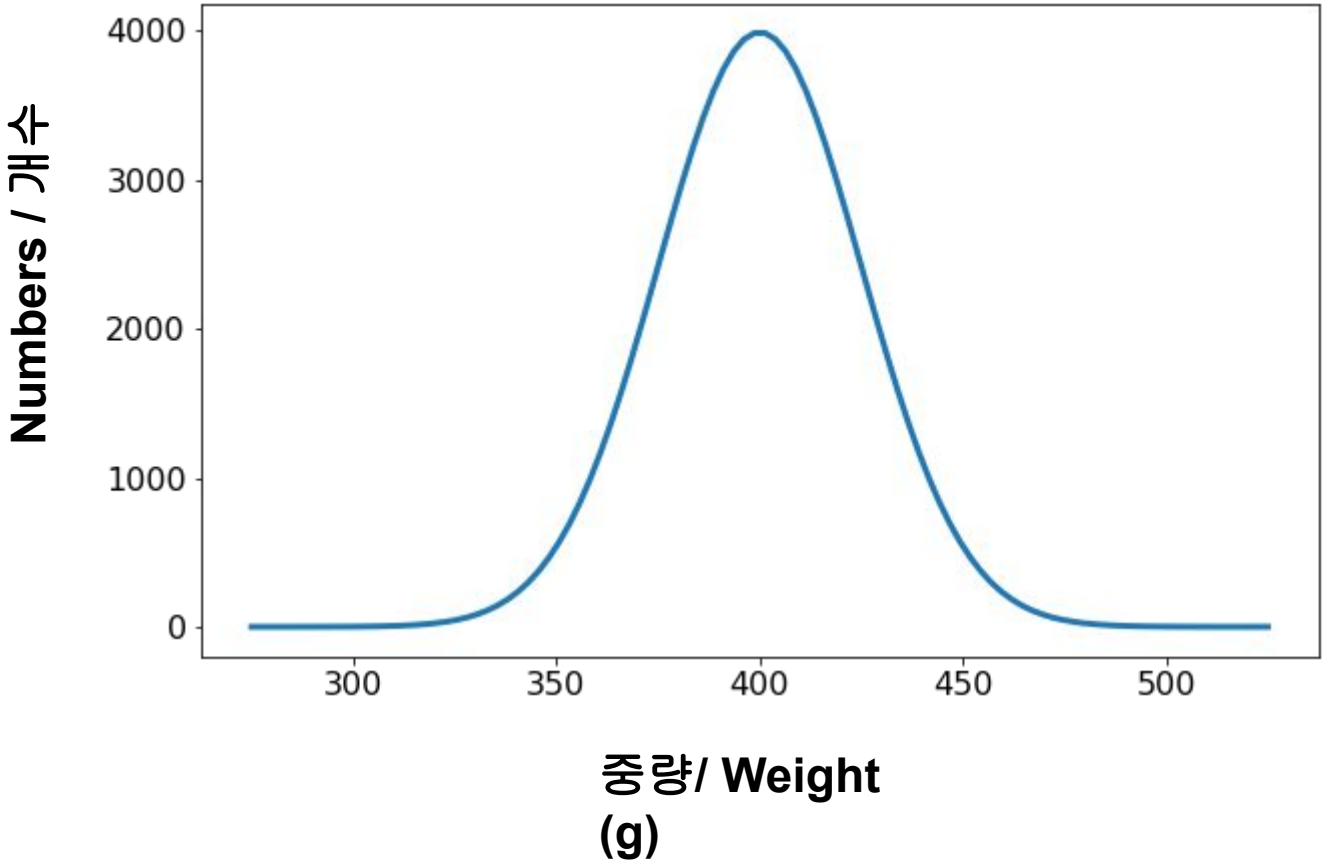


Population (모집단) and Sample (표본)

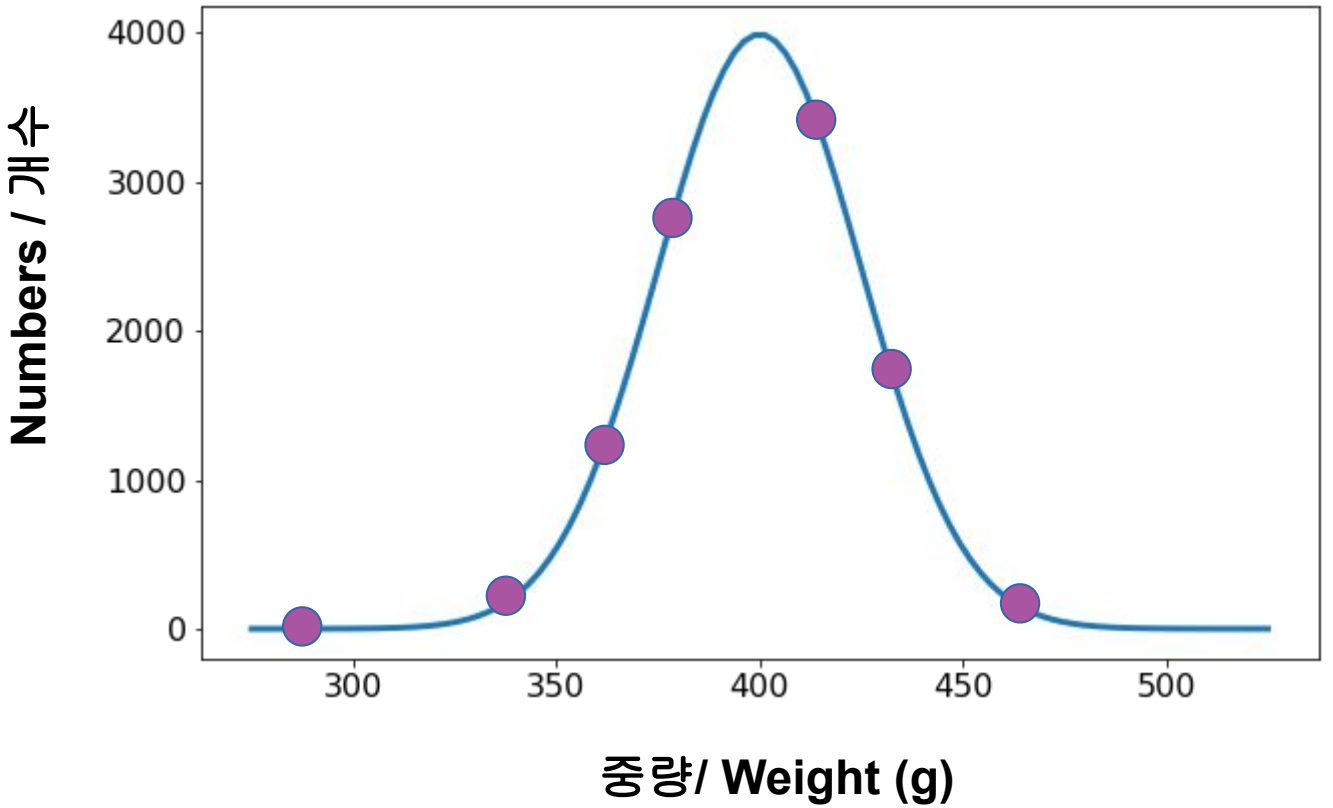
Population (모집단) and Sample (표본)



Where does Probability come in?



표본조사 (Sampling Survey)



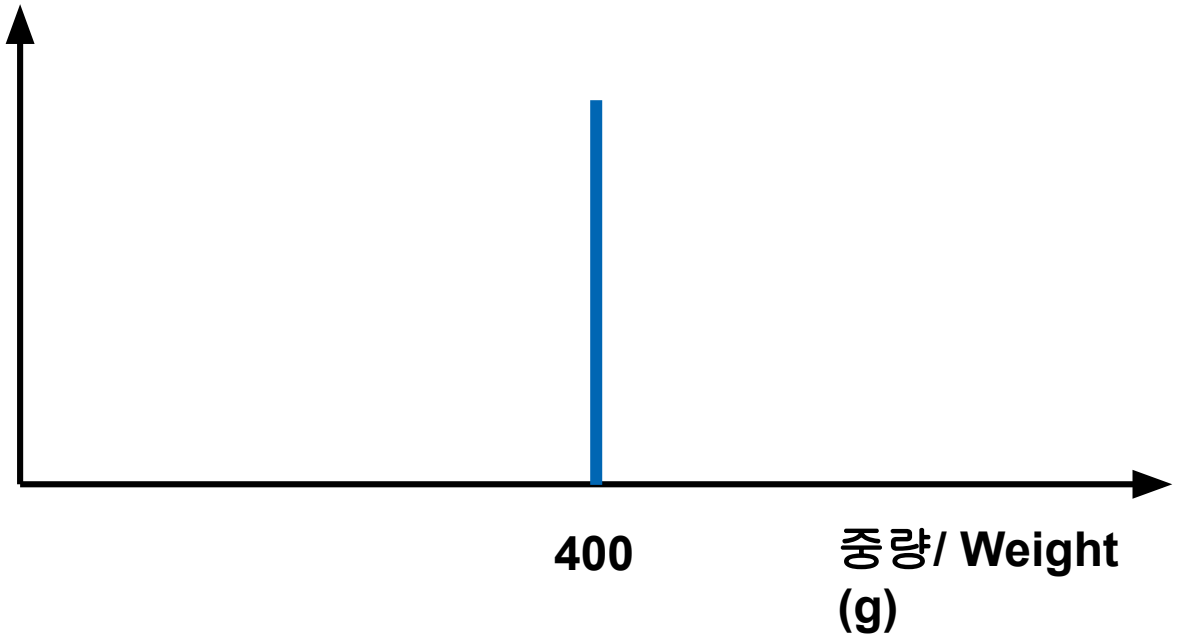
Random Sampling

- Choose seven hamburgers randomly.

Probability and Statistics

Experiment / 실험
(Survey / 조사) → **Outcome (결과)**

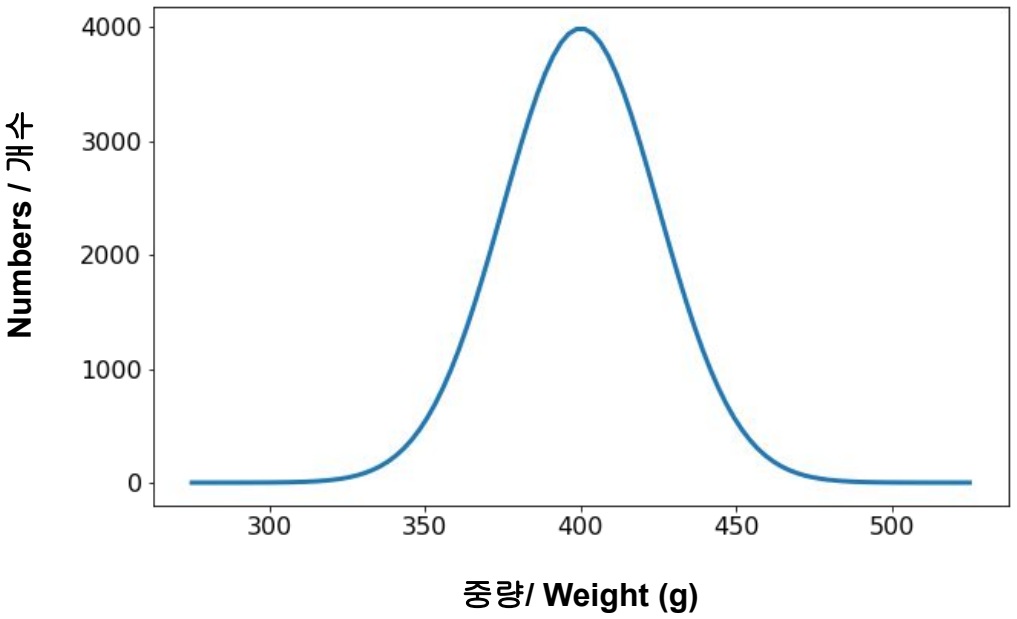
- **Can be predicted with certainty**
- **Boring data, not interest of Statistics**
- 표본조사를 하지 않아도 된다.



Probability and Statistics

(Random) Experiment → **Outcome**

- **Cannot be predicted with certainty**
- **Outcomes with uncertainty**



Probability and Statistics

- **Even though there is uncertainty on sample data from random sampling or random experiments, characteristics of population can be inferred from sample data.**
- **Inferred properties of population is NOT necessarily 100% correct due to the uncertainty of random sampling.**
- **There is a probability that the inference is true, and the inferred conclusion can possibly have some error.**

통계의 종류: Branches of Statistics

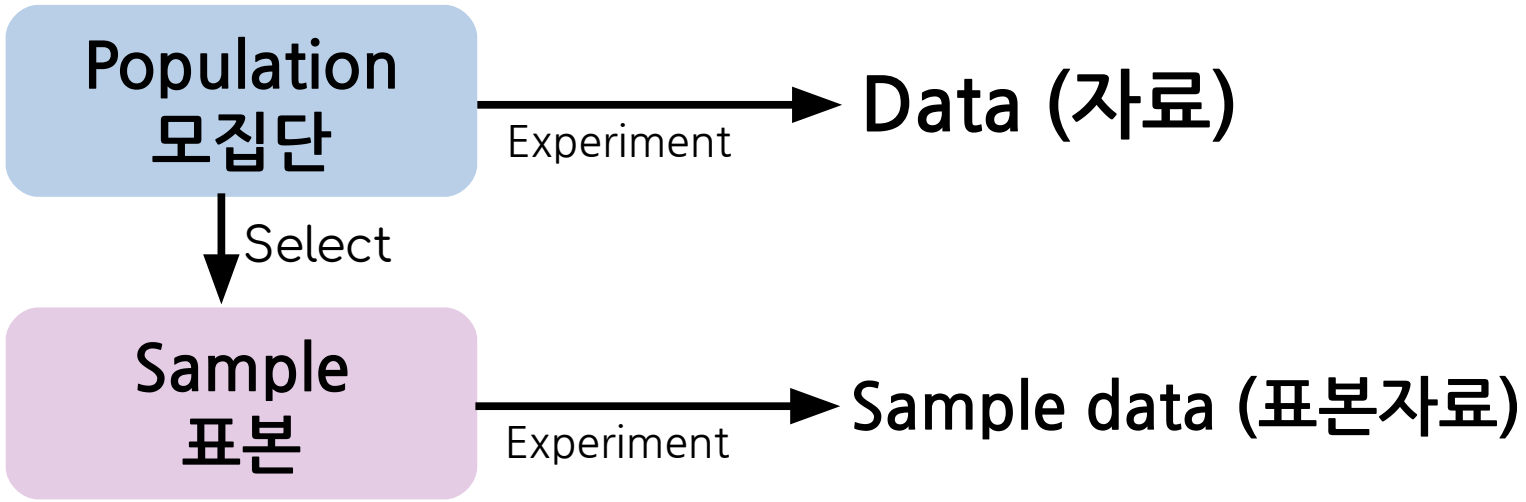
기술통계 Descriptive Statistics

- Organizing data
- Visualizing data
- Quantifying data
- Summarizing data

추론통계 Inferential Statistics

- Drawing conclusions on a population from sampling data
- Generalization from a sample to a population

Descriptive Statistics: 기술통계



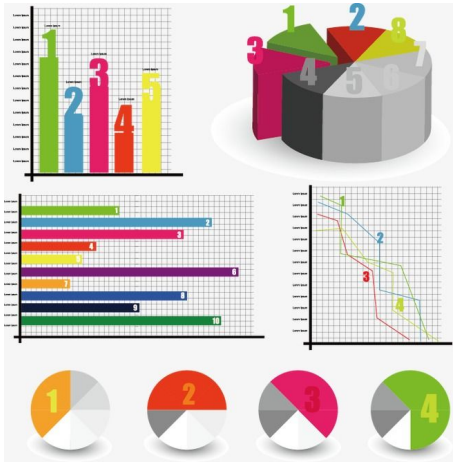
	A	B
1		Hamburger Weight (g)
2	1	325.5
3	2	350
4	3	364
5	4	387
6	5	400

Organizing data
- Listing data
- Sorting data

$$\mu = \frac{1}{N}\sum_{i}^{N} x_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

Summarizing data
- Calculating statistics (통계량)



Visualizing data
- Histogram, bar chart, pie chart,etc.

Parameters (모수) and Statistics (통계량)

Parameter (모수)

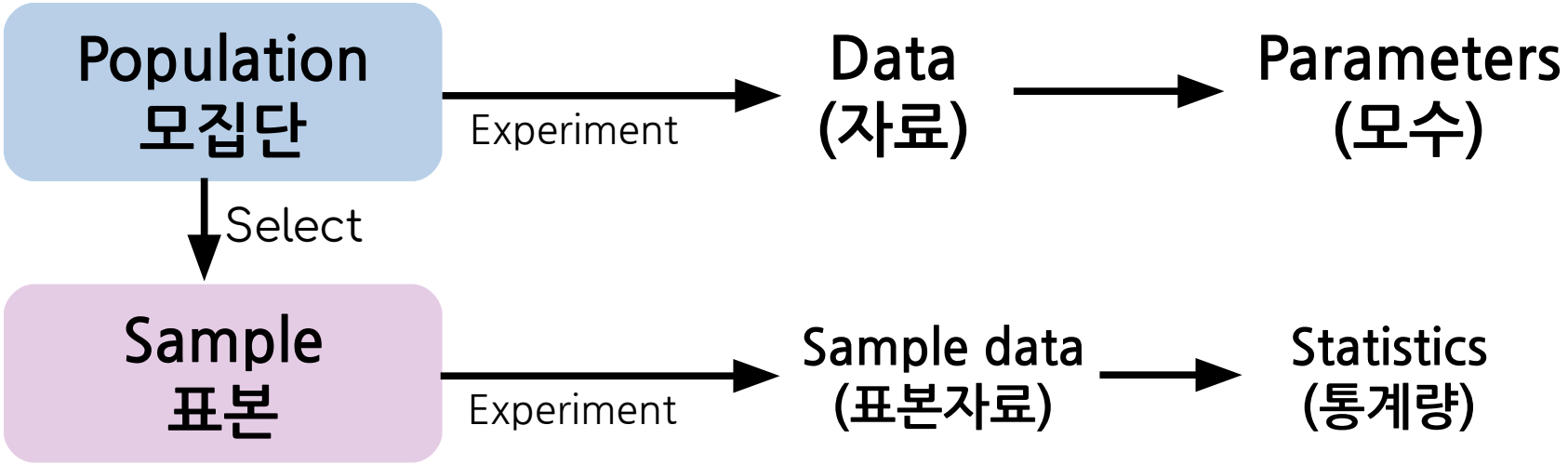
- 모집단 혹은 모집단의 자료로부터 얻을 수 있는 수치적 결과
- Any numbers calculated from a population
- 모집단을 특정하거나 기술하는 수치
  Numeric quantities that describe or specify a population
- 예: 모평균, 모분산, 모표준편차 등
- example: population average, population variance, population standard deviation

Statistic (통계량)

- 모집단으로부터 얻은 표본으로부터 계산하여 얻은 수치적 결과
- Any numbers calculated from sample data
- 표본을 특정하거나 기술하는 수치
  Numeric quantities that describe sample data
- 예: 표본평균, 표본분산, 표본 표준편차
- example: sample average, sample variance, sample standard deviation

Please do not be confused between statistics (통계량) and statistics (통계, 통계학)

Parameters (모수) and Statistics (통계량)



Inferential Statistics/Statistical Inference: 추론통계

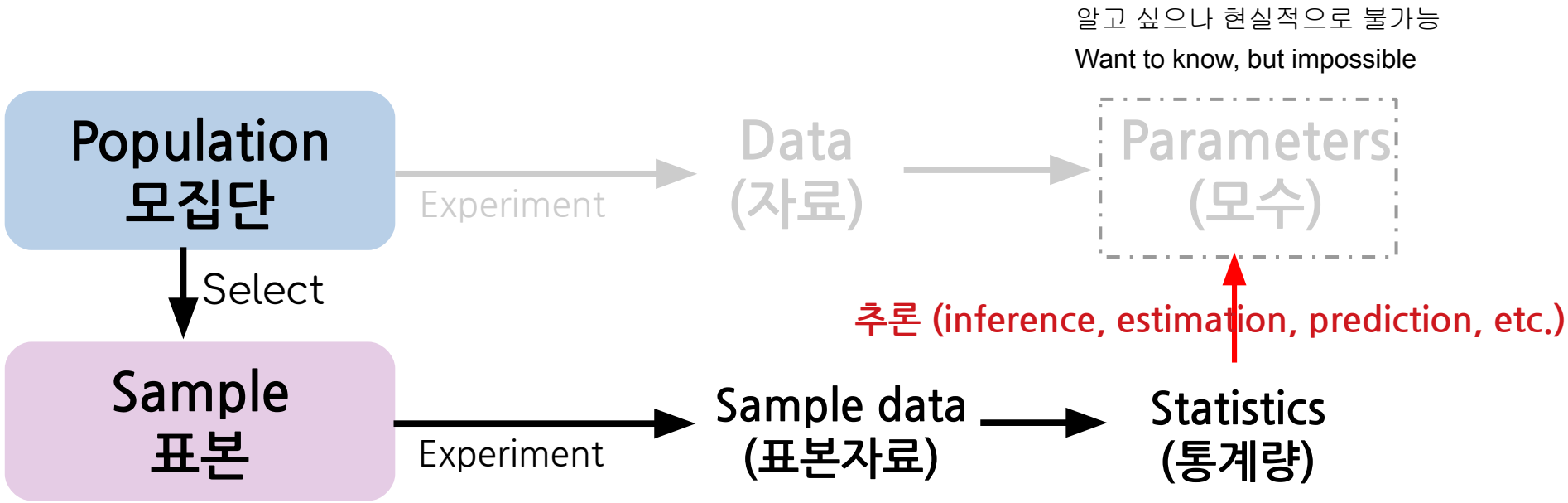
추론통계

- 현실적으로 모집단 전체를 조사하거나 모집단을 대상으로 실험을 하는 것은 불가능하다. 표본을 뽑아 표본조사를 하거나 표본을 대상으로 실험을 하여 얻은 결과를 가지고 모집단의 특징이 무엇인지를 추론할 수 있다. 이를 **추론통계**라고 한다.
- 구체적으로 표본의 통계량으로 모집단의 모수를 추정하는 수치적인 추론을 중심으로 추론통계에 대하여 배우고자 한다.

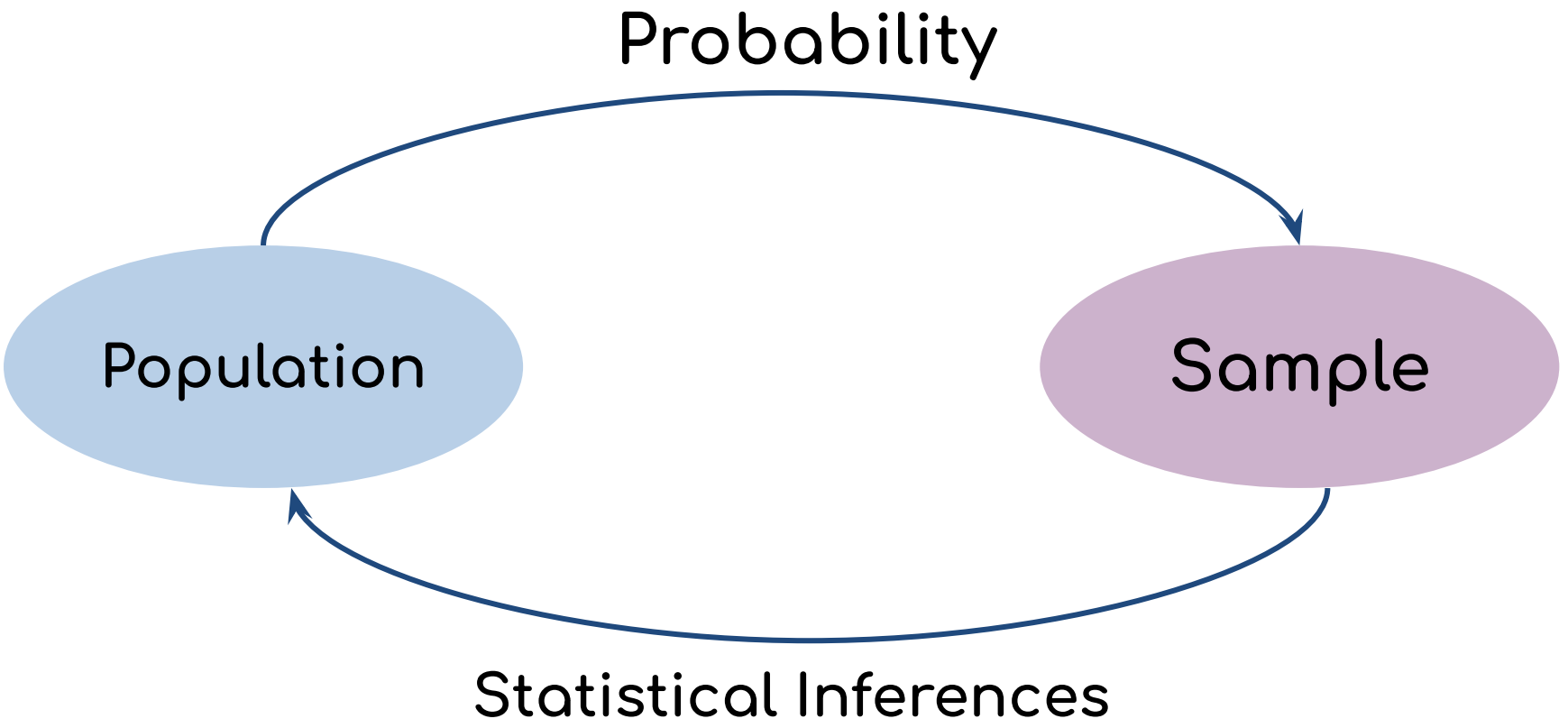
Inferential Statistics

- In reality, it is impossible to investigate a population due to the fact that the number of elements in the population is generally huge. Instead, one can select a sample from the population, and can conduct some experiments on the sample. From those results, one can infer characteristics or features of the population.
- In particular, we would like to focus on numerical inferences where (population properties) parameters can be inferred from statistics of the sample.

Inferential Statistics : 추론통계



Probability and Statistics



Contents : Keywords

I. Population, Sample, Descriptive Statistics

II. Probability, Random Variables, Distributions

III. Statistical Inferences

Probability and Statistics

I. Probability

1. Basics of Probability
2. Random Variables
3. Probability Distributions
4. Sampling

II. Statistical Inferences

1. Point Estimation
2. Interval Estimation
3. Hypothesis Testing
4. Regression

Statistical Analysis Tools: Python

1. Anaconda

   https://www.anaconda.com/products/individual

2. Google Colab

   https://colab.research.google.com/

3. kaggle.com

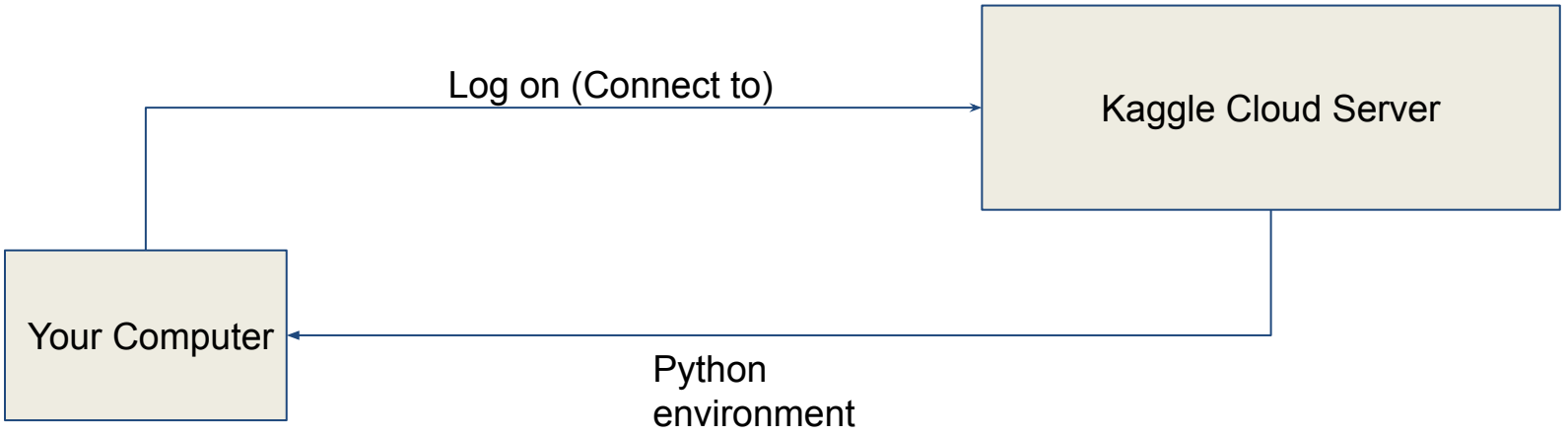
   https://www.kaggle.com/

Kaggle (Cloud Computing)

kaggle.com : https://www.kaggle.com/

1. It is a cloud computing service :

   We can use the same python environment.
2. Kaggle provides many datasets.
3. You can share your output in Kaggle.



Kaggle Notebook

Notebook Name

Connecting to Kaggle Server