

DAY 1 . 28 MAR 2021

✓ Orientation

- 강당별 교수 : nbaumtin@gmail.com
- 관기별 대처 : 010.9024.39ab.

- 2012년 복수 과목 학점 등급
- 'R' 사용 예제

✓ machine learning

- What is learning? => in machine, extracting features & recognizing pattern.

◦ Differences among AI, ML and Deep learning.

- ① AI - mimic human behavior → Decision tree
- ② ML - enable machine to improve with experiences → logistic Regression, SVM
- ③ DL - multi-layer neural network. → CNN, RNN

✓ Geoffrey Hinton (CIFAR image classif.)

- Variadic Embed.
- Deep belief network.

Melting with teacher

(177W & 275W => R 222Mnb. + 13GB)

(337W
429W)
=> GitHub + Python.

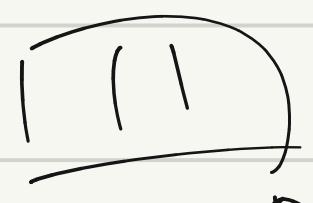
=> Python

tuple list
sklearn.model_selection import train_test_split
train,test = train_test_split(iris, 2)
!w: w1: b1: C: 2

Lecture 1. R Data structure

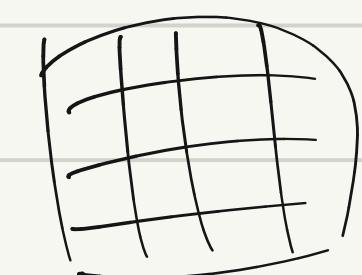
R Data type, re index '1' 2 1 2 3 4

vector



a[3]

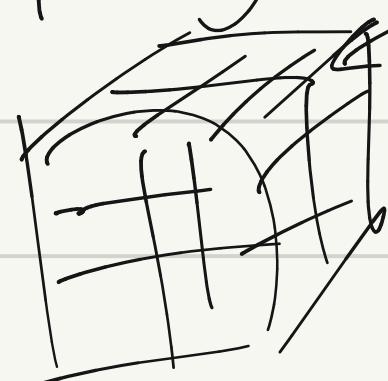
Data type of matrix



a[m, n]

(numeric)
logical
character
complex
(Rational)

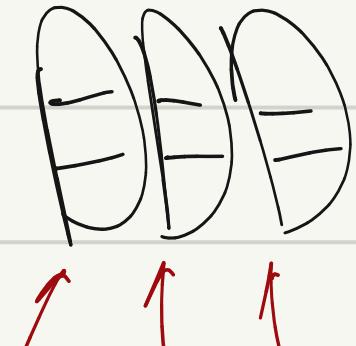
Arry



a[1, 2, 3]

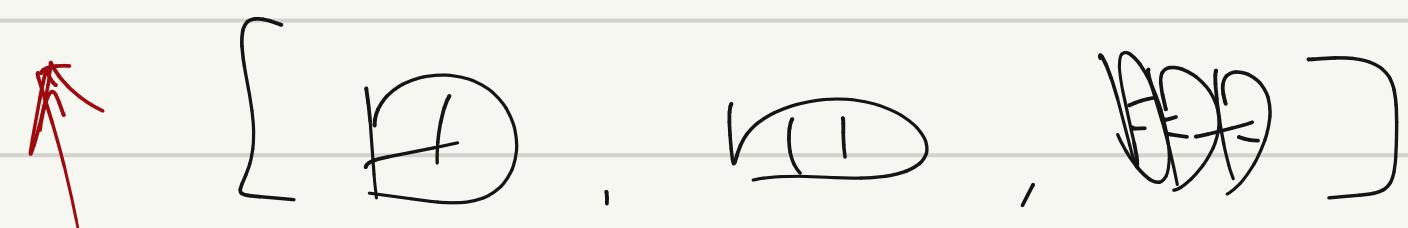
a[m, n, p, dim]

* Data frame (table) = tibble



각각 data type의 차이
인덱스가 가능.

Lists



different structure, but same type

but, Data frame or array, different type.
Index 2 가능.

Lecture 2. Probability Distributions

* 자료형 정리

① 연속형 자료

- { 등간격 (interval) : 절대값 0이거나 > 0. e.g. 올드
 비례형 (ratio) : 절대값 0이거나 > 0. e.g. 1, 2, 3, 4, 5.

사칙연산 가능하지 않음.

② 범주형 자료

- { 명목형 (Nominal) : 속성을 부여, 순서 없음 e.g. 남/여
 순서형 (Ordinal) : 속성간 순서개념, ↗
 ↘ panelist) 2차원 (2x3x1)

* 자료형 분석

- 모집단 & 표본 추출

- ① 모집단 : 03번학번과 같은 집단 전체 e.g. 92번학번 고등학교 1학년
 ② 표본 : 모집단에서 추출한 일부 e.g. 추출된 20명 중 10명.

n	$n/15$
0~2	10
3~6	5

자료를 정리하는 데 사용되는

수학적 도구이다.

자료를 정리하는 데 사용되는 수학적 도구이다.

자료를 정리하는 데 사용되는 수학적 도구이다.

$$\begin{cases} P \\ M, \sigma^2 \\ P_i = 0.1 \end{cases}$$

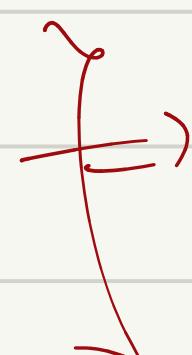
$$\begin{cases} S \\ \hat{P}_i = 0.13 \end{cases}$$

Sampling의 확률 분포

인구.

* 연속형 자료 분석

- histogram
- boxplot
- violin plot



자료를 정리하는 데 사용되는 수학적 도구이다.

자료를 정리하는 데 사용되는 수학적 도구이다.

Sampling의 확률 분포

인구.

continue ..

Lecture 2 . Probabilistic

* 수치를 통한 연속적 확률 분석.

- 중심_trend 추정.

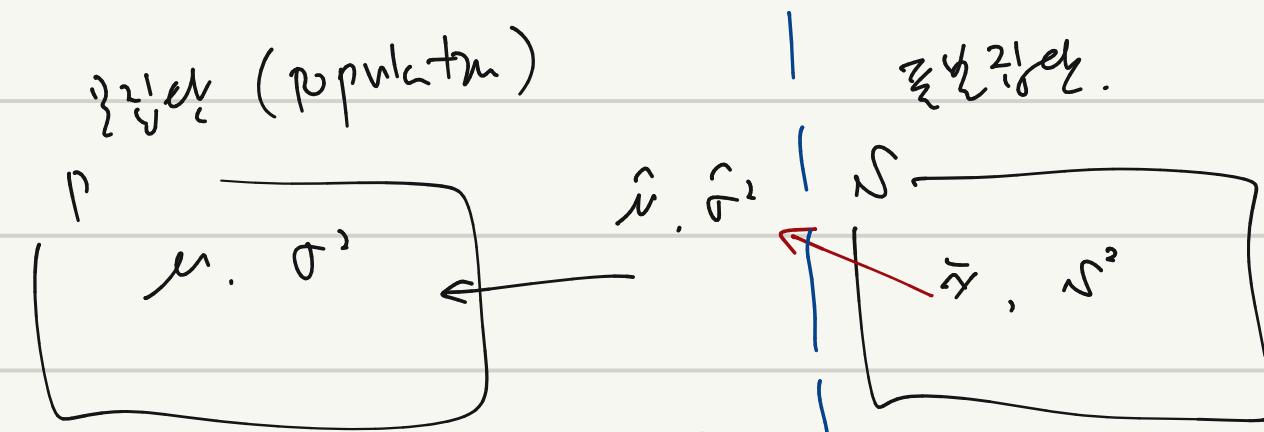
① 표본 평균 (Sample Mean)

- 중심을 나타내는 추정

- 확률 분포 중심 μ

$$\bar{x} = \frac{\sum x_i}{n} = \mathbb{E} x_i$$

- Outlier에 취약한 (Trimmed mean)
influenceless



표본평균.

Universal
notation

표본평균 = 표본 평균

자료는 대체로

(continuous distribution, ...)

② 중앙값 (Median)

- 전자 계산기로 표기되는 이유는 표준화된 표기법

$$\text{Median} : P(X \leq m) = \int_{-\infty}^m f(x) dx = \frac{1}{2}$$

- 확률 누적도 (a)가 odd $\rightarrow \frac{n+1}{2}$

even $\rightarrow \frac{n}{2}, \frac{n+1}{2}$ 의 평균

- Outlier에 영향을 덜 받음.

* 자료가 표준화된 표기법
자료를 표기하는 → Mean

→ Median

ME 표준화된 표기법.

③ 최빈값 (Mode)

- 관찰값 중 가장 많은 값.

* Mean, Median, Mode 비교

- Mean \rightarrow 모든 수를, 모든 경우를 평균

\rightarrow Outlier에 영향을 받음

- Median \rightarrow Outlier 영향을 받지 않음

- Outlier에 영향을 덜 받음.



표본평균은 Outlier에 영향을 받는다.

i.e. 표집단의 균형점.

* 11/26 일정

평균과 편차 분산 표준 편차 Variance & Standard Deviation.

① 평균

- 자료가 중심으로 둘러 열려서 편차 있다.

$$- \bar{x} = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

★ 자료 = $\sum f_i$, 자료의 개수이거나 제작한 것을 뜻.
⇒ 흔히 더 좋은 표현법.

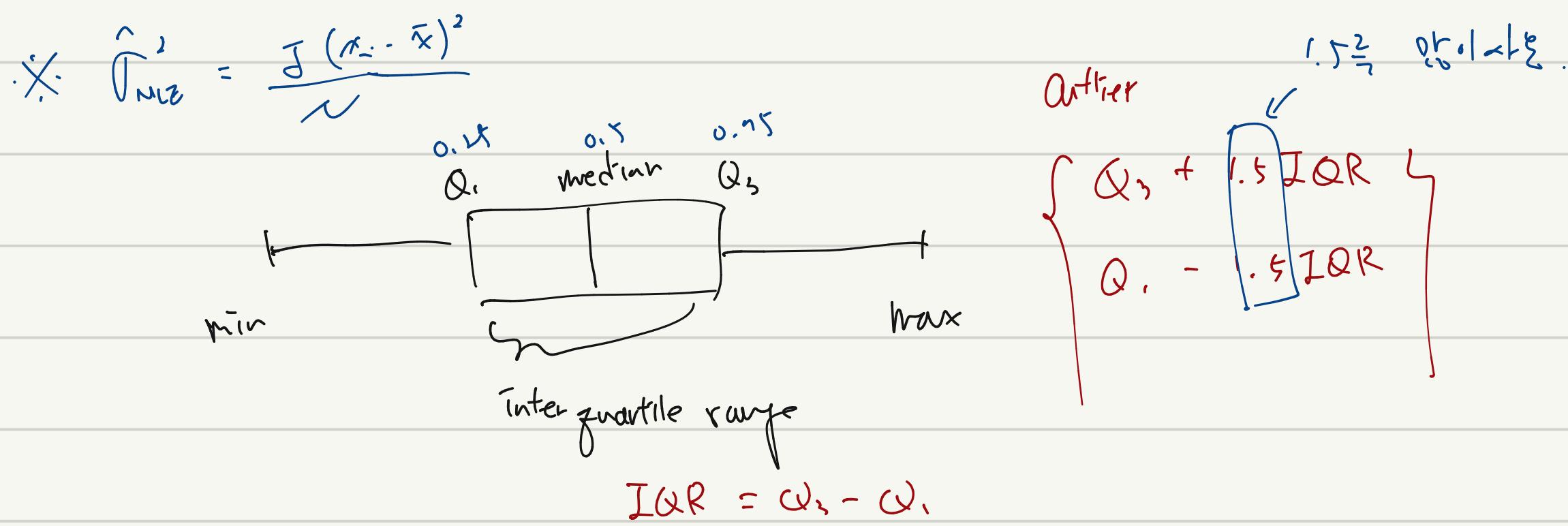
$$- sd = \sqrt{s^2} = s$$

= parameter가 아니라 통계학적 특성의 가로로 사용하는 수식

$$- Range = \text{Max} - \text{Min}$$

('통계학적'은 관점인)

② 사분위수범위 (Inter Quartile Range, IQR) : 3-4주차 - 1사분위수



* Median은 7.21 미만 \rightarrow IQR 미만 \rightarrow Outlier 21(p)

\rightarrow min, max 같은 유형의 수치