밑바닥부터 시작하는 데이터 과학

3회차

- 통계와 확률
- 가설과 추론
- 회귀분석 심화

1. 통계와 확률

통계와 확률

1. 통계와 확률

기술 통계 (descriptive statistics)

수집한 데이터를 요약 묘사 설명하는 통계 기법

1. 평균 (mean)

- 데이터의 대표값을 말해준다.

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

2. 분산 (variance)

- 데이터의 흩어짐 정도를 보여준다.
- 데이터와 평균 간의 차이의 제곱의 평균이다.

$$s^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2$$

1. 통계와 확률

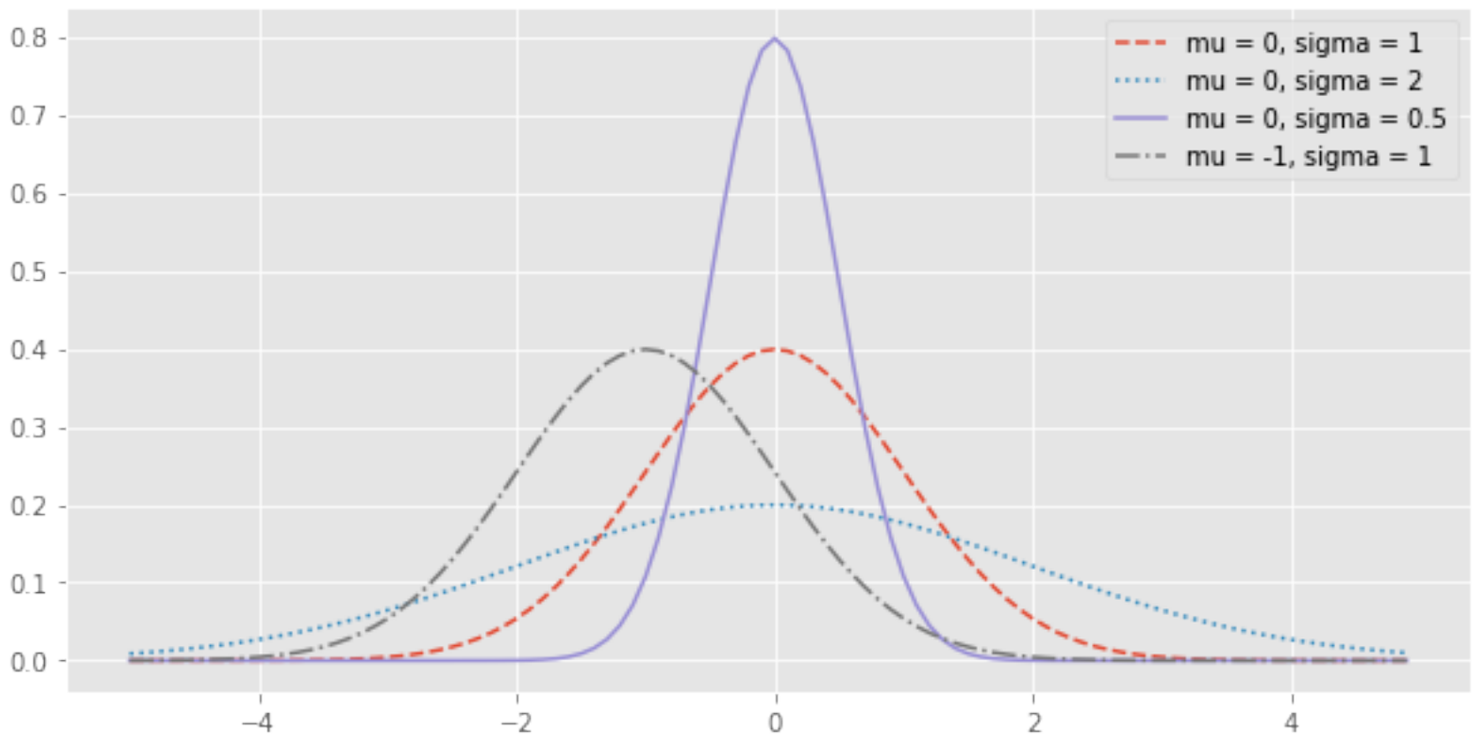
기술 통계 (descriptive statistics)

수집한 데이터를 요약 묘사 설명하는 통계 기법

3. 표준편차 (standard deviation)

- 분산의 제곱근 값이다.
  - 분산으로는 데이터의 분포를 직관적으로 표현하기 어렵기 때문에 표준편차를 많이 사용한다.
    - ex. [2, 4, 6]의 경우
      - 평균 : 4
      - 분산 : 4, 표준편차 : 2

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



1. 통계와 확률

기술 통계 (descriptive statistics)

4. 공분산 (covariance)

- 평균을 중심으로 각 자료들이 어떻게 분포되어 있는지 크기와 방향성을 같이 보여준다

$$s_{xy} = \frac{1}{N} \sum_{i=1}^{N} (x_i - m_x)(y_i - m_y)$$

5. 상관 계수 (correlation)

- 공분산에서 방향성만 분리한 값이다.
- 보통 피어슨 상관계수라고 말하는 값을 사용한다.

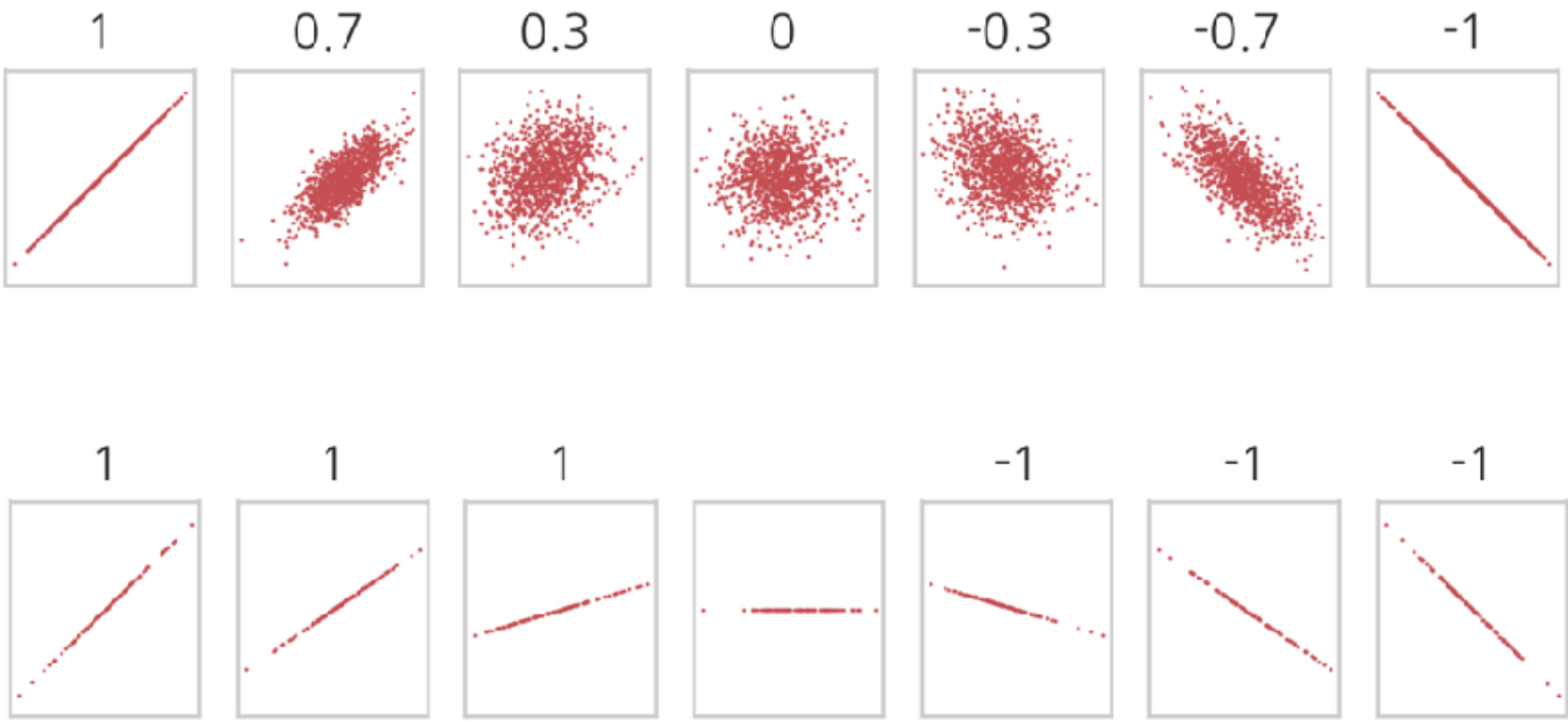
$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}}$$

- 상관계수 = 1 : 완전한 선형관계
- 상관계수 = 0 : 무상관
- 상관계수 = -1 : 완전한 반선형관계

1. 통계와 확률

기술 통계 (descriptive statistics)

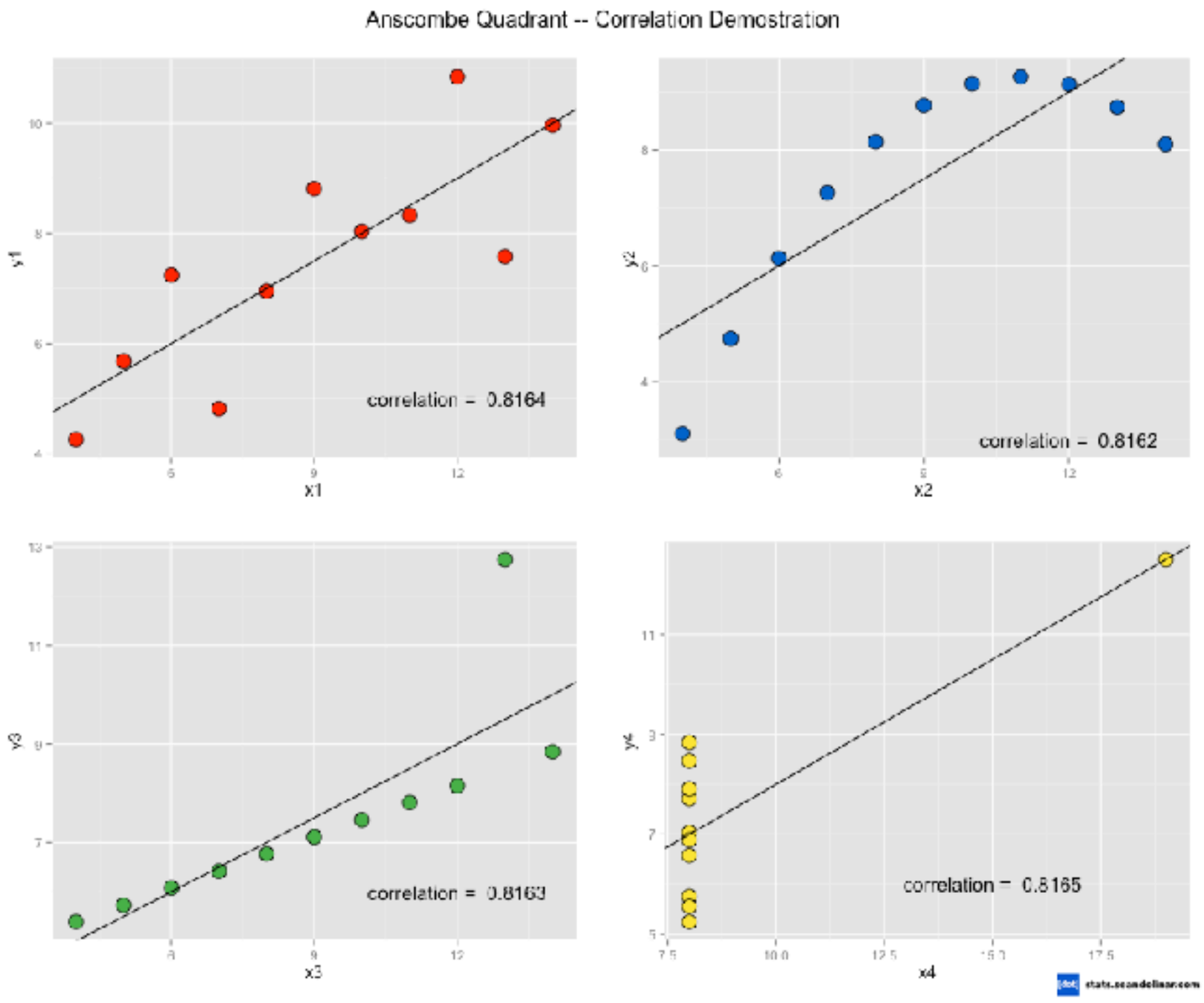
- 선형 관계만을 보여줄 뿐 기울기는 의미가 없다.



1. 통계와 확률

기술 통계 (descriptive statistics)

- 특이값 (이상치)에 영향을 많이 받는다.



1. 통계와 확률

확률 (probability)

어떠한 사건의 공간에서 특정 사건이 선택될 때 발생하는 불확실성을 수치적으로 나타낸 것

1. 생활 속의 확률

- 동전을 한번 던졌다. 동전이 앞면이 나올 것인가 뒷면이 나올 것인가?
- 주사위를 던져 하나의 숫자가 나왔다. 이 숫자는 무엇인가?
- 삼성전자 주식의 가격은 내일 몇 % 오를까?

 # 이 문제들의 공통점은?

- 답을 100% 확신할 수 없다.
- 어떤 문제는 무엇이 답인지 전혀 예측할 수 없는 것도 있고 어떤 문제는 어느 정도 정확도 혹은 범위내에서 예측할 수 있는 것도 있다.
- **확률론**은 이러한 문제가 어떤 답을 가질 수 있고 그 답의 **신뢰성**을 계산하는 **정량적인 방법**을 제시한다.

1. 통계와 확률

확률 (probability)

2. 확률 표본 (표본, sample)

- 우리가 풀고자 하는 확률적 문제에서 선택될 수 있는 혹은 답이 될 수 있는 하나의 경우 혹은 숫자를 말한다.

3. 표본 공간 (sample space)

- 답이 될 수 있는 혹은 선택될 수 있는 모든 표본의 집합

 # 예시

- **동전을 한 번 던지는 문제**
  - 표본 : 앞면 (Head), 뒷면 (Tail)
  - 표본 공간 : {앞면 (Head), 뒷면 (Tail) }

- **주사위를 던져 나오는 숫자를 구하는 문제를 확률적으로 접근할 때, 표본 공간은?**
- **내일 삼성 전자의 주식에 대한 표본 공간은?**

1. 통계와 확률

확률 (probability)

4. 사건 (event)

- 사전 (event)는 표본 공간의 부분 집합
  - 즉, 전체 표본 공간 중에서 우리가 관심을 가지고 있는 일부 표본의 집합을 뜻한다.
  - 보통 대문자 알파벳으로 표기한다.

- 동전 던지기 예시에서 가능한 사건 (부분집합)
  - { }, {H}, {T}, {H, T}

5. 확률 (probability)

- 확률 (probability)이란 사건 (부분 집합)을 입력하면 숫자 (확률값)이 출력되는 함수이다.
  - 사건 (부분집합) -> 숫자

- 각각의 사건 (부분집합)에 어떤 숫자를 할당(allocate)한 것이다. 보통 대문자 알파벳 P로 나타낸다.
  - P는 함수이고, P(A)는 A라는 사건에 할당된 숫자를 뜻한다.

1. 통계와 확률

확률 (probability)

확률의 공리 (Komogrov's axioms)

- 모든 사건에 대해 확률은 실수이고 양수이다.
  - **표본공간**이라는 **사건**에 대한 **확률**은 1이다.
  - 공통 원소가 없는 두 사건의 합집합의 확률은 각각의 사건의 확률의 합이다.

$$A \cap B = \varnothing \;\; -> \;\; P(A \cup B) = P(A) + P(B)$$

예시

- A = { }, B = {H}, C = {T}, D = {H, T}
  - P(A) = 0
  - P(B) = 0.4
  - P(C) = 0.6
  - P(D) = 1 = 표본공간

- **즉, 확률은 표본이 아닌 사건 (부분집합)에 대해 정의되어야 하는 것이다.**

1. 통계와 확률

확률 (probability)

표본의 수가 무한한 경우

- 시계의 시침이 모든 각도에 대해 가능성이 똑같다면, 시침이 12시를 가르킬 확률은?

왜 그럴까 ?

- 모든 각도에 대해 가능성이 똑같으므로 그 확률을 x 라는 값이라고 하자.
- 그런데 각도가 나올 수 있는 경우는 무한대의 경우가 있으므로 만약 x가 0이 아니라면 $x \times \infty = \infty$ 로 전체 표본 집합의 확률이 무한대가 된다.
- 즉, 1이 아니다. 따라서 **표본의 수가 무한**하고 모든 표본에 대해 **표본 하나만을 가진 사건의 확률이 동일하다**면, **표본 하나에 대한 사건의 확률**은 언제나 **0**이다.

1. 통계와 확률

확률 (probability)

6. 결합 확률 (joint probability)

- 사건 A와 B가 동시에 발생할 확률
  - $P(A \cap B)$ 또는 P(A, B)

7. 조건부 확률 (conditional probability)

- B가 사실일 경우, 사건 A에 대한 확률을 사건 B에 대한 사건 A의 조건부 확률이라고 한다.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

조건부 확률이 왜 위와 같이 되는가?

(1) 사건 B가 사실이므로 모든 가능한 표본은 사건 B에 포함되어야 한다.
  - P(A)와 비교하면 표본 공간 Ω -> B

(2) 사건 A의 원소는 모두 사건 B의 원소도 되므로 사실상 사건 $A \cap B$ 가 원소가 된다.
  - A -> $A \cap B$

(3) 따라서 사건 A의 확률 즉, 신뢰도는 원래의 신뢰도 (결합 확률)를 새로운 표본 공간의 신뢰도 (확률)로 정규화(normalize)한 값이라고 할 수 있다.

1. 통계와 확률

확률 (probability)

8. 독립

- $P(A \cap B)=P(A)P(B)$ 관계가 성립하면 두 사건 A와 B는 서로 독립(independent)라고 정의한다.
- 독립인 경우 조건부 확률과 원래의 확률이 같아짐을 알 수 있다. 즉, B 라는 사건이 발생하든 말든 사건 A 에는 전혀 영향을 주지 않는 다는 것이다.

$$P(A|B)=\frac{P(A,B)}{P(B)}=\frac{P(A)P(B)}{P(B)}=P(A)$$

원인과 결과, 근거와 추론, 가정과 조건부 결론

조건부 확률 P(A|B)에서 B, A는 각각
- "원인과 결과" 또는
- "근거와 추론" 또는
- "가정과 그 가정에 따른 조건부 결론" 으로 생각할 수 있다.

사건 B 가 발생한다는 가정(근거, 조건)에서 사건 A의 확률을 계산하기 때문에 사건 B가 주장 사건 A의 근거가 되기 때문이며, 또한, 사건 B가 발생하지 않는다면 사건 A에 대해 고려할 필요도 없을 것이기 때문이다.

- 조건부 확률의 정의는 P(A, B) = P(A|B)P(B)로도 정의가 가능하다.

> - A, B가 모두 발생할 확률은 B라는 사건이 발생할 확률과 그 사건이 발생한 경우 다시 A가 발생한 경우의 곱

1. 통계와 확률

확률 (probability)

9. 베이즈 정리

- 사건 B가 발생함으로써 (사건 B가 진실이라는 것을 알게 됨으로써 즉, 사건 B의 확률 P(B)=1 이라는 것을 알게 됨으로써) 사건 A의 확률이 어떻게 변화하는지 표현한 정리
- 사건 B가 발생하였다는 것은 우리가 찾는 샘플이 사건 B라는 부분집합에 포함되어 있다는 새로운 정보를 취득하였다는 의미이다.
- 즉, 베이즈 정리는 새로운 정보가 기존의 의사 결정에 어떻게 영향을 미치는지 설명한다.

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

- P(A) : 사전 확률 (prior), 사건 B가 발생하기 전에 가지고 있던 사건 A의 확률
- P(A|B) : 사후 확률 (posterior), 사건 B가 발생한 이후 갱신된 사건 A의 확률
- P(B|A) : likelihood, 사건 A가 발생한 경우 사건 B의 확률
- P(B) : 정규화 상수 (normalizing constant), 확률의 크기 조정

1. 통계와 확률

확률 (probability)

베이즈 정리 예시 (동전던지기)

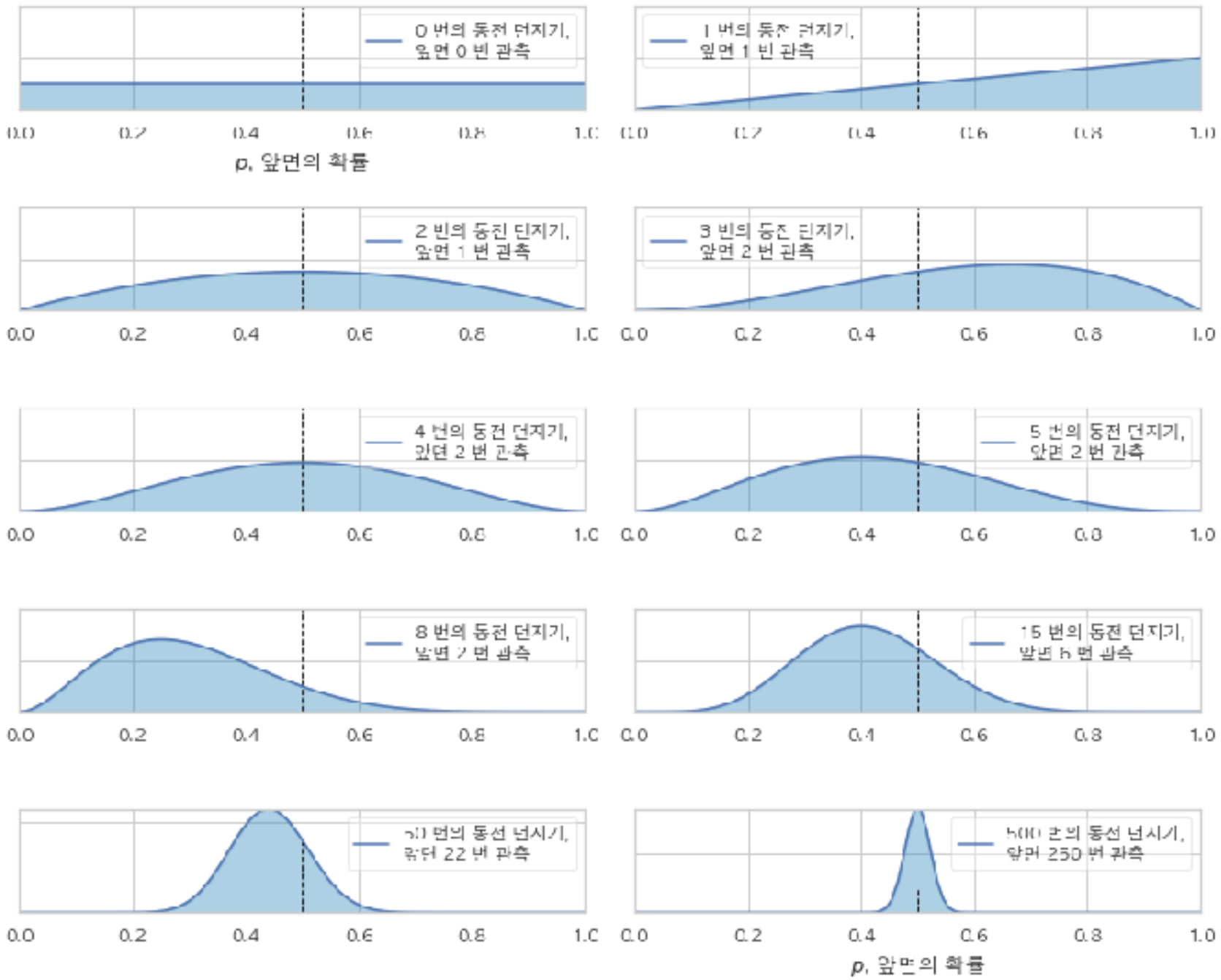
- 동전 던지기에서 앞면이 나올 확률을 확신하지 못하고 있는 상태.
- 실제 앞면이 나올 확률이 존재한다고 믿고 있으나 (이를 p라고 부르자), 다만 p가 무엇인지에 대한 사전적인 견해는 없다.
- 동전을 던지고, 앞면이나 뒷면의 관측 결과를 기록한다.

점점 더 많은 동전을 던지고 관측할수록 p에 대한 우리의 추론은 어떻게 변하는가?

1. 통계와 확률

확률 (probability)

베이즈 정리 예시 (동전던지기)



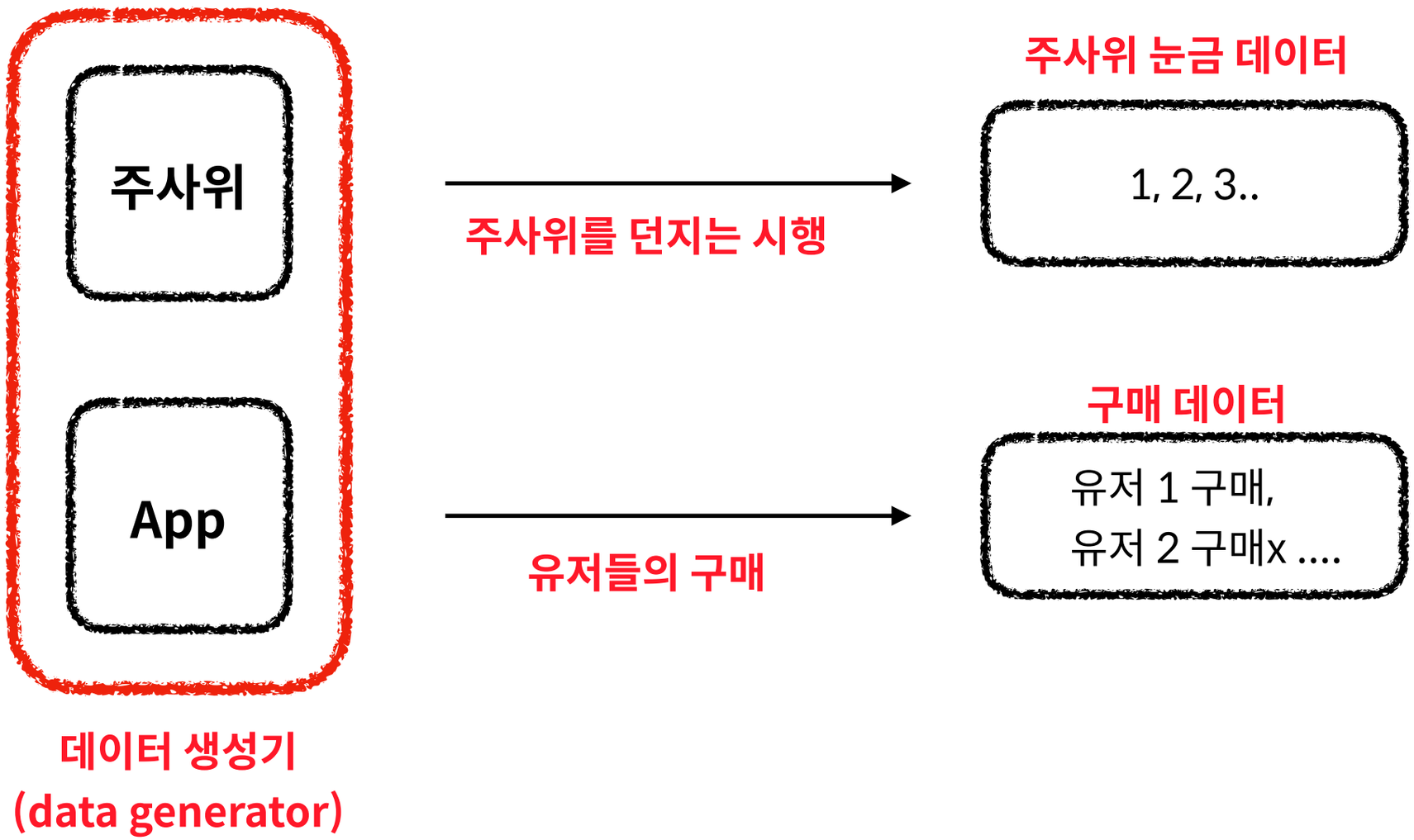
1. 통계와 확률

확률 모형

1. 확률 모형

지금부터 우리가 데이터를 바라봐야하는 관점

- 앞으로 다루게되는 데이터는 **어떤 실험(experiment)** 혹은 **조사 (research)** 행위에 의해 얻을 수 있다.



1. 통계와 확률

확률 모형

1. 확률 모형

- 데이터 분석에 확률론을 적용하면 데이터를 생성한 구체적인 데이터 생성기가 존재하지 않더라도 다음처럼 가상의 데이터 생성기가 존재한다고 생각할 수 있다.

(1) **데이터를 뽑을 수 있는 후보자 집합**: 가능한 모든 데이터값으로 이루어진 **표본 공간** Ω가 존재한다고 가정한다.

(2) **데이터를 뽑는 방법**: 표본 공간의 모든 사건에 대해서 **확률**이 배정되어 있다고 가정한다.

(엄밀하게는 모든 사건이 아니어도 되지만 여기에서는 이렇게 생각하자.)

이렇게 표본 공간과 확률이 정해져 있으면 이 두 가지를 사용하여 데이터를 생성할 수 있다. 이를 **확률 모형**이라고 부른다.

확률 모형은 주사위나 App처럼 내가 원하는 시점에 데이터를 생성하는 일종의 기계(machine)라고 생각하면 된다.

1. 통계와 확률

확률 모형

2. 샘플링, 실현

- 우리가 가진 데이터가 확률모형이라고 하는 가상의 주사위에 의해 생성된 것이라고 할 때, 이 주사위를 던져서 데이터를 생성하는 과정을 **샘플링**(sampling) 또는 **실현**(realization)이라고 한다. 또한 샘플링을 통해 얻어진 데이터를 **표본**이라고 한다.
  - 샘플링은 다른 의미로도 사용되는데 많은 수의 데이터 집합에서 일부 데이터만 선택하는 과정도 샘플링이라고도 한다.

3. 데이터의 특성

- 확률 모형론에서는 **데이터의 개별적인 값 하나 하나에는 의미가 없으며 데이터 전체의 특성만이 중요하다**고 생각한다. 또 특성이 같은 데이터는 실질적으로 동일한 정보를 주는 데이터라고 본다.

1. 통계와 확률

확률 모형

확률 모형과 실제 데이터의 관계

- 확률 모형으로부터 데이터를 여러번 생성하는 경우 실제 데이터 값은 매번 달라질 수 있지만 확률 모형 자체는 변하지 않는다.
- 확률 모형은 우리가 직접 관찰할 수 없다. 다만 확률 모형에서 만들어지는 실제 데이터 값을 이용하여 확률 모형이 이러한 것일 거라고 추정하고 가정할 뿐이다.
- 확률 모형에서 만들어 지는 실제 데이터의 값은 확률 모형이 가진 특성을 반영하고 있다. 다만 데이터의 개수가 적을 수록 부정확하여 확률 모형이 가진 특징을 정확하게 추정할 수 없다.

4. 데이터 분석의 과정

(1) 데이터를 확보한다.
(2) 확보된 데이터를 어떤 **확률 모형의 표본**으로 가정한다.
(3) 데이터의 특성으로부터 **확률 모형의 특성**을 추정한다.
(4) 구해진 확률 모형의 특성으로 해당 확률 모형의 종류를 결정하고 모수를 **추정**한다.
(5) 구해진 확률 모형으로부터 다음에 생성될 데이터나 데이터 특성을 **예측**한다.

1. 통계와 확률

확률 분포

1. 확률 변수 (random variable)

- 표본 공간의 모든 표본에 대해 어떤 실수 값을 붙인 것이다.
  - 특정 확률 분포와 연관되어 있는 변수를 의미하기도 한다.

(1) 이산 확률 변수 (discrete random variable)

  - 확률 변수값이 연속적이지(continuous) 않고 떨어져(discrete) 있는 경우 (ex. 주사위)

(2) 연속 확률 변수 (continuous random variable)

  - 확률 변수의 값이 실수 집합처럼 연속적이고 무한개의 경우의 수를 가진 경우 (ex. 시계 바늘의 위치)

확률 vs 확률 변수

- **확률**은 표본으로 이루어진 집합 즉, 사건에 대해 할당된 숫자이지만, **확률 변수**는 표본 하나 하나에 대해 할당된 숫자이다.

- **확률**은 0부터 1사이의 숫자만 할당할 수 있지만 **확률 변수**는 모든 실수 범위의 숫자를 할당할 수 있다.

1. 통계와 확률

확률 분포

2. 확률 분포

- 확률의 정의에서 확률은 **사건(event)이라는 표본의 집합에 대해 할당된 숫자**라고 하였다. 데이터 분석을 하려면 확률이 구체적으로 어떻게 할당되었는지를 묘사(describe)하거나 전달(communicate)해야 할 필요가 있다. **어떤 사건에 어느 정도의 확률이 할당되었는지를 묘사한 것을 확률 분포(distribution)**라고 한다.

(1) 누적 확률 밀도 함수 (cumulative probability density function)

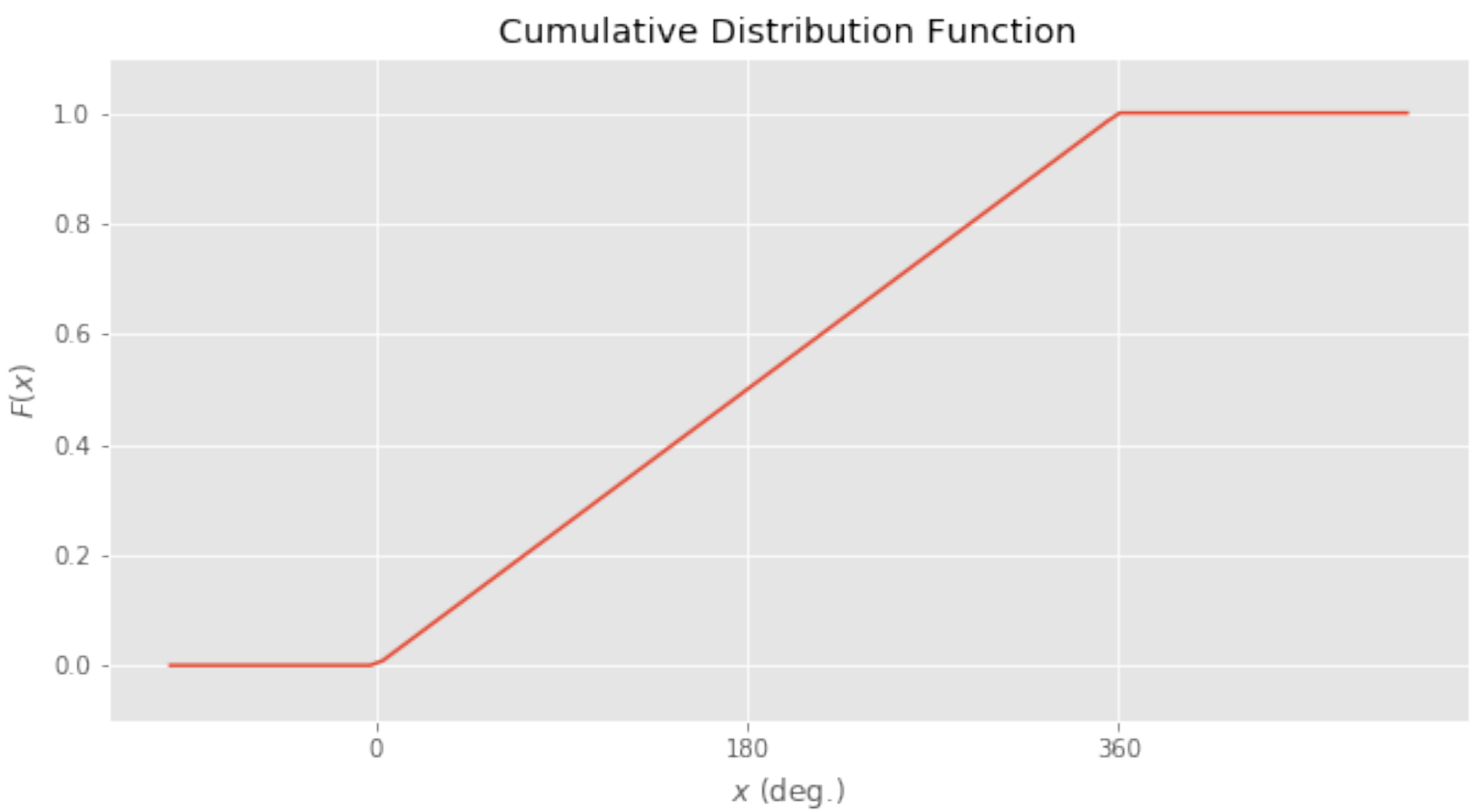
- 연속 분포를 표현하기 위해 사용된다.
- 확률이 어느 사건(event)에 어느 정도 분포되어 있는지 수학적으로 명확하게 표현해 준다.
- 일반적으로 cdf는 대문자를 사용하여 F(x)와 같은 기호로 표시하며 이 때 독립 변수 x는 범위의 끝을 뜻한다.
- 범위의 시작은 일반적으로 음의 무한대(negative infinity, $-\infty$) 값을 사용한다.

1. 통계와 확률

확률 분포

시계 바늘 문제

- 시침이 어떠한 위치에 있을 확률을 cdf를 통해 나타내보자.
  - 시침의 위치는 연속 분포를 이룬다.

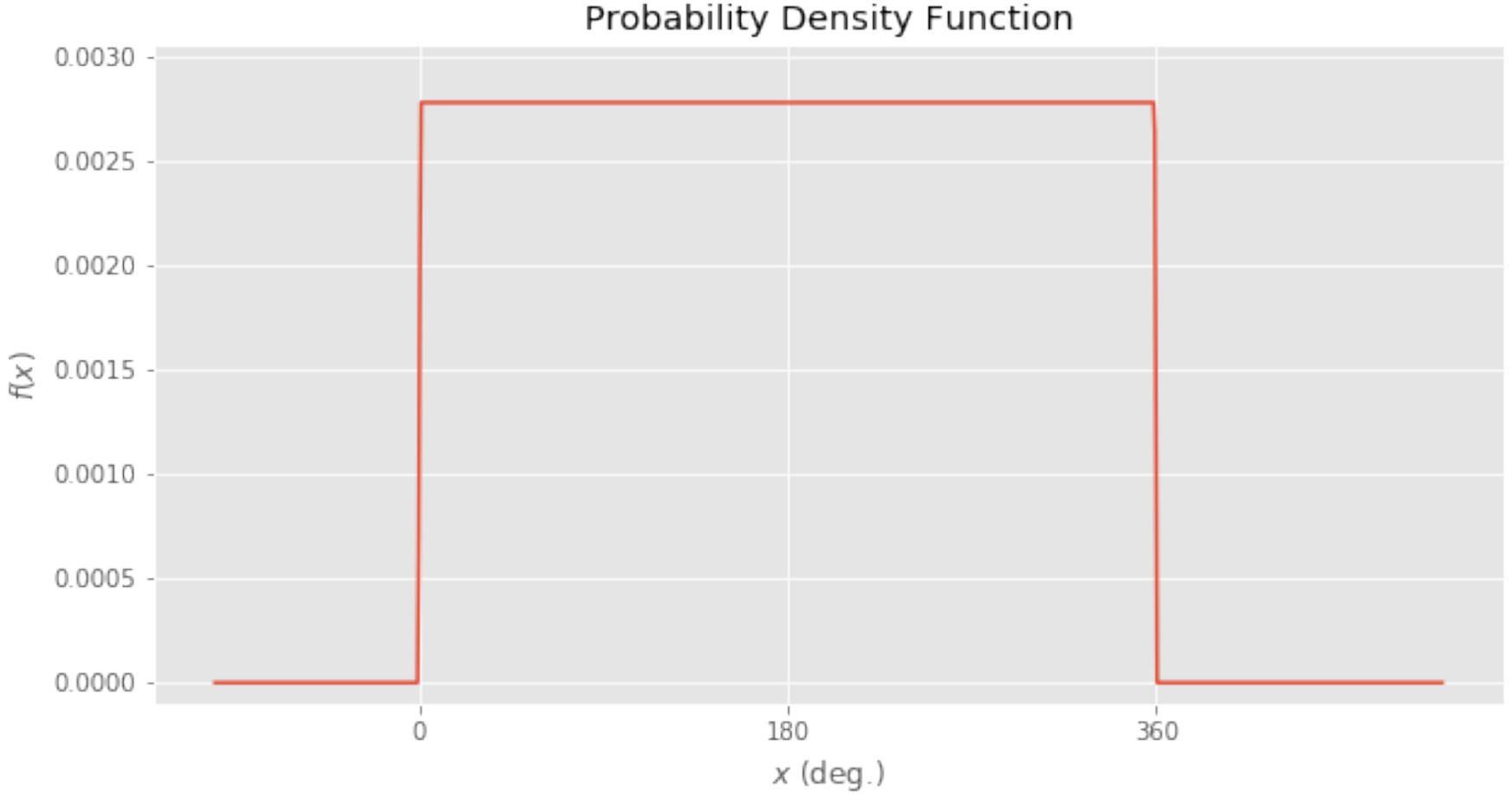


1. 통계와 확률

확률 분포

(2) 확률 밀도 함수 (probability density function, pdf)

- 누적밀도함수는 어떤 값이 더 자주나오든가 혹은 더 가능성이 높은지에 대한 정보를 알기 힘들다.
- 이를 알기 위해서는 전체 구간을 아주 작은 폭을 가지는 구간들로 나눈 다음 각 구간의 확률을 살펴보는 것이 편리하다
  - 각 구간을 나누기 위해 미분 (differentiation)을 사용한다.
- 이를 통해 상대적인 확률 분포를 알 수 있다.

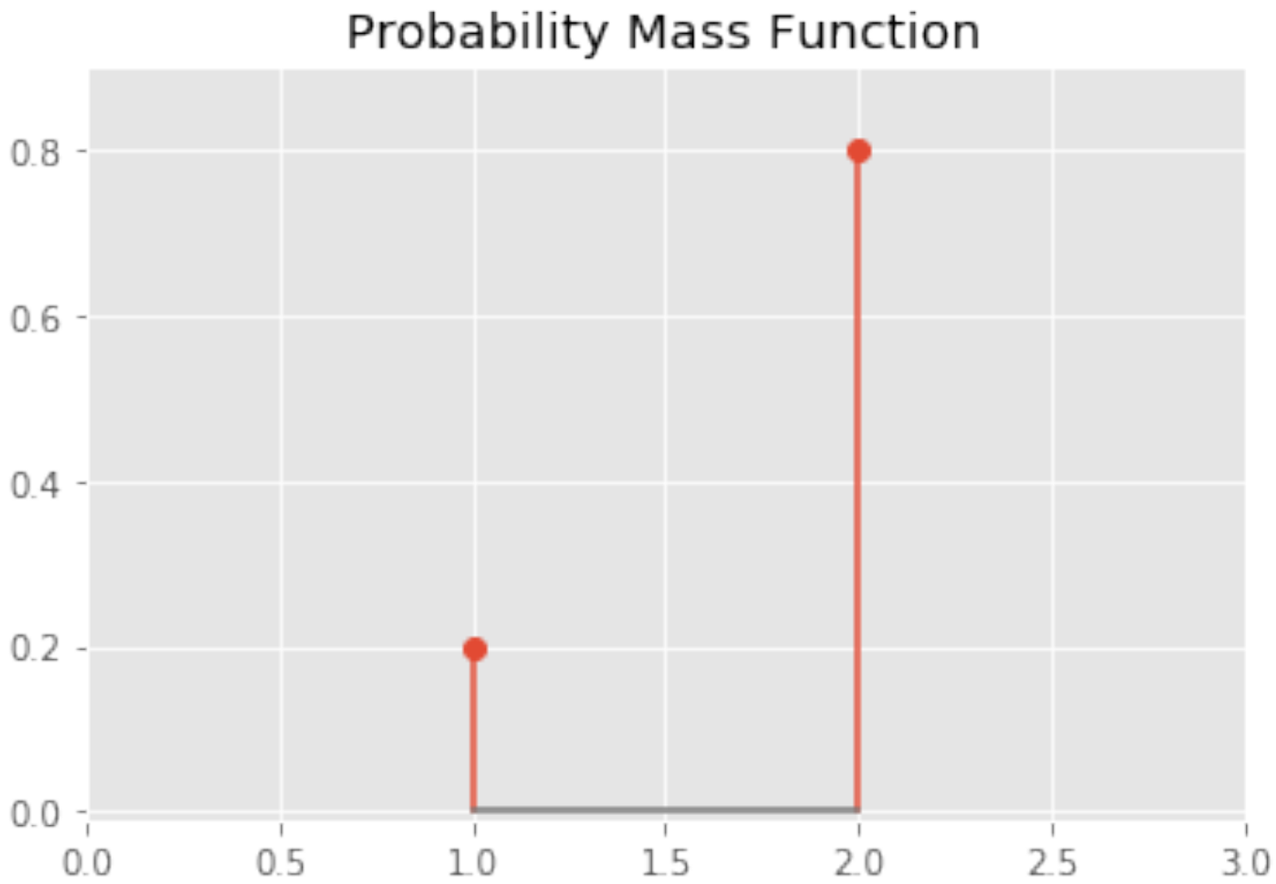


1. 통계와 확률

확률 분포

(3) 확률 질량 함수 (probability mass function, pmf)

- 이산 확률 변수의 가능한 값 하나하나에 대해 확률을 정의한 함수
- 앞면이 나올 확률이 0.8인 동전은 아래와 같은 확률 질량 함수를 가질 수 있다.



1. 통계와 확률

확률 분포

3. 확률 분포의 종류

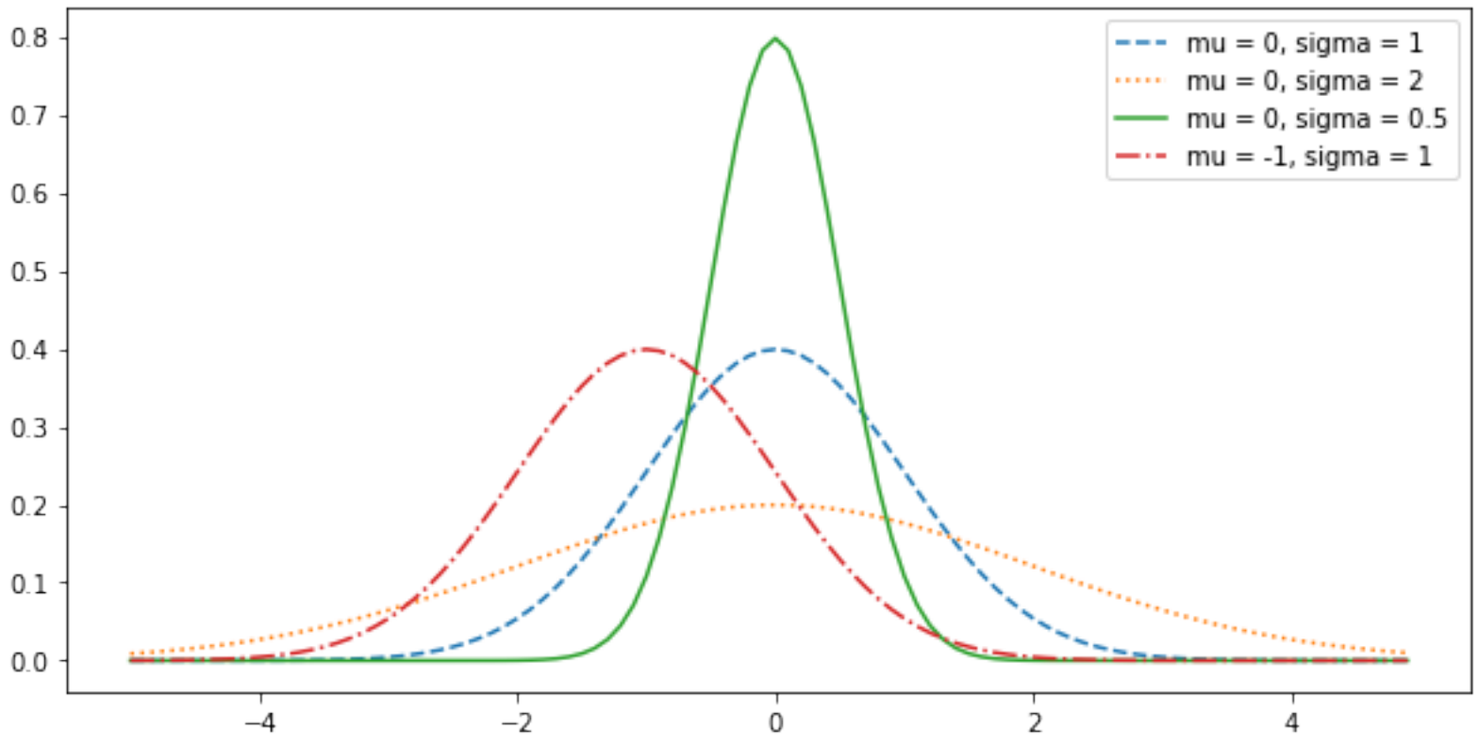
(1) 정규분포 (Gaussian normal distribution)

- 자연 현상에서 많이 나타나는 숫자를 확률 모형으로 모형화할 때 가장 많이 사용된다.
- 종형 곡선 모양의 분포이며, 평균과 표준편차로 정의된다.
  - 평균 : 종의 중심이 어디인지
  - 표준 편차 : 종의 폭이 얼마나 넓은지

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

평균 분산 (표준편차^2)

확률밀도함수 (pdf)



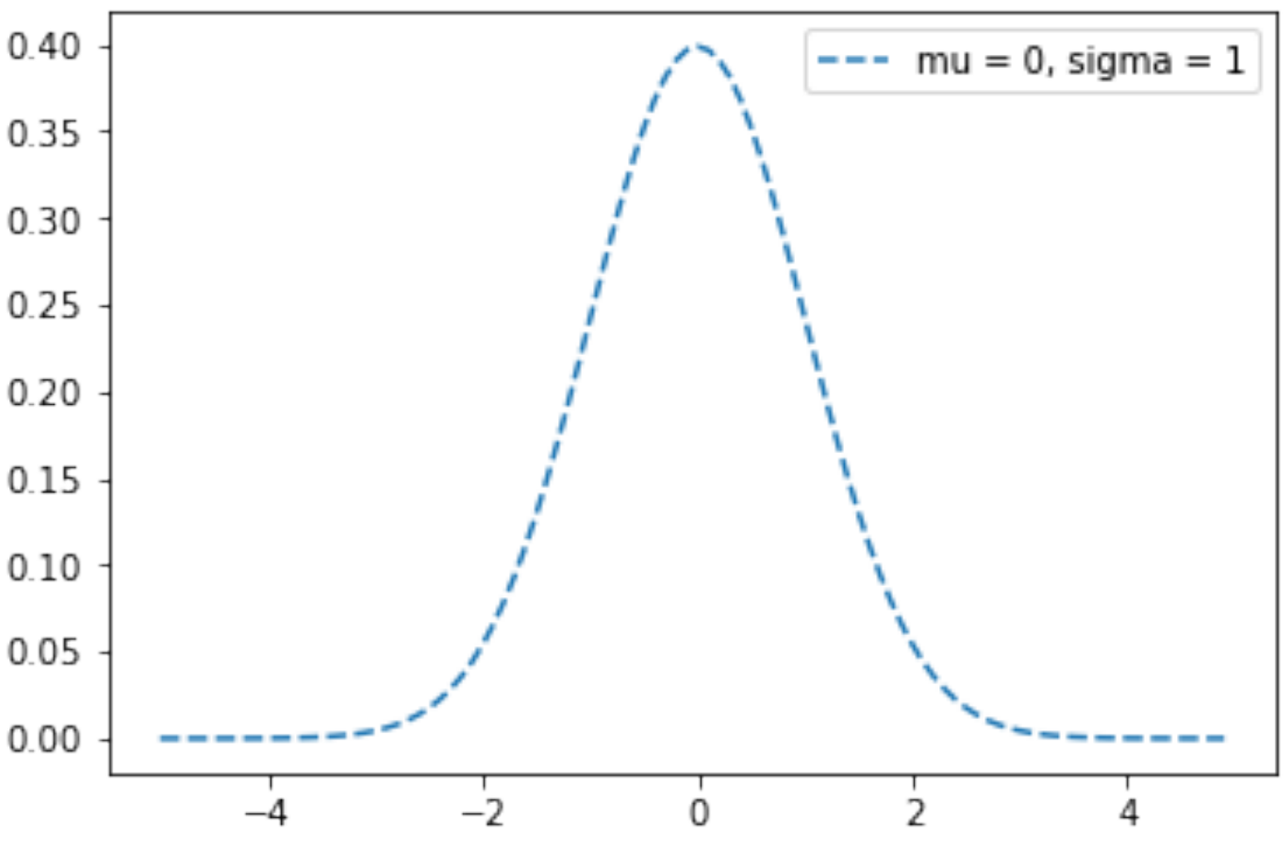
1. 통계와 확률

확률 분포

3. 확률 분포의 종류

(2) 표준 정규분포

- 평균이 0이고, 표준편차가 1인 정규분포 $\mathcal{N}(x; 0, 1)$



1. 통계와 확률

확률 분포

3. 확률 분포의 종류

(3) 베르누이 분포

- 베르누이 시도
  - 결과가 성공(Success) 혹은 실패(Fail) 두 가지 중 하나로만 나오는 것 (ex. 동전)
- 베르누이 확률 변수는 0, 1 두 가지 값 중 하나만 가질 수 있으므로 이산 확률 변수이다. 따라서 확률 질량 함수(pmf: probability mass function)로 정의할 수 있다.

$$Bern(x;\theta)=\begin{cases}\theta & if \quad x=1,\\ 1-\theta & if \quad x=0\end{cases}$$

$$Bern(x;\theta)=\theta^{x}(1-\theta)^{(1-x)}$$

변수 (variable)　　모수 (parameter)

- 어떤 확률 변수 X가 베르누이 분포에 의해 발생된다면 **"확률 변수 X가 베르누이 분포를 따른다"**라고 말하고 다음과 같이 수식으로 쓴다.

$$X\sim Bern(x;\theta)$$

1. 통계와 확률

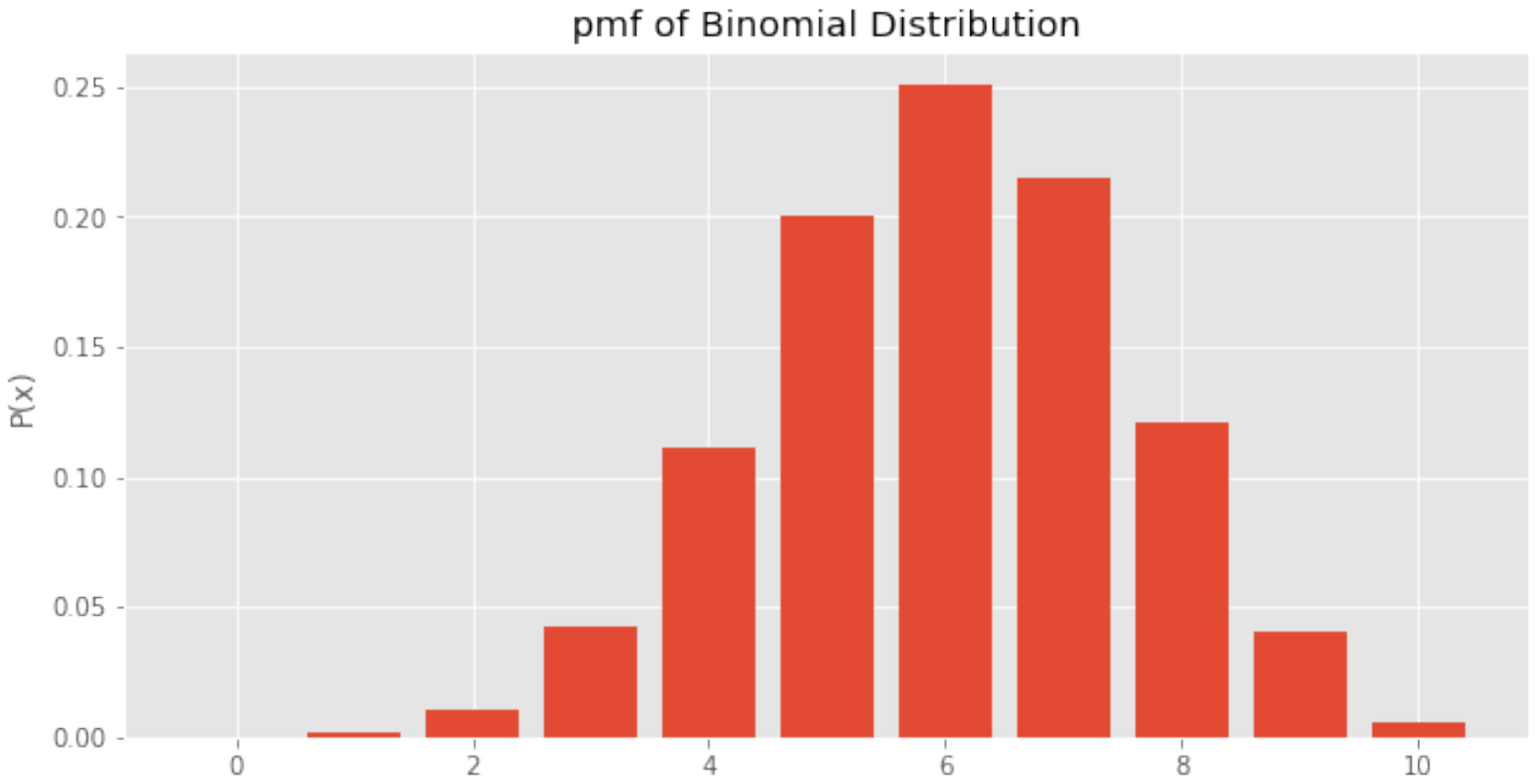
확률 분포

3. 확률 분포의 종류

(4) 이항 분포 (binomial distribution)

- 베르누이 시도를 N번 하는 경우를 생각하자.

$$Bin(x;N,\theta)=\binom{N}{x}\theta^x(1-\theta)^{N-x} \qquad X\sim Bin(x;N,\theta)$$



1. 통계와 확률

확률 분포

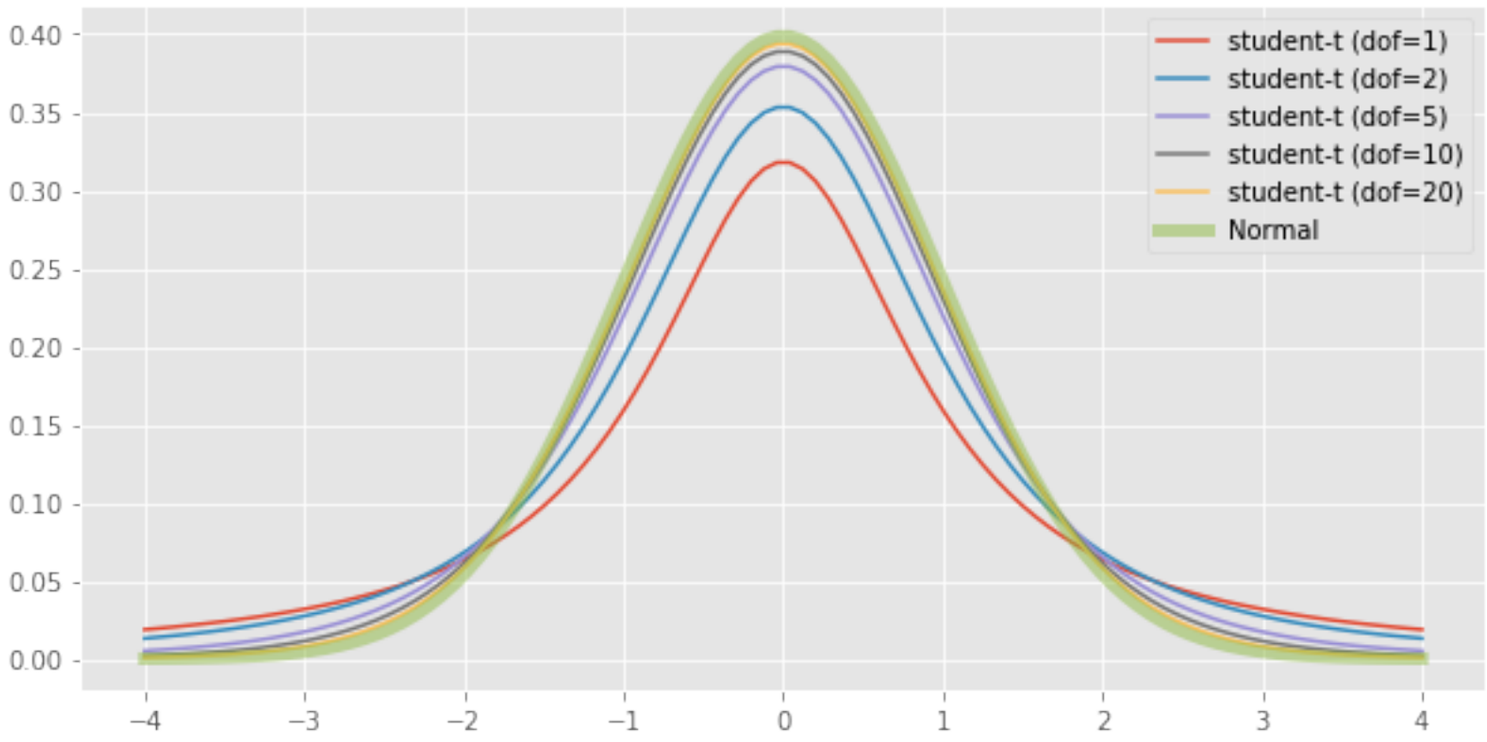
3. 확률 분포의 종류

(5) t 분포

- 우리가 분석해야할 실제 데이터들의 분포는 정규분포보다 양 끝단의 비중이 더 큰것을 알 수 있다. 이를 fat tail 현상이라고 한다.
- Fat tail 을 가진 데이터 모형에 적합한 것이 스튜던트 t 분포이다.

$$X \sim t(x; \mu, \sigma^2, \nu)$$

자유도
(Degree Of Freedom)



2. 가설 검정

> **검정(testing)**
>
> 데이터 뒤에 숨어있는 확률 변수의 분포와 모수에 대한 가설의 진위를 정량적(quantitatively)으로 증명하는 작업.

1. 가설 검정

가설 검정으로 접근할 수 있는 문제들..

문제 1.

어떤 동전을 15번 던졌더니 12번이 앞면이 나왔다.
이 동전은 휘어지지 않은 공정한 동전(fair coin)인가?

문제 2.

어떤 트레이더의 일주일 수익률은 다음과 같다.
-2.5%, -5%, 4.3%, -3.7% -5.6%
이 트레이더는 계속해서 돈을 잃을 사람인가?

1. 가설 검정

가설 검정의 기본 논리

1. 데이터가 어떤 고정된(fixed) 확률 분포를 가지는 **확률 변수**라고 가정한다.

2. 이 확률 분포의 모수값이 특정한 값을 가진다고 가정한다. 이 때 모수가 가지는 특정한 값은 우리가 검증하고자 하는 사실과 관련이 있어야 한다. 이러한 가정을 **귀무 가설(null hypothesis)**이라고 한다.

3. 만약 데이터가 주어진 귀무 가설에 따른 표본이고 이 표본 데이터를 특정한 수식에 따라 계산한 숫자는 특정한 확률 분포를 따르게 된다. 이 숫자를 **검정 통계량(test statistics)**라고 하며 검정 통계량의 확률 분포를 **검정 통계 분포(test statistics distribution)**라고 한다. 검정 통계 분포의 종류 및 모수의 값은 처음에 정한 가설 및 수식에 의해 결정된다.

1. 가설 검정

가설 검정의 기본 논리

> **4.** 주어진 귀무 가설이 맞으면서도 표본 데이터에 의해서 실제로 계산된 검정통계량의 값과 같은 혹은 그보다 더 극단적인(extreme) 또는 더 희귀한(rare) 값이 나올 수 있는 확률을 계산한다. 이를 **유의 확률(p-value)**이라고 한다.

> **5.** 만약 유의 확률이 미리 정한 특정한 기준값보다 작은 경우를 생각하자. 이 기준값을 **유의 수준(significance level)**이라고 하는 데 보통 1% 혹은 5% 정도의 작은 값을 지정한다. 유의 확률이 유의 수준으로 정한 값(예 1%)보다도 작다는 말은 해당 검정 통계 분포에서 이 검정 통계치(혹은 더 극단적인 경우)가 나올 수 있는 확률이 아주 작다는 의미이므로 가장 근본이 되는 가설 즉, 귀무 가설이 틀렸다는 의미이다. 따라서 이 경우에는 귀무 가설을 **기각(reject)**한다.

> **6.** 만약 유의 확률이 유의 수준보다 크다면 해당 검정 통계 분포에서 이 검정 통계치가 나오는 것이 불가능하지만은 않다는 의미이므로 귀무 가설을 기각할 수 없다. 따라서 이 경우에는 귀무 가설을 **채택(accept)**한다.

3. 회귀분석 심화

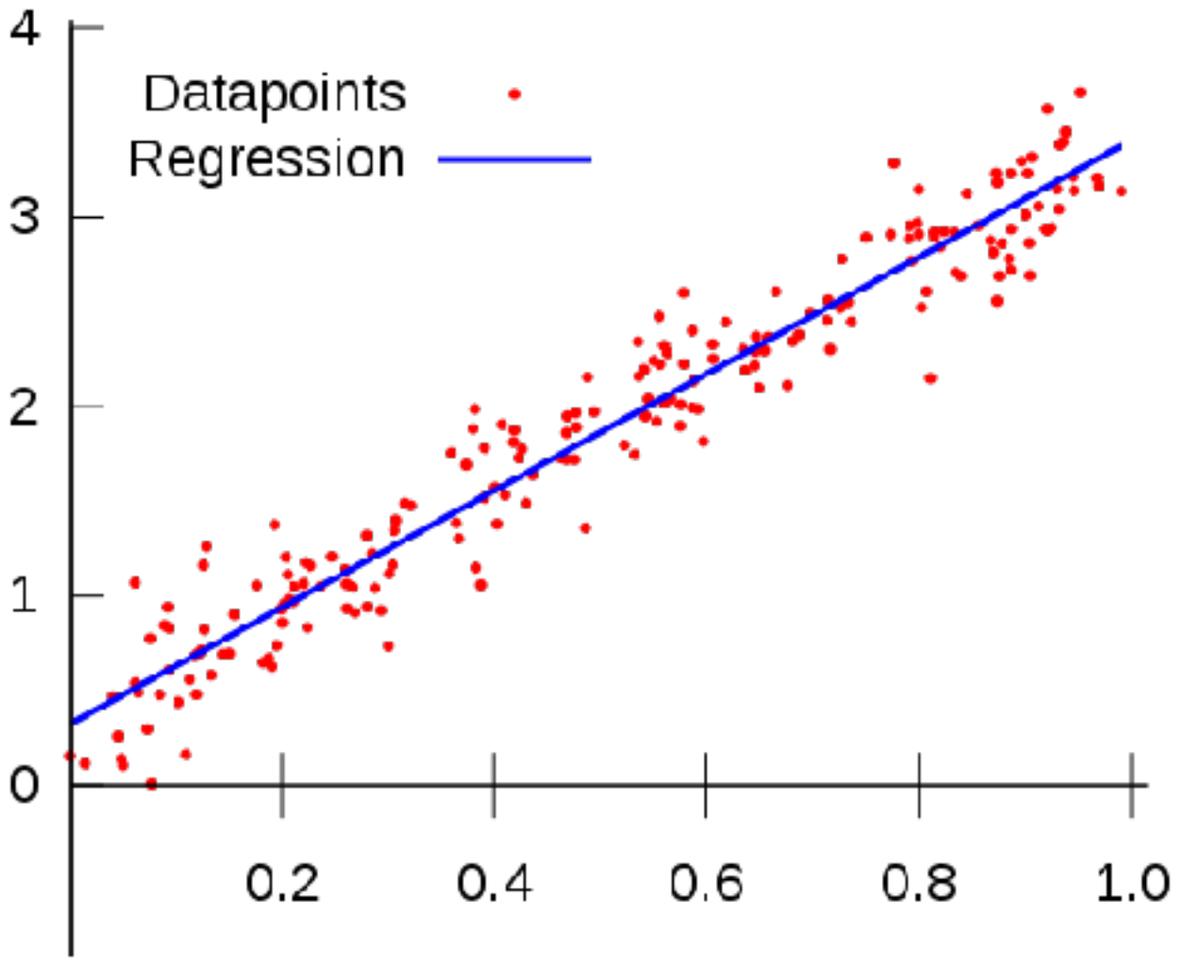
선형회귀 (Linear Regression)

3. 회귀분석 심화

선형회귀

수치형 **설명변수 X**와 **연속형 숫자**로 이뤄진 **종속변수 Y**간의 관계를 **선형**으로 가정하고
이를 가장 잘 표현할 수 있는 **회귀계수**를 데이터로부터 추정하는 모델

- **예시 : 집 크기 (x)에 대한 집 값 (y) 예측**



$$Y = Xw$$

3. 회귀분석 심화

선형회귀식의 표기

$$Y=Xw$$

- 설명 변수 x가 D개가 있고, 데이터가 N개가 있을 때, x에 대한 전체 데이터를 오른쪽과 같이 (D X N) 행렬로 표기가 가능하다.

$$x_i=\begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iD} \end{bmatrix} \rightarrow X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix}$$

- 전체 수식은 설명 변수 벡터(x)와 가중치 벡터(w)의 내적으로 간단하게 나타낼 수 있다.

$$f(x)=w_1x_1+w_2x_2+\cdots+w_Dx_D=\begin{bmatrix} x_1 & x_2 & \cdots & x_D \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}=x_a^Tw_a=w_a^Tx_a$$

3. 회귀분석 심화

목적함수

RSS (Residual Sum of Squares)

- 실제값 - 예측값

$$\hat{y} = Xw$$

$$e = y - \hat{y} = y - Xw$$

$$\begin{aligned} RSS &= e^T e \\ &= (y - Xw)^T (y - Xw) \\ &= y^T y - 2y^T Xw + w^T X^T Xw \end{aligned}$$

- 선형회귀분석도 마찬가지로 RSS 를 minimize 하기 위한 w 를 구하는 것이다.

3. 회귀분석 심화

최적화

- RSS의 최솟값을 구하기 위해 그레디언트 벡터를 구한다. (RSS를 w에 대해 미분한다)

$$\frac{dRSS}{dw} = -2X^T y + 2X^T Xw$$

- RSS가 최소가 되는 최적화 조건은 그레디언트 벡터가 0벡터이어야 하므로 다음 식이 성립한다.

$$\frac{dRSS}{dw} = 0$$

$$X^T Xw^* = X^T y$$

- X^TX의 역행렬이 존재한다면 최적 가중치 벡터 w를 구할 수 있다.

$$w^* = (X^T X)^{-1} X^T y$$

- 이와 같은 방식으로 명시적 해를 단번에 추정할 수 있다.

3. 회귀분석 심화

선형회귀식의 적합성

결정 계수 (R square)
- 종속변수 (y)의 총 변화량 중 모델이 잡아낼 수 있는 변화량의 비율

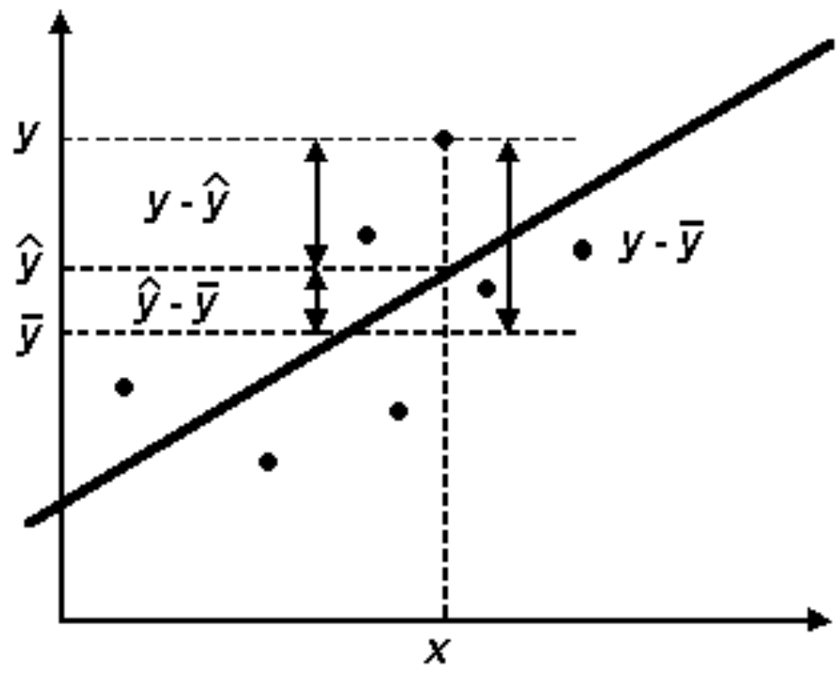
- **결정계수 = 1- RSS (모델이 잡아내지 못한 변화량) / TSS (데이터 전체 변화량)**
  -> 전체 변화량 중 모델이 잡아내지 못하는 변화량의 비율
  - 만들어낸 모델이 실제 관측된 값의 평균값 정도만 예측한다고 하면 결정계수는 0이 된다.

$$RSS(Residual\ Sum\ of\ Square) = \sum (y - \hat{y})^2$$ (실제값 - 예측값)

$$TSS(Total\ Sum\ of\ Square) = \sum (y - \overline{y})^2$$ (실제값 - 평균)

$$ESS(Explained\ Sum\ of\ Square) = \sum (\hat{y} - \overline{y})^2$$ (예측값 - 평균)

$$R^2 = 1 - \frac{RSS}{TSS} \quad (0 \leq R^2 \leq 1)$$



결국, RSS를 줄이면 결정계수는 커지게 된다.

3. 회귀분석 심화

회귀계수의 신뢰성 문제

부트스트래핑(Bootstrapping)

- OLS(Ordinary Least Square) 방법을 사용하면 데이터에 대한 확률론적인 가정없이도 최적의 가중치를 계산할 수 있다. 그러나 이 경우에는 계산한 가중치가 어느 정도의 신뢰도 또는 안정성을 가지는지 확인할 수 있는 방법이 없다.
- 부트스트래핑(bootstrapping)은 회귀 분석에 사용한 데이터가 달라진다면 회귀 분석의 결과는 어느 정도 영향을 받는지를 알기 위한 방법이다.

절차

- 기존 데이터에서 중복 재추출하여 새로운 데이터를 만들어낸다.
- 새로운 만들어진 데이터에 대해 회귀분석을 실시한다.
- 분석 결과를 저장하고, 통계치를 확인한다.

3. 회귀분석 심화

정규화 (Regularization)

변수가 많을 때 발생할 수 있는 문제점

(1) 오버피팅의 가능성

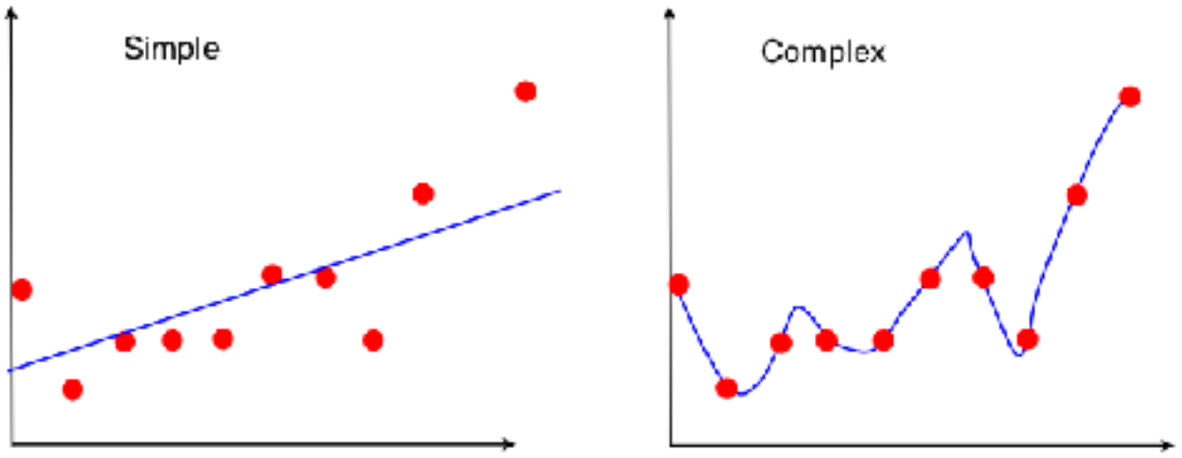
(2) 모델 해석의 어려움

- 회귀계수가 너무 많으면 모델을 해석하기 어려워진다.

- 이 문제를 해결하기 위해, 적은 개수의 변수로 일반화된 모델을 만들기도한다.

정규화 (Regularization)

- 회귀계수가 가질 수 있는 값에 제약조건을 부여하여 일반화 성능을 높이는 기법
- 모델의 설명력은 다소 포기하더라도 안정적인 결과를 내도록 만든다.
  - 안정적인 결과를 만들어내는 변수들을 선택하게된다.



3. 회귀분석 심화

정규화 (Regularization)

Ridge Regression (릿지 회귀)

- 잔차제곱합(RSS)를 최소화하면서 회귀계수 벡터 w에 L2 norm을 제한하는 기법이다.

$$L_2 = \sqrt{\sum_{i}^{n} x_i^2}$$

  - 어떤 크기의 람다를 선택하느냐에 따라 제약의 정도가 결정된다.
- 총 계수의 합을 줄여주는 효과가 있다.

$$\hat{w}^{ridge} = arg\min_{w} \{(Y - Xw)^T(Y - Xw) + \lambda w^T w\}$$

$$\hat{w}^{ridge*} = (X^T X + \lambda I)^{-1} X^T Y$$

3. 회귀분석 심화

정규화 (Regularization)

Lasso Regression (라쏘 회귀)

- 잔차제곱합(RSS)를 최소화하면서 회귀계수 벡터 w에 L1 norm을 제한하는 기법이다.

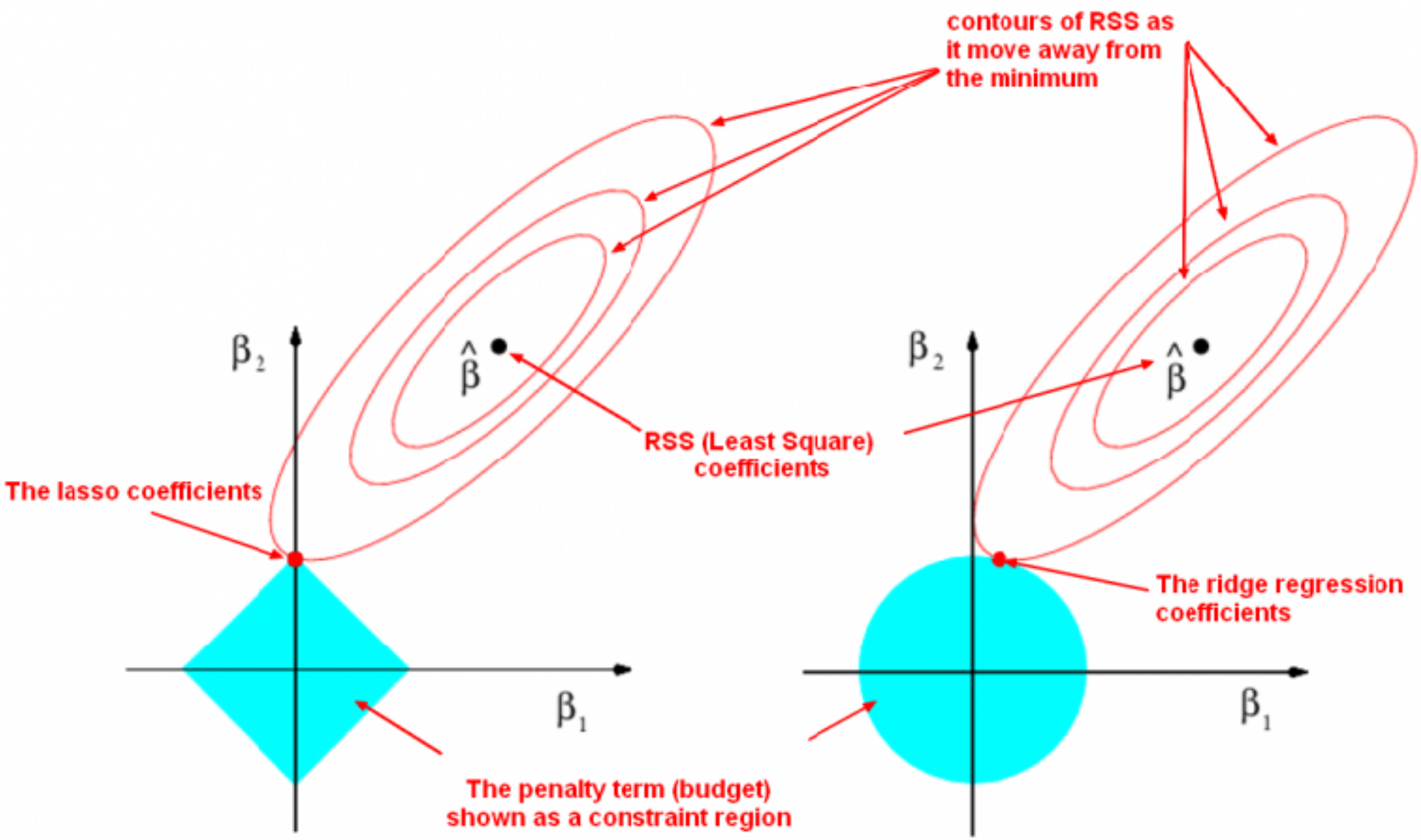
$$L_1 = \sqrt{\sum_{i}^{n} |x_i|}$$

$$\hat{w}^{lasso} = arg\min_{w} \{(Y - Xw)^T(Y - Xw) + \lambda ||w||_1\}$$

- 제약식이 절대값으로 들어가기 때문에 릿지회귀처럼 미분 연산이 불가능하다.
  - 수치적 최적화로 w를 추정함.

3. 회귀분석 심화

정규화 (Regularization)



LASSO

RIDGE REGRESSION

다음에
또!
같이!
만나요!

learnit Flip your study

1. 통계와 확률

확률 (probability)

베이즈정리 확장

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B \cap A) + P(B \cap A^C)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

$$= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)(1 - P(A))}$$

1. 통계와 확률

확률 (probability)

검사 시약 문제 (74p)

- **양성 판정을 받았을 경우 실제로 병에 걸렸을 확률은?**
  - 사건 T : 양성판정
  - 사건 D : 질병에 걸리는 사건
  - 질병에 걸릴 확률 : 0.0001 (P(D))
  - 질병에 걸린 사람이 양성 판정을 받을 확률 : 0.99 (P(T|D))

$$P(D|T)=\frac{P(T|D)P(D)}{P(T)} \longrightarrow P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D)+P(T|D^C)P(D^C)}$$

$$\frac{(0.99\times0.0001)}{(0.99\times0.0001+0.01\times0.9999)}$$

1. 통계와 확률

확률 vs 통계

확률 : 어떤 일이 일어나기 전, 수학적 모델을 바탕으로 어떤 일이 일어날지 예측하는 것.

통계 : 주어진 데이터를 모두 관찰하고, 어떤 수학 모델을 통해 나왔는지 추론하는 것.