3장. 확률과 확률분포

박명현

서울대학교 학부대학

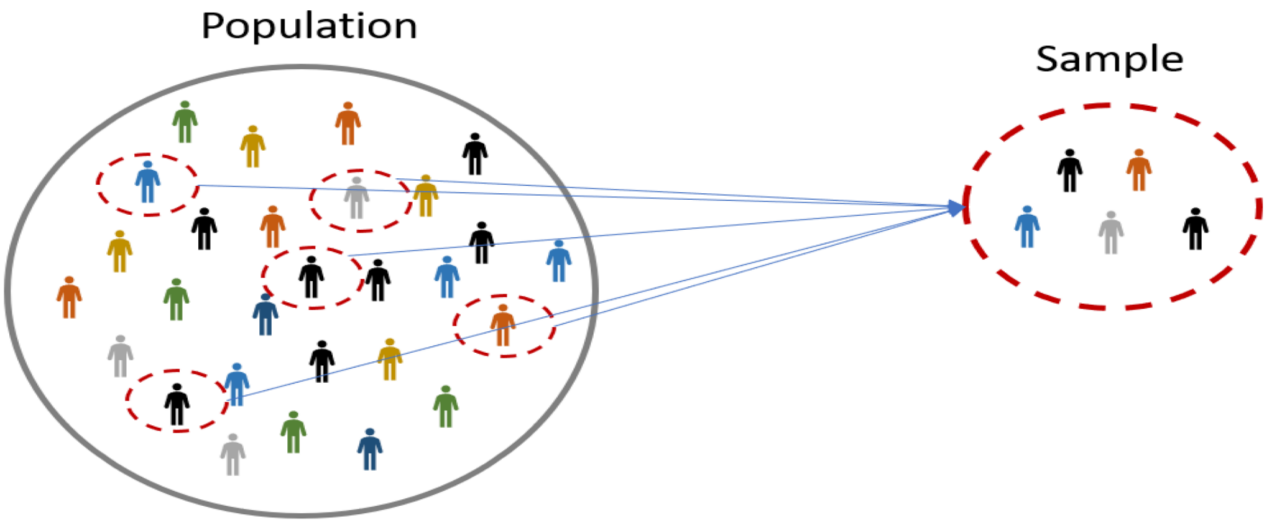
Review

- 확률의 정의
- 확률의 성질
  - 조건부/곱셈/전확률/베이즈/독립
- 확률변수
  - 이산형/연속형
- 확률분포의 대표값
  - 중심/변동/연관성

확률(probability)의 정의

- 표본이 뽑히는 과정
- 통계자료를 분석하기 위해서는 통계이론이 필요
- 통계적 조사에서는 표본을 바탕으로 모집단 전체에 대한 결론을 이끌어 낸다.

  - 이에 대한 논리적 근거가 되는 것이 확률의 개념이다.



확률의 정의

- 확률을 정의하기 위해 다음의 용어를 알아보자.

- 사건(event): 발생 가능한 결과들의 집합
- 단순사건(simple event) 또는 근원사건(elementary event): 오직 한 개의 원소로 이루어진 사건
- 표본공간(sample space): 일어날 수 있는 모든 가능한 단순사건을 모아 집합으로 표시한 것. S로 표시.
  - 모든 원소를 포함(exhaustive)
  - 상호배반(mutually exclusive)

사건과 표본공간의 예

- 주사위를 던지는 경우 표본공간 $S = \{1, 2, 3, 7\text{보다 작은 짝수}\}$?

- 동전과 주사위를 동시에 던지는 실험에 대한 표본공간
  sol) $S = \{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}$
- 어느 공정에서 10개의 불량품이 제조될 때까지 제조된 제품의 개수를 조사할 때,
  sol) $S = \{10, 11, 12, \cdots\}$
- 한 공장에서 생산된 전구의 수명시간을 조사하기 위해 한 개의 전구를 추출하여 수명시간을 관측할 때,
  sol) $S = \{t | t \geq 0\}$

확률의 정의

- 확률의 고전적 정의(라플라스, Laplace, 1749-1827)

: N개의 원소로 구성된 표본공간 $S = \{e_1, \cdots, e_N\}$에서 각각의 근원사건이 일어날 가능성이 같은 경우에 m개의 원소로 구성된 사건 A의 확률은 다음과 같이 정의한다.

$$P(A) = \frac{m}{N}$$

예) 주사위를 던졌을 때 홀수의 눈이 나올 확률:

$$A = \{1, 3, 5\},\ P(A) = \frac{3}{6}$$

- 확률을 가능성에 대한 수리적 측도로 위와 같이 정의
- 표본공간이 유한개의 원소만을 가질 때 적용됨
- 각 결과가 나타날 가능성이 같지 않은 경우에는 적용이 안됨

확률 예제

Example (3.1)

흰 탁구공 3개와 노란 탁구공 1개가 들어 있는 상자에서 탁구공 2개를 단순랜덤추출할 때, 흰 탁구공 1개와 노란 탁구공 1개가 나올 확률을 구하라.

sol) - 흰 탁구공 3객를 각각 a,b,c로, 노란 탁구공을 d로 표기하면 표본공간은 다음과 같다.

$$S = \{\{a,b\},\{a,c\},\{a,d\},\{b,c\},\{b,d\},\{c,d\}\}$$

- 흰 탁구공 1개와 노란 탁구공 1개가 나올 사건은

$$A = \{\{a,d\},\{b,d\},\{c,d\}\}$$

- 색상과 상관없이 탁구공 2개가 뽑힐 가능성은 모두 같다고 할 수 있으므로

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

확률의 정의

- 일반적인 확률의 개념은 동일한 통계적 조사를 반복적으로 시행하면서 누적되는 경험을 바탕으로 이루어짐
- 한 공장에서 생산되는 제품의 불량품에 관심

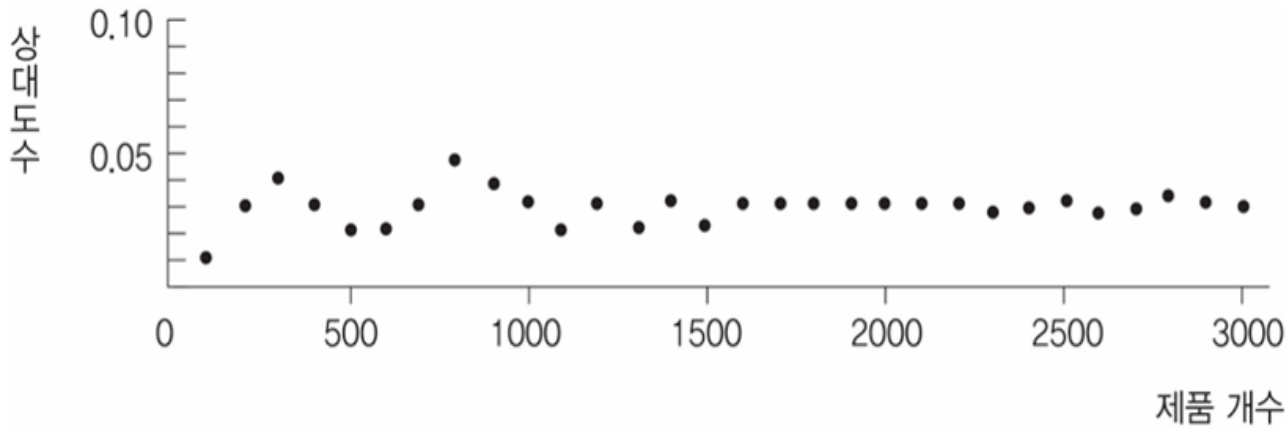


그림 3–1 불량품의 상대도수 변화

- 상대적 비율 접근법:
  불량품의 비율은 제품의 수가 증가함에 따라 어떤 상수에 수렴
  -> 이 상수를 불량품이 생산될 확률로 정의

확률의 정의

- 안드레이 콜모고로프(Andrey Kolmogorov, 1903 1987)는 확률을 상대도수의 극한 개념으로 위와 같이 확률의 공리적 정의를 정립

- 확률의 공리적 정의:

  (1)표본공간 S에서 임의의 사건 A에 대하여 $0 \leq P(A) \leq 1$

  (2) $P(S) = 1$

  (3) 서로 배반인 사건 A_1, $A_2, \cdots$에 대하여

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$$

  를 만족할 때, $P(A)$를 사건 A의 확률이라 한다.

- 상대도수의 극한과 비슷하게 반복 시행에서의 경험을 바탕으로 한 경험적 확률뿐만 아니라 개인의 주관에 따라 확률 값을 정하는 주관적 확률 개념도 포함

확률에 관한 성질

- $A \cup B$: 사건 A, B의 합집합, $A \cap B$: 사건 A, B의 곱집합
- A^c: 사건 A의 여집합

- 확률에 관한 성질
  (1) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  (2) $P(A) = 1 - P(A^c)$
  (3) $A \subset B$이면 $P(A) \leq P(B)$

Figure: $A \cup B$의 분할

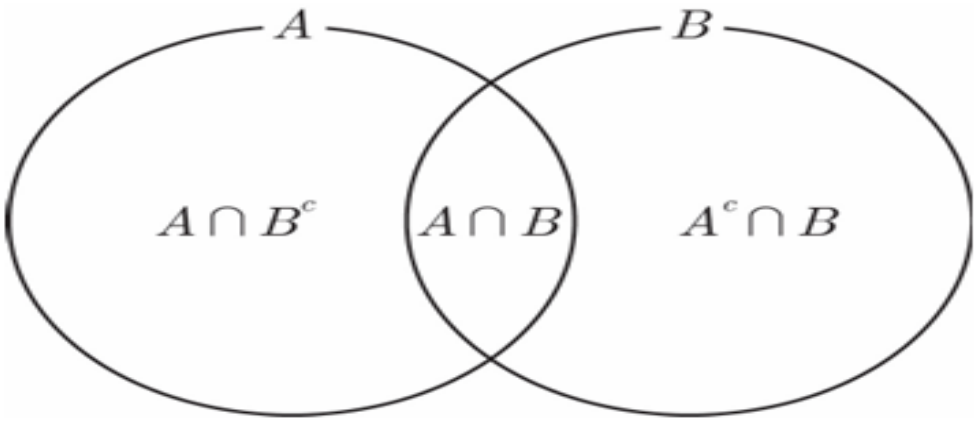


그림 3–2 $A \cup B$의 분할

확률에 관한 성질

Example (3.2)

세 개의 주머니에서 각각 1부터 5까지의 번호가 매겨진 5개의 구슬이 들어 있다. 각 주머니에서 1개씩의 구슬을 꺼낼 때, 꺼낸 구슬의 숫자의 합이 5 이상일 확률을 구하여라.

sol) A: 꺼낸 숫자의 합이 5 이상일 사건

B: 꺼낸 숫자의 합이 3일 사건, C: 꺼낸 숫자의 합이 4일 사건

$$P(A^c) = P(B) + P(C)$$

- 가능한 모든 결과는 $5 \times 5 \times 5 = 125$가지이고

$B = \{(1,1,1)\}$, $C = \{(1,1,2).(1,2,1).(2,1,1)\}$이므로

$$P(B) = 1/125, \quad P(C) = 3/125$$

따라서 $P(A) = 1 - \{P(B) + P(C)\} = 1 - \frac{4}{125} = \frac{121}{125}$.

조건부확률

- 조건부확률(conditional probability): 한 사건이 일어났다는 조건하에 다른 사건이 일어날 확률
- 사건 A가 주어졌을 때, 사건 B가 일어날 조건부확률은 다음과 같이 정의된다.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (P(A) > 0)$$

- 사건 A를 새로운 표본공간으로 간주

- 예) 학점이 좋은 학생의 취업 확률(조건='학점이 좋다')

조건부확률

Example (3.3)

두 개의 주사위를 던지는 경우 첫번째 던진 주사위의 눈이 두번째 주사위의 눈보다 클 때, 두 주사위의 눈의 합이 7일 확률을 구하여라.

sol) A: 첫번째 던진 주사위의 눈이 두번째 주사위의 눈보다 클 사건

B: 두 주사위의 눈의 합이 7일 사건

- 표본공간은 36개, 사건 A는 15개의 원소를 가지므로

$$P(A) = 15/36$$

- $B = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$이고, $A \cap B = \{(4,3),(5,2),(6,1)\}$이므로,

$$P(B) = 6/36, \quad P(A \cap B) = 3/36$$

- 따라서 $P(B|A) = \frac{P(A \cap B)}{P(A)} = 1/5$

곱셈법칙

- 곱셈법칙:
  $P(A) > 0,\ P(B) > 0$이면 다음이 성립한다.

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Example (3.4)

불량품 20개와 양호품 80개로 구성된 로트에서 2개의 제품을 단순랜덤 추출할 때, 2개 모두 불량품일 확률을 구하여라.

sol) A: 첫번째 꺼낸 제품이 불량품일 사건

B: 두번째 꺼낸 제품이 불량품일 사건

- 구하는 확률은 $P(A \cap B) = P(B|A)P(A) = \frac{19}{99} \times \frac{20}{100} = \frac{19}{495}$.

베이즈 정리(Bayes' theorem)

- 조건확률을 구할 때, 조건 상황이 역으로 되어있는 확률을 이용하는 것

Theorem (베이즈 정리)

$A_i \cap A_j = \emptyset (i \neq i)$이고, $A_1 \cup A_2 \cup \cdots \cup A_n = S$일 때 $A_1, A_2, \cdots, A_n$으로 표본공간 S를 분할하면

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

$$= \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n)}$$

- 사전확률(prior probability):
  사건 B가 일어나기 전의 확률 $P(A_i)$
- 사후확률(posterior probability):
  사건 B가 일어난 후의 확률 $P(A_i|B)$

베이즈 정리 예제

Example

어떤 대학의 통계학과 학생의 30%는 1학년, 25%는 2학년, 25%는 3학년, 20%는 4학년 학생이라 하자. 그런데 1학년의 50%, 2학년의 30%, 3학년의 10%, 4학년의 2%가 수학 과목의 수강생이라 한다. 수학 과목을 수강하고 있는 학생 중 통계학과 1학년 학생의 비율을 구하여라.

sol) A_1: 뽑힌 학생이 1학년일 사건, A_2: 뽑힌 학생이 2학년일 사건
A_3: 뽑힌 학생이 3학년일 사건, A_4: 뽑힌 학생이 4학년일 사건
B: 뽑힌 학생이 수학 과목 수강생일 사건

- $P(A_1) = 0.3$, $P(A_2) = 0.25$, $P(A_3) = 0.25$, $P(A_4) = 0.2$
- $P(B|A_1) = 0.5$, $P(B|A_2) = 0.3$, $P(B|A_3) = 0.1$, $P(B|A_4) = 0.02$

- 사건 A_1, A_2, A_3, A_4는 서로 배반이므로,

$$\begin{aligned} P(B) &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) \\ &\quad + P(B|A_3)P(A_3) + P(B|A_4)P(A_4) \\ &= 0.5 \times 0.3 + 0.3 \times 025 + 0.1 \times 0.25 + 0.02 \times 0.2 \\ &= 0.254. \end{aligned}$$

- $P(B) = 0.254$이므로

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{P(A_1)P(B|A_1)}{P(B)} = \frac{0.3 \times 0.5}{0.254} = 0.59$$

- $P(A_1|B) > P(A_1)$

의료진단과 베이즈 규칙

Example

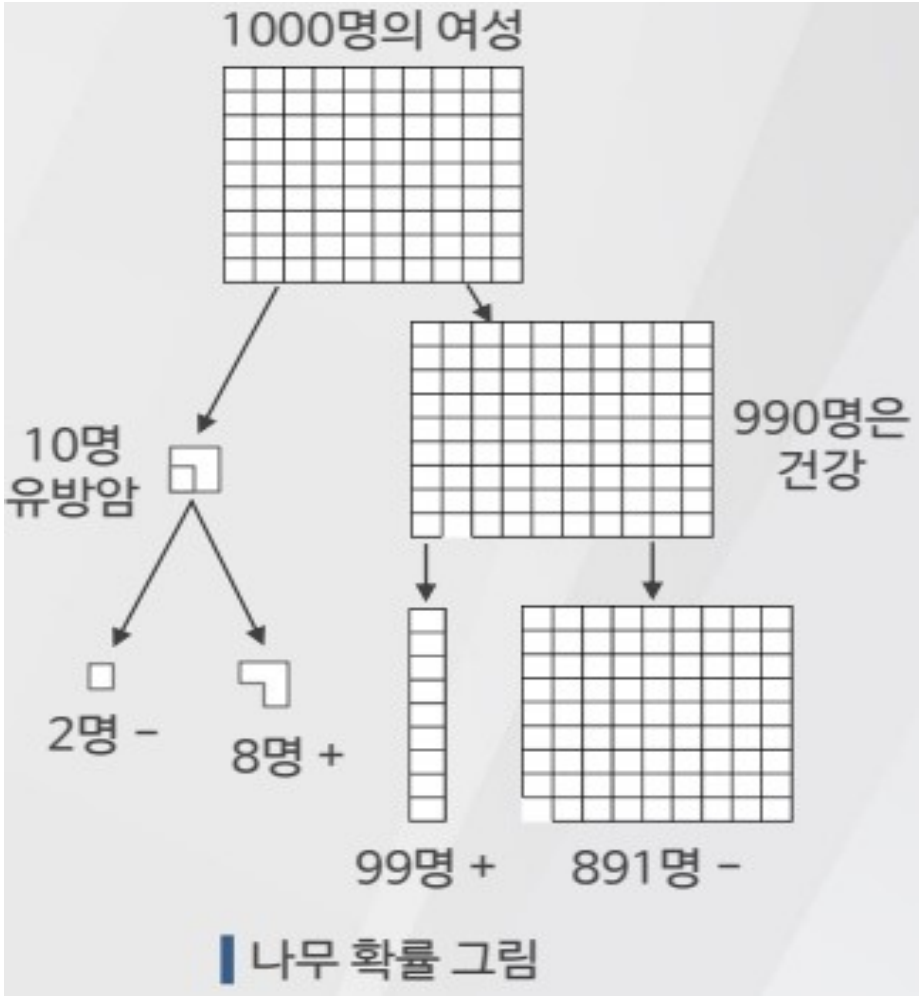
한 여성이 유방조영술 촬영 결과 양성판정을 받았다. 그 여성이 진짜로 유방암에 걸렸을 확률은 어떻게 될까? 유방암이 있는 환자가 유방조영술 검사를 통해 유방암 판정을 받을 확률(양성 반응)은 80%이고, 유방암이 걸리지 않은 여성이 음성 반응을 보일 확률은 90%, 그리고 검사를 받은 여성 중 1%에게 실제 암이 있다고 가정하자.

sol) C: 유방암에 걸린 경우, $+$: 검사 결과 양성반응, $-$: 음성반응

$P(C) = 0.01$, $P(+|C) = 0.8$, $P(-|C^C) = 0.9$

$$\begin{aligned} P(+) &= P(+|C)P(C) + P(+|C^C)P(C^C) \\ &= 0.8 \times 0.01 + 0.1 \times 0.99 = 0.107. \end{aligned}$$

$$P(C|+) = \frac{P(C \cap +)}{P(+)} = \frac{P(C)P(+|C)}{P(+)} = 0.075$$



- $P(C) = 0.01 => P(C|+) = 0.075$
- 암 발생률이 낮음(검사를 받는 여성 대다수가 건강함)
- $P(+|C)$와 $P(C|+)$ 개념 혼용

독립사건

- 독립사건: 사건 A가 일어날 사건이 사건 B가 일어날 확률에 아무런 영향을 미치지 않음

예) 동전이 앞, 뒤가 나오는 것과 주사위가 어떤 값이 나오는지는 서로 무관.

$$P(\text{"}H\text{"} \cap \text{"}1\text{"}) = P(\text{"}H\text{"})P(\text{"}1\text{"})$$

예) 주사위가 무엇이 나오는지 조건을 걸어도 동전이 앞, 뒤 나오는 것은 전혀 영향을 받지 않는다.

$$P(\text{"}H\text{"} \mid \text{"}1\text{"}) = P(\text{"}H\text{"})$$

독립사건

Theorem (독립사건)

사건 A와 사건 B가 독립이면

$$P(A \cap B) = P(A)P(B),$$
$$P(A|B) = P(A), (P(B) > 0 \text{인 경우})$$
$$P(B|A) = P(B), (P(A) > 0 \text{인 경우})$$

이 성립한다.

- 사건 A와 사건 B는 서로독립(mutually independent)이라 한다.
- 독립이 아닌 두 사건을 서로종속(mutually dependent)이라고 한다.

- 상호배반과 독립은 다른 개념
  - 상호배반: $P(A \cap B) = 0$
  - 독립 : $P(A \cap B) = P(A)P(B)$

독립사건

Example (3.6)

Example 3.4에서 두 사건 A와 B는 서로 독립인가? 또한 단순랜덤복원추출할 경우, A와 B는 서로 독립인가?

sol) 전확률공식을 이용하면

$$\begin{aligned} P(B) &= P(B|A)P(A) + P(B|A^c)P(A^c) \\ &= \frac{19}{99} \times \frac{20}{100} + \frac{20}{99} \times \frac{80}{100} = \frac{1}{5}. \end{aligned}$$

따라서 $P(B) \neq P(B|A)$이므로 서로 독립이 아니다.

- 단순랜덤복원추출의 경우에는

$$P(B|A) = \frac{20}{100} = \frac{1}{5}$$

이므로 $P(B) = P(B|A)$가 되어 서로 독립이다.

확률변수

- 표본공간을 수의 집합에 대응시켜 생각해보자.

Example

동전 1개를 던져 나오는 결과를 관측하는 실험이 있다고 하자.

- 표본공간은 $S = \{'H', 'T'\}$
- 확률은 $P('H') = P('T') = \frac{1}{2}$
- 표본공간 S에서 다음과 같이 함수 X를 정의해보자.

$$X('H') = 1,\ X('T') = 0$$

- 확률을 대응시키면,

$$P(X = 1) = \frac{1}{2},\ P(X = 0) = \frac{1}{2}$$

- 표본공간 S와 그 위에서 주어진 확률이 함수 X에 의해 수의 집합 $\{0, 1\}$과 그 위에서 주어진 확률에 대응되는 것.

확률변수

- 확률변수(random variable): 표본공간에서 정의된 실수값 함수

  - 확률변수 -> X, Y 등 대문자로 표현
  - 확률변수가 취하는 값 -> x,y 등 소문자로 표현

$$P(X = x)$$

- 확률변수의 의미: 모집단에서 하나의 특성값을 관측할때, 그 관측값은 어떤것을 관측하느냐에 따라 값이 변하므로 확률변수임.

확률분포

- 확률분포(probability distribution): 확률변수 X의 값에 따라 확률이 어떻게 흩어져 있는지를 합이 1인 양수로 나타낸 것을 X의 확률분포라 한다.
- 동전 1개를 던지는 실험에서 확률변수 X의 확률분포

표 3-1 X의 확률분포표

x	0	1
$P(X=x)$	$\frac{1}{2}$	$\frac{1}{2}$

이산확률변수와 확률질량함수

- 이산확률변수(random variable of discrete type): 취할 수 있는 모든 값이 $x_1, x_2, \cdots$로 셀 수 있는(countable) 값을 갖는 확률변수
- 확률변수 X가 x의 값을 취할 확률은 다음과 같다.

$$p(x) = \begin{cases} P(X = x_i), & when\ x = x_i (i = 1, 2, \cdots) \\ 0, & when\ x \neq x_i \end{cases} \tag{1}$$

- 이때, p(x)를 확률변수 X의 확률질량함수(probability mass function, p.m.f)라고 한다.
- 확률질량함수의 성질:

1. $0 \leq p(x) \leq 1,\ \sum_{all\ x} p(x) = 1$
2. $P(a < X \leq b) = \sum_{a<x\leq b} p(x)$

이산확률변수와 확률질량함수

Example

어느 인플루언서가 올린 게시물 12개 중 4개가 '좋아요' 1만 개를 넘었다고 하자. 게시물 중 3개를 단순랜덤추출했을 때, '좋아요'가 1만 개를 넘는 게시물의 수를 X라고 하자. 확률변수 X의 확률분포를 구하라.

sol) 게시물 12개 중 3개를 택하는 방법의 수는 $\binom{12}{3} = 220$이므로,

$$p(0) = P(X=0) = \binom{8}{3}\binom{4}{0}/\binom{12}{3} = \frac{56\times 1}{220} = \frac{14}{55}$$

$$p(1) = P(X=1) = \binom{8}{2}\binom{4}{1}/\binom{12}{3} = \frac{28\times 4}{220} = \frac{28}{55}$$

$$p(2) = P(X=2) = \binom{8}{1}\binom{2}{2}/\binom{12}{3} = \frac{8\times 6}{220} = \frac{12}{55}$$

$$p(3) = P(X=3) = \binom{8}{0}\binom{4}{3}/\binom{12}{3} = \frac{1\times 4}{220} = \frac{1}{55}$$

이를 정리하면 다음과 같은 X의 확률분포표를 얻는다.

x	0	1	2	3
p(x)	$\frac{14}{55}$	$\frac{28}{55}$	$\frac{12}{55}$	$\frac{1}{55}$

연속확률변수와 확률밀도함수

- 연속확률변수(random variable of continuous type): 취할 수 있는 모든 값이 하나씩 셀 수 없는 경우(uncountable), 즉 어떤 구간 내의 모든 값을 취할 수 있는 확률변수

- 연속확률변수의 성질
  1. 임의의 c에 대해 P(X=c)=0
  2. $P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b)$

- 연속확률변수 X에 대하여 함수 f(x)가 다음을 만족시킬 때,
  1. $f(x) \geq 0$, $\int_{-\infty}^{\infty} f(x)dx = 1$
  2. $P(a \leq X \leq b) = \int_a^b f(x)dx$ (단, $-\infty \leq a < b \leq \infty$)

  => f(x)를 X의 확률밀도함수(probability density function, p.d.f)라고 한다.

연속확률변수와 확률밀도함수

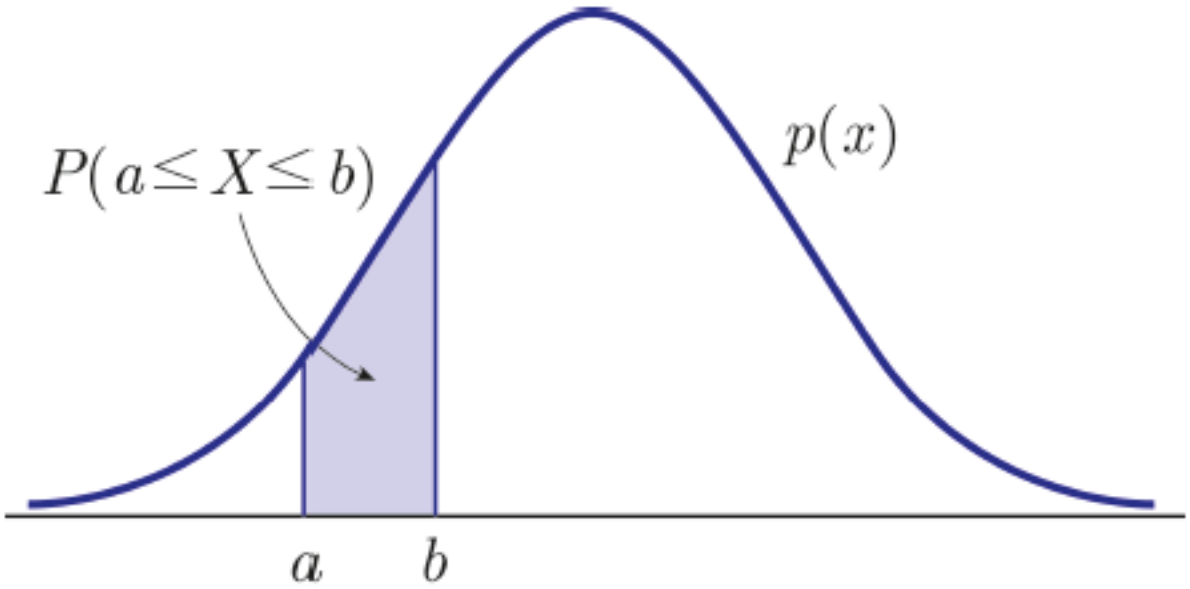


그림 3–5 연속확률변수의 확률밀도함수와 확률

- $P(a \le X \le b)$의 확률은:
  f(x)를 구간 [a,b]에서 적분한 것(빗금친 부분의 넓이)

연속확률변수와 확률밀도함수

Example

아래 그림과 같이 숫자판 중심에 한쪽 끝이 고정된 화살표가 자유롭게 회전할 수 있다고 하자. 이 화살표를 회전시켰을 때, 화살표가 저절로 멈추면서 가리킨 눈금을 X라고 하자. X는 0에서 1까지의 모든 값을 취할 수 있는 연속확률변수이다.

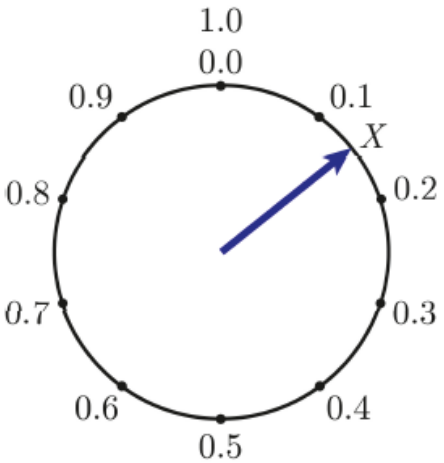


그림 3–3 연속확률변수의 예

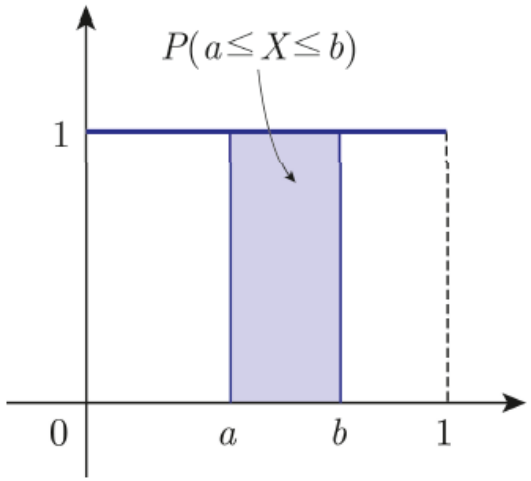


그림 3–4 확률밀도함수의 예

연속확률변수와 확률밀도함수

Example

연속확률변수 X의 확률밀도함수가

$$f(x) = \begin{cases} bx(1-x), & (when\ 0 \le x \le 1) \\ 0, & (when\ x < 0\ or\ x > 1) \end{cases}$$

로 주어질 때, 상수 b와 확률 $P(0 \le X \le 2/3)$를 구하여라.

sol) 확률밀도함수의 정의로부터 $\int_0^1 f(x)dx = 1$이여야 하므로,

$$b\int_0^1 x(1-x)dx = b\left[-\frac{1}{3}x^3 + \frac{1}{2}x^2\right]_0^1 = b\left[-\frac{1}{3} + \frac{1}{2}\right] = \frac{b}{6} = 1$$

따라서 b=6.

- $P(0 \le X \le 2/3) = \int_0^{2/3} 6x(1-x)dx = 20/27.$

확률분포의 대표값

- 중심위치 척도
  - 확률변수의 평균 또는 기대값
- 변동성(산포) 척도
  - 표준편차, 분산
- 연관성 척도
  - 공분산, 상관계수

- 이산확률변수 X와 Y가 각각 취하는 값이 $x_1, x_2, \cdots$와 $y_1, y_2, \cdots$이고 확률질량함수는 $p(x)$, $p(y)$라고 가정하자.
- 연속확률변수 X, Y에 대하여도 확률밀도함수를 각각 $f(x)$, $f(y)$라고 가정하자.

기대값(expected value)

- 확률변수 X의 평균(mean) 또는 기대값(expected value)은 다음과 같다.

$$E(X) = \mu = \begin{cases} \sum_{all\ x} xp(x), & (discrete\ case) \\ \int_{-\infty}^{\infty} xf(x)dx, & (continuous\ case) \end{cases}$$

- 확률의 가중값으로 계산되는 모든 가능한 결과들의 평균값.

Example

확률변수 X의 확률분포가 다음 표와 같을 때 X의 평균을 구하여라.

x	1	2	3	4
$p(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

sol) $E(X) = \sum_{x=1}^{4} xp(x) = 1 \times \frac{1}{8} + 2 \times \frac{3}{8} + 3 \times \frac{3}{8} + 4 \times \frac{1}{8} = 2.5$

기대값(expected value)

- 앞의 예제에서 평균 2.5는 X의 확률분포의 그래프에서 중심과 같은 의미를 가진다.

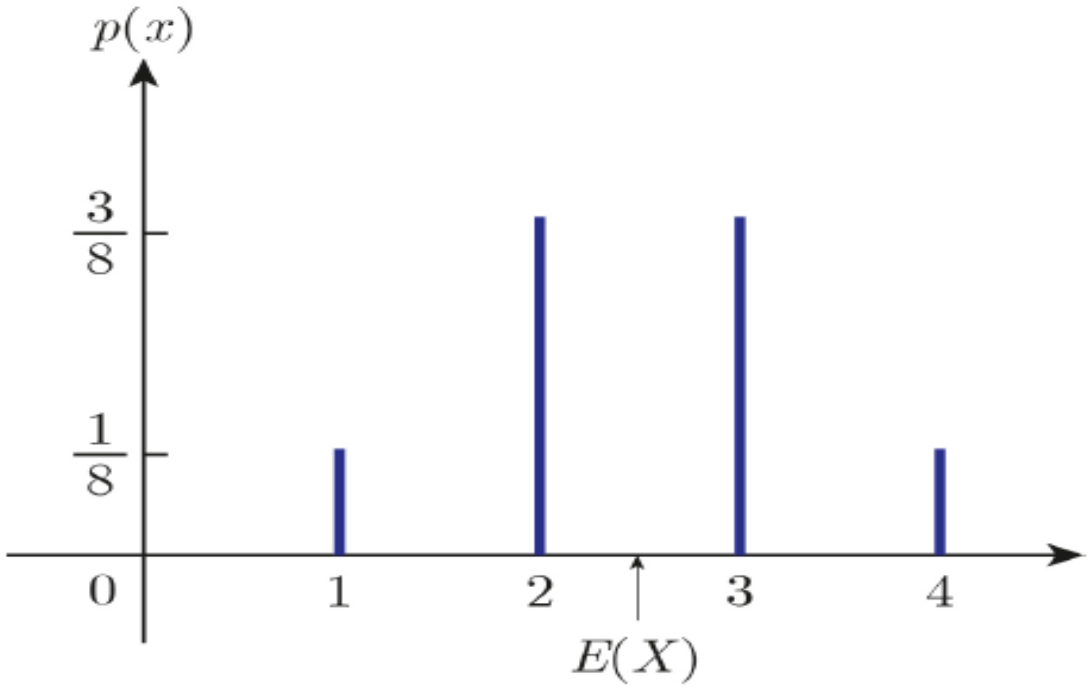


그림 3–6 평균의 역학적 의미

기대값(expected value)

Example

동전을 3회 던지는 실험에서 앞면의 개수를 X라고 할 때, 확률변수 X의 확률분포를 구하고, X의 기대값을 구하여라.

- sol) X의 확률질량함수, $p_X(x)$는 다음 표와 같다.

Table: X의 확률분포

x	0	1	2	3
$p_X(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

- 따라서 X의 기대값은

$$E(X) = \sum_{x=0}^{3} x p_X(x) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{3}{2}$$

기대값(expected value)

- 확률변수의 함수, $g(X)$도 확률 분포를 가지는 확률변수이다.
- $g(X)$의 기대값:

$$E[g(X)] = \begin{cases} \sum_{all\ x} g(x)p(x), & (discrete\ case) \\ \int_{-\infty}^{\infty} g(x)f(x)dx, & (continuous\ case) \end{cases}$$

- 확률변수 X의 함수, $g_1(X)$, $g_2(X)$에 대하여 다음이 성립한다. (c_1, c_2는 상수)

  1. $E[c_1 g_1(X)] = c_1 E[g_1(X)]$
  2. $E[c_1 g_1(X) + c_2 g_2(X)] = c_1 E[g_1(X)] + c_2 E[g_2(X)]$

기대값(expected value)

Example

위 예제에서(동전을 3회 던지는 실험) 확률변수 $Y = X(3 - X)$의 확률분포를 구하고, Y의 기대값을 구하여라.

sol) Y의 확률질량함수, $p_Y(y)$는 다음 표와 같이 주어진다.

Table: $Y = X(3 - X)$의 확률분포

y	0	2
$p_Y(y)$	$\frac{1}{4}$	$\frac{3}{4}$

따라서 $E(Y) = \sum y p_Y(y) = 0 \times \frac{1}{4} + 2 \times \frac{3}{4} = \frac{3}{2}$

- 또는 다음과 같이 직접 계산도 가능

$$E(Y) = E[X(3 - X)] = \sum_{x=0}^{3} X(3 - X) p_X(x)$$
$$= 0(3 - 0)p_X(0) + \cdots + 3(3 - 3)p_X(3) = \frac{3}{2}$$

기대값(expected value)

Example

어느 버스 정류장에서 매시 0분, 15분, 35분에 버스가 각각 1회씩 출발한다. 한 사람이 우연히 이 정거장에 도착했을 때, 버스 출발까지 기다려야 하는 시간의 기댓값을 구하라.

sol) 정류장에 도착한 시각을 X분이라 하면, X의 확률밀도함수는

$$f(x) = \frac{1}{60}, \ 0 \leq x \leq 60$$

- 이 사람이 x분에 왔을 때 기다리는 시간 g(x)는

$$g(x) = \begin{cases} 15 - x, & 0 < x \leq 15 \\ 35 - x, & 15 < x \leq 35 \\ 60 - x, & 35 < x \leq 60 \end{cases} \tag{2}$$

- 따라서 g(X)의 기대값, E(g(X))은

$$\begin{aligned}
&\int g(x)f(x)dx \\
&\quad = \int_0^{15}(15-x)\frac{1}{60}dx + \int_{15}^{35}(35-x)\frac{1}{60}dx + \int_{35}^{60}(60-x)\frac{1}{60}dx \\
&\quad = \frac{125}{10}
\end{aligned}$$

- 즉, 기다리는 시간의 기대값은 10분 25초이다.

분산과 표준편차

- 확률변수의 분산(variance)과 표준편차(standard deviation):
- 산포를 나타내는 대표값
- X의 평균이 μ이고 확률질량/밀도함수가 p(x), f(x)일 때,

$$(1)\ Var(X) = E[(X-\mu)^2] = \begin{cases} \sum_{all\ x}(x-\mu)^2 p(x), & (discrete\ r.v) \\ \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx, & (continuous\ r.v) \end{cases}$$

$$(2)\ sd(X) = \sqrt{Var(X)} = \sqrt{E[(X-\mu)^2]}$$

- Var(X)=σ^2, sd(X)=σ

분산과 표준편차

- 분산의 간편계산법: $Var(X) = E(X^2) - [E(X)]^2$

$$
\begin{aligned}
pf)\ Var(X) &= E[(X-\mu)^2] \\
&= E(X^2 - 2\mu X + \mu^2) \\
&= E(X^2) - 2\mu E(X) + \mu^2 \\
&= E(X^2) - \mu^2
\end{aligned}
$$

분산과 표준편차

Example

동전을 2회 던질 때 나오는 표면의 개수 X의 평균과 표준편차를 구하여라.

sol) X의 확률분포는 다음 표와 같다.

x	0	1	2
$p_X(x)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

- 따라서 X의 기대값과 표준편차는

$$E(X) = \sum_{x=0}^{2} x p_X(x) = 0 \times \frac{1}{4} + 1 \times \frac{2}{4} + 2 \times \frac{1}{4} = 1$$

$$E(X^2) = \sum_{x=0}^{2} x^2 p_X(x) = 0^2 \times \frac{1}{4} + 1^2 \times \frac{2}{4} + 2^2 \times \frac{1}{4} = \frac{3}{2}$$

$$Var(X) = \frac{3}{2} - 1^2, \; sd(X) = \sqrt{Var(X)} = 1/\sqrt{2}$$

기대값과 분산의 성질

- *for any* $a, b \in \mathbb{R}$,
  aX+b의 기대값과 분산은 다음과 같다.

- E(aX+b)=aE(X)+b

  pf) $E(aX+b) = \sum(ax+b)p(x) = a\sum xp(x) + b\sum p(x) = a\mu + b$

- $Var(aX+b) = a^2 Var(X)$, $sd(aX+b) = |a|sd(X)$

$$
\begin{aligned}
pf)\ Var(aX+b) &= E[(aX+b-E(aX+b))^2] \\
&= E(aX+b-a\mu-b)^2 \\
&= a^2 E(X-\mu)^2 \\
&= a^2 Var(X)
\end{aligned}
$$

확률변수의 표준화(standardization)

- 표준화: 확률변수 X에 대하여 그 평균 μ를 뺀 다음 표준편차로 나누는 것.

  - 확률변수 X를 표준화시킨 확률변수 Z는 다음과 같다.

$$Z = \frac{X - E(X)}{sd(X)}$$

  - 평균이나 분산에 의존하지 않고도 데이터의 상대적인 위치 관계를 알 수 있는 지표이다.

- 이때 Z의 평균은 0, 분산은 1이다.

  - $E(Z) = E\left(\frac{X-E(X)}{sd(X)}\right) = E\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma}E(X) - \frac{\mu}{\sigma} = 0$

  - $Var(Z) = Var\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma^2}Var(X) = 1$

두 확률변수의 결합분포

- 두 개 이상의 확률변수를 동시에 고려할 때 이들 사이의 확률적 관계를 나타내는 방법에 대해 학습해 보자.

예) - 아버지의 키와 아들의 키

- 환자의 건강기록을 구성하는 변수로 나이, 혈압, 몸무게 등 고려

Table: 사건 A와 B가 발생할 확률

	B_1	B_2	total
A_1	$P(A_1 \cap B_1)$		$P(A_1)$
A_2		$P(A_2 \cap B_2)$	
	$P(B_1)$		1

- 결합확률(joint probability): 교집합(intersection)의 확률
- 주변확률(marginal probability): 어느 한쪽만의 변수를 생각할 때의 확률

두 확률변수의 결합분포

- 결합확률분포(joint probability distribution):

  두 확률변수 X와 Y에 대해, 이들이 취할 수 있는 값들의 모든 순서쌍에 확률이 어떻게 분포되어 있는지를 합이 1인 양수로 나타낸 것.

Table: 동전(X)와 주사위(Y)를 동시에 던질 때의 예

	1	2	3	4	5	6	합계
앞면	1/12	1/12	1/12	1/12	1/12	1/12	1/2
뒷면	1/12	1/12	1/12	1/12	1/12	1/12	1/2
	1/6	1/6	1/6	1/6	1/6	1/6	1

- P(X=1, Y=2)=
- P(Y=1)=

두 이산형 확률변수의 결합분포

- 이산확률변수 (X,Y)가 순서짝 (x,y)를 취할 확률은 다음과 같고, 이때 $p(x,y)$를 X, Y의 결합확률질량함수(joint p.m.f)라고 한다.

$$p(x,y) = P(X = x_i, Y = y_j), \ for \ (x,y) = (x_i, y_j), (i,j = 1,2,\cdots)$$

- 주변확률질량함수(marginal p.m.f)

  다음 식은 각각 X와 Y의 주변확률질량함수이고, X, Y의 분포를 결정해 준다.

$$p_1(x) = \sum_{all \ y}^{\infty} p(x,y), \quad p_2(y) = \sum_{all \ x}^{\infty} p(x,y)$$

- 결합확률질량함수의 성질:

  (1) $0 \le p(x,y) \le 1, \ \sum_{all \ x} \sum_{all \ y} p(x,y) = 1$

  (2) $P(a < X \le b, \ c < Y \le d) = \sum_{a<X\le b} \sum_{c<Y\le d} p(x,y)$

두 연속형 확률변수의 결합분포

- 연속형확률변수 X,Y의 결합확률밀도함수(joint p.d.f)는 어떤 2차원 집합 A에 대하여 다음을 만족한다.

$$P\{(X,Y) \in A\} = \int \int_A f(x,y)dxdy$$

- X와 Y 각각의 분포를 나타내는 주변확률밀도함수(marginal p.d.f)는 다음과 같다.

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)dy, \quad -\infty < x < \infty$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)dx, \quad -\infty < y < \infty$$

- 결합확률밀도함수의 성질:
  (1) $0 \leq f(x,y)$, for all x and y, $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$
  (2) A가 $(x,y) : a \leq x \leq b,\ c \leq y \leq d$라면,
  $P\{(X,Y) \in A\} = \int_c^d \int_a^b f(x,y)dxdy$

두 확률변수의 결합분포 예제

Example

서로 다른 동전 A, B, C를 동시에 던지는 실험에서 확률변수 X, Y, Z를 다음과 같이 정의하자.

$$X = \begin{cases} 1, & \textit{when } A \textit{ is Heads} \\ 0, & \textit{otherwise} \end{cases}$$

$$Y = \begin{cases} 1, & \textit{when } A, B \textit{ are Heads} \\ 0, & \textit{otherwise} \end{cases}$$

$$Z = \begin{cases} 1, & \textit{when } B, C \textit{ are Heads} \\ 0, & \textit{otherwise} \end{cases}$$

X와 Y, X와 Z의 결합확률분포를 구하여라.

sol) 표본공간과 그 원소에 대응하는 확률변수 X, Y, Z의 값들은 다음과 같다.

표본공간	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	1	0	0	0	1	0	0	0

- 위의 표로부터 X와 Y, X와 Z의 결합확률분포는 다음과 같다.

(1) X와 Y

	x=0	x=1	Sum
y=0	2/4	1/4	3/4
y=1	0	1/4	1/4
Sum	2/4	2/4	1

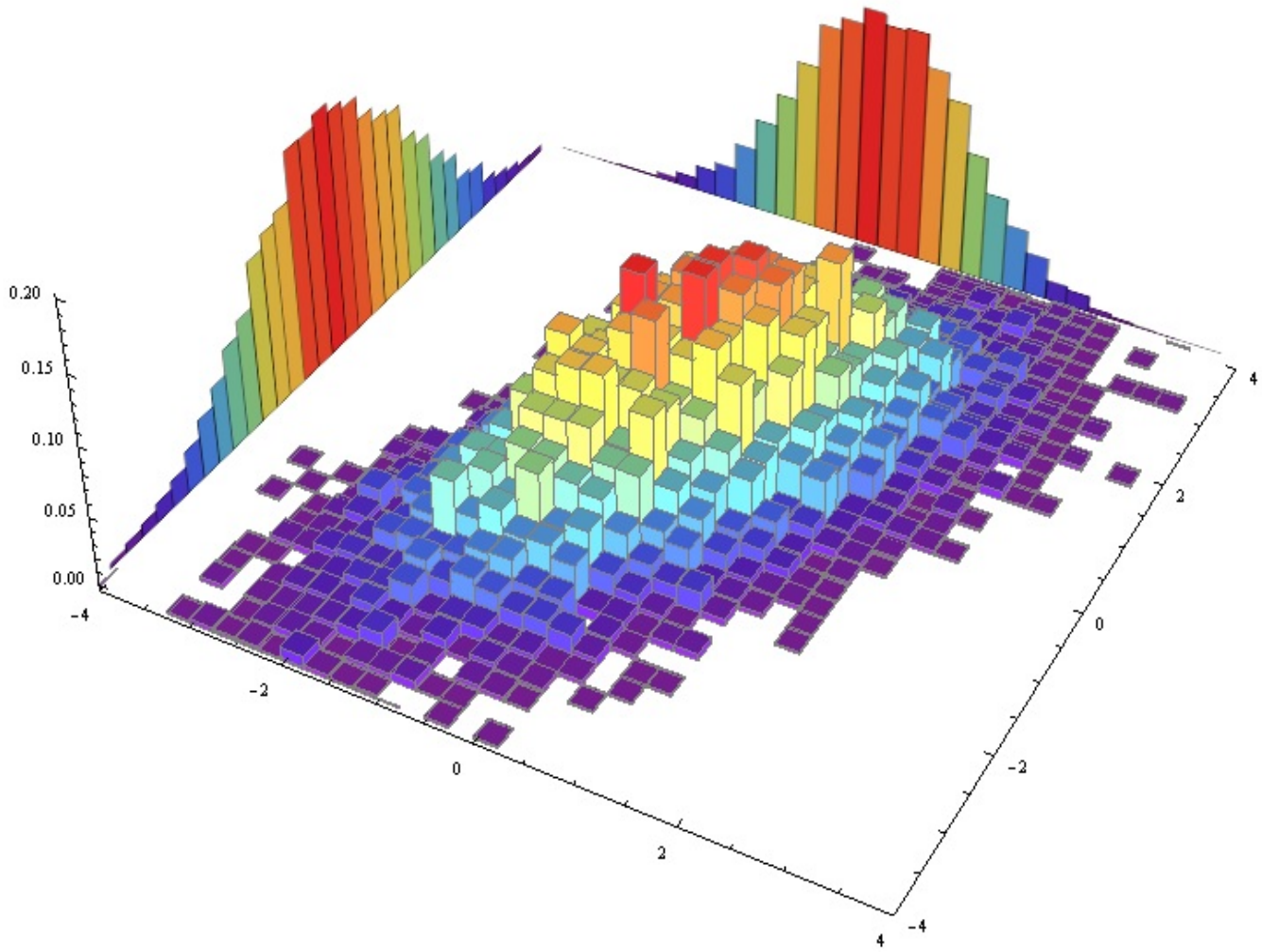
(2) X와 Z

	x=0	x=1	Sum
z=0	3/8	3/8	6/8
z=1	1/8	1/8	2/8
Sum	4/8	4/8	1

- 각 표에서 중간부분은 결합확률질량함수
열의 합, 행의 합은 주변확률질량함수를 나타낸다.

두 확률변수의 결합분포

- Joint probability distribution of two random variables.



두 확률변수의 결합분포

- 두 확률변수의 함수도 자신의 확률분포를 가지는 확률변수이다.
- 두 확률변수의 함수, $g(X, Y)$의 기대값:

$$E[g(X,Y)] = \begin{cases} \sum_{all\ x}\sum_{all\ y} g(x,y)p(x,y), & (discrete\ case) \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy, & (continuous\ case) \end{cases}$$

- X와 Y의 함수 g_1, g_2에 대하여 다음의 사실들이 성립한다.
  (c_1, c_2는 상수)
  $E[c_1 g_1(X,Y) + c_2 g_2(X,Y)] = c_1 E[g_1(X,Y)] + c_2 E[g_2(X,Y)]$
- E(X+Y)=E(X)+E(Y)

두 확률변수의 결합분포

Example

하나의 주사위를 던져서 나오는 눈의 종류에 따라 상금이 걸린 게임이 있다. A, B 두 사람이 각각 다음과 같은 게임을 택했다고 하자.

A: 1 또는 2 가 나오면 100원
3 또는 4 가 나오면 200원
5 또는 6 이 나오면 300원
B: 짝수가 나오면 100원
홀수가 나오면 (눈의 수X100)원

이 경우에 A와 B의 수입의 합 Z에 대한 분포를 구하여라.

sol) A의 수입과 B의 수입을 각각 X와 Y로 나타내면, X와 Y의 결합확률분포는 다음과 같다.

Table: X와 Y의 결합확률분포

x	y 100	300	500	행의 합
100	2/6	0	0	2/6
200	1/6	1/6	0	2/6
300	1/6	0	1/6	2/6
열의 합	4/6	1/6	1/6	1

- $P(X = 100, Y = 100) = P(\text{"1" } or \text{ "2"}) = 2/6$
- $E(X) = 100 \times 2/6 + \cdots = 200$, $E(Y) = 100 \times 4/6 + \cdots = 200$

- $Z = X + Y$라 하면 Z의 확률질량함수 $p(z)$는 다음과 같다.

Table: $Z = X + Y$의 확률분포

z	200	300	400	500	800	합계
$p(z)$	2/6	1/6	1/6	1/6	1/6	1

- $E(Z) = 200 \times 2/6 + \cdots = 400 = E(X) + E(Y)$

독립성

- 두 확률변수 X와 Y가 서로 독립이라는 것은 모든 x와 y에 대하여 $\{X = x\}$인 사건과 $\{Y = y\}$인 사건이 서로 독립(mutually independent)임을 뜻한다.

Theorem

두 이산형 확률변수 X와 Y의 결합확률함수를 $p(x, y)$라 하고, X와 Y의 주변확률함수를 각각 $p_1(x)$, $p_2(y)$라 하자. 이 때,

$$p(x, y) = p_1(x)p_2(y)$$

가 성립할 때, X와 Y는 서로 독립이라고 한다.

- 두 연속형 확률변수 X와 Y의 결합확률질량함수를 $f(x, y)$, X와 Y의 주변확률질량함수를 각각 $f_1(x)$, $f_2(y)$라 하면 다음이 성립할 때 서로 독립이라고 한다. $f(x, y) = f_1(x)f_2(y)$

독립성 예제

Example

하나의 동전을 세 번 던질 때, X=처음 두 번에서 나오는 앞면의 개수, Y=세 번째에 나오는 앞면의 개수 라고 하면 X와 Y는 서로 독립인가?

sol) 먼저 X와 Y의 결합확률분포를 구해보자.

	y=0	y=1	행의 합
x=0	1/8	1/8	1/4
x=1	2/8	2/8	1/2
x=2	1/8	1/8	1/4
열의 합	1/2	1/2	1

- 위의 표에서 모든 (x,y)에 대해서 $p(x,y) = p_1(x)p_2(y)$가 성립함을 알 수 있다. 따라서 X와 Y는 서로 독립이다.

$$p_1(0)p_2(0) = \frac{1}{4} \times \frac{1}{2} = \frac{1}{8} = p(0,0), \cdots, p_1(2)p_2(1) = \frac{1}{4} \times \frac{1}{2} = p(2,1)$$

공분산과 상관계수

- 공분산(covariance): 두 확률변수 X와 Y의 연관관계의 척도

$$\mathrm{Cov}(X, Y) = E[(X - \mu_1)(Y - \mu_2)]$$

- 공분산의 간편계산법: $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$

$$\begin{aligned} pf)\ \mathrm{Cov}(X, Y) &= E[(X - \mu_1)(Y - \mu_2)] \\ &= E(XY - \mu_2 X - \mu_1 Y + \mu_1\mu_2) \\ &= E(XY) - \mu_2 E(X) - \mu_1 E(Y) + \mu_1\mu_2 \\ &= E(XY) - \mu_1\mu_2. \end{aligned}$$

- Var(X)=Cov(X,X)

공분산과 상관계수

- 상관계수(correlation coefficient): 두 확률변수의 직선관계가 얼마나 강하고 어떤 방향인지를 나타냄 (방향과 정도)
- 확률변수 X, Y를 표준화한 것의 곱의 기대값

  (or) 공분산을 σ_1, σ_2로 나눈 것

$$\mathrm{Corr}(X, Y) = E\left[\left(\frac{X-\mu_1}{\sigma_1}\right)\left(\frac{Y-\mu_2}{\sigma_2}\right)\right] = \frac{\mathrm{Cov}(X, Y)}{\mathrm{sd}(X)\mathrm{sd}(Y)}$$

- $-1 \leq \mathrm{Corr}(X.Y) \leq 1$
- X와 Y가 완전한 직선 관계를 가지는 경우: $Y = aX + b$,

  $(a \neq 0, b$는 상수$)$

$$\begin{cases} \mathrm{Corr}(X, Y) = 1, & (a > 0) \\ \mathrm{Corr}(X, Y) = -1, & (a < 0) \end{cases} \tag{3}$$

공분산과 상관계수

Example

X와 Y의 결합분포가 다음과 같이 주어질 때, X와 Y의 공분산과 상관계수를 구하여라.

표 3-14 X와 Y의 결합확률분포표

x \ y	0	1	2	3	행의 합
0	0.05	0.05	0.10	0.00	0.20
1	0.05	0.10	0.25	0.10	0.50
2	0.00	0.15	0.10	0.05	0.30
열의 합	0.10	0.30	0.45	0.15	1.00

sol) X와 Y의 평균과 분산을 먼저 구해보자.

$$\mu_1 = E(X) = \sum x p_1(x) = 0 \times 0.2 + 1 \times 0.5 + 2 \times 0.3 = 1.1$$

$$\mu_2 = E(Y) = \sum y p_2(y) = 0 \times 0.1 + 1 \times 0.3 + 2 \times 0.45 + 3 \times 0.15 = 1.65$$

$E(X^2) = \sum x^2 p_1(x) = 0^2 \times 0.2 + 1^2 \times 0.5 + 2^2 \times 0.3 = 1.7$

$E(Y^2) = \sum y^2 p_2(y) = 0^2 \times 0.1 + 1^2 \times 0.3 + 2^2 \times 0.45 + 3^2 \times 0.15 = 3.45$

$\sigma_1^2 = 1.7 - 1.1^2 = 0.49, \quad \sigma = 0.7$

$\sigma_2^2 = 3.45 - 1.65^2 = 0.7275, \quad \sigma = 0.85$

- 공분산과 상관계수를 구해보자.

Table: E(XY)의 계산

	y=0	y=1	y=2	y=3
x=0	0.05	0.05	0.1	0
	(0)	(0)	(0)	(0)
x=1	0.05	0.1	0.25	0.1
	(0)	(1)	(2)	(3)
x=2	0	0.15	0.1	0.05
	(0)	(2)	(4)	(6)

- $E(XY) = \sum xyp(x, y) = 0 \times 0.05 + \cdots + 6 \times 0.05 = 1.9$
- $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = 1.9 - 1.65 \times 1.1 = 0.085$
- $\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_1 \sigma_2} = \frac{0.085}{0.7 \times 0.85} = 0.14$

공분산과 상관계수의 성질

- $\mathrm{Cov}(aX+b,\ cY+d) = ac\mathrm{Cov}(X,Y)$

$$
\begin{aligned}
pf)\ \ \mathrm{Cov}(aX+b,\ cY+d) &= E[(aX+b-a\mu_1-b)(cY+d-c\mu_2-d)] \\
&= E[ac(X-\mu_1)(Y-\mu_2)] \\
&= ac\mathrm{Cov}(X,Y).
\end{aligned}
$$

- $\mathrm{Corr}(aX+b,\ cY+d) =$

$$
\begin{aligned}
pf)\ \ \mathrm{Corr}(aX+b,\ cY+d) &= \frac{\mathrm{Cov}(aX+b,\ cY+d)}{\sqrt{\mathrm{Var}(aX+b)}\sqrt{\mathrm{Var}(cY+d)}} \\
&= \frac{ac\mathrm{Cov}(X,\ Y)}{\sqrt{a^2\mathrm{Var}(X)}\sqrt{c^2\mathrm{Var}(Y)}} \\
&= \frac{ac}{|ac|}\mathrm{Corr}(X,Y).
\end{aligned}
$$

확률변수의 결합분포

- 확률변수 X와 Y의 선형결합: $aX \pm bY$

$$E(aX \pm bY) = aE(X) \pm bE(Y)$$

$$Var(aX \pm bY) = a^2 Var(X) + b^2 Var(Y) \pm 2abCov(X, Y)$$

- 일반적으로, 확률변수 $X_1, \cdots, X_n$와 상수 $c_1, \cdots, c_n$에 대하여,

$$E\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} [c_i \cdot E(X_i)]$$

$$Var\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$$

확률변수의 결합분포

- Var(X+Y)=Var(X)+Var(Y)+2Cov(X,Y)

$$
\begin{aligned}
pf) \quad Var(X+Y) &= E[X+Y-E(X+Y)]^2 \\
&= E[X+Y-E(X)-E(Y)]^2 \\
&= E[(X-E(X))+(Y-E(Y))]^2 \\
&= E[(X-E(X))]^2 + E[(Y-E(Y))]^2 \\
&\qquad + 2E[(X-E(X))(Y-E(Y))] \\
&= Var(X) + Var(Y) + 2\mathrm{Cov}(X,Y).
\end{aligned}
$$

상관계수 범위 증명

- $-1 \leq \mathrm{Corr}(X.Y) \leq 1$

  pf) $\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right)$의 분산을 고려해 보자.

$$
\begin{aligned}
Var\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right) &= Var\left(\frac{X}{\sigma_X}\right) + Var\left(\frac{Y}{\sigma_Y}\right) \pm 2\mathrm{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \\
&= \frac{1}{\sigma_X^2} Var(X) + \frac{1}{\sigma_Y^2} Var(Y) \pm 2\frac{1}{\sigma_X \sigma_Y}\mathrm{Cov}(X, Y) \\
&= \frac{1}{\sigma_X^2}\sigma_X^2 + \frac{1}{\sigma_Y^2}\sigma_Y^2 \pm 2\mathrm{Corr}(X, Y)
\end{aligned}
$$

  이때, $0 \leq 1 + 1 \pm 2\mathrm{Corr}(X, Y)$ 이므로 다음이 성립한다.

$$-1 \leq \mathrm{Corr}(X.Y) \leq 1$$

두 확률변수의 독립성

- 두 확률변수가 서로 독립인 경우:
  (1) E(XY)=E(X)E(Y)

$$
\begin{aligned}
pf) \quad E(XY) &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} x_i y_j p(x_i, y_j) \\
&= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} x_i y_j p_1(x_i) p_2(y_j) \\
&= \sum_{i=1}^{\infty} x_i p_1(x_i) \sum_{j=1}^{\infty} y_j p_2(y_j) \\
&= E(X)E(Y).
\end{aligned}
$$

  (2) Cov(X,Y)=0, Corr(X,Y)=0

$$pf) \mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0$$

  (3) $Var(X \pm Y) = Var(X) + Var(Y)$