NO.

DATE.

7.1 확률적 데이터와 확률변수

- 이번 데이터가 생년월일처럼 변하지 않는 데이터인지 혹은 측정할 때마다 변할수 있는 데이터인지 어떻게 구분할 수 있는가?
- 혈압과 같이 100% 정확하게 예측할수 없는 데이터가 있을 때, 이 데이터로부터 우리가 얻는 지식은 무엇인가? 2것을 어떻게 표현할수 있는가?
- 이러한 데이터를 수학적으로 표현하는 방법을 알아본다.

확률적 데이터

확률적 데이터	결정적 데이터
(random, probabilistic, stochastic, data)	(deterministic data)

* 우리가 다루는 대부분의 데이터
  - 여러 조건이나 상황에 따라 데이터값이 영향을 받기 때문 } 확실하지 않다, "어느정도" 예측할 수 있다.
  - 측정시 오차가 발생할 수 있기 때문

분포 확률적 데이터에서 어떠한 값이 "자주" 나오고 어떠한 값이 "드물게" 나오는가
⇒ 분포 (distribution)

- 범주형 data → count plot
- 실수형 data → histogram

) 시각화

기술통계

분포를 표현하는 여러가지 방법 — 분포의 특징을 표현하는 숫자(통계치)를 계산하여 그 숫자들로 분포를 나타내기
⇒ 기술통계 (descriptive statistics)

"기술통계에서 다루는 통계량은 '표본'을 불인다."

- 표본평균, 표본중앙값, 표본최빈값
- 표본분산, 표본표준편차
- 표본왜도, 표본첨도

표본평균

Sample mean, Sample average

- 데이터분포의 대표적인 위치를 표시
- 표본평균의 기호
  > 알파벳 m
  > 변수기호 위에 bar를 붙인 $\bar{x}$ 기호

$$m = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

표본중앙값

Sample median

자료를 크기별 정렬했을 때 정중앙에 위치하는 값.

N=odd) sample median = (N+1)/2 th sample value

N=even) sample median = (N/2 th sample value + (N/2+1)th sample value) / 2

NO.

DATE.

표본최빈값

most frequent value, sample mode

- 연속된 값을 갖는 데이터에서는 주의해야 한다.
  - ↳ 구간을 정해서 빈도 count = histogram
    - (BUT) 구간을 어떻게 나누는가에 따라 달라질 수 있어 신뢰하기 어렵다.

〈파이썬을 사용한 대푯값 계산〉

numpy

- ㄴ mean() 표본평균
- ㄴ median() 표본중앙값
- ㄴ argmax() 표본최빈값
- ㄴ histogram() 표본최빈구간

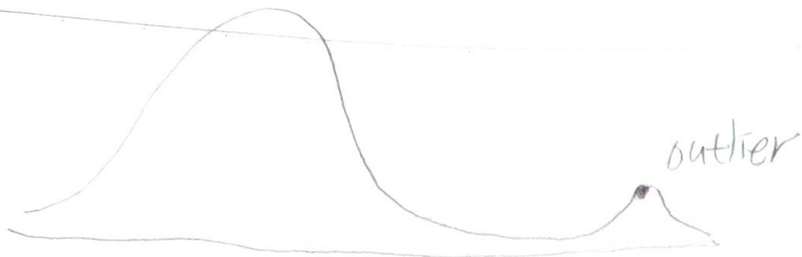
단봉분포와 다봉분포

(uni-modal) (multi-modal)

대칭분포 symmetric

표본평균, 표본중앙값, 표본최빈값은 분포의 모양에 따라 다음과 같은 특성을 보인다.

- 분포가 표본평균을 기준으로 대칭인 대칭분포이면 표본평균은 표본중앙값과 같다.
- 분포가 대칭분포이면서 하나의 최빈값만을 갖는 단봉분포이면 표본최빈값은 표본평균과 같다.
- 대칭분포를 비대칭으로 만드는 데이터가 더해지면, 표본평균이 가장 크게 영향을 받고 표본최빈값이 가장 적게 영향을 받는다.



outlier가 존재할 때
표본통계치의 이동하는 크기비교

mode < median < mean
영향X mean에 비해 덜 움직임 --- outlier 존재할 때 대푯값으로 적합함. 가장 크게 영향받는 값

분산과 표준편차

전체 데이터가 얼마나 변동(variation)해 있는가?

→ 표본분산 (sample variance) & 표본표준편차 (Sample standard deviation)

→ 분포의 폭 (width)을 대표하는 값.

$$\text{표준편차} = \sqrt{\text{분산}}$$

$$\text{Sample variance} = s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

* S = Standard deviation

* 편향이 내포되어 있어 편향오차를 가짐
편향 표본분산 (biased sample variance)

NO.

DATE.

* 비편향 표본분산 (unbiased sample variance)

$$S^2_{unbiased} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

표본비대칭도 sample skewness

평균값의 차이의 세제곱을 이용하여 구한 특징값을 표본비대칭도 (sample skewness) 라고 한다. 표본비대칭도가 0이면 완전 대칭이다. 표본비대칭도가 음수면 표본평균값을 기준으로 왼쪽에 있는 값을 가진 표본이 나올 가능성이 더 많다는 것이다.

$$\text{sample skewness} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}^3}$$

sample skewness < 0 | sample skewness = 0 | sample skewness > 0

표본첨도 sample kurtosis

평균과의 거리의 네제곱을 이용하여 구한 특징값을 표본첨도 (sample kurtosis) 라고 한다. 표본첨도는 데이터 중앙의 몰려있는 정도를 정밀하게 비교하는데 쓰인다.
사람의 눈으로 첨도를 구별하는 것은 어렵다. 표본첨도의 값은 다음에 설명할 정규분포의 첨도와 비교하여 정규분포보다 첨도가 높으면 양수, 정규분포보다 첨도가 낮으면 음수로 정의한다.

$$\text{sample kurtosis} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2\right)^2} - 3$$

표본 모멘트

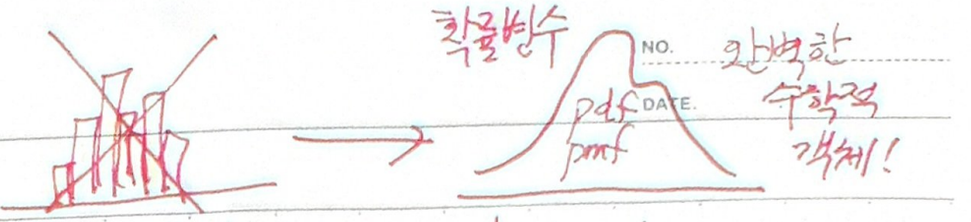
분산, 비대칭도, 첨도를 구하기 위해 제곱, 세제곱, 네제곱을 하는 것처럼 k제곱을 이용하여 구한 모멘트를 k차 표본모멘트 (sample moment) 라고 한다.

$$\underset{(k)}{\text{sample moment}} = \frac{1}{N}\sum_{i=1}^{N} x_i^k$$

2차 표본모멘트 이상은 평균을 뺀 표본중앙모멘트 (sample centered moment) 값을 사용하기도 한다.

$$\text{sample centered moment} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^k$$

∴ 평균은 1차 모멘트, 분산은 2차 모멘트,
비대칭도는 3차 모멘트, 첨도는 4차 모멘트에서 유도된 값.



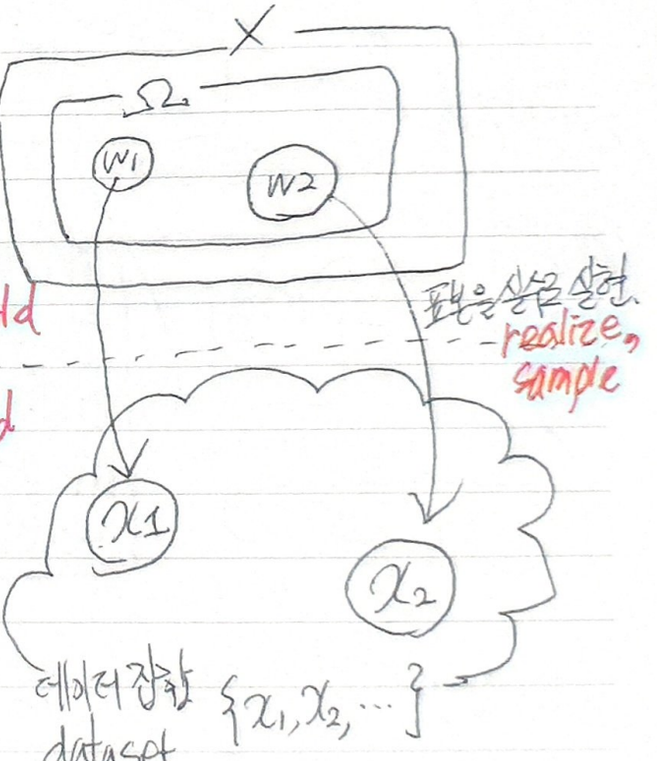
확률변수 는도에 대한 "완전한 정보"를 구하고자 하는 시도.

확률변수는 수학적으로 확률공간의 표본을 입력으로 받아서 실수인 숫자로 바꾸어 출력하는 함수다. 출력되는 실수가 데이터의 값이다. 표본값을 굳이 실수로 바꾸는 이유는 표본이 실수가 아니면 확률분포함수를 정의할 수 없기 때문이다.

$$w \in \Omega \xrightarrow{\text{확률변수 } X} x \in \mathbb{R}$$

$$X(w) = x$$

(notation)
확률변수 = X
확률변수에 의해 할당된 실수값 = x

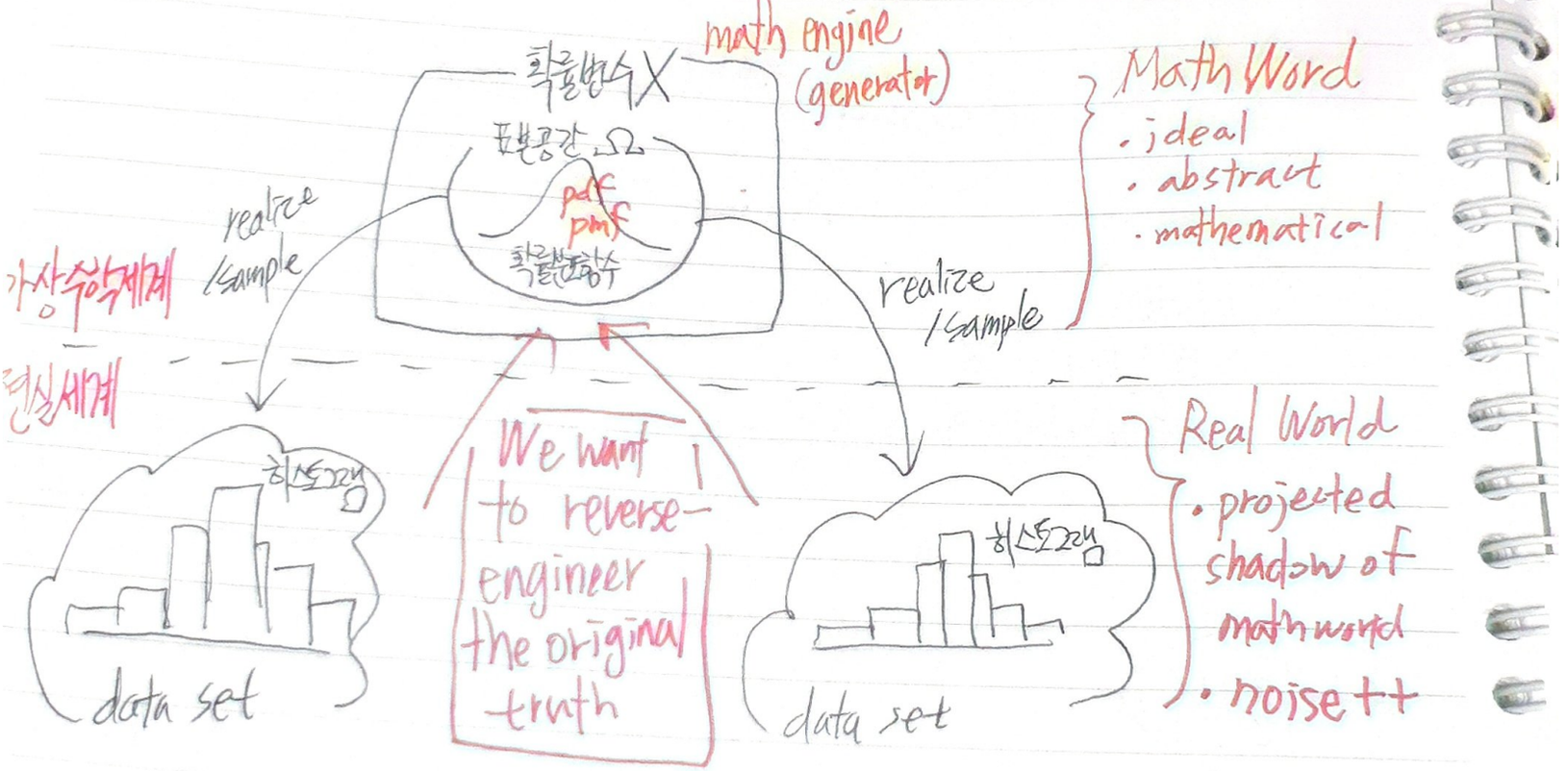


이산확률변수
NOT CONTINUOUS,
DISCRETE (서로 떨어진)
(셀 수 있는)
★ 표본공간 원소가 무한대여도 셀 수 있다면 이산확률변수

연속확률변수 원판의 각도 문제처럼 연속적이고 무한대의 실수표본값을 갖는 확률변수
- 구간사건의 조합으로 표시
- 확률분포를 수학적인 확률분포함수로 나타낼 수 있다.

"확률변수는 data generator"

- rolling dice
- engine output
- blood pressure measure



1. 확률변수로부터 데이터를 여러번 생성하는 경우, 실제 데이터값은 매번 달라질 수 있지만 확률변수 자체는 변하지 않는다.
2. 확률변수의 확률분포함수는 우리가 직접 관찰할 수 없다. 다만 확률변수에서 만들어지는 실제 데이터값을 이용하여 확률분포함수가 어떠한 것일 것이라고 추정할 뿐이다.
3. 확률변수에서 만들어지는 실제 데이터값은 확률변수가 가진 특성을 반영하고 있다. 데이터 갯수가 적어질수록, 확률변수가 가진 특징을 정확하게 표현하지 못하지만 데이터 갯수가 증가하면 보다 정확하게 확률분포함수를 묘사할 수 있게 된다.

NO.

DATE.

확률변수를 사용한 데이터분석

확률변수를 사용하게되면 데이터분석은 보통 다음과 같은 순서로 이루어진다.

1. 데이터를 수집한다.
2. 수집한 데이터가 어떤 확률변수의 표본 데이터라고 가정한다.
3. 데이터를 사용하여 해당 확률변수의 확률분포함수의 모양을 결정한다. ★
4. 결정된 확률분포로부터 다음에 생성될 데이터나 데이터 특성을 예측한다.

이 과정 중 가장 중요한 것이 데이터값에서 확률변수의 확률분포 함수를 역설계 (reverse-engineering)하여 만들어내는 세 번째 단계다.
데이터에서 확률분포함수의 모양을 구하는 방법은 여러가지가 있는데, 가장 간단한 방법은 다음과 같이 기술통계값을 이용하는 것이다.

1. 데이터의 분포가 가지는 표본평균, 표본분산 등의 기술통계값을 구한다.
2. 이 값과 같은 기술통계값을 가지는 확률분포함수를 찾는다.

위 방법을 쓰려면 표본 데이터가 없는 확률분포함수의 기술통계값을 구하는 방법을 알아야 한다. 다음 절부터는 확률분포함수의 기술통계값인 기댓값, 분산 등에 대해 공부한다.

확률변수의 기댓값 expectation.

확률변수의 확률밀도함수를 알면 확률변수의 이론적 평균값을 구할 수 있다. 이러한 이론적 평균값을 확률변수의 기댓값 (expectation) 이라고 한다. 간단히 평균 (mean) 이라 말하기도 한다.

확률변수 X의 기댓값을 구하는 연산자 (operator) $= E[X]$
확률변수 X의 기댓값 $= \mu_X$ 또는 μ

- 이산확률변수의 기댓값은 표본공간의 원소 x_i의 가중평균 (가중치 = 확률질량함수 $p(x)$)

$$\mu_X = E[X] = \sum_i x_i p(x)$$