사회통계
이산확률분포

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 이론적 확률분포
2. 베르누이 분포
3. 이항분포

이론적 확률분포

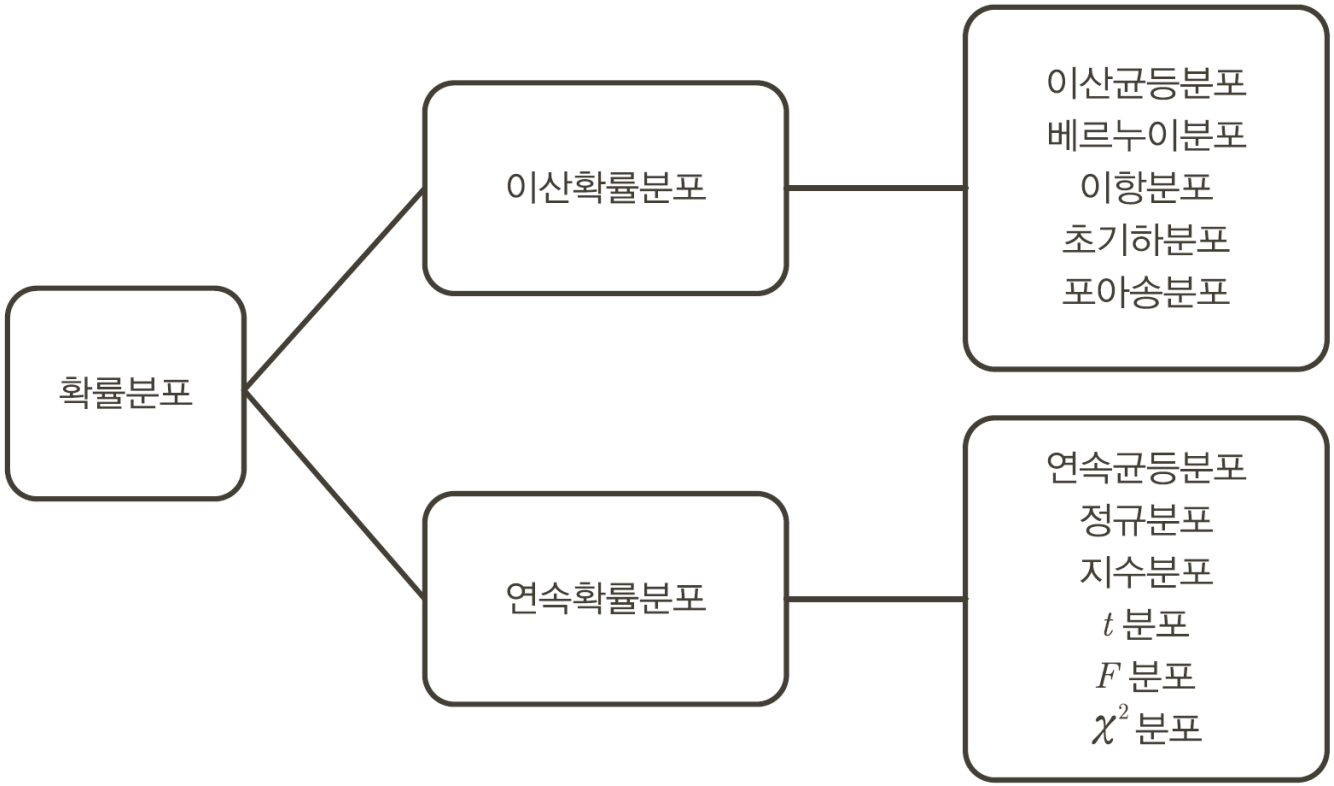
이론적 확률분포

몇몇 확률분포는 수리적으로 아름다울 뿐 아니라 문제 해결에도 유용하다.

- 이런 확률분포에 대해서는 특별히 많은 논의가 이루어졌다.
- 그러므로 이론적 확률분포(theoretical probability distribution)라는 이름 아래 좀 더 신경써서 배워야 할 필요가 있다.
- 일단 확률분포가 주어지면 그것의 평균과 분산을 계산할 수 있다. 이론적 확률분포 역시 마찬가지이다!
- 수리통계학(mathematical statistics)에서는 꽤 공들여 여러 종류의 이론적 확률분포를 공부한 다음, 각각의 평균과 분산을 도출하고 각종 수학적 테크닉을 동원해서 다양하게 증명한다.

이론적 확률분포

- 확률변수가 이산형과 연속형으로 나뉘듯, 확률분포 역시 이산확률분포(discrete probability distribution)와 연속확률분포(continuous probability distribution)로 구분된다.



이론적 확률분포

- 수업시간에서 우리는 오로지 몇 가지 이론적 확률분포만 간략히 공부한다.
- 수업에서 다룬 이론적 확률분포만 시험 범위로 다루어진다.
- 이것만으로는 실무와 연구에서 당연히 한계가 있다.
- 다루지 않은 중요한 이론적 확률분포는 교과서에 잘 설명되어 있으므로 참고하자.

베르누이 분포

베르누이 분포

동전 던지기의 결과를 수학적으로 일반화해보자.

- 동전 던지기처럼 앞면(H)과 뒷면(T)의 두 가지 결과가 예상되는 시행 내지 실험을 상상해보자.
- 어떤 실험이나 도박 등에서 성공 혹은 실패를 구분해볼 수도 있다.
- Jacob Bernoulli (1655-1700)가 처음으로 수학적 상상력을 발휘하여 이 문제를 정리하였고 이를 베르누이 과정(Bernoulli process)이라고 부른다.

베르누이 분포

- 베르누이 과정에서는 딱 한 번의 시행이 이루지고 각 시행의 사건(event) X는 두 가지 뿐이다(e.g., 성공/실패, H/T, $1/0$).
- 두 사건의 확률은 각각 $P(X=1)=p$ 와 $P(X=0)=1-p$ 로 표현할 수 있다.

$$f(X=x)=\begin{cases} p & \text{if } x=1 \\ 1-p & \text{if } x=0 \end{cases}$$

- 좀 더 축약해서 표현할 수도 있다.

$$f(x)=p^x(1-p)^{1-x}$$

- 이때 p는 성공 확률(e.g., 동전 던지기라면 0.5), X는 성공 여부(i.e., $x=1$이 성공, $x=0$이 실패)를 나타낸다.
- 이것을 베르누이 분포(Bernoulli distribution)라고 부른다.

베르누이 분포

- 위 식의 좌변(left-hand side) $f(x)$는 이것이 확률질량함수(PMF) 또는 확률밀도함수(PDF)임을 나타낸다.
- 좀 더 구체적으로 말하면, X는 이산확률변수이므로 $f(x)$ 은 확률질량함수이다.
- 참고로 (p같이) 어떤 확률질량함수 또는 확률밀도함수를 정의하기에 (확률변수 외에도) 꼭 필요한 값들을 패러미터(parameter)라고 부른다.
- $f(x)$ 대신 $f(x;p)$ 라고 패러미터를 일부러 포함해서 적기도 한다.
- 때때로 학생들은 X가 확률변수(random variable)인 반면, p는 상수(constant)라는 점을 이해하지 못하는데 꽤 중요한 기초이므로 신중히 고민해보자.

베르누이 분포

베르누이 분포 역시 평균과 분산을 갖는다.

- 확률분포가 주어지면 확률변수의 평균과 분산을 계산할 수 있다. 베르누이 분포도 그렇다.
- 베르누이 확률분포의 평균(=기대값)은 $E(X) = p$ 이다.

$$\begin{aligned} E(X) &= \sum_{\forall x} x \cdot P(X = x) \\ &= 1 \cdot p + 0 \cdot (1 - p) \\ &= p \end{aligned}$$

- (딱딱한 증명보다) 확률분포표를 그려보고 거기에서 평균을 계산해보자!

베르누이 분포

- 베르누이 확률분포의 분산은 $Var(X) = p(1-p)$ 이다.

$$\begin{aligned} Var(X) &= \sum_{\forall x} (x-\mu)^2 \cdot P(X=x) \\ &= E[(X - E(X)]^2 \\ &= E[X^2 - 2XE(X) + E(X)^2] \\ &= E(X^2) - E(X)^2 \\ &= p - p^2 \\ &= p(1-p) \end{aligned}$$

- (딱딱한 증명보다) 확률분포표를 그려보고 거기에서 분산을 계산해보자!

베르누이 분포

예제 1. 고래와 송이는 논쟁을 벌이다 주사위를 던져 누구 말이 맞는지 결정하기로 했다. 주사위에 관해서라면 자신감이 넘치는 송이는 3 이하의 값이 나오면 자기가 진 걸로 쳐도 좋다고 선언하였다. (고래의 관점에서) 이 내기의 확률분포를 표로 나타내고 평균과 분산을 계산하시오. 이 내기를 표현할 수 있는 이론적 확률분포가 무엇인지 판단하고, 그 공식을 사용하여 평균과 분산을 구하시오.

베르누이 분포

- 이 내기의 승패는 이산확률변수이므로 다음의 확률변수로 표현할 수 있다:

  "주사위에서 3 이하의 값이 나오는 사건(승리)의 확률은 $P(X \leq 3)$는 $3/6$이다."
  "주사위에서 3보다 큰 값이 나오는 사건(패배)의 확률은 $P(X > 3)$는 $3/6$이다."

x	$P(X = x)$	$x \cdot P(X = x)$	$(x - \mu)^2 \cdot P(X = x)$
$0(=$ 패배$)$	$3/6$	$3/6 \cdot 0 = 0$	$(0 - 0.5)^2 \cdot 3/6 = 0.125$
$1(=$ 승리$)$	$3/6$	$3/6 \cdot 1 = 3/6$	$(1 - 0.5)^2 \cdot 3/6 = 0.125$
	합계	0.5	0.25

- 평균은 0.5, 분산은 0.25임을 알 수 있었다.

베르누이 분포

- 이론을 생각한다는 것은 경험을 멈추는 것에서 시작한다.
- 승리와 패배의 일회성 시행 결과와 그 확률 p가 주어져 있으므로 이것은 베르누이 분포를 사용하여 묘사할 수 있다.
- 베르누이 확률분포의 평균(=기대값)은 $E(X) = p = 0.5$ 이고 분산은 $Var(X) = p(1-p) = 0.25$ 이다.
- 이론적인 계산 결과와 앞서 확률분포를 하나하나 계산해 본 결과는 동일하다.

베르누이 분포

벌써 하나의 확률분포를 공부했다.

- 확률분포에 관해 공부할 때는 가장 먼저 (1) 그것이 묘사하고 있는 상황에 관해 수학적 상상력을 습득하는 것이 중요하다!
- 그런 다음, (2) 확률질량함수 또는 확률밀도함수의 구조를 이해하고,
- 마지막으로 (3) 그것의 평균과 분산을 이해하면 된다.
- 수리통계학에서는 (1) 때문에 연습문제를 많이 풀어보도록 권장한다.

이항분포

이항분포

이항분포는 "반복적인 베르누이 과정"을 나타낸 확률분포이다.

- 베르누이 분포는 굉장히 단순하기 때문에 이를 새삼 확률분포로 나타내는 것도 웃기다!
- 그보다는 이항분포(binomial distribution)야말로 가장 대표적인 이산확률분포라고 할 수 있다.
- 앞서 설명한 베르누이 과정은 단지 1번 시행을 염두에 둔 것이였다.
- 이항분포는 베르누이 과정을 n번 독립시행(n times independent trials)하여 얻은 성공 횟수를 확률변수로 삼는다.
- 독립시행이란 n번의 시행들의 결과가 서로에게 영향을 미치지 못하는 경우이다.
- 이항분포는 성공 횟수 X가 확률변수이고 그에 대응하는 확률 p가 있는 시행(또는 실험)을 수학적으로 묘사한다.

이항분포

이항분포를 먼저 직관적으로 이해하자.

- "동전을 세 번 던져서 앞면이 한 번 나올 확률은 무엇인가?"
- 여기서 보려는 사건(event)은 다음과 같다.

$$\{H, T, T\}, \quad \{T, H, T\}, \quad \{T, T, H\}$$

- (각 사건과 결부된) 확률은 각각 다음과 같다.

$$p \cdot (1-p) \cdot (1-p), \quad (1-p) \cdot p \cdot (1-p), \quad (1-p) \cdot (1-p) \cdot p$$

- 그러면 찾고 있는 답은 다음과 같다.

$$\begin{aligned} &= 0.5 \cdot (1-0.5) \cdot (1-0.5) + \\ &\quad (1-0.5) \cdot 0.5 \cdot (1-0.5) + \\ &\quad (1-0.5) \cdot (1-0.5) \cdot 0.5 \\ &= 0.375 \end{aligned}$$

이항분포

곰곰히 생각해보면 이 계산을 일반화할 수 있다.

- 일반화하려면 (여러분이 고교 시절 배운) 경우의 수 계산이라는 수리적 기법이 필요하다.
- 그런데 특히 실무나 연구에서 "가장" 유용한 경우의 수 계산식을 딱 하나 고르라고 한다면 그건 바로 조합(combination)이다.
- 조합은 "순서 없이" 서로 다른 n개 가운데 x개를 선택하는 경우의 수이다.

$$ {}_nC_x = \binom{n}{x} = \frac{n!}{x!(n-x)!} $$

- 여기서 $x!$는 x 팩토리얼(factorial)이라고 읽으며, $x \cdot (x-1) \cdot (x-2) \cdots \cdot 1$로 계산한다.

이항분포

- 그 유용성을 확인할 수 있는 굉장히 많은 현실 사례가 있다.
- "파티에 3명이 모였다. 둘 씩 서로 악수를 한다면 모두 몇 번 소개해야 할까?"
- "$\{A, B\}$, $\{A, C\}$, $\{B, C\}$. 모두 3번이다."

$$ {}_3C_2 = \binom{3}{2} = \frac{3!}{2!(3-2)!} = 3 $$

- "어떤 수업에 38명이 모였다. 3명 씩 한 조를 짠다면 모두 몇 가지의 가능성이 있는가?"

$$ \binom{38}{3} = \frac{38!}{3!(38-3)!} = 8436 $$

- 엑셀에서는 COMBIN(n, x)를 함수로 입력하여 조합을 계산할 수 있다.

이항분포

- 동전을 한 번 던진 결과는 베르누이 분포로 나타낼 수 있다.
- 동전을 세 번 던지고, 앞면이 한 번만 나오는 "경우의 수"를 계산할 때 이항분포가 사용된다.
- 그러므로 베르누이 분포의 확률질량함수를 살짝 수정하고, 그 앞머리에 "어떤 조합"을 가져다 붙이면 이항분포의 확률질량함수를 나타낼 수 있을 것이다!

$$\binom{n}{x} \cdot p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} \cdot p^x (1-p)^{n-x}$$

- 그 앞머리를 이항계수(binomial coefficient) 또는 파스칼의 삼각형(Pascal's Triangle)이라고 부른다.
- 이 식과 베르누이 분포의 확률질량함수가 어떻게 다른지 면밀히 살펴보자.

이항분포

조합을 사용하여 아까 문제를 다시 계산해보자.

- "동전을 세 번 던져서 앞면이 한 번 나올 확률은 무엇인가?"
- 세 번 시행하여($n = 3$) 1회 성공이라는 "경우의 수"이므로($x = 1$),

$$\frac{3!}{1!(3-1)!} \cdot 0.5^1 \cdot (1-0.5)^{3-1} = 0.375$$

- 물론 이론적으로 계산한 답과 확률분포를 하나하나 그려 계산한 답은 같다.

이항분포

일단 확률분포가 주어지면 그것의 평균과 분산을 계산할 수 있다.

- 이론적 확률분포를 공부한다는 것은 (1) 그것의 확률질량함수/확률밀도함수가 어떻게 구성되고, (2) 그 확률분포의 평균이 어떻게 계산되고, (3) 그 확률분포의 분산이 어떻게 계산되는지 아는 것이다.
- 평균과 분산 계산의 증명은 응용통계학에서는 사실 그렇게까지 중요하지는 않다.
- 보다 중요한 것은 이 문장 자체를 기억하는 것이다!

이항분포

이항분포 역시 확률분포이므로 평균과 분산을 갖는다.

- 이항분포의 평균은 다음과 같다.

$$E(X) = np$$

- 엄밀한 수학적 증명은 내버려두고 개념적으로만 살펴보자!
- 베르누이 분포는 "단 한 번"의 성공 횟수를 확률로 나타낸다. 그 평균은 p이다.
- 반면 이항분포는 베르누이 과정을 "n번 독립 시행"했을 때 성공 횟수를 확률로 나타낸다.
- 그러므로 이항분포의 평균은 베르누이 분포의 평균 p에 n을 곱한 값과 같다.

이항분포

- 이항분포의 분산은 다음과 같다.

$$Var(X) = np(1-p)$$

- 여기서도 까다로운 수학적 증명은 수학도들에게 맡기고 그저 직관적으로만 살펴보자.
- 베르누이 분포는 "단 한 번"의 성공 횟수를 확률로 나타낸다. 그 분산은 $p(1-p)$이다.
- 반면 이항분포는 베르누이 과정을 "n번 독립 시행"했을 때 성공 횟수를 확률로 나타낸다.
- 그러므로 이항분포의 분산은, 개별 베르누이 시행의 분산 $p(1-p)$이 서로 얽히지 않고 독립적으로 나타나므로, 여기에 n을 곱한 값과 같다.

이항분포

베르누이 분포와 이항분포를 수학적으로 표현할 수 있다.

- 확률변수 X가 베르누이 분포를 따른다면 이렇게 표현할 수 있다(Why?).

$$X \sim \mathcal{B}(1, p)$$

- 확률변수 X가 (시행횟수가 n이고 확률이 p인) 이항분포를 따른다면 이렇게 표현한다.

$$X \sim \mathcal{B}(n, p)$$

이항분포

이항분포를 그래프로도 나타낼 수 있다.

- 아까 사례("동전을 세 번 던져서 앞면이 한 번 나올 확률은 무엇인가?")에서는 "특정 경우의 수"만 물어보았다.
- 이제는 동전을 세 번 던져 앞면이 나올 수 있는 "모든 경우의 수"를 각각 이항분포에 따라 계산할 수도 있다.

$$P(X=0) = \frac{3!}{0!(3-0)!} \cdot .5^0 \cdot (1-0.5)^{3-0} = 0.125$$

$$P(X=1) = \frac{3!}{1!(3-1)!} \cdot .5^1 \cdot (1-0.5)^{3-1} = 0.375$$

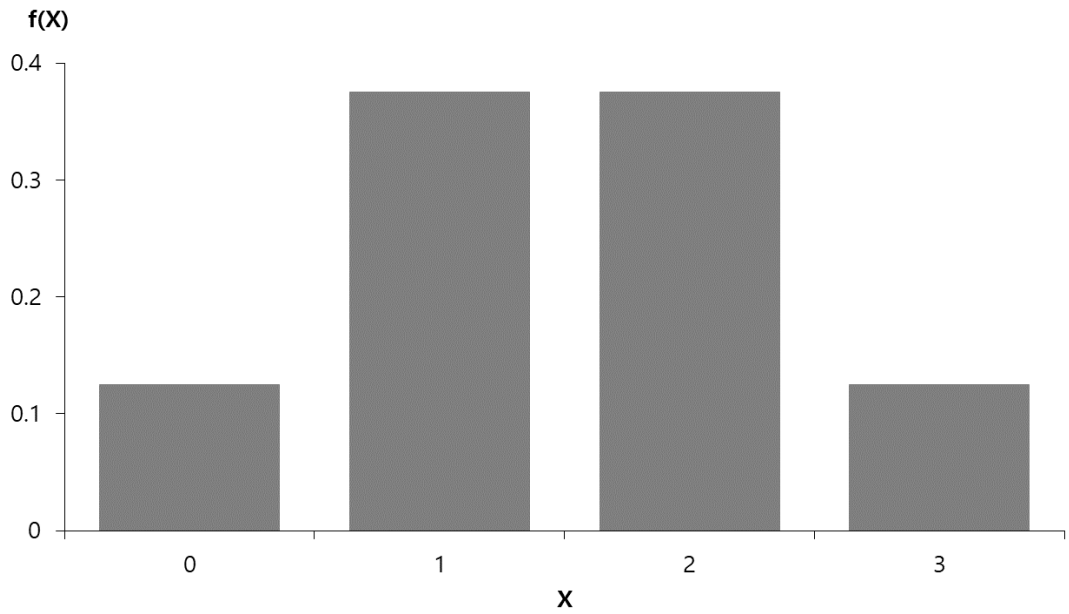
$$P(X=2) = \frac{3!}{2!(3-2)!} \cdot .5^1 \cdot (1-0.5)^{3-2} = 0.375$$

$$P(X=3) = \frac{3!}{3!(3-3)!} \cdot .5^1 \cdot (1-0.5)^{3-3} = 0.125$$

이항분포

- 이 모든 경우의 수를 다시 확률분포로 표현할 수 있다.
- x축은 확률변수 X이고, y축은 이항분포를 따르는 확률질량함수 $f(x)$인 그래프를 그릴 수 있다.

x	$P(X=x)$
0	0.125
1	0.375
2	0.375
3	0.125



이항분포

- 엑셀에서 이 그림을 연습해보자.
- 먼저 n, x, p의 열을 준비하고, 조합은 COMBIN(n, x) 함수로 쉽게 계산하자.
- 각각의 시나리오 별로 베르누이 분포를 따르는 확률질량함수를 계산하자.
- 참고로 제곱은 ^ (hat) 기호를 사용한다.

n	x	p	$f(x)$
3	0	0.5	=COMBIN(3,0)*(0.5^0)*(0.5^(3-0))
3	1	0.5	=COMBIN(3,1)*(0.5^1)*(0.5^(3-1))
3	2	0.5	=COMBIN(3,2)*(0.5^2)*(0.5^(3-2))
3	3	0.5	=COMBIN(3,3)*(0.5^3)*(0.5^(3-3))

이항분포

- 엑셀에서 이항분포를 따르는 확률질량함수를 계산할 때마다 매번 이렇게 공식을 사용해야 할까?
- 지금까지 숨겨서 미안하지만 사실 쉬운 함수가 있다. 그것은 BINOM.DIST(x, n, p, FALSE)이다.
- 다시 한 번 각각의 시나리오 별로 베르누이 분포를 따르는 확률질량함수를 계산하자.

n	x	p	$f(x)$
3	0	0.5	=BINOM.DIST(0, 3, 0.5, FALSE)
3	1	0.5	=BINOM.DIST(1, 3, 0.5, FALSE)
3	2	0.5	=BINOM.DIST(2, 3, 0.5, FALSE)
3	3	0.5	=BINOM.DIST(3, 3, 0.5, FALSE)

이항분포

예제 2. 동전을 열 번 던져서 앞면이 세 번 이하로 나올 확률을 구하시오.

이항분포

- 엑셀을 사용해 확률분포를 나타내는 표와 그래프를 그려보고 x에 따라 상이한 확률질량함수를 직접 계산해보자.
- 참고로 답은 $0.0009765625 + 0.009765625 + 0.043945313 + 0.1171875$ 이다.
- 그러나 더 쉽게 생각할 수도 있다. 바로 지난 주에 배운 누적분포함수(CDF)를 활용하는 것이다(이항분포에도 색칠공부가 있다!).
- 엑셀 함수가 BINOM.DIST(x, n, p, FALSE)라고 했는데 이때 FALSE 대신 TRUE를 입력하면 누적분포함수를 바로 계산할 수 있다.
- 다시 말해 답은 BINOM.DIST(3, 10, 0.5, TRUE) 로 더 쉽게 계산할 수 있었던 셈이다.