

## 확률을 수학적으로 정의하기 위한 개념들

- 실험
- 확률표본
- 표본공간
- 사건

### 0) 실험

- 실험 : 어떤 현상의 관찰 결과를 얻기 위한 과정

### 1) 표본공간과 확률표본

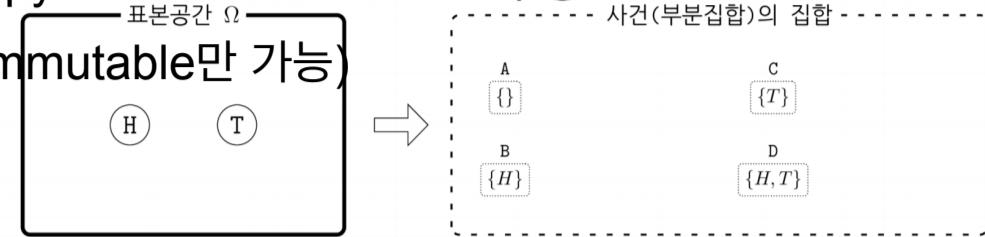
- 표본(sample) : 관찰 가능한 결과
- 표본공간(표본집합)(sample space) : 모든 관찰 가능한 결과의 집합

## 2) 사건

- 사건(event) : 표본공간의 부분집합 (사건 = event = 부분집합)

\*부분집합의 집합을 구현 시, python - frozenset 으로 구성

<== dict의 key로 넣기위해(immutable만 가능)



- 사건의 갯수 :  $2^{**n}$  (n : 표본공간 원소 갯수)

- 예시) 주사위를 1회 던져 나오는 눈의 수를 관찰하는 실험

1) 표본공간 = 모든 관찰 가능한 결과의 집합 = S : {1,2,3,4,5,6}

2) 사건 A = 짹수의 눈이 나오는 경우 = A : {2,4,6}

사건 B = 3의 배수의 눈이 나오는 경우 = B : {3,6}

... 무수히 많은 조합을 만들어 낼 수 있다.

\*6개의 원소를 가지고.\* 하지만, 그 조합의 최대 갯수는  $2^{**n}$ .

3) 사건의 갯수 = 가능한 모든 부분집합의 갯수 =  $2^6$

<== 원소의 갯수가 6개니까, 가능한 조합은  $2^6$  개 까지 가능.

ex)  $P(A) = P(\{2,4,6\}) = 2,4,6$  란 선택지를 갖는 사건(부분집합) A의 확률

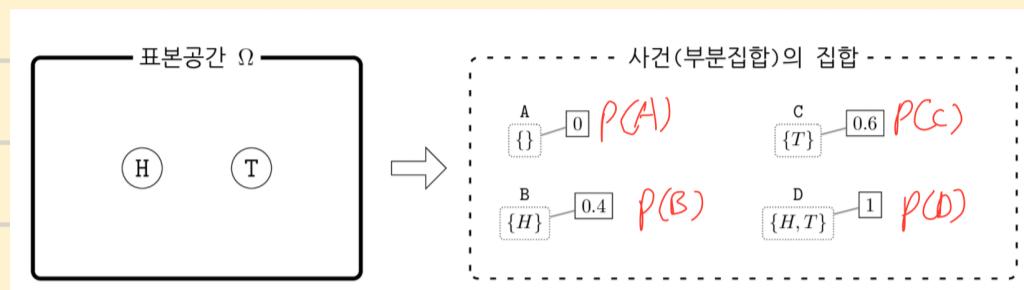
$\neq$  1번씩 총 3번 던져서 각각 2,4,6이 나올 확률

## 3) 확률

↑ 단순사건이 아니라면, 다변수  $\Sigma$ 칼라 함수와 같다.

수학적 정의와 의미

- 확률(probability) : 사건(부분집합)을 입력하면 숫자(확률값)이 출력되는 함수.
  - \* 함수의 정의역(domain) : 모든 사건(sample space의 모든 부분집합)
  - \*  $P(A) \Rightarrow P$ 는 함수,  $P(A)$ 는 A라는 사건(부분집합)에 할당된 숫자를 뜻함
- \*(06.02 5page)



- 콜모고로프의 공리

- 1)  $P(A) \geq 0$

- 2)  $P(\text{표본공간}) = 1$  \*'표본공간'이라는 사건(부분집합)에 대한 확률은 1이다.

- 3) 서로소의 가산가법성

: 상호배반인 사건의 합집합 확률은 각 사건 확률의 합과 같다.

\*  $P(A \cup B) = P(A) + P(B)$ , if A와 B의 교집합이 공집합 뿐

\* A와 B의 교집합이 공집합인 경우,

A와 B는 상호배반(서로소)이라고 한다. (A와 B의 공통원소가 없다.)

\* 공집합은 모든 부분집합과 교집합

- 확률은 표본이 아닌 사건을 입력으로 갖는 함수

\* 잘못된 표현 :  $P(1) = 1/6$ , 제대로된 표현 :  $P(\{1\}) = 1/6$

(1은 표본일 뿐, 사건이 아니다. 사건은 부분'집합'이다.)

## 4) 확률의 의미

06.02 확률의

수학적 정의와 의미

1) 빈도주의적 관점(현실) :

$P(A)$  = 반복적으로 선택된 '표본'이 사건A의 '원소'가 될 경향

ex) $P(\{2,4,6\})$  = 수 많이 숫자 선택 시, 2,4,6 중 하나가 나올 경향

2) 베이지안 관점(믿음) :

$P(A)$  = 선택된 '표본'이 사건A에 속한다는 "가설, 문제, 주장의 신뢰도"  
(반복 개념X)

ex) $P(\{2,4,6\})=1$  이라면, 주사위를 던지면 2,4,6 중 하나 나온다는 주장의 신뢰도가 100% (항상 맞다)

ex) $P(\{2,4,6\})=0.5$  라면, 주사위를 던지면 2,4,6 중 하나 나온다는 주장의 신뢰도가 50%. (몇번 해보면 그 중 반쯤 맞다)

3) 베이지안 관점에서의 사건( $\{2,4,6\}$  or A) :

가능한 후보의 집합 = 원하는 답(선택된 표본)이 포함되어있을 가능성이 있는 후보의 집합

"사건이 발생했다" : 베이지안은 "그 주장이 진실임을 알게 되었다.

= 그 전에 몰랐던 추가적인 정보가 들어옴"

ex)주사위의 눈금이 짹수가 나오는 사건이 발생했다

= 주사위의 눈금이 짹수다. 라는 사실을 알게됨

\*빈도주의적 관점에서의 사건 : 모든 관찰가능한 결과의 집합 중 부분집합

4) 정리 (베이지안적 관점)

1. 사건( $\{2,4,6\}$ ) : 발생 가능한 후보의 집합 (빈도주의=> 사건 : 부분집합)

2. 확률( $P(\{2,4,6\})$ ) : 어떤 주장의 신뢰도

(빈도주의=> 확률 : 반복적으로 시행 시, 사건의 원소가 될 경향)

\*베이지안은 사전지식을 갖고, 사후확률을 계산한다.

여기서 사후확률은 궁극적으로, 사전지식이 맞을 확률이 된다.

(사후확률 = 가능성 \* 사전확률(분포, 불확실))

\*06.03 4page

#### 확률의 성질 요약

- 공집합의 확률

$$P(\emptyset) = 0 \quad (6.3.18)$$

- 여집합의 확률

$$P(A^C) = 1 - P(A) \quad (6.3.19)$$

$$0 \leq P(A) \leq 1 \quad (6.3.20)$$

- 포함-배제 원리

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (6.3.21)$$

- 전체 확률의 법칙

$$C_i \cap C_j = \emptyset \quad (i \neq j) \quad (6.3.22)$$

$$C_1 \cup C_2 \cup \dots = \Omega \quad (6.3.23)$$

$$P(A) = \sum_i P(A, C_i) \quad (6.3.24)$$

1)  $P(\text{공집합})$ , 공집합인 사건의 확률은 0이다.

2)  $P(\text{여집합})$ , 어떤 사건의 여집합인 사건의 확률은  $(1 - \text{원래 사건의 확률})$  같다.

3) 포함-배제 원리

: 두 사건의 합집합의 확률은 각 사건의 확률의 합에서 두 사건의 교집합의 확률을 뺀 것과 같다.

4) 전확률 법칙 (가장 중요)

: 06.03 3page

- 조건 2개 : 교집합이 없고, 모든 합집합 = 1

\*앞으로, 교집합은 ', '로 표시한다. ex)  $P(A,B) = A\text{와 }B\text{의 교집합 확률}$

06.04

확률분포함수

- 확률밀도함수,  $f(x)$ , pdf, pmf : 확률질량함수(이산형), 확률밀도함수(연속형)
- 확률분포함수,  $F(x)$ , cdf : 누적분포함수 or 누적확률분포함수,  $F(a) = f(X \leq a)$

- 확률밀도함수의 조건 (질량, 밀도함수 모두)

- 1)  $f(x) \geq 0$
- 2)  $\sum [f(x)] = 1$

1) 단순사건과 확률질량함수 (pmf)

- 단순사건 : elementary event. 표본이 하나인 사건. ex)  $A = \{1\}$  or  $B = \{3\}$
- 확률질량함수 :

유한 개의 사건이 존재할 때, 각 단순사건의 확률을 정의하는 함수  $p(a) = P(\{a\})$

ex)  $P(\{1\}) = 0.2$ ,  $p(1) = 0.2$ ,  $p(1,2) =$  정의되지 않음.

- 확률을 단순사건으로 간단화 하기 때문에, 연산과 파악이 간단함.

\*전체 확률 = 1 =  $\sum(p(\text{단순사건}))$

2) 표본의 수가 무한한 경우 (pdf)

- 표본의 수가 무한할 경우, 단순사건에 대해선, 확률값이 0이 된다. (적분값 0)
- 따라서,  $P(\{1\})$  이 아닌,  $P(\{ < < \})$  와 같이, 범위를 준다.

06.04

### 3) 구간

확률분포함수

- $P(A) = P(\{a < x < b\}) = P(a, b)$

### 4) 누적분포함수 (cdf)

- 확률밀도함수의 변수를 2개에서 1개로

- $F(x) = P(\{X < x\}) = P(\{-\infty < X < x\})$

- $F(b) - F(a) = P(\{a < X < b\})$

- 변수가 하나가 되니, 그래프를 그리기 매우 쉬워진다. 1차원의 그래프를 그릴 수 있음 (정보 전달 용이)

### 5) pdf는 cdf의 1차 도함수

- pdf에서 확률은 '면적'이다.

- pdf의 높이는 단지, cdf의 그 점에서의 기울기 (pdf는 도함수니까)

1) 결합확률 (joint probability) :  $P(A, B)$ , 사건 A와 B가 동시에 발생할 확률

- 주변확률 (marginal probability) :

$P(A)$ ,  $P(B)$ , 결합확률과 대비되는 개념으로 결합되지 않은 개별사건의 확률

2) 조건부확률 : B가 사실인 경우, A의 확률

-  $P(A|B) = P(A, B)/P(B)$

- 06.05 3page, 2page

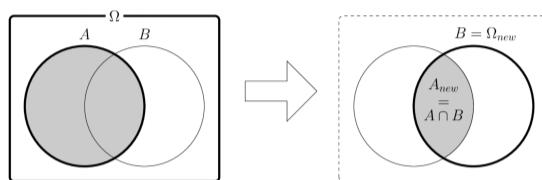
• 조건부확률  $P(A|B)$

- 사건 B가 발생한 경우의 사건 A의 확률
- 표본이 이벤트 B에 속한다는 새로운 사실을 알게 되었을 때,
- 이 표본이 사건 A에 속한다는 사실의 정확성(신뢰도)이 어떻게 변하는지를 알려준다.

조건부확률이 위와 같이 정의된 근거는 다음과 같다.

1. 사건 B가 사실이므로 모든 가능한 표본은 사건 B에 포함되어야 한다. 즉, 새로운 실질적 표본공간은  $\Omega_{\text{new}} \rightarrow B$  된다.
2. 사건 A의 원소는 모두 사건 B의 원소로 되므로 사실상 사건  $A \cap B$ 의 원소가 된다. 즉, 새로운 실질적  $A_{\text{new}} \rightarrow A \cap B$  된다.
3. 따라서 사건 A의 확률 즉, 신뢰도는 원래의 신뢰도(결합확률)를 새로운 표본공간의 신뢰도(확률)로 정규화(normalize)한 값이라고 할 수 있다.

$$P(A|B) = \frac{P(A_{\text{new}})}{P(\Omega_{\text{new}})} = \frac{P(A, B)}{P(B)} \quad (6.5.9)$$



3) 독립

- 06.05 5page, 6.5.12 - 13

독립

수학적으로는 사건 A와 사건 B의 결합확률의 값이 다음과 같은 관계가 성립하면 두 사건 A와 B는 서로 독립 (independent)이라고 정의한다.

$$P(A, B) = P(A)P(B) \quad (6.5.12)$$

독립인 경우 조건부확률과 원래의 확률이 같아짐을 알 수 있다. 즉, B라는 사건이 발생하든 말든 사건 A에는 전혀 영향을 주지 않는다는 것이다.

$$P(A|B) = \frac{P(A, B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A) \quad (6.5.13)$$

4) 조건부확률 => 원인/결과, 근거/추론, 가정/조건부결론

-  $P(A|B) = B$ 는 원인, 근거, 가정 / A는 결과, 추론, 조건부결론

- 6.5.14, 필기

현대  
논리학

① B는 진실 B  $P(B) = 0.8$

② B가 진실이면,  $B \rightarrow A$   $P(A|B) = 0.9$

③ A 진실

$P(A|B) = 0.72$

$$P(A, B) = P(A|B) \cdot P(B)$$

조건부결론 전제

- 6.5.16

위 식을 응용하면 다음과 같은 수식도 성립한다.

$$P(A, B, C) = P(A|B, C)P(B, C) \quad (6.5.15)$$

확률표기에서 쉼표(comma)가 교집합을 뜻한다는 것을 기억하면 이 식은 쉽게 증명할 수 있다.

$$\begin{aligned} P(A, B, C) &= P(A \cap B \cap C) \\ &= P(A \cap (B \cap C)) \\ &= P(A|B \cap C)P(B \cap C) \\ &= P(A|B, C)P(B, C) \end{aligned} \quad (6.5.16)$$

## 5) Chain rule (사슬법칙)

- 6.5.21

## 사슬 법칙

조건부확률과 결합확률의 관계를 확장하면 복수의 사건  $X_1, X_2, \dots, X_N$ 에 대한 조건부 확률을 다음처럼 쓸 수 있다. 이를 사슬 법칙(chain rule)이라고 한다.

$$\begin{aligned}
 P(X_1, X_2) &= P(X_1)P(X_2|X_1) \\
 P(X_1, X_2, X_3) &= P(X_3|X_1, X_2)P(X_1, X_2) \\
 &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \\
 P(X_1, X_2, X_3, X_4) &= P(X_4|X_1, X_2, X_3)P(X_1, X_2, X_3) \\
 &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3) \\
 &\vdots \\
 P(X_1, \dots, X_N) &= P(X_1) \prod_{i=2}^N P(X_i|X_1, \dots, X_{i-1})
 \end{aligned} \tag{6.5.21}$$

## 6) 확률변수

- 확률적인 숫자값을 출력하는 변수
- 사실은 함수다. 특정한 높이값이 출력되는 함수.

$P(X(w) = a) = 0.1$  이라면,  $p(a) = 0.1$

\*(높이로)  $a$ 를 출력하는 함수  $X$ 가 있는 세계( $w$ )의 면적은 0.10이다.

## 7) pgmpy 패키지

# 1) 베이즈 정리

06.06

## - 6.6.1

### 베이즈 정리

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

베이즈 정리

데이터가 주어졌을 때, 그것에서  $\theta$ 를 추론해내는!

$$P(\text{Data}) \Rightarrow P(\theta | \text{Data})$$

조건부 확률을 구하는 다음 공식을 베이즈 정리(Bayesian rule)라고 한다.

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

$P(A)$ 가  $P(A|B)$ 에 가는

비례는 데이터에 주어질 때,

(증명)

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

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

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

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

- $P(A|B)$ : 사후확률(posterior). 사건 B가 발생한 후 경신된 사건 A의 확률
- $P(A)$ : 사전확률(prior). 사건 B가 발생하기 전에 가지고 있던 사건 A의 확률
- $P(B|A)$ : 가능도(likelihood). 사건 A가 발생한 경우 사건 B의 확률
- $P(B)$ : 정규화 상수(normalizing constant) 또는 증거(evidence). 확률의 크기 조정

사후확률  $\propto$  가능도  $\times$  사전확률

## 2) 베이즈 정리의 확장

- "사건 A가 서로 배타적이고 완전하다" = A가 서로소 + 완전(합집합이 표본공간)

전체 확률의 법칙을 이용하여 다음과 같이 베이즈 정리를 확장할 수 있다.

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B)} \\ &= \frac{P(B|A_1)P(A_1)}{\sum_i P(A_i, B)} \\ &= \frac{P(B|A_1)P(A_1)}{\sum_i P(B|A_i)P(A_i)} \end{aligned} \quad (6.6.9)$$

## - Multi-class Classification (6.6.10)

이 식은 멀티-클래스 분류(multi-class classification) 문제에서 베이즈 정리가 어떻게 사용되는지를 보여주는 수식이다. 멀티-클래스 분류 문제는 여러 베이스적이고 완전한 사건 중에서 가장 확률이 높은 하나의 사건을 고르는 문제다. 예를 들어 B라는 힌트를 주고 4번부터 4번까지의 보기 중 하나를 골라야 하는 4지선다면 문제는 4개의  $A_1, A_2, A_3, A_4$  중 B에 대한 조건부 확률이 가장 높은 사건을 고르는 것과 같다. 이 문제를 풀기 위해서는 위의 베이즈 정리 확장은 사용하여 4개의 조건부 확률값을 비교하면 된다.

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4)} \quad (6.6.10)$$

$$P(A_2|B) = \frac{P(B|A_2)P(A_2)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4)} \quad (6.6.11)$$

$$P(A_3|B) = \frac{P(B|A_3)P(A_3)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4)} \quad (6.6.12)$$

$$P(A_4|B) = \frac{P(B|A_4)P(A_4)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4)} \quad (6.6.13)$$

그런데 분모에 있는  $\sum_i P(B|A_i)P(A_i)$  식은 /값이 바뀌어 항상 같은 값이므로  $A_1, A_2, A_3, A_4$  중 B에 대한 조건부 확률이 가장 높은 사건을 고르는 것이 목적이라면 분자와 값만 비교하면 된다. 다음 식에서  $\propto$  기호는 비례한다는 뜻이다.

$$P(A_1|B) \propto P(B|A_1)P(A_1) \quad (6.6.14)$$

$$P(A_2|B) \propto P(B|A_2)P(A_2) \quad (6.6.15)$$

$$P(A_3|B) \propto P(B|A_3)P(A_3) \quad (6.6.16)$$

$$P(A_4|B) \propto P(B|A_4)P(A_4) \quad (6.6.17)$$

## 3) pgmpy를 사용한 베이즈 정리 적용

## 4) 베이즈 정리의 확장 2

### - 6.6.23 증명

#### 베이즈 정리의 확장 2

조건전에 공통적으로 들어가 있는 것들은  
다 제외해도 된다. ex 'B'  
공통 조건 제외하고도 베이즈를 이면, 맞는 것.

베이즈 정리는 사건 A의 확률이 사건 B에 의해 갱신(update)된 확률을 계산한다. 그런데 만약 이 상태에서 또 추가적인 사건 C가 발생했다면 베이즈 정리는 다음과 같이 쓸 수 있다.

$$P(A|B, C) = \frac{P(C|A, B)P(A|B)}{P(C|B)} \quad (6.6.23)$$

- 중요한 것은, 6.6.23 식을 보고 맞는 건지 아닌 건지 바로 알 수 있어야 한다.

\*우변의 모든 항목에 공통으로 들어간 조건을 제외해도 베이즈룰이 맞다면,

맞는 식이다.

- 연습문제 6.6.1-3 풀기

## 5) 몬티홀 문제

-  $P(C_1 | X_1) = 1/3$

-  $P(C_0 | X_1, H_2)$  계산 해보기! (참가자가 선택을 바꾸면 이길 확률)

기본 사실 1) 자동차의 위치와 참가자의 선택은 서로 독립이다.

$$P(C, X) = P(C)*P(X)$$

기본 사실 2) 진행자가 어떤 문을 여는 지는 조건부 확률이다.

(참가자의 선택에 따라 달라진다.)

## 확률적 데이터와 확률변수

### 1) 확률적 데이터

- 결정론적 데이터 : 생년월일처럼 언제든지 항상 같은 값이 나오는 데이터
- 확률적 데이터 : 혈압처럼 예측할 수 없는 값이 나오는 데이터
  - 범주형 / 실수형 데이터

### 2) 분포

- 분포 : 확률적 데이터의 분포
  - count plot : 범주형 데이터의 시각화
  - histogram : 실수형 데이터의 시각화
  - 기술통계(descriptive statistics) : 표본평균, 표본중앙값, 표본최빈값, 표본분산, 표본표준편차, 표본왜도, 표본첨도

일반적으로 부르는 평균(mean, average)의 정확한 명칭은 표본평균(sample mean, sample average)이다. 표본평균은 데이터 분포의 대략적인 위치를 나타낸다. 표본평균의 기호로는 알파벳  $m$  또는 데이터를 나타내는 변수 기호 위에  $\bar{x}$ 를 붙인  $\bar{x}$  기호를 사용한다.

$$m = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (7.1.1)$$

### 3) 표본평균 (7.1.1)

- 데이터 분포의 대략적인 위치를 나타냄

### 4) 표본중앙값 (sample median)

- 전체 자료를 크기 별로 정렬 시, 가장 중앙에 위치하는 값
- 중앙값 사용 시, 아웃라이어 영향을 줄일 수 있음  
(소득 데이터 시, 평균은 아웃라이어로 영향을 받지만, 중앙값은 아웃라이어가 단지 1개의 표본일 뿐.)

### 5) 표본최빈값

- 가장 빈번하게 나오는 값 (범주형 데이터에만 가능. 실수형 데이터는 연속형이라 엄밀한 의미에서 최빈값이 없음)

07.01

확률적 데이터와  
확률변수

## 6) 단봉분포와 다봉분포 (uni-modal, multi-modal)

- 분포의 모양에서 봉우리가 하나면 uni-modal distribution

## 7) 대칭분포

- 표본평균을 기준으로 대칭이라면, 표본평균 = 표본중앙값
- 대칭분포면서 단봉분포라면, 표본평균 = 표본최빈값
- 대칭분포를 비대칭으로 만드는 데이터가 더해지면, mean > median > mode 순으로 영향을 받는다.

so, 대칭 + uni-modal ==> mean = median = mode

## 8) 분산과 표준편차

- 7.1.3

- 비편향 분산 : 7.1.4

표본분산은 다음처럼 구한다. 식에서  $\bar{x}$ 은 표본평균이다.

$$s^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \quad (7.1.3)$$

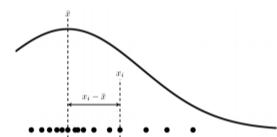


그림 7.1.1 : 표본분산과 표본표준편차

위 식에서 구한 표본분산은 정확하게 말하면 편향오차를 가진 편향 표본분산(biased sample variance)이다. 이와 대조되는 비편향 표본분산(unbiased sample variance)은 다음과 같이 구한다.

$$s_{\text{unbiased}}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \quad (7.1.4)$$

## 9) 파이썬을 사용한 표본분산 및 표본표준편차 계산

- ddof(자유도, degree of freedom)를 활용해서 np.var(x, ddof=1), np.std(x, ddof=1)로 계산

## 10) 표본비대칭도 (왜도, skewness)

- 세제곱

- 대칭분포 (왜도 = 0)

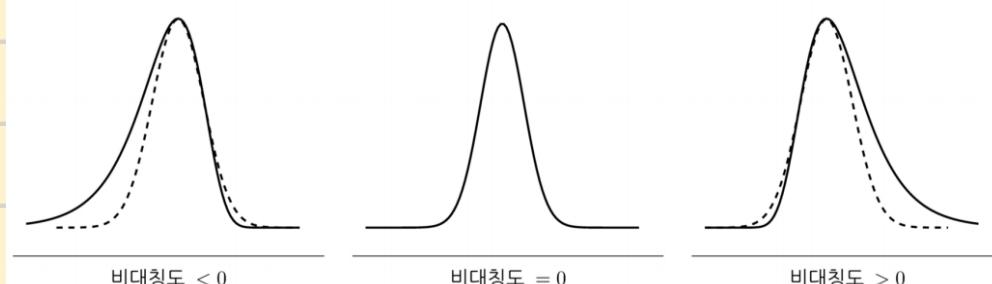
- 평균보다 왼쪽 몰림 (왜도 (-))

- 7.1.5, 07.01 8page

- sp.stats.skew(x)

평균과의 거리의 세제곱을 이용하여 구한 특징값을 표본비대칭도(sample skewness)라고 한다. 표본비대칭도가 0이면 분포가 대칭이다. 표본비대칭도가 음수면 표본평균값을 기준으로 왼쪽에 있는 값을 가진 표분이 나올 가능성이 더 많다는 뜻이다.

$$\text{표본비대칭도} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^3}{\sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}^3} \quad (7.1.5)$$



07.01

확률적 데이터와  
확률변수

## 12) 표본모멘트

- 표본모멘트 / 표본중앙모멘트
- 7.1.7 / 7.1.8
- 평균, 분산, 왜도, 첨도 ==> 1,2,3,4차 모멘트에서 유도된 값
- sp.stats.moment(x, k) (k차 모멘트)

## 13) 확률변수

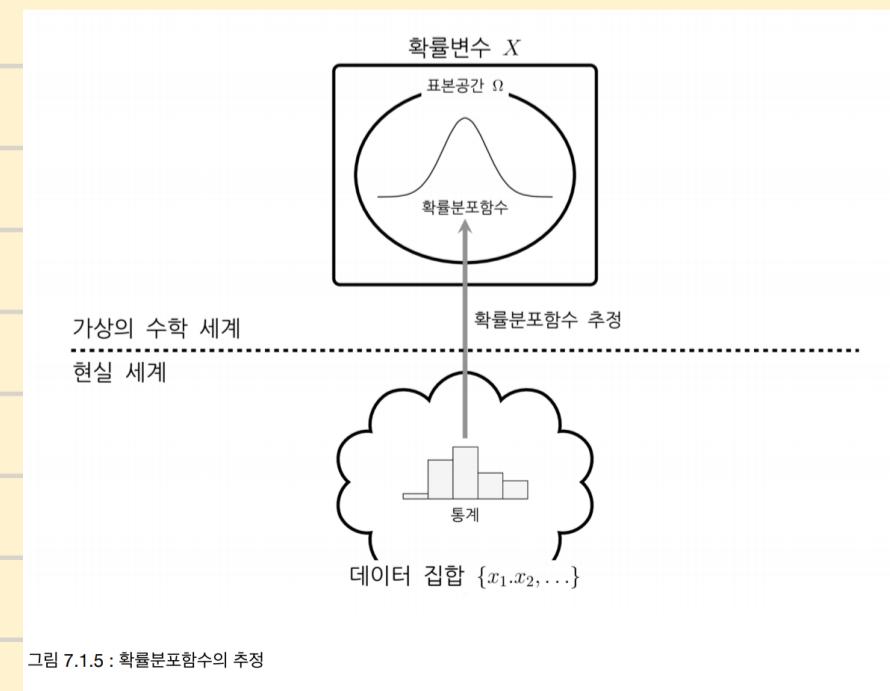
- 확률변수는 데이터 생성박스 (누르면 샘플링!)
- 이산확률변수, 연속확률변수
- 그림 7.1.4

## 14) 확률변수를 사용한 데이터분석 <— 확률분포함수(pdf)를 추출하는 과정

- 확률분포가 어떻게 생겼는지 찾아내는 과정!
- 표본평균, 표본표준편차, 왜도, 첨도를 활용! (기술통계량 활용)

### [데이터분석의 과정]

- 데이터수집 -> 수집한 데이터의 숨겨진 확률분포함수의 모양 결정  
(결정 = 추정, reverse engineering)
- 확률분포함수로부터 우린, 데이터 생성기(확률변수)를 갖게된 것.
- 이를 활용해 다음에 생성될 데이터나 데이터 특성을 예측



## 기댓값과

## 확률변수의 변환

## 1) 확률변수의 기댓값

- 기댓값 : 확률변수의 가중평균

“확률값이 가중치이다. 따라서, 봉우리(확률값이 가장 큰) 중심으로 가중평균 된다.

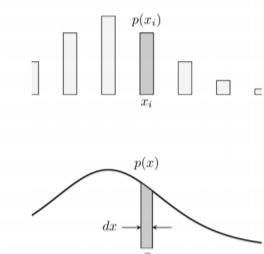
확률분포의 중심이 어디인지를 기대값이 알려준다.”

## - 7.2.5

연속확률변수의 기댓값은 확률밀도함수  $p(x)$ 를 가중치로 하여 모든 가능한 표본  $x$ 를 적분한 값이다.

$$\mu_X = E[X] = \int_{-\infty}^{\infty} xp(x)dx$$
(7.2.5)

$$E[X] = \sum_{x_i \in \Omega} x_i \overbrace{p(x_i)}$$



$$E[X] = \int_{-\infty}^{\infty} x \overbrace{p(x)} dx$$

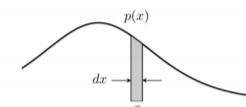


그림 7.2.1 : 기댓값 계산

## 2) 확률변수의 변환

- 우리가 얻은 데이터 값을, 각각 함수에 집어넣어 변화시켜 새로운 데이터 집합 get (7.2.7) (그림 7.2.2)

우리가 얻은 데이터의 값을 어떤 함수  $f$ 에 넣어서 변화시킨다고 가정하자. 그러면 새로운 데이터 집합이 생긴다.

$$\{x_1, x_2, \dots, x_N\} \rightarrow \{f(x_1), f(x_2), \dots, f(x_N)\}$$
(7.2.7)

이 새로운 데이터를  $\{y_i\}$ 라고 부르자.  $\{y_i\}$ 는 기존의 데이터와 다른 새로운 데이터이므로 다른 확률변수라고 볼 수 있다. 예를 들어 데이터  $\{x_i\}$ 를 만드는 확률변수가  $X$ 라면 데이터  $\{y_i\}$ 를 만드는 데이터는  $Y$ 라는 새로운 확률변수가 된다.

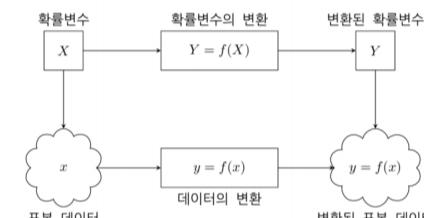


그림 7.2.2 : 확률변수의 변환

## 3) 기댓값의 성질 (07.02 6page)

기댓값은 다음과 같은 성질을 가진다는 것을 수학적으로 증명할 수 있다. 변환된 확률변수의 기댓값을 계산할 때는 기댓값의 성질을 이용한다.

- 확률변수가 아닌 상수  $c$ 에 대해

$$E[c] = c$$
(7.2.12)

- 선형성

$$E[cX] = cE[X]$$
(7.2.13)

$$E[X + Y] = E[X] + E[Y]$$
(7.2.14)

$$E[c_1X + c_2Y] = c_1E[X] + c_2E[Y]$$
(7.2.15)

## 4) 통계량 (statistics) (07.02 6page)

확률변수  $X$ 로부터 데이터 집합  $\{x_1, x_2, \dots, x_N\}$ 을 얻었다고 하자. 이 데이터 집합의 모든 값을 정해진 어떤 공식에 넣어서 하나의 숫자를 구한 것을 통계량(statistics)이라고 한다. 예를 들어 표본의 합, 표본평균, 표본중앙값, 표본분산 등은 모두 통계량이다. 통계량도 확률변수의 변환에 포함된다.

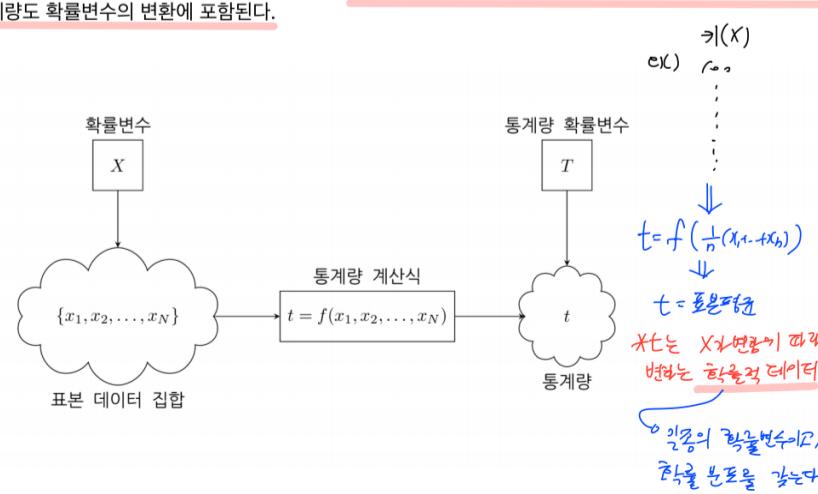
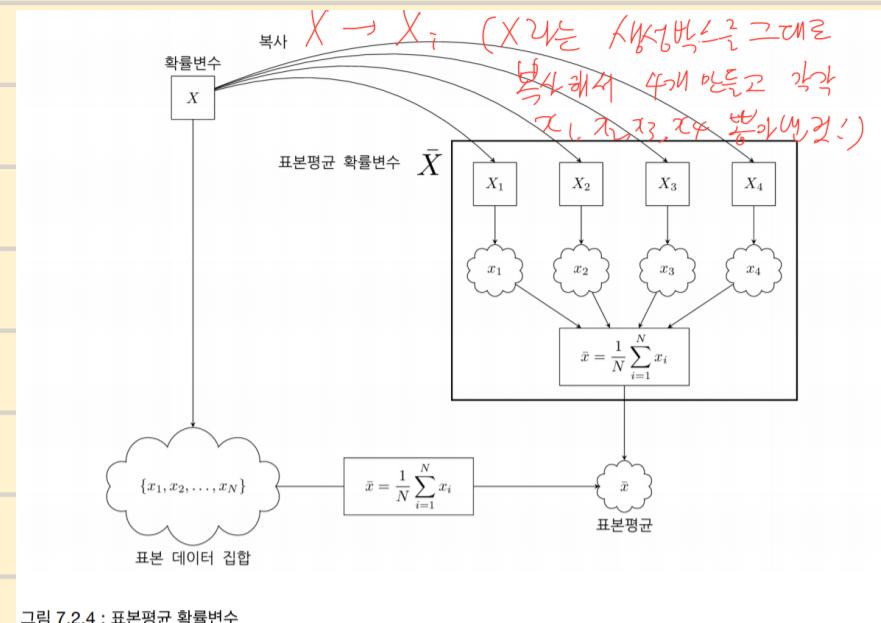


그림 7.2.3 : 통계량

## 기댓값과

## 확률변수의 변환

## 5) 표본평균 확률변수 (통계량 중 하나) (그림 7.2.4)



## 6) 기댓값과 표본평균의 관계

- 7.2.17식

$$E[\bar{X}] = E[X]$$

- 표본평균은 확률변수의 기댓값 근처에서 분포한다.

- 표본평균의 기댓값 = 확률변수의 기댓값

- 모집단의 기댓값(확률변수의 기댓값) 근처에서 표본평균이 주로 나온다.

(예 : 공정한 주사위 코멘트)

예를 들어 공정한 주사위의 기댓값은 3.5이다. 이 주사위를 던져 나온 값의 평균 즉 표본평균은 3.62346 또는 3.40987처럼 항상 3.5 근처의 값이 나오게 된다.

- 왜 중요하냐면, 확률변수의 분포를 알기위해  $E[X]$ 를 알고 싶다.

이 때, 샘플링으로 얻은 표본평균의 기댓값이 곧 확률변수의 기댓값임을 알고, 이를 통해 분포를 유추하는 데 도움이 된다.

## 7) 중앙값 (median)

-  $F(\text{중앙값}) = 0.5$  (그림 7.2.5)

확률변수의 중앙값(median)은 중앙값보다 큰 값이 나올 확률과 작은 값이 나올 확률이 0.5로 같은 값을 뜻한다. 따라서 다음과 같이 누적확률분포  $F(x)$ 에서 중앙값을 계산할 수 있다.

$$0.5 = F(\text{중앙값}) \quad (7.2.19)$$

$$\text{중앙값} = F^{-1}(0.5) \quad (7.2.20)$$

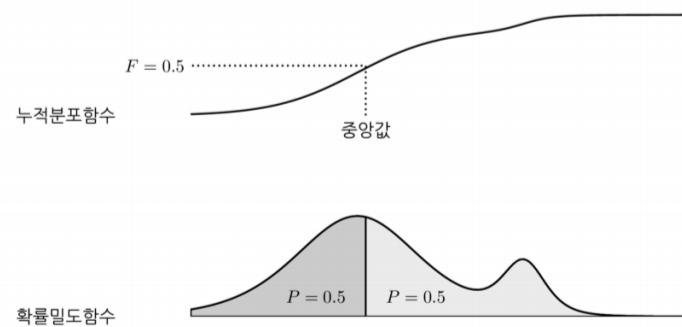


그림 7.2.5 : 중앙값

## 8) 최빈값 (mode)

- 최빈값 =  $\operatorname{argmax} p(x)$  = 확률값이 가장 클 때의 확률변수 = 가장 많이 나오는 수

## 1) 확률분포(X)의 분산(7.3.1)

확률밀도함수  $p(x)$ 의 수식을 알고 있다면 이론적인 분산을 구할 수 있다. 분산을 구하는 연산은 영어 Variance의 앞글자를 따서  $\text{Var}[\cdot]$ 로 표기하고 이 연산으로 계산된 분산값은  $\sigma^2$ 으로 표기한다.

$$\sigma^2 = \text{Var}[X] = E[(X - \mu)^2] \quad (7.3.1)$$

이산확률변수의 분산은 평균으로부터 표본 데이터까지 거리의 제곱을 확률밀도함수  $p(x)$ 로 가중하여 더한 값이다.

$$\sigma^2 = \sum_{x_i \in \Omega} (x_i - \mu)^2 p(x_i) \quad (7.3.2)$$

연속확률변수의 분산은 평균으로부터 표본 데이터까지 거리의 제곱을 확률밀도함수  $p(x)$ 로 가중하여 적분한 값이다.

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx \quad (7.3.3)$$

## 2) 분산의 성질(7.3.4)

분산은 다음과 같은 성질을 만족한다.

- 분산은 항상 0 또는 양수이다.

$$\text{Var}[X] \geq 0 \quad (7.3.4)$$

- 확률변수가 아닌 상수 값  $c$ 에 대해 다음 식이 성립한다.

$$\text{Var}[c] = 0 \quad (7.3.5)$$

$$\text{Var}[cX] = c^2 \text{Var}[X] \quad (7.3.6)$$

또한 기댓값의 성질을 이용하여 다음 성질을 증명할 수 있다.

$$\text{Var}[X] = E[X^2] - (E[X])^2 = E[X^2] - \mu^2 \quad (7.3.7)$$

3) 두 확률변수 합의 분산,  $\text{Var}(X+Y)$  (7.3.12)

$$\begin{aligned} \text{Var}[X+Y] &= E[(X+Y - (\mu_X + \mu_Y))^2] \\ &= E[((X - \mu_X) + (Y - \mu_Y))^2] \\ &= E[(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)] \\ &= E[(X - \mu_X)^2] + E[(Y - \mu_Y)^2] + 2E[(X - \mu_X)(Y - \mu_Y)] \\ &= \text{Var}[X] + \text{Var}[Y] + 2E[(X - \mu_X)(Y - \mu_Y)] \end{aligned} \quad (7.3.12)$$

## 4) 확률변수의 독립

- 독립이면, 공분산 = 0

- 공분산 = 0 이면,  $\text{Var}(X+Y) = \text{Var}(X)+\text{Var}(Y)$

5) 표본평균의 분산 (  $\text{Var}(\bar{X})$  )

- 07.03, 3page

표본평균  $\bar{X}$ 의 분산  $\text{Var}[\bar{X}]$ 은 원래 확률변수  $X$ 의 분산  $\text{Var}[X]$ 과 다음 관계를 가진다.

$$\text{Var}[\bar{X}] = \frac{1}{N} \text{Var}[X] \quad (7.3.16)$$

따라서 표본평균을 계산한 표본 개수가 커지면 표본평균의 값의 변동은 작아진다. 표본의 수가 무한대가 되면 표본평균의 값은 항상 일정한 값이 나온다. 즉 확률적인 값이 아니라 결정론적인 값이 된다.

$$\text{Var}(\bar{X}) = \frac{1}{N} \text{Var}(X)$$

-  $\text{Var}(\bar{X}) = 1/n * \text{Var}(X)$  증명 꼭 해보기

\*E, 기대값은 선형성으로 summation에선 안으로 들어가도 무방!

- 07.03 4page

$N \rightarrow \infty, \bar{X} \approx E(X)$

$$\begin{aligned} \text{Sampling으로 이런 앙수있지?} \\ \text{③ } E(\bar{X}) &= E(X) \\ \text{Var}(\bar{X}) &= \frac{1}{N} \text{Var}(X) \\ \text{④ } \text{증명이 어렵거나!} \\ \{x_1, x_2, \dots, x_n\} \\ \text{① } \text{증명이 어렵거나!} \\ \text{N} \rightarrow \infty \text{ (표본 수 많아 봄는 경우)} \\ \bar{X} = E(X) \end{aligned}$$

07.03

분산과 표준편차

6) 표본분산의 기대값 ( $E[S^2]$ )

- $\text{var}(X_{\bar{}}) \neq S^2$  (표본평균의 분산  $\neq$  표본의 분산)

## - 7.3.20

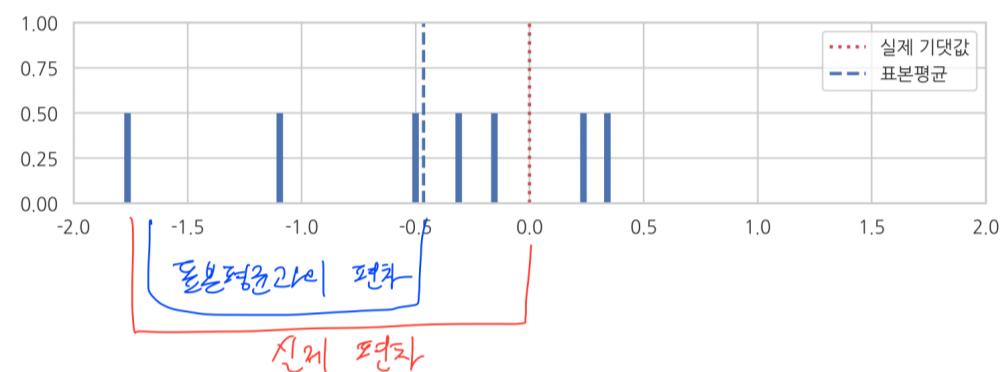
그런데 표본분산  $S^2$ 의 기대값을 구하면 이론적인 분산  $\sigma^2$ 과 같아지는 것이 아니라 이론적인 분산값의  $\frac{N-1}{N}$  배가 된다. 즉 표본분산값이 이론적인 분산값보다 더 작아진다.

$$E[S^2] = \frac{N-1}{N} \sigma^2 \quad (7.3.20)$$

- 표본분산이 실제 분산보다 작아지는 이유

: 표본평균이 데이터가 몰린 쪽으로 편향. 따라서, 표본들의 분산 값도 작아짐

(07.03 9page)



## 7) 비대칭도와 첨도

- 왜도 : 3차 모멘트 값에서 계산 (7.3.32)
- 첨도 : 4차 모멘트 값에서 계산 (7.3.32)

비대칭도(skewness)는 3차 모멘트 값에서 계산하고 확률밀도함수의 비대칭 정도를 가리킨다. 비대칭도가 0이면 확률분포가 대칭이다.

$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} \quad (7.3.32)$$

첨도(kurtosis)는 4차 모멘트 값에서 계산하며 확률이 정규분포와 대비하여 중심에 모여있는지 바깥으로 퍼져있는지를 나타낸다.

$$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} \quad (7.3.33)$$

## 8) 모멘트

- 기댓값, 분산 모두 확률분포의 모멘트 중 하나
- 두 확률분포  $X, Y$ 의 1차부터 계속해서 모멘트값이 서로 같다면, 두 확률분포는 같은 확률분포이다.

## (7.3.35)

앞서 구한 기댓값이나 분산은 확률분포의 모멘트(moment)의 하나다.

$$\mu_n = E[(X-\mu)^n] = \int (x-\mu)^n p(x) dx \quad (7.3.34)$$

모멘트는 확률분포에서 계산한 특징값이다. 만약 두 확률분포  $X, Y$ 가 있고 1차부터 무한대 차수에 이르기까지 두 확률분포의 모든 모멘트값이 서로 같다면 두 확률분포는 같은 확률분포다.

$$\begin{aligned} E[X] &= E[Y] \\ E[(X-\mu_X)^2] &= E[(Y-\mu_Y)^2] \\ E[(X-\mu_X)^3] &= E[(Y-\mu_Y)^3] \\ E[(X-\mu_X)^4] &= E[(Y-\mu_Y)^4] \\ E[(X-\mu_X)^5] &= E[(Y-\mu_Y)^5] \\ &\vdots \end{aligned} \quad (7.3.35)$$

이면

$$X \stackrel{d}{=} Y \quad (7.3.36)$$

이다.  $\stackrel{d}{=}$  는 두 확률변수가 같은 분포(distribution)를 가진다는 것을 표시하는 기호다.

## 1) 결합 질량함수 (joint pmf) (7.4.3)

$$p_{XY}(x, y)$$

즉,  $p_{XY}(2, 3)$ 은  $\{x = 2, y = 3\}$ 이라는 특정한 숫자 쌍으로만 이루어진 사건의 확률이다.

## 2) 주변 질량함수 (marginal pmf) (7.4.5)

$$p_X(x) = \sum_{y_i} p_{XY}(x, y_i)$$

$$p_Y(y) = \sum_{x_i} p_{XY}(x_i, y)$$

## 3) 조건부 확률질량함수 (conditional pmf) (7.4.8)

- 결합 pdf의 단면을 잘라

조건부 확률질량함수(conditional probability mass function)는 다변수 확률변수 중 하나의 값이 특정 값으로 고정되어

상수가 되어 버린 경우, 나머지 변수에 대한 확률질량함수를 말한다. 조건부 확률질량함수는 다음과 같이 정의된다.

$$p_{X|Y}(x | y) = \frac{p_{XY}(x, y)}{p_Y(y)} \quad (7.4.8)$$

$$p_{Y|X}(y | x) = \frac{p_{XY}(x, y)}{p_X(x)} \quad (7.4.9)$$

조건부 확률질량함수의 모양은 결합질량함수  $p_{XY}(x, y)$ 에서  $y$  값이 고정된 함수, 즉, 결합질량함수의 단면과 같아진다. 다만 조건부 확률질량함수의 합은 1이 된다.  $\Leftarrow$  주변 pmf는 스케일링!

## 4) 다변수 연속확률변수

: 연속확률분포에선, 확률 정의를 단순사건으로 할 수 없음.

: 따라서, cdf를 먼저 정의한 후, 미분하여 pdf를 구한다.

## 1) 결합 cdf (7.4.10)

두 연속 확률변수  $X, Y$ 에 대한 결합누적확률분포함수  $p_{XY}(x, y)$ 는 다음과 같이 정의한다.

$$F_{XY}(x, y) = P(\{X < x\} \cap \{Y < y\}) = P(\{X < x, Y < y\}) \quad (7.4.10)$$

## 2) 결합 pdf (7.4.14)

단변수 확률변수의 경우처럼 결합누적확률분포함수를 미분하여 결합확률밀도함수(joint probability density function)를 정의할 수 있다. 독립 변수가 2개이므로 각각에 대해 모두 편미분(partial differentiation)해야 한다.

$$p_{XY} = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y} \quad (7.4.13)$$

결합확률밀도함수를 특정 구간에 대해 적분하면 해당 구간에 대한 확률이 된다.

$$\int_{x_1}^{x_2} \int_{y_1}^{y_2} p_{XY}(x, y) dx dy = P(\{x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2\}) \quad (7.4.14)$$

## 3) 주변 pdf (7.4.16-17)

주변확률밀도함수(marginal probability density function)는 결합확률밀도함수를 특정한 하나의 변수에 대해 가중평균한 값을 말한다. 따라서 결합확률밀도함수를 하나의 확률변수에 대해서만 적분하여 구한다.

가중평균(적분)으로 인해 차원이 한 개 줄어들기 때문에 2차원 확률변수의 주변 확률 밀도 함수는 1차원 함수가 된다.

$$p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x, y) dy \quad \text{는 } \text{전략적 공식} \quad (7.4.16)$$

$$p_Y(y) = \int_{-\infty}^{\infty} p_{XY}(x, y) dx \quad (7.4.17)$$

주변 pmf  
는 전략적 공식

## 4) 결합 pdf와 주변 pdf (07.04 14page)

↑  
구조부

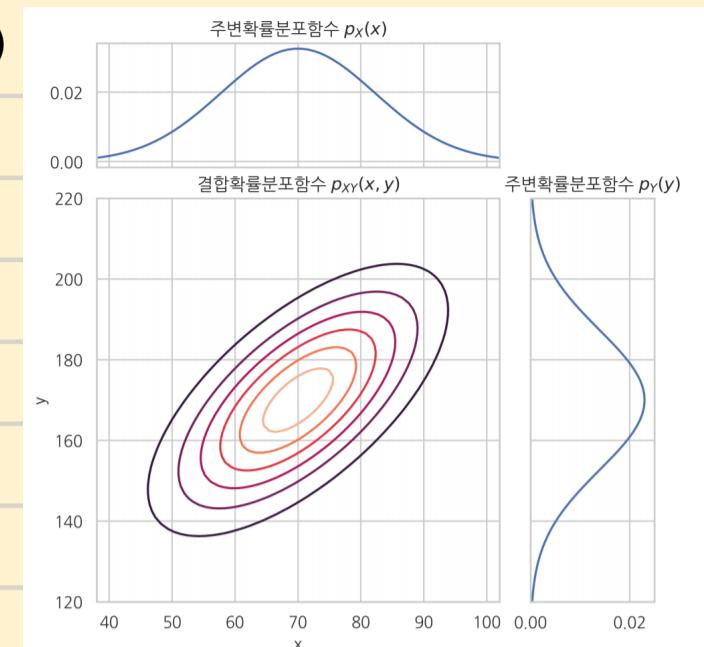
자른  
→ 단면을 스케일링

## 5) 조건부 pdf

- 7.4.18

- 결합 pdf의 단면을 각각 전체 sum이 1이 되도록 normalize!
- 결합 pdf의 단면의 부피(sum) = 주변 pdf, 결합 pdf의 단면(한 변수를 상수로 고정한 단면)을 면적이 1이 되도록 스케일링(주변 pdf로)

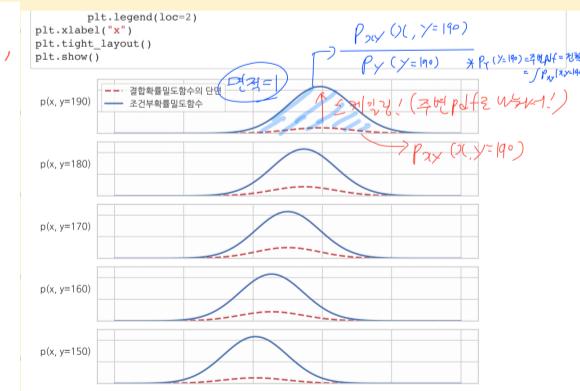
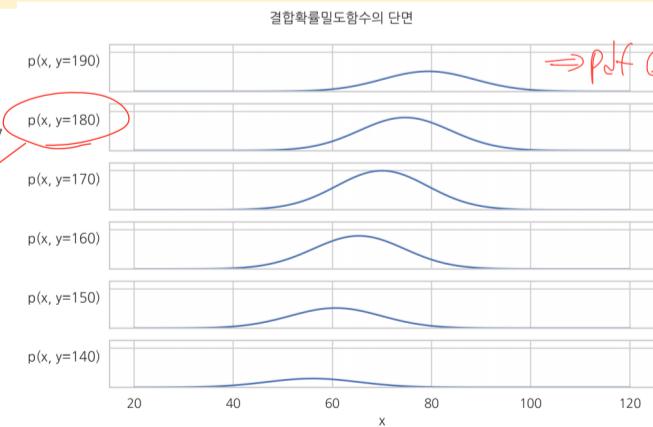
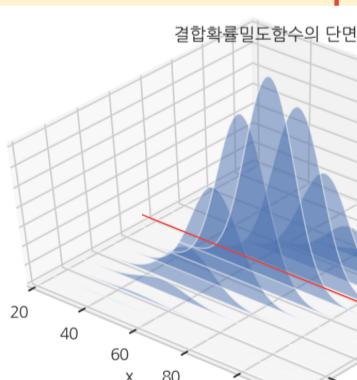
(07.04 5page, 14page, 15page, 17page)



조건부 확률밀도함수(conditional probability density function)는 다변수 확률변수 중 하나의 값이 특정 값이라는 사실이 알려진 경우, 이러한 조건(가정)에 의해 변화한 나머지 확률변수에 대한 확률밀도함수를 말한다.

$$p_{X|Y}(x | y) = \frac{p_{XY}(x, y)}{p_Y(y)} \quad (7.4.18)$$

$$p_{Y|X}(y | x) = \frac{p_{XY}(x, y)}{p_X(x)} \quad (7.4.19)$$



## 07.04 다변수

## 확률변수

## 6) 독립과 상관

(07.04 18page), 7.4.22, (4.1.40식 + 04.01 25page), 07.04 19page

- 상관 : 한 확률변수의 표본 값이 달라질 때, 다른 확률변수의 조건부분포가 달라지면 서로 상관관계에 있다고 한다.

- 독립 : 상관이 아닌 관계

두 확률변수가 있을 때, 한 확률변수의 표본 값이 달라지면 다른 확률변수의 조건부 분포가 달라질 때 서로 상관 관계가 있다 고 한다. 반대로 두 확률변수가 상관 관계가 아니면 서로 독립(independent)이라고 한다. 확률변수의 독립을 수학적으로 정 의하면 다음과 같다.

두 확률변수  $X, Y$ 의 결합확률밀도함수(joint pdf)가 주변확률밀도함수(marginal pdf)의 곱과 같으면 서로 독립(independent)이다.

$$p_{XY}(x, y) = p_X(x)p_Y(y) \quad (7.4.20)$$

이 정의는 확률변수가 두 개 보다 많을 때도 적용된다. 예를 들어 세 개의 확률변수  $X, Y, Z$ 의 결합확률밀도함수가 각각의 주변확률밀도함수(marginal pdf)의 곱과 같으면 세 확률변수는 서로 독립이다.

$$p_{XYZ}(x, y, z) = p_X(x)p_Y(y)p_Z(z) \quad (7.4.21)$$

이 때  $X, Y, Z$  중 어느 두 확률변수를 골라도 서로 독립이 된다.

$$\begin{aligned} p_{XY}(x, y) &= \sum_{z \in \Omega_z} p_{XYZ}(x, y, z) \\ &= \sum_{z \in \Omega_z} p_X(x)p_Y(y)p_Z(z) \\ &= p_X(x)p_Y(y) \sum_{z \in \Omega_z} p_Z(z) \\ &= p_X(x)p_Y(y) \end{aligned} \quad (7.4.22)$$

사실, 독립 관계는 다변수함수를 단변수함수의 곱으로 표현하는 것.

다변수함수 중에는 단변수함수의 곱으로 표현 가능한 다변수함수도 있다.

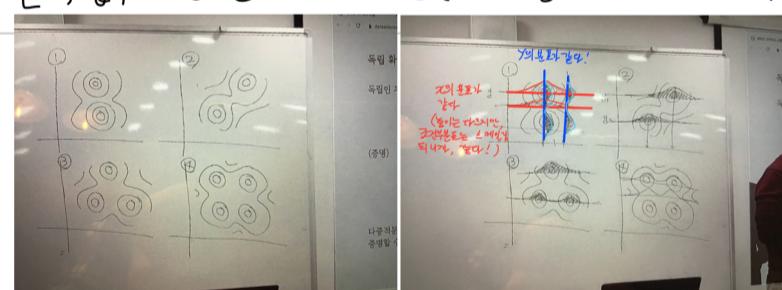
$$f(x, y) = f_1(x)f_2(y) \quad (4.1.37)$$

그런데 분리 가능 다변수함수는 단면의 모양이 모두 같다. 예를 들어  $x = x_0$ 로 고정하면 단면 함수는

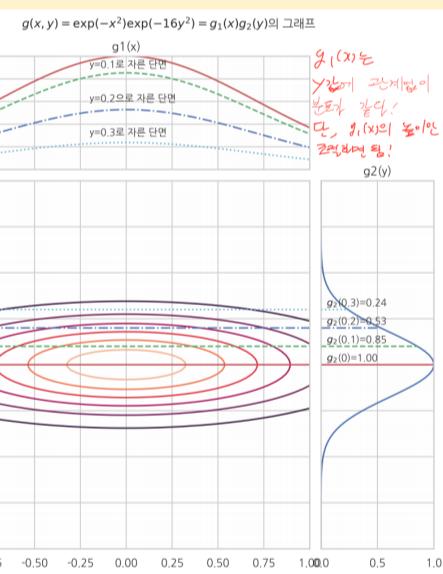
$$g(x_0, y) = g(x_0)g(y) = k_0g(y) \quad (4.1.40)$$

이므로  $x_0$ 의 값과 상관없이  $g(y)$ 의 높이만 조절한 모양이 된다.

[독립, 상관 해석] Q. 다음 중 독립인 경우는?



1. 4번



## 7) 독립 확률변수의 기대값 증명

-  $E[XY] = E[X]E[Y]$

- 증명  $\Rightarrow P(X,Y)=P(X)P(Y)$ , 푸비니 정리 활용

$\hookrightarrow$  2차적분 = 1차적분 2번한 것!

\*독립 확률변수의 특징

1) 기댓값

2) 공분산

3) 분산

다면수 확률변수 간 상관관계를 숫자로 나타낸 것 = 공분산, 상관계수

### 1) 표본공분산 (7.5.1 + 그림 7.5.1)

공분산:

상관관계의 방향

분산 정도의 크기

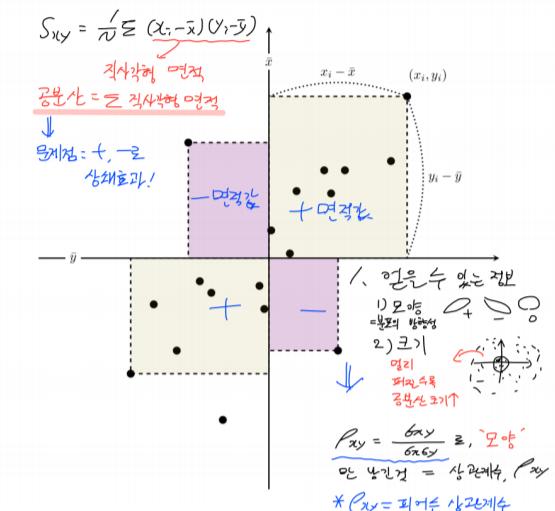
Scatter의 모양, 크기는 알 수 있다.

#### 표본공분산 떨어진 정도 + 부호

표본공분산(sample covariance)은 다음처럼 정의한다. 여기에서  $x_i$ 와  $y_i$ 는 각각  $i$  번째의  $x$  자료와  $y$  자료의 값을 가리키고,  $\bar{x}$ 와  $\bar{y}$ 는  $x$  자료와  $y$  자료의 표본평균을 가리킨다.

$$s_{xy} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \quad (7.5.1)$$

표본분산과 마찬가지로 표본공분산도 자료가 평균값으로부터 얼마나 떨어져 있는지를 나타낸 것이다. 공분산은 평균값 위치와 표본 위치를 연결하는 사각형의 면적을 사용한다. 다만 공분산의 경우에는 자료의 위치에 따라 이 값의 부호가 달라진다. 데이터가 1사분면이나 3사분면에 있는 경우에는 양수가 되고 데이터가 2사분면이나 4사분면에 있는 경우에는 음수가 된다. 따라서 공분산의 부호는  $X, Y$  데이터가 같은 부호를 가지는지 다른 부호를 가지는지에 대한 지표라고 할 수 있다.



### 2) 표본상관계수 (피어슨 상관계수) 7.5.2

표본공분산은 평균을 중심으로 각 자료들이 어떻게 분포되어 있는지 크기와 방향성을 같이 보여준다. 그런데 분포의 크기는 공분산이 아닌 분산만으로도 알 수 있기 때문에 대부분의 경우 자료 분포의 방향성만 분리하여 보는 것이 유용하다. 이 때 필요한 것이 표본상관계수(sample correlation coefficient)다.

표본상관계수는 다음과 같이 공분산을 각각의 표본표준편차값으로 나누어 정규화(normalize)하여 정의한다.

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} \quad (7.5.2)$$

이와 다르게 정의한 상관계수도 있기 때문에 다른 종류의 상관계수와 비교하여 말하는 경우에는 피어슨(Pearson) 상관계수라고 하기도 한다.

### 3) 데이터(표본)이 아닌, 확률변수의 공분산, 상관계수 (07.05 3page, 4page)

- 피어슨 상관계수 (= 선형 상관계수, 선형적으로 얼마나 상관관계가 있는지 측정)
- ~~로우 = 0 이 곧 독립을 의미하지 않음. (독립이면 로우=0)~~
- ~~선형상관만 가능한(비선형상관가능x)~~
- 상관계수는 스캐터플롯의 기울기와 상관 없음 (단지 상관계수는 선형의 정도를 의미, 기울기와는 무관)

상관계수: 선형의 정도만

마찬가지로 두 확률변수  $X$ 와  $Y$ 의 상관 계수도 다음과 같이 정의한다.

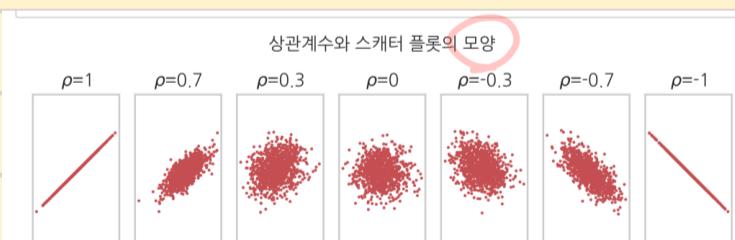
$$\rho[X, Y] = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X] \cdot \text{Var}[Y]}} \quad (7.5.4)$$

확률변수의 상관계수는 다음과 같은 성질을 가진다.

$$-1 \leq \rho \leq 1 \quad (7.5.5)$$

또한  $\rho$ 가 -1, 0, 1인 경우를 각각 다음과 같이 부른다.

- $\rho = 1$ : 완전선형 상관관계
- $\rho = 0$ : 무상관 (독립과는 다른) 독립  $\cancel{\rho=0}$
- $\rho = -1$ : 완전선형 반상관관계



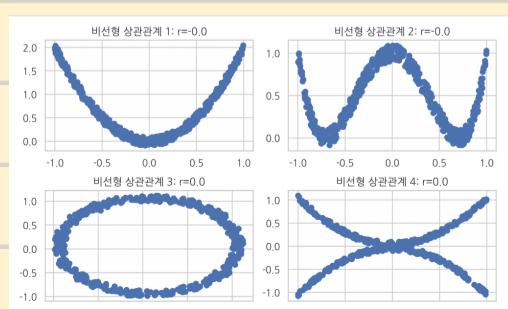
또한 상관계수는 스캐터 플롯의 기울기와는 아무런 상관이 없다.



또한 상관계수는 스캐터 플롯의 기울기와는 아무런 상관이 없다.

### 4) 비선형 상관관계 (로우=0 이어도 상관관계 가능. 단, 비선형으로!) 07.05 5page

- X값을 알면, Y값을 알 수 있는 힌트가 주어지는 데, 대신 그게 선형은 아닌 것

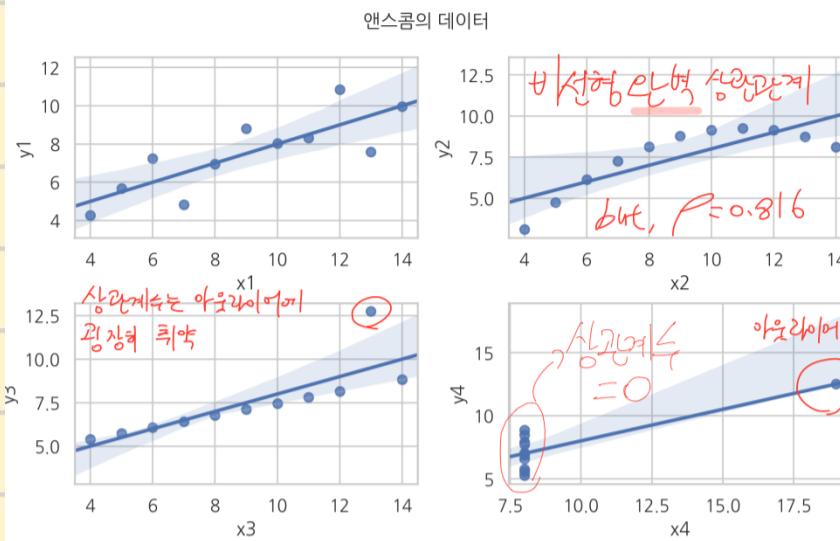


## 5) 상관계수의 약점 = 아웃라이어, 비선형관계

- 앤스콤 데이터 (07.05 7page)
- 아웃라이어, 비선형 상관관계의 경우,

독립이거나 완전한 비선형상관관계여도 선형 상관계수는 이를 제대로 표현하지 못

함



## 6) 다변수 확률변수의 표본공분산 = 표본공분산 행렬

- 안타깝게도, 공분산은 변수 2개에 대해서만 직접 한번에 구할 수 있음
- 대신, M개의 확률변수가 있다면, 2개씩 짹어서 공분산을 구해 이를 행렬로 표현!
- 표본공분산 행렬 (07.05 8page S식\_수기, 밑의 필기)

$((2,3) = 2\text{번째 확률변수와 } 3\text{번째 확률변수의 공분산}, (2,2) = 2\text{번째 확률변수의}$

분산)

M개의 서로 다른 확률변수의 모든 조합에 대한 공분산을 한꺼번에 표기하기 위해 다음처럼 표본공분산행렬(Sample Covariance Matrix)을 정의한다. 대각성분은 각각의 확률변수의 분산, 비대각성분은 서로 다른 두 확률변수의 공분산으로 정의되는 행렬이다. 예를 들어 두번째 행, 세번째 열의 원소  $s_{2,3}$ 은 두번째 확률변수와 세번째 확률변수의 공분산이다.

$$S = \begin{bmatrix} s_{x_1}^2 & s_{x_1 x_2} & \cdots & s_{x_1 x_M} \\ s_{x_1 x_2} & s_{x_2}^2 & \cdots & s_{x_2 x_M} \\ \vdots & \vdots & \ddots & \vdots \\ s_{x_1 x_M} & s_{x_2 x_M} & \cdots & s_{x_M}^2 \end{bmatrix} \quad (7.5.7)$$

$$\Sigma = \frac{1}{N} X_o^\top X_o$$

= 분산행렬 (양수 고유값 + 징역)

$$\begin{array}{c}
 \mathcal{X}_i \\
 \left[ \begin{array}{c} x_{i,1} \\ \vdots \\ x_{i,M} \end{array} \right] \\
 i = \text{표본 개수} \\
 (i=1 \sim N \text{개}) \\
 x_{i,M} = \text{번째 표본의 } M \text{번째 특징}
 \end{array}
 \rightarrow
 \begin{array}{c}
 \mathcal{X}_i^\top \\
 \boxed{\times} \\
 \Rightarrow
 \begin{bmatrix} x_{i,1} - \bar{x}_j \\ \vdots \\ x_{i,M} - \bar{x}_M \end{bmatrix} \\
 \boxed{\times} = X_o = X - \bar{X}
 \end{array}$$

$M$  특징 개수

## 7) 다변수 확률변수의 공분산 (07.05 9page)

- 다변수 확률변수 공분산 행렬은 '시그마'로 표현

다변수 확률변수의 공분산

M개의 다변수 확률변수 벡터

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix} \quad (7.5.16)$$

의 이론적 공분산행렬은  $\Sigma$ 로 표기하며 다음과처럼 정의한다.

$$\begin{aligned}
 \Sigma = \text{Cov}[X] &= \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1 x_2} & \sigma_{x_1 x_3} & \cdots & \sigma_{x_1 x_M} \\ \sigma_{x_1 x_2} & \sigma_{x_2}^2 & \sigma_{x_2 x_3} & \cdots & \sigma_{x_2 x_M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_1 x_M} & \sigma_{x_2 x_M} & \sigma_{x_3 x_M} & \cdots & \sigma_{x_M}^2 \end{bmatrix} \\
 &= E \begin{bmatrix} (X_1 - E[X_1])(X_1 - E[X_1])^T & \cdots & (X_1 - E[X_1])(X_M - E[X_M]) \\ (X_1 - E[X_1])(X_2 - E[X_2]) & \cdots & (X_2 - E[X_2])(X_M - E[X_M]) \\ \vdots & \ddots & \vdots \\ (X_1 - E[X_1])(X_M - E[X_M]) & \cdots & (X_M - E[X_M])^2 \end{bmatrix}
 \end{aligned}$$

다음과 같이 표기 수도 있다.

$$\begin{aligned}
 \Sigma &= E[(X - E[X])(X - E[X])^T] = E[\boxed{P}] \\
 &= E \left[ \begin{bmatrix} X_1 - E[X_1] \\ X_2 - E[X_2] \\ \vdots \\ X_M - E[X_M] \end{bmatrix} \begin{bmatrix} X_1 - E[X_1] & X_2 - E[X_2] & \cdots & X_M - E[X_M] \end{bmatrix} \right] \quad (7.5.18)
 \end{aligned}$$

07.06

조건부 기댓값과

예측 문제

기대값의 가중치로 pdf가 아닌,

기대값의 가중치로 조건부 pdf를 사용하면, 조건부 기댓값을 구할 수 있다.

(식 7.6.1)

확률변수  $Y$ 의 기댓값을 구할 때 주변 확률밀도함수  $p_Y(y)$ 를 사용하여 가중치를 계산하지 않고 조건부 확률밀도함수  $p_{Y|X}(y|x)$ 를 이용하여 기중치를 계산하면 조건부기댓값(conditional expectation) 혹은 조건부평균(conditional mean)이 된다.

$$E_Y[Y|X] = \int_{y=-\infty}^{y=\infty} y p_{Y|X}(y|x) dy \quad (7.6.1)$$

또는 간단히 다음처럼 쓴다.

$$E[Y|X] = \int y p(y|x) dy \quad (7.6.2)$$

기댓값은 결정론적 데이터

조건부 기댓값은 확률적 데이터로, 함수이다.

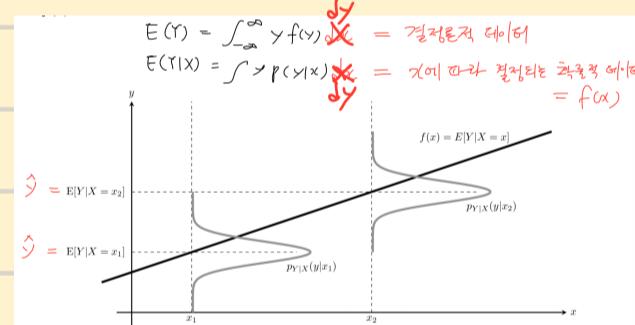
0) 조건부 기댓값 : 조건부 분포의 기댓값 (밑의 그림 참고)

1) 예측문제

-  $X$ 값을 알 때,  $Y$ 값을 알아내는 것 = 예측(Y가 이산  $\rightarrow$  분류, Y가 연속  $\rightarrow$  회귀분석)

- $X$ 의  $x$ 값을 알면, 그때의  $y$ 값(예측값)으로 조건부확률값( $P(y|x)$ ) 제시
- 단, 조건부확률값 중 가장 대표적인 값을 제시하기 위해,  $E[y|x]$ , 즉, 기댓값을

예측값으로 제시한다. (7.6.5, 그림 7.6.1)



$$x \xrightarrow{\text{예측}} \hat{y} = E[y|x] = f(x)$$

2) 조건부 기댓값의 성질 (07.06 2page)

- 조건부 기댓값은 상수 값이 될 수도, 확률변수가 될 수도 있다.
- $Y$ 와  $X$ 의 관계가 완벽한 선형으로,  $X$ 를 아는 순간 그것의 함수값  $Y$ 도 안다면, 조건부기댓값은 더 이상 확률적 데이터가 아닌 결정론적 상수가 된다.

결정, 완전한 상관관계  
 $Y = g(X) \Rightarrow E(Y|X) =$  상수, 결정론적 데이터  $= E(g(X)|X) = g(X) =$  상수

$Y \neq g(X) \Rightarrow E(Y|X) =$  확률적 데이터

3) 전체기댓값의 법칙

- 조건부 기댓값의 기댓값을 취하면, 조건부가 벗겨진다. (식 7.6.9)

조건부기댓값은 확률변수이므로 조건이 되는 확률변수에 대해 다시 기댓값을 구할 수 있다. 이렇게 반복하여 구한 조건부기댓값의 기댓값은 원래 확률변수의 댓값과 같다.

$$E_X[E_Y[Y|X]] = E_Y[Y] \quad (7.6.9)$$

## 4) 조건부분산

- 식 7.6.13

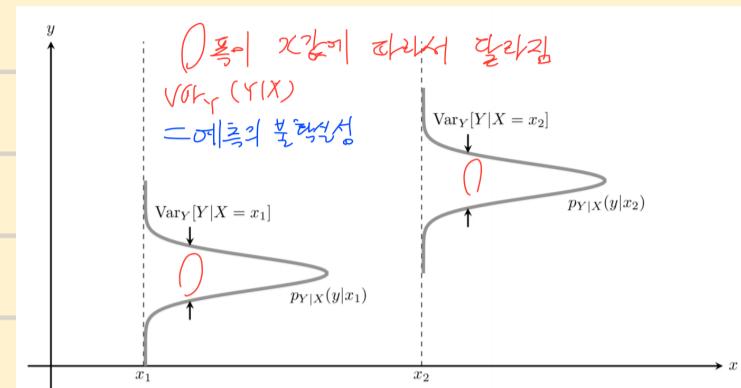
조건부기댓값을 정의한 것처럼 조건부분산(conditional variance)도 다음처럼 정의할 수 있다.

$$\text{Var}_Y[Y|X] = E_Y[(Y - E_Y[Y|X])^2 | X] = \int (Y - E_Y[Y|X])^2 f_{Y|X}(y|x) dy$$

07.06

조건부 기대값과  
예측 문제

- 조건부분산은 예측의 불확실성 정도를 말한다. 조건부분산이 클수록, 예측값(조건부 기댓값)이 불확실한 것(그림 7.6.2)



## 5) 전체 분산의 법칙 (식 7.6.14, 07.06 4page)

확률변수의 분산은 조건부분산의 기댓값과 조건부기댓값의 분산의 합과 같다. 이를 전체 분산의 법칙(law of total variance)라고 한다.

$$\text{Var}[Y] = E[\text{Var}[Y|X]] + \text{Var}[E[Y|X]] \quad (7.6.14)$$

전체 기댓값의 법칙을 사용하여 증명할 수 있다.

$$\text{Var}[Y] = E[Y^2] - (E[Y])^2$$

↳ 편향(오차)  
↳ 예측값의 일관성

↳ 예측값의 변동  
↳ 예측값의 평균

### \* 편향 - 분산 상충

- 평균적인 편향(오차) 와 예측값의 분산 은 서로 상충 관계이다.

- 편향(오차)줄이려고 모형복잡하게 하다보면, 예측값의 분산이 커지고, overfitting이 되곤 한다.

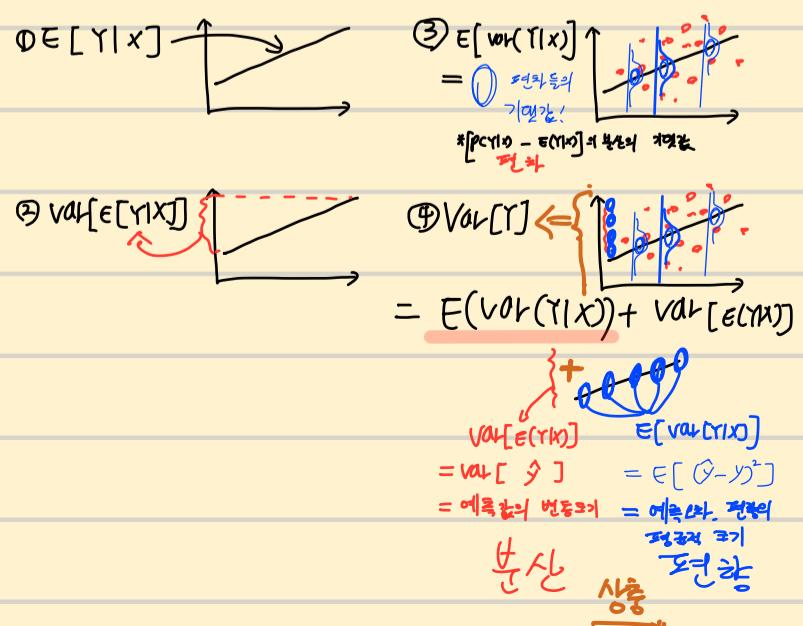
- overfitting을 우려해, 모형 간단하게 하면, 반대로 편향(오차)가 커진다.
- 결국, 적절한 절충 최적점을 찾아야 한다. 보통의 편향 과 분산 정도가 적절한. "항상 그 합은 일정하니까."

- 예측오차의 크기 + 예측값의 변동 =  $\text{Var}[Y]$ , 일정한 값!

- "편향(오차)줄이고 싶다" -> "모형 복잡하게" -> 과최적화, 비선형적, 분산이 커짐 (예측값(조건부기댓값)이 들쭉날쭉)

- "과최적화 줄이기 위해 -> 모형 간단하게 -> 편향(오차)가 커짐

- 결국, 편향(오차)와 분산은 상충관계!



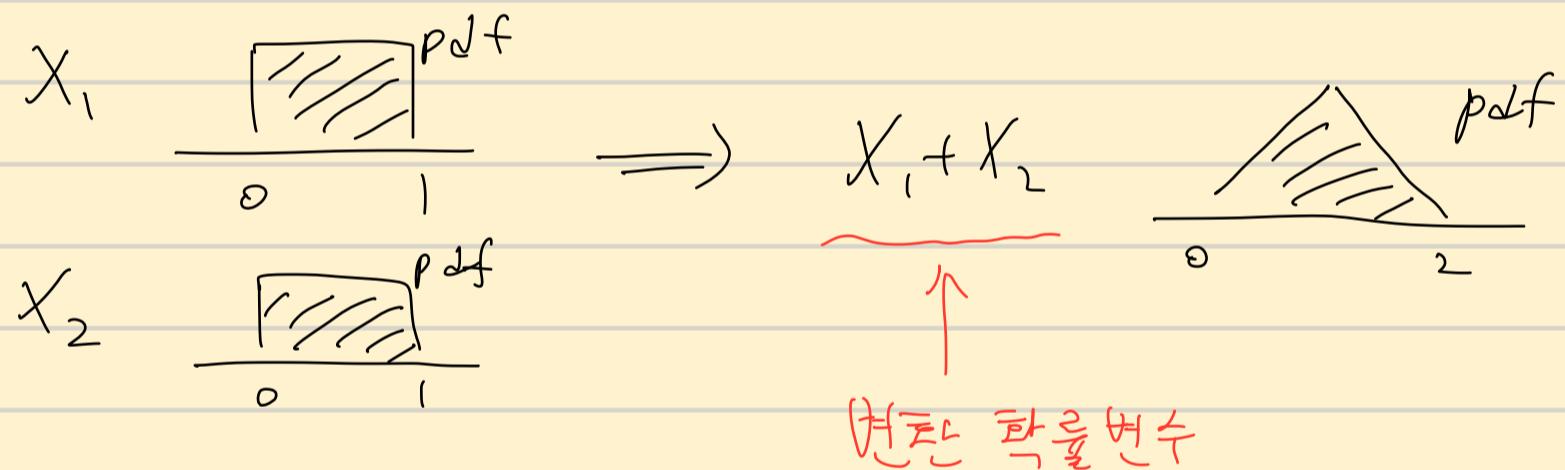
08.01

사이파이를 이용한  
확률분포 분석

\*pdf의 높이는 확률값이 아니다. 단지 cdf의 미분값(기울기)일 뿐.  
어디에 많은 관찰값들이 있는지 그 밀도를 보여주는 것뿐  
(실제로 pdf의 높이는 1이 넘는 값들이 많다.)  
(cdf는 그 높이가 확률값이 된다.)

[사이파이에서의 확률분포 기능 사용 순서]

- 1) 확률분포 클래스 설정 <= norm, beta, bernoulli, uniform..
- 2) 모수지정 <= 평균, 표준편차 지정
- 3) 확률분포 메소드 <= pmf, pdf, cdf ..
- 4) 무작위 표본 생성 (rvs 메소드) <= 확률변수를 생성하는 것
- 5) 변환확률변수의 시뮬레이션 <= 확률변수를 변환해보는 것 (08.01 5page)



08.02

베르누이분포와

이항분포

$$X \sim \text{Bern}(x; \mu)$$

↳ 모수! (분포를 설명할 수 있는)

은 다음과 같다.

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

## 2) 사이파이를 사용한 베르누이 확률변수의 시뮬레이션

- 시뮬레이션 : 확률변수의 표본을 얻는 과정(동전을 한번 던져보는 것)

## 3) 베르누이분포의 모멘트 식(8.2.5)

- 기댓값 :

$$E[X] = \mu$$

- 분산 :

$$\text{Var}[X] = \mu(1 - \mu)$$

## 4) 이항분포 : 베르누이 시행을 n번해서 성공할 횟수 = 이항분포 확률변수

- 이항분포의 확률변수는 베르누이시행을 n번 반복해서 성공한 횟수가 된다. (베르누이 분포 확률변수는 1번 시행)
- 이항분포의 모수는 시행횟수(N), 성공확률(μ)

## - 식 8.2.11

성공확률이  $\mu$ 인 베르누이 시행을  $N$ 번 반복하는 경우를 생각해보자. 가장 운이 좋을 때에는  $N$ 번 모두 성공할 것이고 가장 운이 나쁜 경우에는 한 번도 성공하지 못할 것이다.  $N$ 번 중 성공한 횟수를 확률변수  $X$ 라고 한다면  $X$ 의 값은 0부터  $N$ 까지의 정수 중 하나가 될 것이다.

이런 확률변수를 이항분포(binomial distribution)를 따르는 확률변수라고 하며 다음과 같이 표시한다.

$$X \sim \text{Bin}(x; N, \mu)$$

(8.2.11)

## 5) 사이파이를 사용한 이항분포의 시뮬레이션

## 6) 이항분포의 모멘트

- 기댓값 :

$$E[X] = N\mu$$

- 분산 :

$$\text{Var}[X] = N\mu(1 - \mu)$$

## 7) 베르누이분포와 이항분포의 모수추정 (μ)

- 모수추정 (μ) =  $\hat{\mu} = \frac{\sum_{i=1}^N x_i}{N} = \frac{N_1}{N}$  (8.2.21)

$$\hat{\mu} = \frac{\sum_{i=1}^N x_i}{N} = \frac{N_1}{N}$$

↳ 성공한 횟수

## 8) 베르누이분포의 활용

- 1) (베이지안적) 분류예측 시, 출력데이터가 2개로 나뉘는 카테고리값인 경우, 이를 표현하는데 사용

- 2) (빈도주의적) 입력데이터가 2개로 나뉘는 카테고리값인 경우, 두 종류의 값이 나타나는 비율을 표현

\*스팸메일 분류 예제 -&gt; BOW 인코딩된 벡터 생성 -&gt; 특징행렬 생성

- 특징행렬의 컬럼 = 베르누이 확률변수

(특징행렬 컬럼의 요소 = 베르누이 시행 결과 얻은 샘플 1 or 0)

각 데이터에 베르누이 시행  
 ↓  
 특징 행렬 열음  
 (정령 = 확률변수)  
 가 1 ~ 가 6 ⇒ 모수 추정  
 ε 1 ~ ε 6 ⇒  $\frac{N_1}{N}$

이번에는 스팸 메일 데이터에서 베르누이분포를 활용하는 방법을 알아보자. 스팸 메일은 특정한 단어(키워드)를 가지고 있을 확률이 높다. 스팸 메일을 구분하기 위한 키워드가 여러 개라면 다음과 같이 BOW(Bag of Words) 인코딩된 벡터로 나타낼 수 있다. 이 예제에서는 4개의 키워드 4개를 사용하였다. 만약 어떤 메일이 첫 번째와 세 번째 키워드를 포함하고 있으며 두 번째와 네 번째 키워드를 포함하지 않으면 다음과 같은 특징 벡터로 표시할 수 있다.

$$x = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

(8.2.23)

아래의 메일이 있으면 다음처럼 특징 행렬로 표시된다. 특징 행렬에서는 벙 벡터가 메일을, 열 벡터가 키워드를 나타낸다.

$$X_{\text{spam}} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

(8.2.24)

아래, 스팸 메일의 특성은 4개의 베르누이 확률변수의 류  $(X_1, X_2, X_3, X_4)$ 로 나타낼 수 있다.

$$X_1: \text{메일이 첫 번째 키워드를 포함하면 } 1, \text{ 아니면 } 0 \text{ 되는 확률변수}$$

(8.2.25)

$$p(X_1 = 1 | Y = 1) = \text{Bern}(x_1; \mu_{\text{spam}, 1})$$

(8.2.26)

$$X_2: \text{메일이 두 번째 키워드를 포함하면 } 1, \text{ 아니면 } 0 \text{ 되는 확률변수}$$

(8.2.27)

$$p(X_2 = 1 | Y = 1) = \text{Bern}(x_2; \mu_{\text{spam}, 2})$$

(8.2.28)

$$X_3: \text{메일이 세 번째 키워드를 포함하면 } 1, \text{ 아니면 } 0 \text{ 되는 확률변수}$$

(8.2.29)

$$p(X_3 = 1 | Y = 1) = \text{Bern}(x_3; \mu_{\text{spam}, 3})$$

(8.2.30)

$$X_4: \text{메일이 네 번째 키워드를 포함하면 } 1, \text{ 아니면 } 0 \text{ 되는 확률변수}$$

(8.2.31)

$$p(X_4 = 1 | Y = 1) = \text{Bern}(x_4; \mu_{\text{spam}, 4})$$

(8.2.32)

특징 행렬의 각 열로부터 각 베르누이 확률변수의 모수의 추정값을 구하면 다음과 같다.

$$\hat{\mu}_{\text{spam}, 1} = \frac{5}{6}, \hat{\mu}_{\text{spam}, 2} = \frac{4}{6}, \hat{\mu}_{\text{spam}, 3} = \frac{3}{6}, \hat{\mu}_{\text{spam}, 4} = \frac{3}{6}$$

(8.2.33)

## [ 학률 분포 표기 ]

## 1. 베르누이

$$X \sim \text{Bern}(\pi; \mu)$$

x  
↓  
0  
1

연수: 학률변수  $X$ 의 출현값  
(함수) (Sample)

## 2.4: Shape factor

## 2. 이항분포

$$X \sim \text{Bin}(n; N, \mu)$$

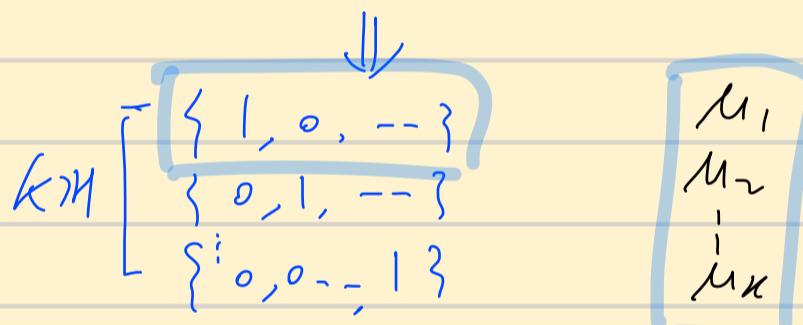
부수      제각각

X의 출현값

## 3. 카페고리

$$X \sim \text{Cat} \left( x_1, \dots, x_k ; \frac{\mu_1, \dots, \mu_k}{\text{모수, 벡터}} \right)$$

(한국, 변수 함수) 출력 벡터



$$x = \boxed{\phantom{000}} \text{ 전제조건}$$

$$\mu = \boxed{\phantom{00}} \text{ 전자조각}$$

$$1) x_i = 0 \text{ or } 1$$

2) [ ] 진짜 sum은 '1' (즉, 주사위 1번던지면

2) [ ] 진짜 SUM은 '1' (즉, 주사위 1번던지면 나오는 수는 단하나)

#### 4. 다음부호

$$X \sim \mathcal{M}_n(\vec{x}; \vec{N}, \vec{\mu})$$

↓  
 출력 벡터

$(4, 2, 3, 4, 3, 2)$  : 18번 던져서, 1014원, 2가 2번, ---

\* (at (0,0,/,--)) : 1번 더져서, 3이 1번!

08.03

카테고리분포와  
다항분포

## 1) 카테고리 확률변수

- 주사위를 던졌을 때의 경우 (베르누이는 2가지 경우, 즉 동전을 던졌을 때)
  - 1부터 K까지 K개의 정수값 중 하나가 나오는 것. 이 정수값을 범주값, 카테고리, 클래스라고 부른다.
  - $K=2$ 라면, 카테고리 확률변수 = 베르누이 확률변수
- K개의 값 중, 실험 결과 a값이 나오는 사건의 확률은 뮤 $a$ 이고, 확률변수는 원핫인코딩-다차원 벡터로 표현한다.  
(확률변수를  $x=2$ 로 표현하지 않고,  $(0,1,0,0,0,0)$ 으로 다차원벡터로 표현)  
(확률변수의 값도 벡터로 표현한다.)

### - 카테고리를 원핫인코딩 -> 다차원 벡터로 출력 (8.3.1)

주의할 점은 원래 카테고리는 스칼라값이지만 카테고리 확률변수는 다음과 같이 1과 0으로만 이루어진 다차원 벡터를 출력한다. (벡터는 원래 세로 열로 표시해야하지만 여기에서는 편의상 가로 행으로 표시하였다.)	
$x = 1 \rightarrow$	$x = (1, 0, 0, 0, 0, 0)$
$x = 2 \rightarrow$	$x = (0, 1, 0, 0, 0, 0)$
$x = 3 \rightarrow$	$x = (0, 0, 1, 0, 0, 0)$
$x = 4 \rightarrow$	$x = (0, 0, 0, 1, 0, 0)$
$x = 5 \rightarrow$	$x = (0, 0, 0, 0, 1, 0)$
$x = 6 \rightarrow$	$x = (0, 0, 0, 0, 0, 1)$

숫자를 이렇게 변형하는 것을 원핫인코딩(One-Hot-Encoding)이라고 한다.

### - 카테고리 확률변수의 값을 표현하는 다차원 벡터의 제한조건 (8.3.4 - 5)

따라서 확률변수의 값도 다음처럼 벡터로 표시한다.

$$x = (x_1, x_2, x_3, x_4, x_5, x_6) \quad (8.3.2)$$

이 벡터를 구성하는 원소  $x_1, x_2, x_3, x_4, x_5, x_6$ 에는 다음과 같은 제한 조건이 붙는다.

$$x_i = \begin{cases} 0 \\ 1 \end{cases} \quad (8.3.3)$$

$$\sum_{k=1}^K x_k = 1 \quad (8.3.4)$$

첫 번째 제한 조건은  $x_k$  값으로 0 또는 1 만 가능하다는 것이고, 두 번째 제한 조건은 여러  $x_k$  중 단 하나만 1일 수 있다는 것이다.

### - 카테고리 확률변수의 모수를 표현하는 다차원 벡터의 제한조건 (8.3.6 - 7)

- 모수는 카테고리가 6개면, 6개의 원소가 각각 1이 나올 확률을 의미

(6개의 동전을 던지는 경우라고 생각)

원솟값  $x_k$ 는 베르누이 확률변수로 볼 수 있기 때문에 각각 1이 나올 확률을 나타내는 모수  $\mu_k$ 를 가진다. 따라서 전체 카테고리분포의 모수는 다음과 같이 벡터로 나타낸다.

$$\mu = (\mu_1, \dots, \mu_K) \quad (8.3.5)$$

이 모수 벡터도 다음과 같이 제한 조건을 가진다.

$$0 \leq \mu_i \leq 1 \quad (8.3.6)$$

$$\sum_{k=1}^K \mu_k = 1 \quad (8.3.7)$$

## 2) 카테고리 확률분포

카테고리분포와

## - 8.3.8-9

다항분포

## 카테고리 확률분포

카테고리 확률변수의 확률분포인 카테고리 확률분포는  $\text{Cat}(x_1, x_2, \dots, x_K; \mu_1, \dots, \mu_K)$

$$\text{Cat}(x_1, x_2, \dots, x_K; \mu_1, \dots, \mu_K) \quad \text{※ 벡터!} \quad (8.3.8)$$

로 표기하거나 출력 벡터  $x = (x_1, x_2, \dots, x_K)$ , 모수 벡터  $\mu = (\mu_1, \dots, \mu_K)$ 를 사용하여  
 $\text{Cat}(x; \mu)$

$$\text{Cat}(x; \mu) \quad (8.3.9)$$

로 간단히 표기할 수 있다.

## - 카테고리 pmf 8.3.10 - 11

확률질량함수는 다음처럼 표기한다.

$$\text{Cat}(x; \mu) = \begin{cases} \mu_1 & \text{if } x = (1, 0, 0, \dots, 0) \\ \mu_2 & \text{if } x = (0, 1, 0, \dots, 0) \\ \mu_3 & \text{if } x = (0, 0, 1, \dots, 0) \\ \vdots & \vdots \\ \mu_K & \text{if } x = (0, 0, 0, \dots, 1) \end{cases} \quad (8.3.10)$$

위 식은 다음과 같이 쓸 수 있다. 이 간략한 표현은 원핫인코딩을 사용한 덕분이다.

$$\text{Cat}(x; \mu) = \mu_1^{x_1} \mu_2^{x_2} \cdots \mu_K^{x_K} = \prod_{k=1}^K \mu_k^{x_k} \quad (8.3.11)$$

## 2) 카테고리분포의 모멘트(기댓값, 분산)

## - 표본값이 벡터. 따라서, 기댓값과 분산 모두 벡터 (벡터함수) (8.3.12-13)

카테고리분포는 표본값이 벡터이므로 기댓값과 분산도 벡터이다. 기댓값과 분산을 구하는 공식은 다음과 같다.

- 기댓값

$$\mathbb{E}[x_k] = \mu_k \quad (8.3.12)$$

- 분산

$$\text{Var}[x_k] = \mu_k(1 - \mu_k) \quad (8.3.13)$$

## 3) 다중분류문제

- 이진분류가 아닌, 3개 이상의 범주값을 분류하는 문제 -> 카테고리분포를 사용해 범주값 데이터 모형을 만들 수 있음
- 이후 decision tree -> random forest 방식으로 확장

08.03

카테고리분포와

다항분포

## 4) 다항분포

- 주사위를 한 번 던지는 것 = 카테고리분포 ( $x$ 는 1~6(k)까지. 대신 원핫인코딩으로 벡터화)
- 주사위를 여러번 던지는 것 = 다항분포 (베르누이 : 이항 / 카테고리 : 다항)
- $k = 2$ , 베르누이 = 카테고리.. 따라서, 이항 = 다항
- 다항분포의 확률변수 벡터 표기 (8.3.14 상단)

다항분포는 카테고리가  $K$ 개인 카테고리 확률변수의 표본 데이터를  $N$ 개 얻었을 때, 각각의 카테고리  $k$  ( $k = 1, \dots, K$ )가 각각  $x_k$  번 나올 확률분포 즉, 표본값이 벡터  $x = (x_1, \dots, x_K)$ 가 되는 확률분포를 말한다.

예를 들어  $x = (1, 2, 1, 2, 3, 1)$ 은 6개의 숫자가 나올 수 있는 주사위를 10번 던져서 1인 면이 1번, 2인 면이 2번, 3인 면이 1번, 4인 면이 2번, 5인 면이 3 번, 6인 면이 1번 나왔다는 뜻이다.

### - 다항분포와 카테고리분포의 pmf

$$\text{Mu}(x; N, \mu) = \binom{N}{x} \prod_{k=1}^K \mu_k^{x_k} = \binom{N}{x_1, \dots, x_K} \prod_{k=1}^K \mu_k^{x_k}$$

$$\text{Cat}(x; \mu) = \mu_1^{x_1} \mu_2^{x_2} \cdots \mu_K^{x_K} = \prod_{k=1}^K \mu_k^{x_k}$$

## 5) 다항분포의 모멘트 (8.3.16 - 17)

다항분포의 기댓값과 분산은 다음과 같다.

- 기댓값

$$E[x_k] = N\mu_k \stackrel{\text{def}}{=} X \sim \text{Bin}(N, p) \quad (8.3.16)$$

- 분산

$$\text{Var}[x_k] = N\mu_k(1 - \mu_k) \stackrel{\text{def}}{=} \text{Var}(X) = np(1-p) \quad (8.3.17)$$

## 6) 다항분포와 이항분포 (8.3 11page)

```
In [8]:
N = 30
mu = [0.1, 0.1, 0.1, 0.1, 0.3, 0.3]
rv = sp.stats.multinomial(N, mu)
```

```
np.random.seed(0)
X = rv.rvs(100)
X[:10]
```

(100 줄 중)

이항분포는 30번 던지는 행위를 100번 했어!

이항분포는 30번 던지는 행위를 100번 했어!

[이항분포: [30번 던지는 행위]  $\xrightarrow{\text{100번 했어!}}$  100세트] \* 단지, 이항분포는  $X$  출력값이 이진, 다항분포는 다중!

$X \sim \text{Bin}(30, \frac{1}{6})$

[다항 분포: [30번 던지는 행위]  $\xrightarrow{\text{100번 했어!}}$  100세트]

$X \sim \text{Mu}(\bar{x}, 30, \bar{\mu}_x)$

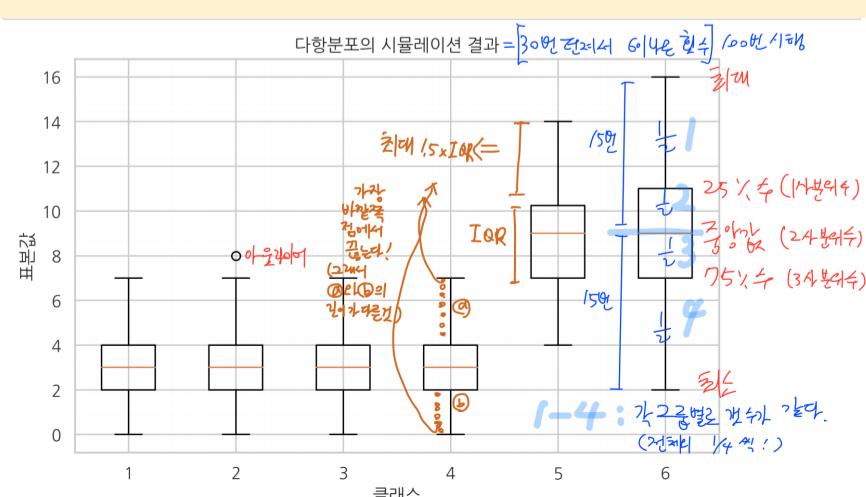
$\hat{X} \sim \text{Bin}(30, \frac{1}{6})$

6이 나오는 확률

1세트: 6이 10번  
2세트: 6이 8번  
해보니 9번 나오는데 100세트:

평균!

## 7) box plot (8.3 12page)



정규분포 : 자연현상에서 나타나는 숫자를 확률모형으로 모형화할 때 사용

정규분포 pdf : 8.4.1 (계산 편의성을 위해, 분산의 역수를 정밀도(precision)이라 하여, 베타로 치환해 사용하기도 함)

- $x = \text{평균}$ 에서 확률밀도가 최대가 됨

- 항상 양수값이 나오는 데이터(예: 거래량)는 로그를 취해 로그정규분포로 활용

### 하기도 함

정규분포는 평균  $\mu$ 와 분산  $\sigma^2$ 이라는 두 모수만으로 정의되며 확률밀도함수(pdf: probability density function)는 다음과 같은 수식으로 표현된다.

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (8.4.1)$$

분산의 역수를 정밀도(precision)  $\beta$ 라고 부르기도 한다.

$$\beta = \frac{1}{\sigma^2} \quad (8.4.2)$$

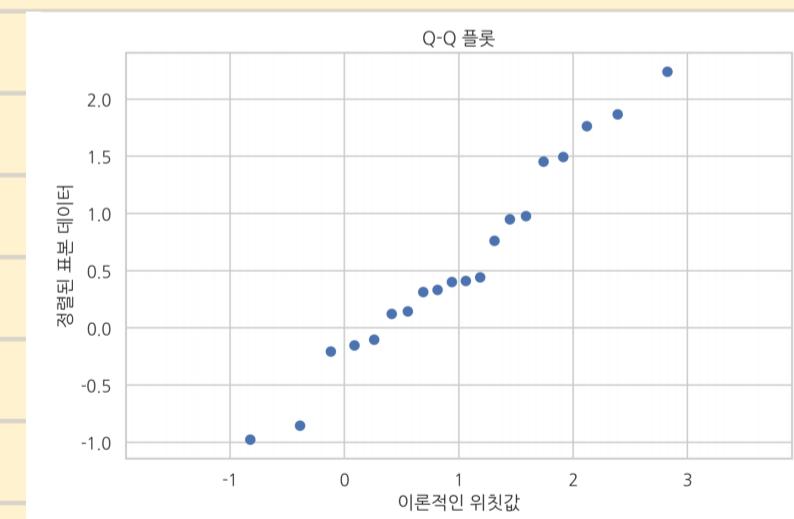
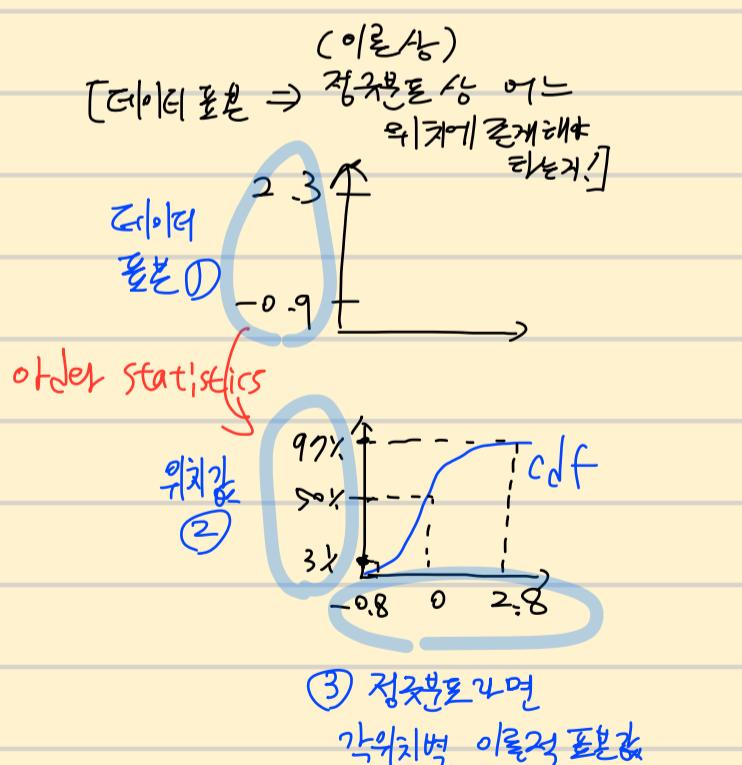
### 1) 표준정규분포

- 정규분포 중, 평균=0, 분산=1인 정규분포

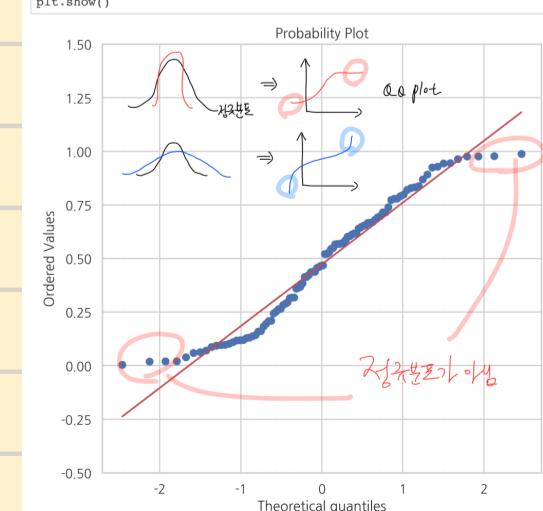


### 2) Q-Q 플롯 (Quantile - Quantile 플롯) 8, 9, 11page

- 어떤 확률변수의 분포가 정규분포인지 아닌지 확인하는 방법 (분포를 비교)



```
np.random.seed(0) # Uniform 한 값! (0 ≤ 1)
x = np.random.rand(100)
plt.figure(figsize=(7, 7))
sp.stats.probplot(x, plot=plt)
plt.ylim(-0.5, 1.5)
plt.show()
```



## 3) 중심극한정리

- 중심극한정리 : 여러 확률변수의 합이 정규분포와 비슷한 분포를 이루는 현상

- 8.4.3 - 5

$X_1, X_2, \dots, X_N$  가 기댓값이  $\mu$ 이고 분산이  $\sigma^2$ 으로 동일한 분포(기댓값과 분산의 값이 동일할 뿐이며 분포의 모양은 달라)도 된다. 서로 독립인 확률변수들이라고 하자. 분포가 어떤 분포인지는 상관없다.

$$\bar{x}_N = \frac{1}{N}(x_1 + x_2 + \dots + x_N) \quad (8.4.3)$$

도 마찬가지로 예측할 수 있는 확률변수다. 이 확률변수를  $\bar{X}_N$  이라고 하자.

중심극한정리는 다음과 같다.

$N$  개의 임의의 분포로부터 얻은 표본의 평균은  $N$  이 증가할수록 기댓값이  $\mu$ , 분산이  $\frac{\sigma^2}{N}$  인 정규분포로 수렴한다.

$$\bar{X}_N \xrightarrow{d} \mathcal{N}\left(x; \mu, \frac{\sigma^2}{N}\right) \quad (8.4.4)$$

→ 기호는 표본 개수  $N$  이 커질수록 분포의 모양이 특정한 분포에 수렴한다는 것을 뜻한다. 이 표본 평균의 평균이 0, 분산이 1이 되도록 다음처럼 정규화를 하면 다음과 같이 쓸 수도 있다.

$N$  개의 임의의 분포로부터 얻은 표본의 평균을 정규화하면  $N$  이 증가할 수록 표준정규분포로 수렴한다.

$$\frac{\bar{X}_N - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(x; 0, 1) \quad (8.4.5)$$

## 4) 정규분포의 통계량 분포 ( z 통계량 ~ 표준정규분포 )

- 모든 분포로부터의 추출을 상정하는 중심극한정리와 달리

- 정규분포로부터 추출한 표본의 합(통계량)은 몇개를 추출했던 간에 모두 정규분포를 따른다.

- z 통계량 : 정규분포 표본의 평균을 정규화한 통계량

=>> z 통계량은 표준정규분포를 따른다. (8.4.6 - 7)

그렇다면 임의의 분포가 아닌 복수의 정규분포로부터 얻은 표본 데이터로 구한 표본평균은 어떤 분포를 가지게 될까?

$N$  개의 정규분포로부터 얻은 표본의 합은  $N$  과 상관없이 기댓값이  $N\mu$ , 분산이  $N\sigma^2$ 인 정규분포다.

$$x_i \sim \mathcal{N}(\mu, \sigma^2) \rightarrow \sum_{i=1}^N x_i \sim \mathcal{N}(N\mu, N\sigma^2) \quad (8.4.6)$$

정규분포의 표본에 상수를 빼거나 곱해도 정규분포다. 이 경우에도 위와 같이 기댓값이 0, 표준편차가 1이 되도록 정규화를 하면 다음과 같이 쓸 수 있다.

$$x_i \sim \mathcal{N}(\mu, \sigma^2) \rightarrow z = \frac{\bar{x} - \mu}{\sigma} \sim \mathcal{N}(x; 0, 1) \quad (8.4.7)$$

정규분포 표본의 평균을 정규화한 통계량은 z 통계량이라고 한다. 중심극한정리와 다른 점에 주의해야 한다. 중심극한정리에서는 표준정규분포로 점점 다가갈 뿐이고 표본 개수가 무한대가 되기 전에는 정확한 정규분포가 아니지만 z 통계량은 개수  $N$ 에 상관없이 항상 정확하게 표준정규분포이다.

## 5) 선형회귀모형과 정규분포

- 선형회귀모형 : 잔차, 잡음  $e$  == 기댓값이 0인 정규분포를 따른다고 가정

- 영향을 미치는 수많은 현실의 다른 변수들의 분포들 영향을 모두 더해서  $e$ 로 놓고, 이것들은 정규분포를 따른 것으로 가정.

(서로 다른 분포의 확률별수를 모두 합한 통계량은 정규분포를 따르기 때문에)

- 8.4.12 ~ 13, 그림 8.4.1

중심극한정리에 의해 임의의 확률변수의 합은 정규분포와 비슷한 형태가 된다. 또한  $e$ 의 기댓값이 0이 아니라면 다음처럼 상수항  $w_0 = E[e]$ 을 추가하는 대신  $e$ 의 기댓값이 0이라고 할 수 있기 때문에

$$y = w_0 + w_1 x_1 + \dots + w_N x_N + e \quad (8.4.12)$$

잡음  $e$ 이 기댓값이 0인 정규분포라고 가정하는 것은 합리적이다.

$$e \sim \mathcal{N}(0, \sigma^2) \quad (8.4.13)$$

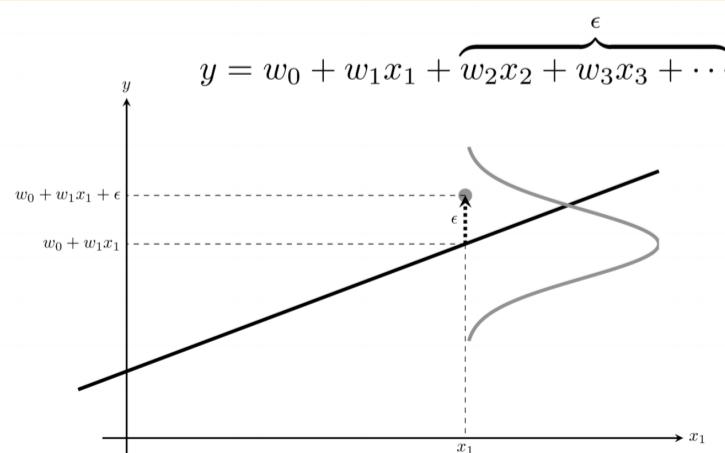


그림 8.4.1 : 선형회귀모형과 정규분포

