사회통계연습

확률과 확률분포

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

October 1, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 지난 주 리뷰
2. 확률이론의 기초
3. 베이즈 정리
4. 확률분포
5. 누적분포함수

지난 주 리뷰

지난 주 리뷰

퀴즈 #3 코멘트

- 괜찮은 상관관계 해석의 예: "측정한 결과 상관계수는 0.038입니다. 0과 1사이를 4등분해서 (0, 0.25, 0.5, 0.75, 1)보면 0.038은 큰 값이 아닙니다. 즉, 두 변수는 깊은 관계가 아니며..."
- 상관관계든 crosstab이든 "두 변수 간에 유의미한 관계가 있다" 라고 결론을 내려선 안된다고 했는데 고대로 하네!! 유의성 검정(significance test) 없이 그런 결론에 도달할 수 없다.
- 솔직히 crosstab 피벗 테이블을 당연히 첨부할 것이라고 예상했으나 소수의 학생만 첨부해서 좀 놀랐음. 3주차 과제까지는 테이블이 없어도 감점 없음. 시험에서 표 문제를 냈는데 표를 내지 않으면 당연히 감점임.
- 괜찮은 crosstab 해석의 예: "취업여부를 행, 장래결혼계획 여부를 열로 놓고 피벗테이블을 사용하여 교차표를 만든 결과, 미취업자, 취업자 모두 장래 결혼을 할 생각이 있다고 답한 비율이 각각 71%, 78%로 높았다. 다만 취업자의 경우 미취업자보다 장래 결혼 계획이 있다고 답한 비율이 높았다. 이는 상대적으로 미취업자에 비해 안정적인 소득이 있는 취업자의 경우, 결혼과 양육 등의 비용을 지출할 경제적 능력이 있기 때문인 것으로 보인다."

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

지난 주 리뷰

퀴즈 #4 코멘트

- 숙제를 너무 늦게 제출해서 도저히 채점을 마칠 수 없었다...
- 어제(9월 30일) 오후 11시 이후에만 7명의 학생이 제출했다...
- 코멘트는 다음 주에...

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

지난 주 리뷰

총평

- 숙제 퀄리티에 점차 양극화 경향이 보입니다. 잘하는 학우는 계속 잘하고, 잘하던 학우가 점점 못해가는 모습이 안타깝습니다!
- 좀 더 신경을 써주세요. “대충 해도 모르겠지”가 아니라 여러 사람의 숙제를 한꺼번에 채점하다보니 티가 확 납니다.
- 강의안과 책을 통해 꼼꼼히 복습을 해야 합니다. 숙제를 하면서 잘 모르겠으면 관련된 부분을 좀 더 철저히 공부하거나 질문하여 반드시 알고 넘어가야 합니다.
- 사회통계-사회통계연습-사회조사방법론 등으로 이어지는 긴 길입니다. 중간에 이해가 부족하면 뒷부분을 감당할 수 없게 됩니다.
- 숙제 파일 이름에 최소한 주차와 학번은 써주세요(e.g., 3주차 20201234.docx). 학번 빼지 말아주세요. 파일이 겹쳐서 저장될 위험이 있습니다

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

확률이론의 기초

확률이론의 기초

확률이론(probability theory)의 역사적 발달은 도박과 밀접한 연관을 갖고 있었다.

- 17세기 프랑스 도박사였던 앙투안 공보(Antoine Gombaud)는 도박과 판돈의 분배에 관한 문제에 관해 고민하다가 결국 해법을 찾는데 실패하고, 당대의 천재 수학자였던 블레즈 파스칼(Blaise Pascal)과 피에르 드 페르마(Pierre de Fermat)에게 이에 관해 물어보았다.
- 이에 파스칼과 페르마는 서신을 교환하면서 확률이론의 토대를 쌓기 시작했다.
- 이 두 사람 덕택에 여러분이 이 고생을 하는 거다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

왜 경험과학(empirical science)에서 확률이론을 필요로 하나?

- (나중에 더 자세히 다루겠지만) 우리는 모집단(population)에서 표본(sample)을 골라낼 것이다.
- 그 뒤, (모집단은 내버려두고) 표본만을 관찰하여 어떤 성질을 발견해 낸다.
- 우리는 그 표본의 성질로부터 모집단의 성질을 유추해 내고(make inference) 싶다.
- 이 표본으로부터 유추된 성질이 (모집단과는 상관없이) 그저 표본만에서 우연히 나타났을 확률을 알고 싶기 때문에 필요하다.

확률이론의 기초

"내가 동전을 두 번 던져서 두 번 모두 앞면이 나올 가능성은?"

- 일단 동전을 (한 번) 던졌을 때, 있을 수 있는 모든 결과(all possible outcomes)는 암만 생각해도 두 가지 뿐: 앞면(head)과 뒷면(tail).
- 앞면(head)이 나올 확률은 1/2, 뒷면(tail)이 나올 확률도 1/2.
- 이제 두 번 던졌을 때, 나올 수 있는 모든 결과(outcomes)들의 집합(set)은 {HH, HT, TH, HH}.
- "응, 1/4이야."
- 이 아이디어를 동전 던지기를 넘어 모든 현상에 일반화 해보자.

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

이 사고실험은 우리에게 몇 가지 유용한 개념이 있음을 시사한다.

- 확률(probability)이란 “어떤 사건(event)이 발생할 가능성을 나타난 값”이다.
- 사건(event)이란 “표본공간(sample space)의 특정 부분집합”을 의미한다. 앞에서 사건은 “두 번 모두 앞면(HH)”.
- 표본공간(sample space; S)이란 “어떤 시행(trials) 또는 실험에서 가능한 모든 결과(outcomes)의 집합”을 의미한다. 앞에서 표본공간은 $S = \{HH, HT, TH, TT\}$.
- 동전을 두 번 던진 결과들(outcomes)은 HH, HT, TH, TT 모두 4개가 있다.
- 결과(outcomes)와는 달리 사건(events)은 연구자의 관심에 따라 달라질 수 있다.
- 앞에서는 “두 번 모두 앞면”이라는 사건에 관심을 두었는데, 만일 “적어도 한 번은 앞면”이라는 사건을 본다면 $\{HH, HT, TH\}$가 모두 사건이 된다.

확률이론의 기초

고전적 확률(classical probabillity)

- "논리적으로 생각하여" 알 수 있는 가능성을 확률로 정의하는 입장이다. 이 때문에 분석적(analytic) 확률로도 불리운다.
- 표본공간 $S = \{s_1, s_2, \ldots, s_n\}$에서 각각의 결과(outcomes)가 일어날 가능성이 동등하다면, m개의 결과(outcomes)를 포함하는 사건(event)의 확률은 다음과 같다: $P(E) = m/n$

경험적 확률(empirical probabillity)

- "장기간에 걸쳐 사건의 발생을 관찰하여(in a long run of observations) 얻은 상대적인 빈도"로 확률을 정의하는 관점!
- 주사위를 단 몇 차례만 던졌다면, 6이 나올 확률은 1/6이 아닐 수도 있다. 그러나 주사위를 제법 여러 차례 던졌다면, 6이 나올 확률은 1/6에 제법 근접할 것이다.
- 주사위를 무한히 많이 던진다면, 6이 나올 확률은 1/6에 무한히 근접할 것이다.

고전적 접근은 확률을 분석적으로 연역하고 있지만, 경험적 접근은 확률을 자료로부터 귀납한다.

확률이론의 기초

확률은 사건(event)에 대응하여 정의된다.

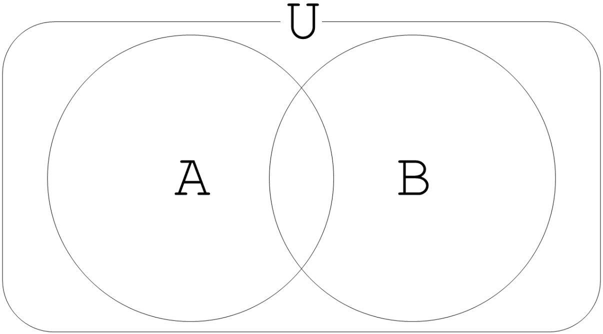
- 경험과학에서 개념의 내포(connotation)를 밝히는 방법을 정의라고 부르고, 개념의 외연(denotation)을 밝히는 방법을 구분이라고 부른다.
- “오늘 비가 내릴” 가능성이나 “동전을 열 번 던져 모두 뒷면이 나올” 가능성과 같이 사건(event)이 먼저 정의되어 있어야 한다.
- 일단 정의했다면 구분도 해야 하는데, 구분할 때는 반드시 두 법칙을 따라야 한다.
- 첫째, (구분한 후에 얻어진) 자항의 총합이 모항과 같아야 한다(exhaustive).
- 예컨대 “10대”, “20대’, “30대 이상”은 자항의 총합이 모항과 같지 않은 구분이다.
- 둘째, (구분한 후에 얻어진) 자항들은 서로 배제해야 한다(mutually exclusive).
- 예컨대 “군필”, “미필”, “장교 전역”은 서로 배제하지 못하는 구분이다.
- 전체포괄적(exhaustive)이고 서로 배제하는(mutually exclusive) 사건들의 확률의 합은 1이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

확률이론의 기초

두 사건의 확률을 나타낼 때는 흔히 벤 다이어그램(Venn Diagram)을 그리곤 한다.

- A와 B의 합사건(union)은 $A \cup B$ 로 표시한다
- A와 B의 교사건(intersection)은 $A \cap B$ 로 표시한다.
- A의 여사건(complement)은 A^C 로 표시한다



확률이론의 기초

확률에는 몇 가지 기본법칙들이 성립한다.

- 덧셈 법칙(addition rule): $P(A \cup B) = P(A) + (B) - P(A \cap B)$
- 여기서 P(A)는 이제 곧 설명할 결합확률(joint probability)이다.
- 그런데 A와 B가 독립적(independent)인 사건들이라면 덧셈 법칙은 더욱 단순해진다: $P(A \cup B) = P(A) + (B)$.
- 여사건 법칙(complement rule): $P(A^C) = 1 - P(A)$
- 곱사건 법칙(multiplication rule): $P(A \cap B) = P(A|B) \cdot P(B)$
- 여기서 $P(A|B)$ 는 이제 곧 설명할 조건부확률(conditional probability)이다.

확률의 기본법칙을 이해하면 독립 사건(independent event)과 종속 사건(dependent event)을 이해할 수 있다.

- 두 개의 사건 A와 B가 있을 때, $P(A|B) = P(A)$ 이거나 $P(B|A) = P(B)$ 이면 두 사건은 독립이다(Why?).
- (두 사건은 독립이 아니면 종속이므로) 위 식이 성립하지 않으면 종속이다.

베이즈 정리

베이즈 정리

조금 특별한 형태의 확률 개념으로 세 가지를 기억해두자: 결합확률, 조건부확률, 그리고 한계확률

결합확률(joint probability)이란 “두 사건이 동시에 일어날 확률”을 의미한다

- $P(A \cap B) = P(A\&B) = P(A|B) \cdot P(B)$

조건부확률(conditional probability)이란 “다른 사건(B)이 이미 있어났다는 전제 아래, 한 사건(A)이 일어날 확률”을 의미한다.

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

$P(A \cap B) = P(B \cap A)$ 이지만, $P(A|B) \neq P(B|A)$ 이다.

베이즈 정리

한계확률(marginal probability)은 “연관된 결합확률(joint probability)의 합(sum)”이다.

$$
\begin{aligned}
P(A) &= P(A \cap B) + P(A \cap \neg B) && (1a)\\
&= P(A|B) \cdot P(B) + (A|\neg B) \cdot P(\neg B) && (1b)
\end{aligned}
$$

(1b)는 조건부확률 P(A|B)의 정의를 가지고 (1a)에서 도출된 것이다.

얼핏 보면 결합확률, 조건부확률, 그리고 한계확률은 되게 추상적인 개념처럼 보이는데, 여러분은 이미 예전 과제에서 실제로 써먹었다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

베이즈 정리

	종교인	비종교인	합계
여자	97	124	221
남자	68	232	300
합계	165	356	521

	종교인	비종교인
여자	0.44	0.56
남자	0.23	0.77
합계	0.67	1.33

데이터로부터 왼쪽 표(crosstab)를 만든 뒤, row total을 기준으로 표준화하여 오른쪽 표를 만들었다.

- 이 표를 들여다보면 $P(\text{종교인}|\text{여자}) = 0.44$, $P(\text{비종교인}|\text{여자}) = 0.56$, $P(\text{종교인}|\text{남자}) = 0.23$, $P(\text{비종교인}|\text{남자}) = 0.77$ 임을 알 수 있다.
- 이것이 바로 조건부확률(conditional probability)이다.

만약 column total을 기준으로 표준화한 경우 다른 해석을 할 수 있다.

- 그 경우에는 $P(\text{여자}|\text{종교인})$, $P(\text{남자}|\text{종교인})$, $P(\text{여자}|\text{비종교인})$, $P(\text{남자}|\text{비종교인})$을 구할 수 있다.
- 이것들도 물론 조건부확률(conditional probability)이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

베이즈 정리

	종교인	비종교인	합계
여자	97	124	221
남자	68	232	300
합계	165	356	521

	종교인	비종교인	합계
여자	0.19	0.24	0.42
남자	0.13	0.45	0.58
합계	0.32	0.68	1

데이터로부터 왼쪽 표(crosstab)를 만든 뒤, grand total을 기준으로 표준화를 하여 오른쪽 표를 만들었다.

- 이 표를 들여다보면 $\mathrm{P}(\text{종교인} \cap \text{여자}) = 0.19$, $\mathrm{P}(\text{비종교인} \cap \text{여자}) = 0.24$, $\mathrm{P}(\text{종교인} \cap \text{남자}) = 0.13$, $\mathrm{P}(\text{비종교인} \cap \text{남자}) = 0.45$ 임을 알 수 있다.
- 이것이 바로 결합확률(joint probability)이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

베이즈 정리

지금까지 배운 교차표(crosstab), 결합확률(joint probability), 조건부확률(conditional probability) 개념을 동시에 곰곰히 생각해보면,

- $\mathrm{P}(\text{종교인}) = \mathrm{P}(\text{종교인} \cap \text{여자}) + \mathrm{P}(\text{종교인} \cap \text{남자})$ 이므로,
  $\mathrm{P}(\text{종교인}) = \mathrm{P}(\text{종교인}|\text{여자}) \cdot \mathrm{P}(\text{여자}) + \mathrm{P}(\text{종교인}|\text{남자}) \cdot \mathrm{P}(\text{남자})$ 이다.
- $\mathrm{P}(\text{비종교인}) = \mathrm{P}(\text{비종교인} \cap \text{여자}) + \mathrm{P}(\text{비종교인} \cap \text{남자})$ 이므로,
  $\mathrm{P}(\text{비종교인}) = \mathrm{P}(\text{비종교인}|\text{여자}) \cdot \mathrm{P}(\text{여자}) + \mathrm{P}(\text{비종교인}|\text{남자}) \cdot \mathrm{P}(\text{남자})$ 이다.
- $\mathrm{P}(\text{여자}) = \mathrm{P}(\text{여자} \cap \text{종교인}) + \mathrm{P}(\text{여자} \cap \text{비종교인})$ 이므로,
  $\mathrm{P}(\text{여자}) = \mathrm{P}(\text{여자}|\text{종교인}) \cdot \mathrm{P}(\text{종교인}) + \mathrm{P}(\text{여자}|\text{비종교인}) \cdot \mathrm{P}(\text{비종교인})$ 이다.
- $\mathrm{P}(\text{남자}) = \mathrm{P}(\text{남자} \cap \text{종교인}) + \mathrm{P}(\text{남자} \cap \text{비종교인})$ 이므로,
  $\mathrm{P}(\text{남자}) = \mathrm{P}(\text{남자}|\text{종교인}) \cdot \mathrm{P}(\text{종교인}) + \mathrm{P}(\text{남자}|\text{비종교인}) \cdot \mathrm{P}(\text{비종교인})$ 이다.
  이것이 바로 한계확률(marginal probability)이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

베이즈 정리

교차표(crosstab)가 2×2 가 아니라 3×3 라도 조금 더 길어지지만 본질은 똑같다.

	보수적 성역할 태도	중립적 성역할 태도	진보적 성역할 태도	합계
소득: 상	486	761	279	1526
소득: 중	1024	1080	967	3071
소득: 하	662	822	434	1918
합계	2172	2663	1680	6515

- 여기서도 여러분은 표준화 방식에 따라 조건부확률(conditional probability)과 결합확률(joint probability)을 구할 수 있다.
- 한계확률(marginal probability)도 마찬가지다. 다만 2×2 방식이 아니므로 B와 $\neg B$가 아니라 B_1, B_2, B_3 같이 표현하면 편리하다.
- $P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)$

베이즈 정리

잠깐! 이 챕터의 제목은 베이즈 정리(Bayes Theorem)인데 지금까지 내내 결합확률, 조건부확률, 한계확률만 실컷 이야기했다.

이제부터 베이즈 정리를 이야기한다.

$$
\begin{aligned}
P(A|B) &= \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(B|A) \cdot P(A)}{P(B)} \quad (\text{if } P(B) \neq 0) \\
&= \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)}
\end{aligned}
$$

베이즈 정리 끝.

베이즈 정리

베이즈 정리는 대체 왜 배우나? 이걸로 무엇을 할 수 있나?

- 몇 가지 흥미로운 논리적 문제를 풀 수 있다. 몬티 홀(Monty Hall) 문제, 다운증후군 임신 테스트 신뢰성 문제 등등...
- 그 밖에도 몇 가지 실생활에서 할 수 있는게 있다. 사람의 말과 글을 알아듣는 기계나 자율주행하는 자동차를 만든다던가 등등...
- (R이나 Python 등을 통해) 20줄-30줄 정도의 코드로도 꽤 쓸모있는 인공지능을 구현할 수 있다.

확률분포

확률분포

확률 표현을 좀 더 일반화하기 위해 편리한 개념을 개발하자!

- X라는 빈 주머니를 상상하자. 이때 X 안에 개별 사건을 집어넣을 수 있다고 하자.
- 동전 던지기에서 앞면이 나오는 사건(H)을 집어넣으면 $\mathrm{X} = \mathrm{H}$ 이고 $\mathrm{P}(\mathrm{X} = \mathrm{H}) = 1/2$ 이다. 뒷면이 나오는 사건(T)도 집어넣을 수 있다. 그러면 $\mathrm{X} = \mathrm{T}$ 이고 $\mathrm{P}(\mathrm{X} = \mathrm{T}) = 1/2$이다.

이런 주머니 X를 이제부터 확률변수(random variable)라고 부르자.

- 확률변수 X에는 표본공간(sample space) 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!

확률분포

주사위를 던진 결과를 확률변수(random variable)로 생각한다면,

- 주사위 던지기에서 1이 나오는 사건(X=1)의 P(X=1)는 1/6이다.

  ...
- 주사위 던지기에서 6이 나오는 사건(X=6)의 P(X=6)는 1/6이다.

이렇게 확률변수 X를 표로 정리해보면,

X_i	$P(X_i)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

- 이렇게 확률변수(X)가 어떤 값을 가질 확률을 쭉 나열한 것을 확률분포(probability distribution)라고 한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

확률분포(probability distribution)에서 기댓값(=평균)을 계산해 낼 수도 있다.

- 기댓값(expected value)은 $\sum X_i \cdot P(X = x_i)$ 로 나타낼 수 있다.
- 나타난 결과와 그에 대응하는 확률을 곱한 다음, 이를 모두 더하면 기댓값이다.

X_i	$P(X_i)$	$X_i \cdot P(X = x_i)$
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
		3.5

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

예제

- 어떤 도박에서 여러분이 이겨서 10만 원을 상금으로 따낸 확률은 30분의 1이다. 지면 1만 원을 잃는다. 여러분의 이 도박을 통해 거둘 수익의 기댓값은 얼마인가?

X_i	$P(X=x_i)$	$X_i \cdot P(X = x_i)$
10	1/30	10/30
-1	29/30	-29/30
		-1.633

확률분포(probability distribution)에서 분산(variance)도 계산해 낼 수 있다.

- 분산(variance)은 $\sum(X_i - \mu)^2 \cdot P(X = x_i)$ 로 나타낼 수 있다.

X_i	$P(X_i)$	$X_i \cdot P(X = x_i)$	$(X_i - \mu)^2 \cdot P(X = x_i)$
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	$(6-3.5)^2 \cdot 1/6$
		3.5	2.917

확률분포

- 예제: 어떤 도박에서 여러분이 이겨서 10만 원을 상금으로 따낸 확률은 30분의 1이다. 지면 1만 원을 읽는다. 여러분의 이 도박을 통해 거둘 수익의 분산은 얼마인가?

X_i	$P(X{=}x_i)$	$X_i \cdot P(X = x_i)$	$(X_i - \mu)^2 \cdot P(X{=}x_i)$
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

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

"일단 확률분포가 주어지면 확률변수의 평균과 분산을 계산할 수 있다."

- 지금 이 말은 매우 중요하기 때문에 반드시 기억해 두어야 한다.
- 확률변수의 기댓값은 E(X)로, 분산은 Var(X)로 표시할 수 있다.

확률변수의 기댓값의 성질

- $\mathrm{E}(\mathrm{aX}) = \mathrm{aE}(\mathrm{X})$
- $\mathrm{E}(\mathrm{X} + \mathrm{b}) = \mathrm{E}(\mathrm{X}) + \mathrm{b}$
- $\mathrm{E}(\mathrm{aX} + \mathrm{b}) = \mathrm{E}(\mathrm{aX}) + \mathrm{b} = \mathrm{aE}(\mathrm{X}) + \mathrm{b}$

확률변수의 분산의 성질

- $\mathrm{Var}(\mathrm{aX}) = \mathrm{a}^2\mathrm{Var}(\mathrm{X})$
- $\mathrm{Var}(\mathrm{X} + \mathrm{b}) = \mathrm{Var}(\mathrm{X})$
- $\mathrm{Var}(\mathrm{aX} + \mathrm{b}) = \mathrm{Var}(\mathrm{aX}) = \mathrm{a}^2\mathrm{Var}(\mathrm{X})$
- $\mathrm{Var}(\mathrm{X} + \mathrm{Y}) = \mathrm{Var}(\mathrm{X}) + \mathrm{Var}(\mathrm{Y}) + 2\mathrm{Cov}(\mathrm{X}, \mathrm{Y})$

X는 확률변수(random variable), a와 b는 임의의 상수(constant)임.

확률분포

지금까지 숫자형(numerical) 척도와 범주형(categorical) 척도만으로 구분했지만 가만히 보면 숫자형 척도 안에서도 차이가 있다.

- 위 예제는 동전 던지기나 주사위 던지기처럼 사건이 H, T 또는 1, 2, 3, 4, 5, 6로 값이 딱딱 떨어져서 셀 수 있는 경우만을 언급하였다.
- 이런 자료유형을 특별히 이산형(discrete)이라고 부르고(이산가족의 離散이다), 그런 확률변수를 이산확률변수(discrete random variable)라고 부른다.
- 또다른 예는 분기별 목표를 달성한 사회복지사의 수, 회사를 떠나는 직원 수, 여성별로 출산한 아이의 수, 특정 달에 파산을 신청한 기업의 수 등이 있다.
- 반면에 딱딱 떨어지지 않아서 하나 둘 이렇게 셀 수 없는 경우를 연속형(continuous) 이라고 부르고, 또 연속확률변수(continous random variable)라고 부른다.

그러므로 새삼 다시 정리하자면,

- 척도(scale)는 숫자형(numerical)과 범주형(categorical)로 일차적으로 구분되고, 숫자형은 다시 연속형(continuous)과 이산형(discrete)으로 구분된다.
- 솔직히 말해, 전통적 척도(명목, 서열, 등간, 비율)보다 이 구분법이 훨씬 더 중요하다!

확률분포

연속형(continuous) 척도는 어떤 예가 있을 수 있나?

- 가령 사람의 키나 체중, 소득액/세액, 펀드의 수익률, 지역별 태어난 아이의 수, 특정 작업을 완료하기까지 걸리는 시간 등은 연속확률변수가 될 수 있다.
- 왜 "여성별로 출산한 아이의 수"는 이산형(discrete)인데, "지역별 태어난 아이의 수"는 연속형(continuous)인가?
- 사람은 0.5421... 명을 낳을 수 없지만, 지역별로는 평균 따위를 계산하다보면 그런 숫자가 나올 수 있기 때문이다.

이산확률변수와 연속확률변수에서 확률의 표현이 조금씩 다르다.

- 만일 여성별로 출산한 아이의 수가 X라면, P(X=1)=0.4와 같은 표현이 가능하다.
- 만일 청주시 출산율이 X라면, $P(X=1) \approx 0$이다. 정확히 1이라는 숫자로 떨어질 확률은 무한히 작기 때문이다.
- 대신 청주시 출산율에 대해서는 다음의 표현이 가능하다:
  $P(0.8<X<1)=P(0.8 \leq X<1)=P(0.8<X \leq 1)=P(0.8 \leq X \leq 1)=0.4$
- 이렇게 연속형 변수에 대한 확률분포를 정의하는 확률밀도함수(probability density function; P.D.F)를 상상할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

확률질량함수와 확률밀도함수

- 이산형 변수에 대한 확률분포를 정의하는 함수를 확률질량함수(probability mass function; P.M.F)라고 부른다.
- 연속형 변수에 대한 확률분포를 정의하는 함수를 확률밀도함수(probability density function; P.D.F)라고 부른다.

누적분포함수

누적분포함수

처음에 사용했던 단순계산 방식을 놔두고 왜 구태여 확률분포 개념을 사용하나?

- 처음에 사용했던 단순계산 방식과는 달리 확률분포에서는 확률변수라는 일반화된 개념을 사용한다.
- 확률변수가 가질 수 있는 값들에 "조건을 부여하고" 그에 맞는 확률을 계산할 수도 있다.
- 따라서 다양한 방식으로 사건을 정의하고 그에 따른 확률을 계산하는데 탄력적으로 쓰일 수 있다.

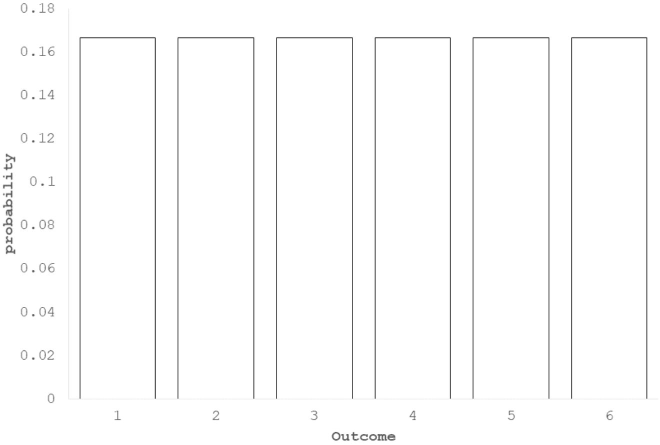
누적분포함수

누적분포함수(cumulative distribution function; C.D.F)는 확률분포 개념의 색다른 유용성을 보여준다.

- 보통의 확률분포함수는 P(X=x)를 보여주지만, 누적분포함수는 P(X〈x)를 보여준다.
- 누적분포함수는 수학적으로 적분(integration)을 이해해야 하는데 직관적으로 보면 걍 색칠 공부다.
- "주사위에서 나온 숫자가 4보다 작을 확률은?"
  $P(X < 4)$
- "주사위에서 나온 숫자가 2보다 크고 5보다 작을 확률은?"
  $P(2 < X < 5) = P(X < 5) - P(X < 2)$
- "주사위에서 나온 숫자가 5 또는 그보다 클 확률은?"
  $P(X \geq 5) = 1 - P(X < 5)$

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

누적분포함수



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

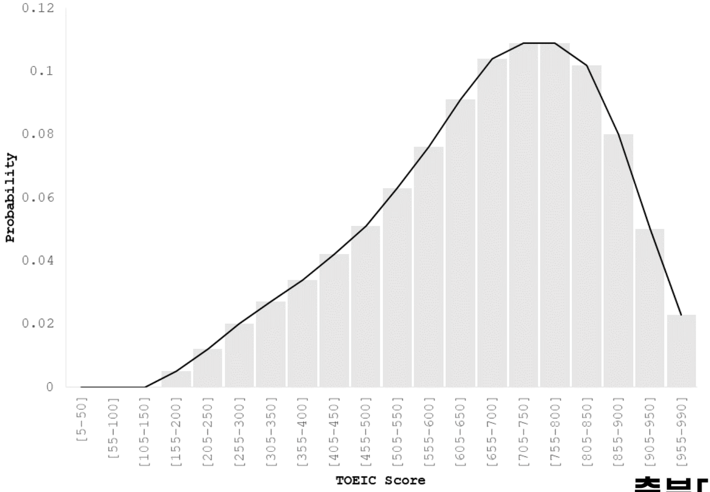
누적확률분포(Cumulative Probability Distribution)

물론 연속형(continuous) 변수에 대해서도 누적분포함수를 그릴 수 있다.

- 앞서 주사위 결과에서의 P(X⟨x)는 이산형(discrete) 누적분포함수다.
- "우리나라 직장인 토익 점수' 처럼 연속형 변수라면 연속형 누적분포함수를 그려야 한다.
- "토익 점수가 750 점보다 낮을 확률은?"
  $P(X < 750)$
- "토익 점수가 500 점보다 크고 650 점보다 낮을 확률은?"
  $P(500 < X < 650) = P(X < 650) - P(X < 500)$
- "토익 점수가 900 점보다 높을 확률은?"
  $P(X > 900) = P(X \geq 900) = 1 - P(X < 900)$

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

누적확률분포(Cumulative Probability Distribution)



충북대학교
CHUNGBUK NATIONAL UNIVERSITY