Introduction to Statistics (통계학 개요)

Chanseok Park (박찬석)

Applied Statistics Laboratory
Department of Industrial Engineering
Pusan National University

August 5, 2020

Hosted by SEC

Overview

1. Definition of Statistics (통계의 정의)
2. Statistical model (통계 모형) and Estimation (추정)
3. Software for Statistics (통계 소프트웨어)

Overview

1. Definition of Statistics (통계의 정의)
2. Statistical model (통계 모형) and Estimation (추정)
3. Software for Statistics (통계 소프트웨어)

Overview

1. Definition of Statistics (통계의 정의)
2. Statistical model (통계 모형) and Estimation (추정)
3. Software for Statistics (통계 소프트웨어)

1. Definition of Statistics (통계학/統計學의 정의)

Google

수학의 한 부문으로, 사회 현상을 통계에 의하여 관찰 · 연구하는 학문.

집단에 관한 자료를 정리하여 그 특징을 나타내는 여러 가지 수치(數値)를 산출하고 그 자료가 가리키는 것을 알려고 하는 기술(記述) 통계학과 집단의 상태를 그로부터 추출(抽出)한 표본에서 수리적(數理的)으로 추측하는 추측 통계학으로 나뉨.

1. Definition of Statistics (통계학/統計學의 정의)

Wikipedia

7/15/2018 통계학 - 위키백과, 우리 모두의 백과사전

WIKIPEDIA

통계학

위키백과, 우리 모두의 백과사전.

통계학(統計學, 영어: statistics)은 수량적 비교를 기초로 하여, 많은 사실을 통계적으로 관찰하고 처리하는 방법을 연구하는 학문이다. 근대 과학으로서의 통계학은 19세기 중반 벨기에의 케틀레가 독일의 "국상학(國狀學, Staatenkunde, 넓은 의미의 국가학)"과 영국의 "정치 산술(Political Arithmetic, 정치 사회에 대한 수량적 연구 방법)"을 자연과학의 "확률이론"과 결합하여, 수립한 학문에서 발전되었다.[1][2]

1. Definition of Statistics (통계학/統計學의 정의)

My Definition

Data Reduction with information.

참조: Talk-1 at ▸ 2018/Seminar

Statistical procedure:

- Collect Data (to analyze data)
- Analyze Data (to make a decision)

Note: Data reduction is essential for this procedure.

2. Statistical Model and Parameter Estimation

Statistical model

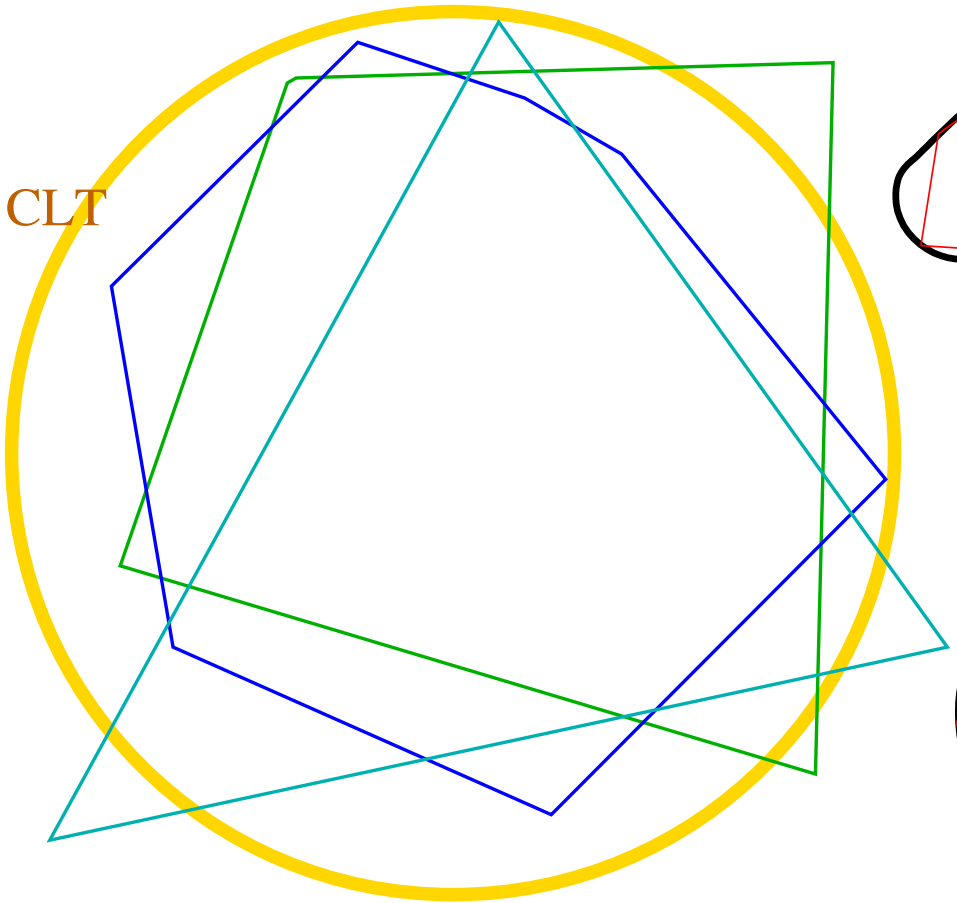
- 통계학을 이용하기 위해서는 statistical model을 정하는 것이 가장 우선적으로 필요로 한다.
- Parameter의 관점에서 보면
  - ◊ parametric: (most commonly used)
  - ◊ non-parametric
  - ◊ semi-parametric으로 분류할 수 있다.

Parameter estimation

- Calculation-based: **ML(우도)**, Method of moments, etc.
- Simulation-based: Bootstrap, MCMC (Markov chain Monte Carlo), **Multiple imputation**, etc.

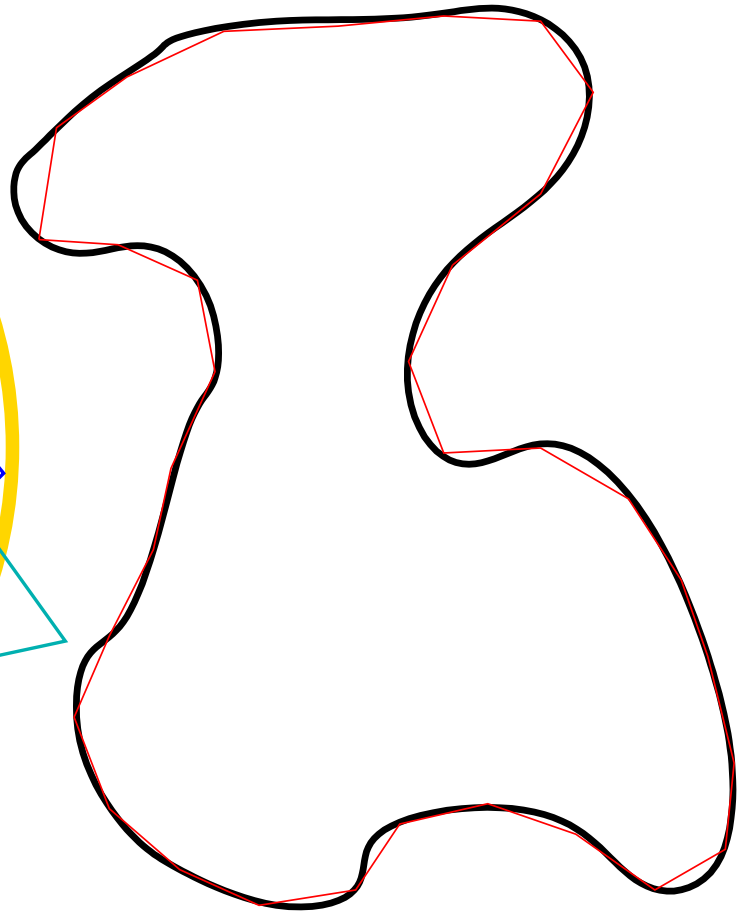
즉, 어떤 model을 사용하고, 어떻게 estimation하나가 큰 issue.

2. Model: Parametric (母數) versus Nonparametric (非母數)



Parametric (Fixed Frame)

(Fixed # of parameters)

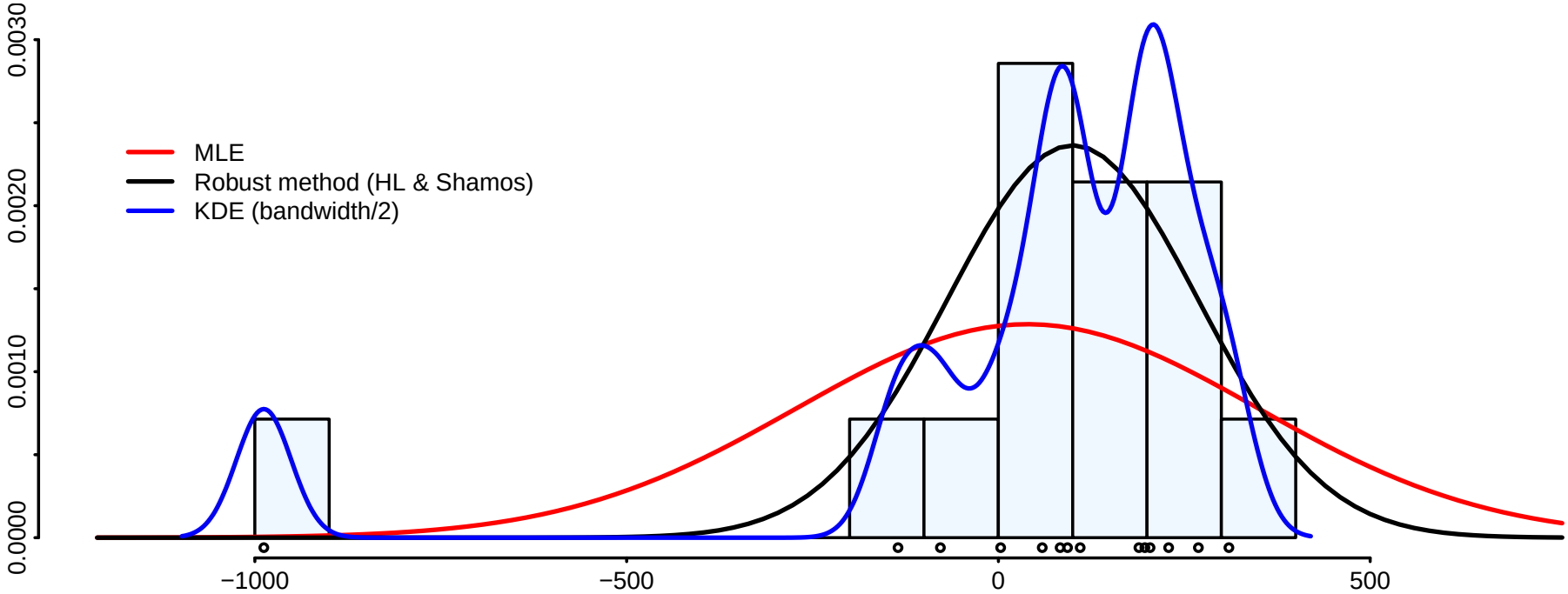


Nonparametric (Non–Fixed Frame)

(Non–fixed # of parameters)

2. Model: Parametric (母數) versus Nonparametric (非母數)

Example from Talk-4 at ▸ 2018/Seminar



The difference between the two faults rate (test and control phone-lines) from Welch (1987)[1]

- Histogram: nonparametric (# of bins $\sim n$).
- KDE (kernel density estimation): nonparametric (# of kernels $\sim n$).
- PDF with MLE or Robust estimation: parametric (μ and σ).

과공 비례(過恭 非禮)! 과모수 비모수 (過母數 非母數)!

[1] Welch, W. J. (1987). Rerandomizing the median in matched-pairs designs. Biometrika, 74:609–614.

2. Model: Parametric (母數) versus Nonparametric (非母數)

Description	Parametric	Nonparametric	Note
Example	mean	median	in general
Robustness	bad	good	in general
Power of test	good	bad	
Robust to model departure	not bad (NB:CLT)	very good	
Sample size	needs big size	small is OK	
Calculation	easier	more complex	
# of parameters	fixed	non-fixed	
Information loss	minor	can be serious	NB: median
Examples	MLE, MME, Bayesian pdf with parameter	Range, KDE, empirical likelihood, bootstrap, plots (hist, etc)	

2. Estimation (MLE)

- **Parametric:**
  Let $f(x|\theta)$ with $\theta \in \Theta$ be pdf/pmf. Then MLE is obtained as

$$\hat{\theta} = \arg\min_{\theta \in \Theta} L(\theta), \quad \text{where} \quad L(\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

- **Nonparametric (CDF example):** n observations $(x_1, x_2, \ldots, x_n)$, but k distinct observations with frequencies $f_1, f_2, \ldots, f_k$.

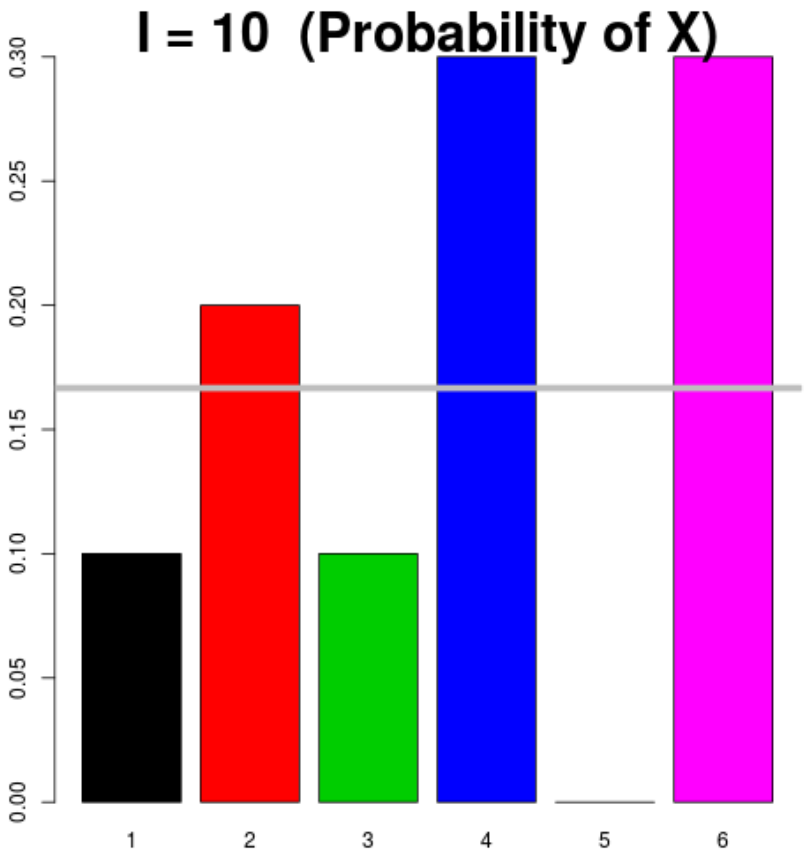
$$L(p_1, p_2, \ldots, p_k) \propto \prod_{j=1}^{k} {p_j}^{f_j}, \text{ with } \sum_{j=1}^{k} p_j = 1$$

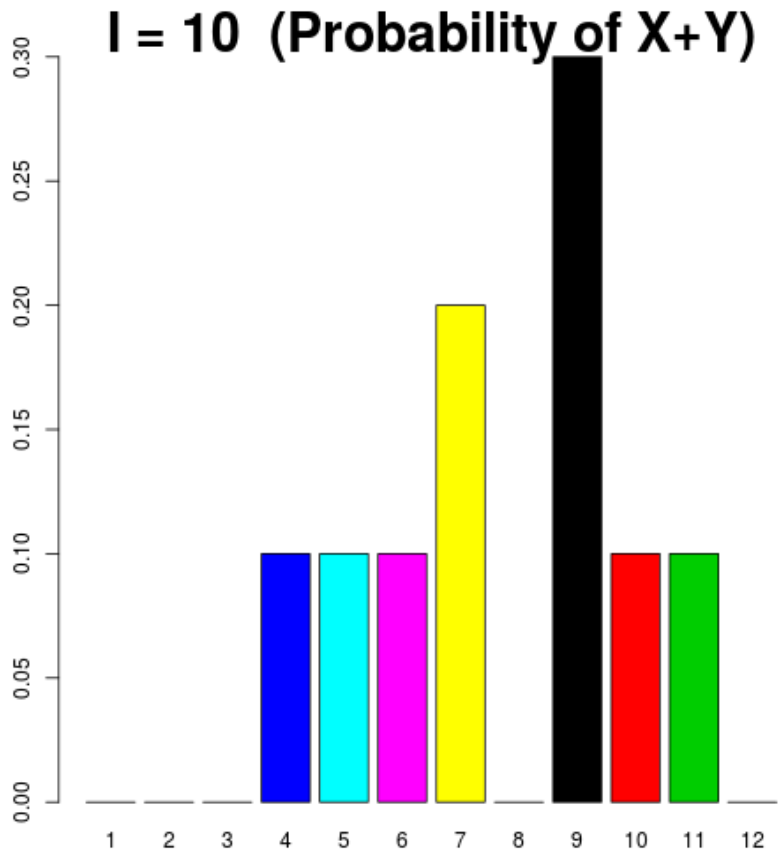
$$\hat{p}_j = \frac{f_j}{n} \quad \Longrightarrow \quad \hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \leq t),$$

  where $\mathbb{I}(\cdot)$ is an indicator function.

2. Estimation (simulation-based)

1. pmf when tossing a die.
2. pmf of the sum of two dice when tossing two dice (with p_1 and p_2).





NOTE: Refer to `Talk-R.r` at ▸ 2020/Talk-R (one can increase `ITER` in the R code). Also, see Talk-5 at ▸ 2018/Seminar

3. Software for Statistics (통계 소프트웨어)

Software for Statistics (통계 소프트웨어)

- SAS, SPSS, Minitab, Matlab, BMDP, S, Splus, etc.
- **R language**: 5th place in the world.
  `http://blog.revolutionanalytics.com/2016/07/r-moves-up-to-5th-place-in-ieee-language-rankings.html`
- Python with statistics additions (recently).