1. 확률, 확률변수 및 확률분포

데이터의 불확실성 측정의 도구

임채영

서울대학교 통계학과

이번 강의에서 다룰 내용

- 확률 - 불확실성을 다루기 위한 도구
- 확률변수와 확률분포 - 데이터의 불확실성을 표현하는 도구

데이터란?

- 관찰, 측정, 실험 등의 과정을 통해 얻어진 정량적(숫자형) 혹은 정성적(문자형) 기록
- 의사결정, 문제해결, 예측 등 목적에 따른 질문과 그 답을 구하기 위한 출발점
- 어떤 종류의 질문들?
  - 커피는 심장질환을 완화 시키는가?
  - 서울의 아파트 매매가격과 관련이 있는 요인은 무엇인가?
  - 내일 최저기온은 얼마일까?
- 데이터는 관찰시점, 측정 방법, 실험의 조건등 다양한 상황에 따라 다른 값이 기록되므로 불확실성을 가진다.

불확실성의 측정

- 불확실성을 과학적으로(수학적으로) 다루는 도구가 확률(Probability)이다.
  - 동전 던지기 에서 앞면이 나올까? 뒷면이 나올까? 앞면이 나올 가능성과 뒷면이 나올 가능성이 같다면?
  - 우리나라 성인 남성의 BMI가 20-25사이에 있을 가능성과 25-30사이 있을 가능성 중 어디가 높을까?
- 어떤 사건의 결과가 우연성을 가지고 있어 미리 그 결과를 정확하게 예측할 수 없지만 가능한 결과들의 가능성에 대해선 믿을 만한 합리적 지식을 가지고 있는 경우가 있음.
- 예를 들어 주사위를 던져서 게임을 하는 경우 6개의 숫자가 나올 가능성은 모두 같을 것이라 믿음.
- 이러한 가능성을 나타내는 개념을 통계학에서 확률이라고 함.

데이터의 불확실성을 측정하는 도구

- 데이터의 불확실성을 표현하기 위해 데이터를 확률변수(Random Variable, RV)의 실현값(또는 관측값)으로 본다.
- 확률에 기반하여 데이터의 불확실성(또는 가능성)을 계량화 해주는 도구가 확률분포(Probability Distribution)이다.
- 데이터의 분포란 데이터가 어떤 곳에 얼마나 집중되어 있는지, 데이터가 퍼진 정도가 얼마나 되는지를 나타내는 것이다.
- 데이터의 특성에 따라 분석의 방향이 달라질 수 있고 특성을 반영하여야 유용한 결과를 얻을 수 있다.

확률

확률이란?

- 어떤 일이 일어날 가능성의 정도를 나타냄.
- 0과 1사이의 값으로 표시되며 확률이 0이면 사건이 일어나지 않는 경우이고 1이면 무조건 일어나는 것을 말함.

확률의 계산

- 확률을 계산하는 두가지 방법
  - 논리적 확률 : 일어날 수 있는 모든 가능한 결과를 생각하고 모든 결과의 가능성을 논리적인 방법으로 유추
  - 경험적 확률 : 어떤 사건이 일어난 상대적인 횟수

논리적(수학적) 확률을 정의하기 위한 개념들

- 표본공간(Sample Space, $\mathcal{S}$) : 어떤 시행 (Experiment)에서 얻을 수 있는 가능한 모든 결과(Outcome)들의 집합
  예 : 하나의 주사위를 던지고, 나오는 눈의 수를 관찰할 때 표본공간
  $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$
- 사건(event, 사상) : 표본공간의 부분집합으로 보통 집합 $A, B, C, \cdots$ 등으로 표현

표본공간과 사건의 예

- 두 개의 동전을 동시에 던져서 나오는 면의 순서쌍 (앞면 H, 뒷면 T)
  - 표본공간, $\mathcal{S} = \{(H,H),(H,T),(T,H),(T,T)\}$
  - 앞면이 적어도 한번 나오는 사건을 A 라 하면,

$$A = \{(H,H),(H,T),(T,H)\},\ A \subset \mathcal{S}$$

- 고객센터에 전화를 했을때 기다려야 하는 시간을 조사하기 위해 한 명의 고객이 기다린 시간(분)을 관측할때,
  - 표본공간 $\mathcal{S} = \{t | t \geq 0\}$
  - 기다린 시간이 3분 이상인 사건을 B 라고 하면,

$$B = \{t \mid t \geq 3\}$$

확률의 계산

논리적 (수학적) 확률

- 사건 A가 일어날 확률 (equally likely outcomes)

$$P(A) = \frac{\text{사건 A에 속하는 원소의 개수}}{\text{표본공간 전체의 원소의 개수}}$$

예) 두 개의 동전을 동시에 던졌을 때, 앞면이 적어도 한 번 나올 확률

경험적 확률

- 사건 A가 일어날 확률

$$P(A) = \frac{\text{사건}A\text{ 가 일어난 횟 수}}{\text{전체 실험 또는 시행 횟수}}$$

확률측도를 통한 수학적 확률의 정의

다음과 같은 성질을 만족하는 $P(\cdot)$를 확률측도(Probability Measure) 라고 한다.

(1) 표본공간 $\mathcal{S}$에서 임의의 사건 A에 대하여 $0 \leq P(A)$

(2) $P(S) = 1$

(3) 서로 배반인 사건 $A_1, A_2, \ldots$에 대하여

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

위의 정의로부터 나오는 성질

- $P(\varnothing) = 0$
- $A \subset B$ 이면 $P(A) \leq P(B)$
- $0 \leq P(A) \leq 1$
- $P(A^C) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

조건부 확률(Conditional Probability)

- 사건 A가 주어졌을 때 사건 B의 조건부확률은 $P(B|A)$로 나타내고 $P(A) > 0$이라는 가정하에 다음과 같이 정의

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

- 사건 A를 새로운 축소된 표본공간으로 간주했을 때, 사건 B가 일어날 확률

  예제: 세 개의 동전을 차례로 던지는 경우, 앞면이 나온 수가 2 (A)일때, 첫번째 던지기에서 앞면이 나올 (B) 확률은?

곱셈법칙(Multiplication Rule)

$P(A) > 0, P(B) > 0$이면

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

예제: 불량품 10개와 정상품 90개가 들어있는 상자에서 2개를 단순 랜덤추출(무작위추출, 비복원)할 때, 2개 모두 불량품일 확률을 구하여라.

전확률공식(Law of Total Probability)

어떤 사건 B 의 확률 $P(B)$을 구할때, 표본공간의 분할정보를 이용하는 공식

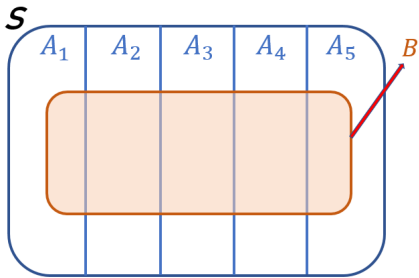
- 표본공간 $\mathcal{S}$의 분할 $\{A_1, \cdots, A_n\}$을 생각하자. 표본공간의 분할은 다음을 만족한다.

  $A_i \cap A_j = \emptyset \ (i \neq j)$, $A_1 \cup A_2 \cup \cdots \cup A_n = \mathcal{S}$

  이때, 전확률공식은

$$P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n)$$

$n = 5$일때



$$
\begin{aligned}
P(B) &= P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) + P(B \cap A_4) + P(B \cap A_5) \\
&= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) \\
&\quad + P(B|A_4)P(A_4) + P(B|A_5)P(A_5)
\end{aligned}
$$

전확률공식 예제

냉장고/김치냉장고 제조 회사는 총 4개의 공장 (A,B,C,D) 을 가지고 있다. 각 공장의 생산량(%)은 다음과 같다. 30% (A), 25% (B), 25% (C) 그리고 20% (D). 그런데 A공장의 50%, B공장의 30%, C공장의 10%, D공장의 2%가 김치냉장고 생산 비율이라고 하자. 하나의 제품을 단순랜덤추출했을 때 그 제품이 김치냉장고일 확률을 구하여라.

독립사건

두 사건 A, B가 서로 독립(mutually independence)

- 사건 A가 일어났다고 하더라도 사건 B가 일어날 확률에 아무런 영향을 미치지 않는 것
- $P(B|A) = P(B)$ 또는 $P(A \cap B) = P(A)P(B)$
- 두 사건 A와 B가 독립이 아니면 종속이라고 한다.

  $<$참고 $>$ $A \cap B = \varnothing$인 두 사건 A와 B는 서로 배반(mutually disjoint), 즉 두 사건이 동시에 일어날 수 없음을 의미하고 A와 B는 종속 사건이다.

두 개 이상의 사건들의 독립

- 사건 $A_1, \ldots A_n$이 서로 독립(mutually independent)이다
  $\iff$

$$\forall 1 \leq i_1 < \cdots < i_k \leq n \; (2 \leq k \leq n)$$

$$P(\cap_{j=1}^{k} A_{i_j}) = \prod_{j=1}^{k} P(A_{i_j})$$

- $A,\ B$가 독립사건이면 $A^C,\ B$ (또는 $A,\ B^C$ 또는 $A^C,\ B^C$)가 독립

예제

불량품 20개와 정상품 80개로 이루어진 100개들이 박스에서 2개의 제품을 단순랜덤추출할 때, 첫번째 제품이 불량품일 사건을 A, 두번째 제품이 불량품일 사건을 B라 하면 A와 B는 독립인가?

독립시행

- 동전을 두 번 연속으로 던졌을 때와 두 개의 동전을 동시에 던졌을 때 하나의 결과가 다른 결과에 영향을 주지 않는다고 생각한다.
- 이 때 두 결과는 서로 독립적으로 일어난다고 한다.
- 이렇게 두 개 이상의 사건이 독립적으로 일어나는 것을 독립 시행이라고 한다.
- 두 사건이 서로 독립이면, 표본 공간은 각각의 사건에 대한 표본 공간으로 부터 얻어지는 모든 가능한 조합으로 구성된다.
- 두 사건이 서로 독립인 경우 두 개의 사건의 결과에 대한 확률은 각각의 사건의 확률을 곱해주면 된다.

베이즈 정리**(Bayes' Theorem)**

- $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$
- 두 사건 A, B의 확률 P(A),P(B), 조건부 확률 $P(B|A) = \frac{P(A \cap B)}{P(A)}$ 를 알고 있을때, $P(A \mid B)$를 구함.
- $P(B)$ 는 전확률공식을 이용하여 $P(B|A)P(A) + P(B|A^c)P(A^c)$로 바꿀수 있다.

베이즈 정리 예제

- 전체 국민의 0.1% 가 걸리는 특정 질병에 대해 질병에 걸렸는지 판정하는 검사법이 있다고 하자. 질병이 없는데 있다고 틀리게 판정할 확률이 5%, 질병이 있는데 질병이 있다고 맞게 판정할 확률은 100% 라고 할때, 검사법에 의해 질병이 있다고 판정받은 사람이 실제로 질병이 걸린 사람의 확률은?

확률변수와 확률분포

확률변수란?

- 관찰, 측정, 실험 등의 과정을 통해 얻어진 우연성을 가진 결과들을 우리가 다루기 쉬운 숫자로 표현한 것
  - 동전을 던지는 실험에서 결과인 '앞면' 과 '뒷면'을 각각 1과 0으로 표현

확률변수의 수학적 정의

- 앞의 예시에서 동전을 던지는 실험에서 나오는 결과는 표본공간으로 설명된다. 즉, $\mathcal{S} = \{H, T\}$ (앞면 : H, 뒷면 : T)
- 앞면은 1이고 뒷면을 0으로 표현한다는 것은 표본공간의 각 원소를 숫자로 대응시킨다는 뜻이다.
- 즉, 확률변수는 표본공간의 각 원소를 하나의 숫자로 대응하는 함수라고 할 수 있다.

X를 확률변수라고 하면, $c \in \mathcal{S}$에 대해

$$X(c) = x \in \mathbb{R}$$

- 앞의 예시에서 1과 0으로 대응시키는 확률변수를 X라고 하면, $X(H) = 1,\ X(T) = 0$ 이라고 쓸 수 있다.
- 이때, 함수 X의 정의역은 $\mathcal{S} = \{H, T\}$, 치역은 $\{0, 1\}$이 된다.

확률변수의 확률

확률변수가 특정 값을 가지는 가능성은 어떻게 표현할까?

- 앞의 예시에서 $X = 1$일 확률은?
- 동전 1개를 던졌을 때, 앞면이 나올 확률과 같다.
- 따라서, 확률변수 X가 1의 값을 가질 확률을 다음과 같이 표현한다.
- $P(X = 1) = P(\{c \in \mathcal{S} \mid X(c) = 1\}) = P(\{H\})$

확률분포

- 데이터를 확률변수의 실현값(또는 관측값)으로 본다면 특정 값이 관측되는 가능성(확률)을 확률변수가 특정 값을 가지는 확률로 설명할 수 있다.
- 그렇다면 확률변수가 어떤 정의역과 치역으로 구성되는 함수인지 어떤 확률구조(확률측도)를 가지는지를 알아야 한다.
- 주 관심사는 데이터의 값과 확률이므로 확률변수와 확률측도의 구체적인 정보를 모르더라도 확률변수의 값 (함수값)과 그에 대응하는 확률을 구할 수 있는 정보만 있어도 충분하다. 이를 확률변수의 분포, 확률분포라고 한다.

확률분포

확률변수 X의 확률분포란: 확률변수 X가 가질 수 있는 값과 해당하는 확률에 대해 나타낸 것으로, 확률을 계산 할 수 있는 정보를 제공.

- X 가 취할 수 있는 값들의 종류에 따라서 일반적으로 두 가지로 나눌 수 있다.
- 이산확률변수(Discrete r.v.): X가 취할 수 있는 값이 $x_1, x_2, x_3, \ldots$와 같이 이산 일 때
  - 해당 값과 대응하는 확률을 제공
- 연속확률변수(Continuous r.v.): X의 취할 수 있는 값이 셀 수 없이 많을 때
  - 특정 구간에 속하는 확률을 계산할 수 있는 정보를 제공.
- 관측한 데이터가 이산이면 이산확률변수의 실현값, 연속이면 연속확률변소의 실현값으로 본다.

이산확률변수의 확률분포

확률분포는 다음과 같은 확률질량함수 (Probability Mass Function, pmf) $p(x)$로 표현 가능

$$p(x) = P(X = x) = \begin{cases} P(X = x_i) & , x = x_i\text{일 때}(i = 1, 2, \dots) \\ 0 & , otherwise. \end{cases}$$

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

예제

15개의 상품 중 5개가 불량품이다. 3개를 단순랜덤추출하였을 때, 그 중 불량품의 개수를 X라 하자. 확률변수 X의 확률분포를 구하여라.

연속확률변수의 확률분포

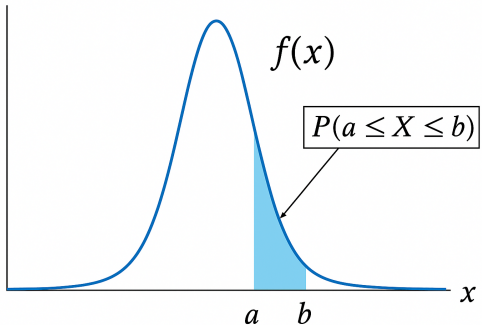
확률분포는 확률밀도함수 (Probability Density Function, pdf) $f(x)$를 도입하여 X의 값이 $a \leq X \leq b$일 확률로 표현

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

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

연속확률변수의 성질

- 연속확률변수의 한 점에서의 확률은 0이다. $P(X = a) = 0$
- $P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b)$
  $= P(a < X < b)$



예제

연속확률변수 X의 확률밀도함수가

$$f(x) = \begin{cases} bx(1-x) & (0 \le x \le 1\text{일 때}) \\ 0 & (x < 0 \text{ 또는 } x > 1\text{일 때}) \end{cases}$$

로 주어질 때, 상수 b와 확률 $P(0 \le X \le \frac{3}{4})$를 구하여라.

누적확률분포함수

- pmf, pdf 외에 확률분포를 나타내는 또 다른 함수로 누적확률분포함수(Cumulative Distribution Function, CDF)가 있다.
- $F_X(x) = P(X \leq x)$로 정의되며 x까지의 누적된 확률을 주는 함수이다.
- 이산확률변수, 연속확률변수에 상관없이 정의 된다.
- Non-decreasing 함수
- 연속확률변수의 경우: $\frac{d}{dx}F(x) = f(x)$

확률분포의 요약

- 데이터가 어떤 확률변수의 실현값인지 알면, 해당 확률변수의 확률분포를 통해 데이터의 특징을 파악할 수 있다.
- 데이터로 구한 분포 (경험적 분포)를 통해 어떤 확률분포를 가지는 확률변수의 관측값으로 생각할지를 파악할 수 도 있다.
- 한편, 확률분포의 특징을 요약한 정보가 유용한 경우도 많다. 이러한 요약한 정보로 평균(기댓값)과 분산이 있다.

기댓값

- 기댓값(Expected value)은 확률변수 X가 가질 수 있는 값들의 중심을 나타내는 값으로 평균(Mean)이라고도 부른다.

$$\mu = E(X) = \begin{cases} \sum_x xp(x) & \text{(이산확률변수)} \\ \int_{-\infty}^{\infty} xf(x)dx & \text{(연속확률변수)} \end{cases}$$

  예제: 동전을 2회 던지는 실험에서 앞면의 개수를 X라고 할 때, X의 기댓값을 구하여라.

- 1차 적률(First Moment)이라고도 부른다.

기댓값의 성질

- 확률변수 X의 함수 $g(X)$의 기대값은 다음과 같이 정의된다.

$$E(g(X)) = \begin{cases} \sum_x g(x)p(x) & \text{(이산확률변수)} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{(연속확률변수)} \end{cases}$$

- 따라서 기댓값이 선형성을 가짐을 보일 수 있다.
  - $E(aX + b) = aE(X) + b$ (a, b는 상수)
  - $E[ag(X) + bh(X)] = aE(g(X)) + bE(h(X))$ (a, b는 상수)

분산과 표준편차

- 분산(Variance)은 확률변수 X가 가질 수 있는 값들이 얼마나 퍼져있는지를 그 변동성을 나타내는 값으로 다음과 같이 정의한다.

$$\sigma^2 = Var(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 p(x) & \text{(이산확률변수)} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx & \text{(연속확률변수)} \end{cases}$$

- 분산의 양의 제곱근을 표준편차(Standard Deviation)라고 한다.

$$SD(X) = \sqrt{Var(X)}$$

- 분산의 성질은 다음과 같다.

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

예제

X의 확률밀도함수가 $f(x) = \begin{cases} 1 & , 0 \leq x \leq 1 \\ 0 & , otherwise \end{cases}$ 일 때,
X의 평균, 분산, 표준편차를 구하여라.

The End