사회통계
확률변수와 확률분포

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

확률변수

확률변수

확률을 더 쉽게 표현하기 위해 편리한 도구를 개발하자!

- X라는 빈 주머니를 상상하자. 이 X 안에 개별 사건(event)을 집어넣을 수 있다고 하자.
- 동전 던지기에서 앞면이 나오는 사건(H)을 집어넣으면 $X = H$ 이고 $P(X = H) = 1/2$이다. 뒷면이 나오는 사건(T)도 집어넣을 수 있다. 그러면 $X = T$ 이고 $P(X = T) = 1/2$이다.
- 이런 주머니 X를 이제부터 확률변수(random variable)라고 부르자.
- 확률변수 X에는 표본공간(sample space) S 안에 있는 어떤 "사건"이든 집어넣을 수 있고 "그 사건에 결부된 확률"을 표현할 수 있다.
- 확률변수 개념을 제대로 상상할 수 있으면 어려운 부분은 거의 끝난다!

확률변수

- 확률변수를 표기하는 수학적 방식에 조금 친숙해질 필요가 있다.

$$P(X = x) = p$$

- 여기서 X가 바로 확률변수이고 x는 특정 사건이다. p는 그 특정 사건이 발생할 확률이다.
- x에는 (특정 사건을 대표하는) 어떤 값이 주어져야 한다.
- 대문자와 소문자는 이제부터 전략적으로 사용되므로 주의를 기울여야 한다!

확률변수

자료유형에 따라 확률변수도 구분되어야 한다.

- 지금까지 자료유형을 양적 변수와 질적 변수로 나누었다. 그러나 가만히 들여다보면 양적 변수 안에서도 차이가 있다.
- 동전 던지기나 주사위 던지기는 H, T 또는 1, 2, 3, 4, 5, 6으로 값이 "딱딱 떨어져" 셀 수 있는 사건들로 이루어졌다.
- 이런 자료유형을 이산형(discrete)이라고 부르고(이산가족의 離散이다), 그러한 확률변수를 이산확률변수(discrete random variable)라고 부른다.
- 또다른 예로는 회사를 떠나는 직원 수, 여성별로 출산한 아이의 수, 특정 달에 파산을 신청한 기업의 수 등이 있다.

확률변수

- 반면에 (딱딱 떨어지지 않아서) 하나 둘 이렇게 셀 수 없는 경우를 연속형(continuous)이라고 부르고, 그러한 확률변수를 연속확률변수(continuous random variable)라고 부른다.
- 가령 사람의 키나 체중, 소득액/세액, 펀드의 수익률, 지역별 태어난 아이의 평균 숫자, 특정 작업을 완료하기까지 걸리는 시간 등은 연속확률변수가 된다.
- "여성별로 출산한 아이의 수"는 이산형이지만 "지역별 태어난 아이의 평균 숫자"는 연속형이다(Why?).

확률변수

이산확률변수와 연속확률변수에서 확률의 표현이 조금씩 다르다.

- 만일 여성별로 출산한 아이의 수가 확률변수 X 라면, $P(X = 1) = 0.4$와 같은 표현이 가능하다.
- 만일 청주시 출산율이 확률변수 X 라면 (정확히 1이라는 값으로 딱 떨어질 확률은 무한히 작기 때문에) $P(X = 1) \approx 0$이다(Why?).
- 대신 연속확률변수에 대해서는 다음과 같이 표현해야 한다(Why?).

$$P(0.8 < X < 1) = P(0.8 \leq X < 1) = P(0.8 < X \leq 1) = P(0.8 \leq X \leq 1) = 0.4$$

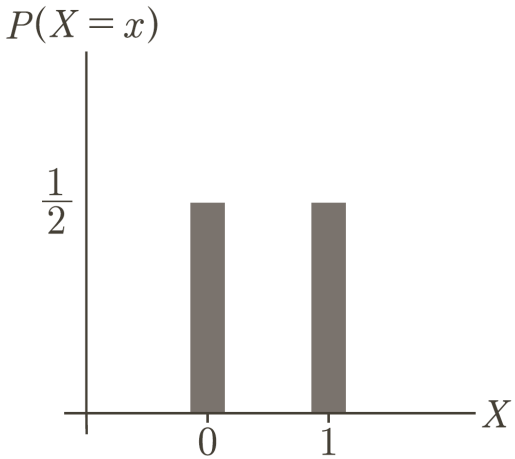
확률분포

확률분포

확률변수를 표와 그림으로 나타낼 수 있다.

- 먼저 동전 던지기의 결과를 이산확률변수로 다음과 같이 나타낼 수 있다:
  "동전을 던져서 앞면(H)이 나오는 사건의 확률 $P(X = H)$은 $1/2$이다."
  "동전을 던져서 뒷면(T)이 나오는 사건의 확률 $P(X = T)$은 $1/2$이다."
- H을 1로, T를 0으로 명목변수를 설정하자!
- 확률분포(probability distribution)는 확률변수 X가 취할 수 있는 모든 사건들 x와 그에 대응한 확률 $P(X = x)$을 나열하여 보여준다.

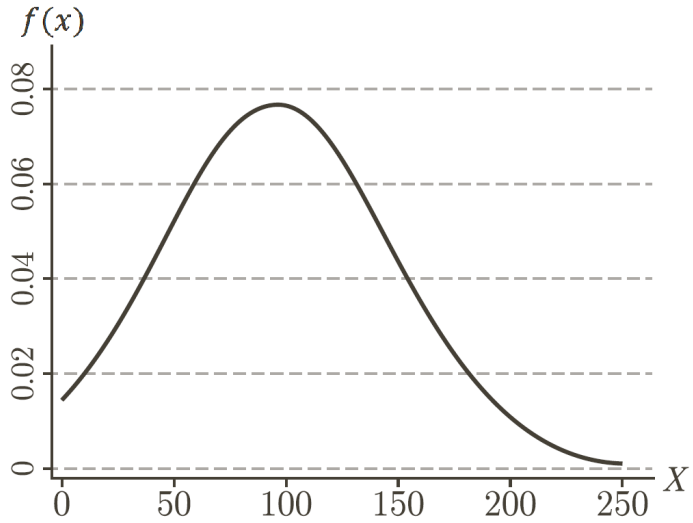
x	$P(X = x)$
0	$1/2$
1	$1/2$



확률분포

- 연속확률변수 역시 확률분포를 표와 그림으로 나타낼 수 있다.
- 어느 카페의 평균 고객 체류시간을 확률변수 X로 하여 조사하니 확률분포는 다음과 같았다.

X	$P(X=x)$
[0–50)	0.17
[50–100)	0.35
[100–150)	0.33
[150–200)	0.13
[200–250)	0.02
[250–300)	0.00



- 이 때 100분에서 150분 사이로 체류할 확률은 다음과 같이 표현할 수 있다.

$$P(100 \leq X < 150) = 0.33$$

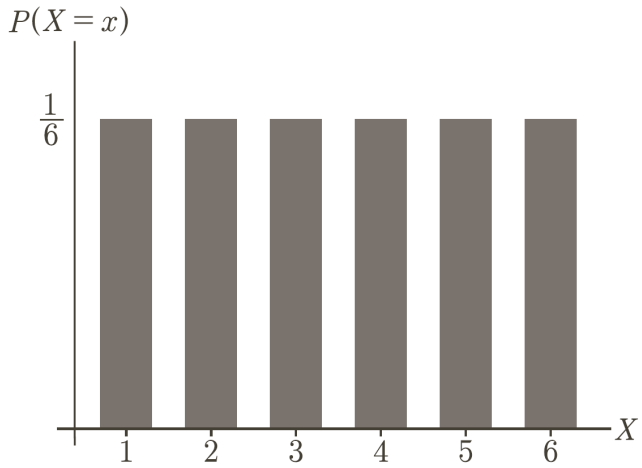
확률분포

예제 3. 주사위 던지기의 결과를 확률변수로 나타낸다면 자료유형은 무엇인가? 적절한 표와 그림을 작성하시오.

확률분포

- 주사위를 던진 결과는 이산확률변수이므로 다음과 같이 확률변수로 표현할 수 있다:
  "주사위 던지기에서 1이 나오는 사건$(X = 1)$의 확률 $P(X = 1)$은 $1/6$이다."
  ...
  "주사위 던지기에서 6이 나오는 사건$(X = 6)$의 확률 $P(X = 6)$은 $1/6$이다."

x	$P(X = x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6



확률질량함수와 확률밀도함수

확률질량함수와 확률밀도함수

확률분포를 수학적 함수로 표현할 수도 있다.

- 앞서 주사위 던지기 사례에서는 $P(X = 1) = 1/6$ 부터 $P(X = 6) = 1/6$ 까지 하나하나 가능성을 열거하였다.
- 그러나 수학자들은 이런 개별적(idiosyncratic) 나열 대신 하나의 함수(function)로 모든 가능성을 일관하여 표현할 수 있을 때 더 아름답다(elegant)고 생각한다.
- 일단 함수를 만들면, 어떤 사건이 주어졌을 때 그에 따른 확률을 계산하는데 사용될 수도 있다!

확률질량함수와 확률밀도함수

- 이산확률변수라면 확률질량함수(probability mass function; PMF)로 나타낸다.
- 확률질량함수 $f(x)$는 확률 $P(X = x)$와 동일하며 다음이 성립한다.

$$f(x) \geq 0, \quad \forall x \in S$$

$$\sum_{\forall x \in S} f(x) = 1$$

- 이때 $\forall x \in S$는 "표본공간 S에 속하는 모든($\forall$) 사건들 x에 대하여"의 줄임이다. $\forall$는 turned A 또는 for all이라고 읽는다.

확률질량함수와 확률밀도함수

- 연속확률변수라면 확률밀도함수(probability density function; PDF)로 나타낸다.
- 확률밀도함수와 관련하여 다음이 성립한다.

$$f(x) \geq 0, \quad \forall x \in S$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

- 이때 상하로 길게 늘어진 기호는 integral이라고 읽으며, 적분(integration)을 의미한다.
- 종종 사람들은 자료유형과 상관없이 확률밀도함수로 통칭하기도 한다.

확률질량함수와 확률밀도함수

- (확률분포를 개별적으로 나열했을 때와 마찬가지로) 확률질량함수와 확률밀도함수를 그래프로 나타낼 수 있다.
- 이렇게 그래프로 나타냈을 때 확률질량함수 혹은 확률밀도함수의 값 $f(x)$는 막대 혹은 곡선의 높이를 나타낸다.
- 확률질량함수라면 $P(X = x)$와 $f(x)$가 같은 의미이지만, 확률밀도함수라면 $P(X = x)$와 $f(x)$는 같지 않다(Why?).

누적분포함수

누적분포함수

확률질량함수를 일정한 범위에 걸쳐 "누적하여" 계산할 수 있다.

- 주사위를 던졌을 때 6이라는 개별 사건이 발생할 확률은 1/6이다.

$$P(X = 6) = \frac{1}{6}$$

- 그러면 주사위가 3 이하의 값이 나올 확률은 얼마인가?
- 여기에는 1, 2, 3의 가능성이 있으므로 1/2이다.

$$P(X \leq 3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

누적분포함수

예제 4. 주사위에서 나온 숫자가 4보다 작을 확률은 얼마인가? 주사위에서 나온 숫자가 2보다 크고 5보다 작을 확률은 얼마인가? 주사위에서 나온 숫자가 5 또는 그보다 클 확률은 얼마인가?

누적분포함수

- “주사위에서 나온 숫자가 4보다 작을 확률은?”

$$P(X < 4)$$

- “주사위에서 나온 숫자가 2보다 크고 5보다 작을 확률은?”

$$P(2 < X < 5) = P(X < 5) - P(X \leq 2)$$

- “주사위에서 나온 숫자가 5 또는 그보다 클 확률은?”

$$P(X \geq 5) = 1 - P(X < 5)$$

누적분포함수

확률질량함수의 누적 계산을 일반화할 수 있다.

- 확률질량함수가 $P(X = x)$에 관한 것이라면 누적분포함수(cumulative distribution function; CDF)는 $P(X \leq x)$에 관한 것이다.
- 만일 a보다 크고 b보다 작은 값이 나올 이산확률을 누적분포함수로 나타낸다면 다음과 같다.

$$P(a \leq X \leq b) = \sum_{x=a}^{b} P(X = x)$$

- 확률질량함수라면 구할 수 있는 개별 사건들의 확률을 모두 더해 누적분포함수를 계산할 수 있다.

누적분포함수

확률밀도함수의 적분은 색칠 공부나 마찬가지다.

- 확률질량함수의 누적 계산과는 달리 확률밀도함수의 누적 계산은 다소 복잡하다.
- 확률질량함수에서 이산확률변수 X는 특정 값을 갖는다. 하지만 확률밀도함수에서 연속확률변수 X는 무한히 작게 나뉠 수 있는 값이다.
- 그러므로 단순히 더할 수 없고, 적분이라는 조금 특별한 개념을 사용해야 한다.
- 이때 적분이란 잘게 나눈(分) 것을 쌓는다(積)는 것을 의미하고, 직관적으로 보면 색칠 공부와 다를 바 없다.

누적분포함수

- a와 b 사이의 값이 나올 연속확률을 누적분포함수로 나타낸다면 다음과 같다.

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

- 이때 연속확률변수의 경우 $P(X = a) = P(X = b) = 0$ 이므로 a와 b를 포함하는가 여부는 아무래도 상관없다.

$$P(a \leq X \leq b) = P(a < X < b) = P(a < X \leq b) = P(a \leq X < b)$$

확률변수의 평균과 분산

확률변수의 평균과 분산

확률분포가 주어지면 확률변수의 평균을 계산할 수 있다.

- 이산확률분포라면 평균 또는 기대값(expected value)을 다음과 같이 나타낼 수 있다.

$$\mu = E(X) = \sum_{\forall x} x \cdot P(X = x)$$

- 나타난 결과와 그에 대응하는 확률을 곱한 다음, 이를 모두 더하면 기대값이다.
- 가령 주사위 던지기 결과의 평균은 다음과 같이 계산할 수 있다.

x	$P(X = x)$	$x \cdot P(X = x)$
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6
	합계	3.5

확률변수의 평균과 분산

확률분포가 주어지면 확률변수의 분산도 계산할 수 있다.

- 분산은 다음과 같이 나타낼 수 있다.

$$\sigma^2 = Var(X) = \sum_{\forall x} (x-\mu)^2 \cdot P(X=x)$$

- 가령 주사위 던지기 결과의 분산은 다음과 같이 계산할 수 있다.

x	$P(X=x)$	$x \cdot P(X=x)$	$(x-\mu)^2 \cdot P(X=x)$
1	1/6	1/6	$(1-3.5)^2 \cdot 1/6$
2	1/6	2/6	$(2-3.5)^2 \cdot 1/6$
3	1/6	3/6	$(3-3.5)^2 \cdot 1/6$
4	1/6	4/6	$(4-3.5)^2 \cdot 1/6$
5	1/6	5/6	$(5-3.5)^2 \cdot 1/6$
6	1/6	6/6	$(6-3.5)^2 \cdot 1/6$
	합계	3.5	2.917

확률변수의 평균과 분산

예제 5. 새우는 지금 도박장 앞에서 망설이고 있다. 이 도박에서 10만 원을 상금으로 따낼 확률은 30분의 1이다. 그러나 지면 1만 원을 잃는다.
새우가 이 도박을 통해 거둘 수익의 기대값과 분산은 얼마인가?

확률변수의 평균과 분산

- 먼저 주어진 조건을 서술하자(이것을 잘하는 것이 중요하다!).

$$P(X = \text{승리}) = P(X = 10) = 1/30$$
$$P(X = \text{패배}) = P(X = -1) = 29/30$$

- 수익의 기대값과 분산은 다음과 같이 계산한다.

x	$P(X = x)$	$x \cdot P(X = x)$	$(x - \mu)^2 \cdot P(X = x)$
10	$1/30$	$10/30$	$(10 + 0.633)^2 \cdot 1/30$
-1	$29/30$	$-29/30$	$(-1 + 0.633)^2 \cdot 29/30$
		-0.633	3.90

확률변수의 평균과 분산

- "확률분포가 주어지면 확률변수의 평균과 분산을 계산할 수 있다!"
- 지금 이 말은 매우 중요하기 때문에 반드시 기억해 두어야 한다!
- 확률변수의 기대값은 다음과 같은 수학적 성질을 갖는다(단 X는 확률변수이고 a와 b는 임의의 상수이다).

$$
\begin{aligned}
E(aX) &= aE(X) \\
E(X+b) &= E(X)+b \\
E(aX+b) &= E(aX)+b = aE(X)+b
\end{aligned}
$$

- 확률변수의 분산은 다음과 같은 수학적 성질을 갖는다.

$$
\begin{aligned}
Var(aX) &= a^2 Var(X) \\
Var(X+b) &= Var(X) \\
Var(aX+b) &= Var(aX) = a^2 Var(X) \\
Var(X+Y) &= Var(X)+Var(Y)+2Cov(X,Y)
\end{aligned}
$$