3. 확률변수와 확률분포

이경재

인하대학교 통계학과

March 17, 2019

- Random variable(X): a (measurable) function from the sample space S (or $\mathcal{X}$) to $\mathbb{R}$.
- Observation(x): the data we have observed, i.e., a realization value of X.
- Parameter(θ): the value describing (or characterizing) the distribution.

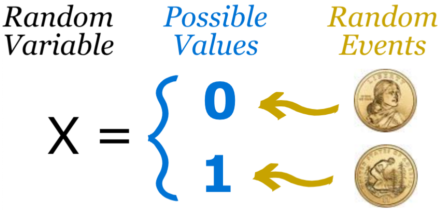

Figure: 출처: https://www.mathsisfun.com/

Discrete Distribution (이산분포)

- 이산 확률변수: 표본공간이 셀 수 있는(countable) 집합인 확률변수. (e.g.) 동전의 앞(1), 뒤(0). 교통사고 건수.
- 각 $x \in \mathcal{X}$에 대해, $f(x) := P(X = x)$를 X의 확률밀도(또는 질량) 함수라 하고, 이는 아래 성질을 만족한다:
  1. $0 \leq f(x) \leq 1$
  2. $\sum_{x \in \mathcal{X}} f(x) = 1$
  3. 임의의 배반사건 A와 B에 대하여,

$$\begin{aligned} P(X \in A \cup B) &= P(X \in A) + P(X \in B) \\ &= \sum_{x \in A} f(x) + \sum_{x \in B} f(x). \end{aligned}$$

Review: Poisson Distribution (포아송 분포)

$$f(x \mid \theta) = \frac{e^{-\theta}\theta^{x}}{x!}, \quad x = 0, 1, \ldots, \quad \theta > 0$$

- $X \sim Poi(\theta)$로 표기.
- 단위 시간(또는 공간)당 특정 사건의 발생 건수에 대한 확률분포
- θ는 단위 시간당 평균적으로 발생하는 사건의 수
- (e.g.) 하루 동안 걸려오는 전화 통화 수
- $\mathrm{E}(X) = \mathrm{Var}(X) = \lambda$

Review: Negative Binomial Distribution (음이항 분포)

$$f(x \mid \theta) = \binom{x+r-1}{x} \theta^r (1-\theta)^x, \quad x = 0, 1, \ldots, \quad 0 < \theta < 1$$

- $X \sim NB(r, \theta)$로 표기.
- 베르누이 시행을 독립적으로 반복할 때, r번째 성공전까지 총 실패 횟수의 확률분포
- $\mathrm{E}(X) = \frac{\theta r}{1-\theta}$
- $\mathrm{Var}(X) = \frac{\theta r}{(1-\theta)^2}$

Continuous Distribution (연속분포)

- 연속 확률변수: 일정 구간의 모든 값을 가질 수 있는 확률변수

  (e.g.) 스마트폰 배터리가 0% 될 때까지 걸린 시간

- 확률밀도함수는 (존재한다면) $f(x) = \frac{d}{dx}F(x) = \frac{d}{dx}P(X \leq x)$로 정의

- 이산분포의 경우와 달리, $f(x)$는 x에서의 확률을 나타내지 않는다.

- $f(x)$는 다음의 성질을 만족한다:

  1. $f(x) \geq 0, \quad \forall x \in \mathbb{R}$
  2. $\int_{-\infty}^{\infty} f(x)dx = 1$
  3. 배반 집합 A와 B에 대하여,

$$\begin{aligned} P(X \in A \cup B) &= P(X \in A) + P(X \in B) \\ &= \int_{x \in A} f(x)dx + \int_{x \in B} f(x)dx. \end{aligned}$$

Review: Gamma Distribution (감마분포)

$$f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0,\ \alpha, \beta > 0$$

- $X \sim Gamma(\alpha, \beta)$로 표기
- $\mathrm{E}(X) = \alpha/\beta$, $\mathrm{Var}(X) = \alpha/\beta^2$
- (지수분포) $Exp(\beta) = Gamma(1, \beta)$
- (카이제곱분포) $\chi^2_\nu = Gamma(\nu/2, 1/2)$

Review: Inverse gamma Distribution (역감마분포)

$$f(x \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}, \quad x > 0, \alpha, \beta > 0$$

- $X \sim IG(\alpha, \beta)$로 표기
- $Y \sim Gamma(\alpha, \beta) \implies X = 1/Y \sim IG(\alpha, \beta)$
- $\mathrm{E}(X) = \beta/(\alpha - 1)$ if $\alpha > 1$

  $\mathrm{Var}(X) = \beta^2/\{(\alpha - 1)^2(\alpha - 2)\}$ if $\alpha > 2$

Review: Beta Distribution (베타분포)

$$f(x \mid \alpha, \beta) \quad = \quad \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1,\ \alpha, \beta > 0$$

- $X \sim Beta(\alpha, \beta)$로 표기
- $\Gamma(\cdot)$는 감마함수: $\Gamma(z) := \int_0^\infty x^{z-1} e^{-x} dx$
- $\mathrm{E}(X) = \alpha/(\alpha+\beta)$, $\mathrm{Var}(X) = (\alpha\beta)/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$
- 만약 $X \sim Gamma(\alpha, \theta)$, $Y \sim Gamma(\beta, \theta)$이고 서로 독립이라면, $X/(X+Y) \sim Beta(\alpha, \beta)$이다.

Review: Normal Distribution (정규분포)

$$f(x \mid \mu, \sigma^2) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \quad x \in \mathbb{R},\ \mu \in \mathbb{R},\ \sigma > 0$$

- $X \sim N(\mu, \sigma^2)$로 표기
- $\mathrm{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$

Review: Student *t*-distribution (스튜던트 *t* 분포)

$$f(x \mid \nu, \mu, \sigma^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left\{ 1 + \frac{(x-\mu)^2}{\nu\sigma^2} \right\}^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R}, \mu \in \mathbb{R}, \nu, \sigma > 0$$

- $\mathrm{E}(X) = \mu$ if $\nu > 1$
- $\mathrm{Var}(X) = \sigma^2\nu/(\nu-2)$ if $\nu > 2$
- (코쉬 분포) $\nu = 1$인 스튜던트 t 분포

베이즈 정리: 연속분포

- $X \in \mathbb{R}$ and we observe $X = x$.
- $\theta \in \mathbb{R}$: parameter of interest (unknown)
- $X \sim f(X \mid \theta)$ and $\theta \sim \pi(\theta)$
- Then the posterior is

$$\begin{aligned}
\pi(\theta \mid x) &= \frac{f(x, \theta)}{f(x)} \\
&= \frac{\pi(\theta) f(x \mid \theta)}{f(x)} \\
&= \frac{\pi(\theta) f(x \mid \theta)}{\int \pi(\theta) f(x \mid \theta) d\theta}.
\end{aligned}$$

변수의 조건부 독립성(Conditional independence)

- $X_1, \ldots, X_n$: random variables from the same sample space
- θ: parameter of interest (unknown)
- If the following equality

$$P(X_1 \in A_1, \ldots, X_n \in A_n \mid \theta) \quad = \quad P(X_1 \in A_1 \mid \theta) \times \cdots \times P(X_n \in A_n \mid \theta)$$

holds for any subsets $A_1, \ldots, A_n \in S$, we say that $X_1, \ldots, X_n$ are conditionally independent given θ.

변수의 조건부 독립성(Conditional independence)

- If $X_1, \ldots, X_n$ are conditionally independent given θ and have the same distribution,

$$f(x_1, \ldots, x_n \mid \theta) \quad = \quad \prod_{i=1}^{n} f(x_i \mid \theta).$$

- Then we say that $X_1, \ldots, X_n$ are conditionally independent and identically distributed (conditionally iid).
- Note that this is different from iid, which means

$$f(x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^{n} f(x_i).$$

변수의 조건부 독립성(Conditional independence)

- If $X_1, \ldots, X_n$ are conditionally independent given θ and have the same distribution,

$$f(x_1, \ldots, x_n \mid \theta) \quad = \quad \prod_{i=1}^{n} f(x_i \mid \theta).$$

- Then we say that $X_1, \ldots, X_n$ are conditionally independent and identically distributed (conditionally iid).
- Note that this is different from iid, which means

$$f(x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^{n} f(x_i).$$

교환가능성(Exchangeability)

- Sometimes, independence between the variables is hard to obtain.
  - (e.g.) sampling from a finite population without replacement
- Exchangeability of random variables is less restrictive than independence, but still useful.

교환가능성(Exchangeability)

- Sometimes, independence between the variables is hard to obtain.
  - (e.g.) sampling from a finite population without replacement
- Exchangeability of random variables is less restrictive than independence, but still useful.

교환가능성(Exchangeability)

- Let $X_1, \ldots, X_n \sim f(x_1, \ldots, x_n)$ and τ be a permutation of $\{1, \ldots, n\}$. If

$$f(x_1, \ldots, x_n) = f(x_{\tau(1)}, \ldots, x_{\tau(n)})$$

for any permutation τ, then $X_1, \ldots, X_n$ is said exchangeable.

Example: Exchangeability

- $\theta \in (0, 1)$: 전체 인원 중 통계학에 관심 있는 사람의 비율
- $X_i \in \{0, 1\}$: 랜덤으로 택한 사람 중 i번째 사람의 통계학 관심여부
- 비복원추출을 하더라도, 전체 인원이 충분히 많다고 가정하면 X_i들은 θ가 주어졌을 때 서로 조건부 독립일 것이다:

$$P(X_i = x_i \mid \theta, X_j, j \neq i) = P(X_i = x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}.$$

- 따라서, 조건부 결합밀도함수는 다음과 같다:

$$\begin{aligned} f(x_1, \ldots, x_n \mid \theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}. \end{aligned}$$

Example: Exchangeability

- $\theta \in (0, 1)$: 전체 인원 중 통계학에 관심 있는 사람의 비율
- $X_i \in \{0, 1\}$: 랜덤으로 택한 사람 중 i번째 사람의 통계학 관심여부
- 비복원추출을 하더라도, 전체 인원이 충분히 많다고 가정하면 X_i들은 θ가 주어졌을 때 서로 조건부 독립일 것이다:

$$P(X_i = x_i \mid \theta, X_j, j \neq i) = P(X_i = x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}.$$

- 따라서, 조건부 결합밀도함수는 다음과 같다:

$$\begin{aligned} f(x_1, \ldots, x_n \mid \theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}. \end{aligned}$$

Example: Exchangeability

- $\theta \in (0,1)$: 전체 인원 중 통계학에 관심 있는 사람의 비율
- $X_i \in \{0,1\}$: 랜덤으로 택한 사람 중 i번째 사람의 통계학 관심여부
- 비복원추출을 하더라도, 전체 인원이 충분히 많다고 가정하면 X_i들은 θ가 주어졌을 때 서로 조건부 독립일 것이다:

$$P(X_i = x_i \mid \theta, X_j, j \neq i) = P(X_i = x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}.$$

- 따라서, 조건부 결합밀도함수는 다음과 같다:

$$\begin{aligned} f(x_1, \ldots, x_n \mid \theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}. \end{aligned}$$

Example: Exchangeability

- $X_1, \ldots, X_n$의 주변 결합밀도함수는 다음과 같다:

$$\begin{aligned} f(x_1, \ldots, x_n) &= \int_0^1 f(x_1, \ldots, x_n \mid \theta)\pi(\theta)d\theta \\ &= \int_0^1 \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}\pi(\theta)d\theta. \end{aligned}$$

- 만약 $n = 10$이고, $x = (1, 0, 0, 1, 0, 1, 1, 0, 0, 0)$, $x' = (0, 1, 1, 0, 0, 1, 0, 1, 0, 0)$이라 하자. 즉, $x \neq x'$이지만 $\sum_i x_i = \sum x_i'$이다.
- 위의 주변 결합밀도함수 식에 의해,

$$f(x) = f(x'),$$

즉 $X_1, \ldots, X_n$는 교환가능하다.

Example: Exchangeability

- $X_1, \ldots, X_n$의 주변 결합밀도함수는 다음과 같다:

$$\begin{aligned} f(x_1, \ldots, x_n) &= \int_0^1 f(x_1, \ldots, x_n \mid \theta)\pi(\theta)d\theta \\ &= \int_0^1 \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}\pi(\theta)d\theta. \end{aligned}$$

- 만약 $n = 10$이고, $x = (1, 0, 0, 1, 0, 1, 1, 0, 0, 0)$, $x' = (0, 1, 1, 0, 0, 1, 0, 1, 0, 0)$이라 하자. 즉, $x \neq x'$이지만 $\sum_i x_i = \sum x_i'$이다.
- 위의 주변 결합밀도함수 식에 의해,

$$f(x) = f(x'),$$

즉 $X_1, \ldots, X_n$는 교환가능하다.

Example: Exchangeability

- $X_1, \ldots, X_n$는 독립일까?
- $\pi(\theta) = 1, \theta \in (0,1)$, 즉 $\theta \sim Unif(0,1)$이라 하자.
- (X_i의 주변 밀도함수) $P(X_i = 1) = \int_0^1 \theta(1-\theta)^0 d\theta = \frac{1}{2}$
- ($X_1, \ldots, X_n$의 결합 밀도함수)

$$
\begin{aligned}
P(X_1 = 1, \ldots, X_n = 1) &= \int_0^1 \theta^n (1-\theta)^0 d\theta = \frac{1}{n+1} \\
&\neq \frac{1}{2^n} = \prod_{i=1}^{n} P(X_i = 1)
\end{aligned}
$$

- 따라서, $X_1, \ldots, X_n$는 독립이 아니다.

Example: Exchangeability

- $X_1, \ldots, X_n$는 독립일까?
- $\pi(\theta) = 1, \theta \in (0,1)$, 즉 $\theta \sim Unif(0,1)$이라 하자.
- (X_i의 주변 밀도함수) $P(X_i = 1) = \int_0^1 \theta(1-\theta)^0 d\theta = \frac{1}{2}$
- ($X_1, \ldots, X_n$의 결합 밀도함수)

$$
\begin{aligned}
P(X_1 = 1, \ldots, X_n = 1) &= \int_0^1 \theta^n (1-\theta)^0 d\theta = \frac{1}{n+1} \\
&\neq \frac{1}{2^n} = \prod_{i=1}^n P(X_i = 1)
\end{aligned}
$$

- 따라서, $X_1, \ldots, X_n$는 독립이 아니다.

Example: Exchangeability

- $X_1, \ldots, X_n$는 독립일까?
- $\pi(\theta) = 1, \theta \in (0,1)$, 즉 $\theta \sim Unif(0,1)$이라 하자.
- (X_i의 주변 밀도함수) $P(X_i = 1) = \int_0^1 \theta(1-\theta)^0 d\theta = \frac{1}{2}$
- ($X_1, \ldots, X_n$의 결합 밀도함수)

$$\begin{aligned} P(X_1 = 1, \ldots, X_n = 1) &= \int_0^1 \theta^n (1-\theta)^0 d\theta = \frac{1}{n+1} \\ &\neq \frac{1}{2^n} = \prod_{i=1}^n P(X_i = 1) \end{aligned}$$

- 따라서, $X_1, \ldots, X_n$는 독립이 아니다.

Example: Exchangeability

- $X_1, \ldots, X_n$는 독립일까?
- $\pi(\theta) = 1, \theta \in (0,1)$, 즉 $\theta \sim Unif(0,1)$이라 하자.
- (X_i의 주변 밀도함수) $P(X_i = 1) = \int_0^1 \theta(1-\theta)^0 d\theta = \frac{1}{2}$
- ($X_1, \ldots, X_n$의 결합 밀도함수)

$$\begin{aligned} P(X_1 = 1, \ldots, X_n = 1) &= \int_0^1 \theta^n (1-\theta)^0 d\theta = \frac{1}{n+1} \\ &\neq \frac{1}{2^n} = \prod_{i=1}^n P(X_i = 1) \end{aligned}$$

- 따라서, $X_1, \ldots, X_n$는 독립이 아니다.

Question

- 이 예제를 일반화시키면, $\theta \sim \pi(\theta)$이고 $X_1, \ldots, X_n$가 θ가 주어졌을 때 조건부 독립이며 동일한 분포 $f(x \mid \theta)$를 따르면, $X_1, \ldots, X_n$은 교환가능하다.
- 위의 역도 성립하는가? 즉, $X_1, \ldots, X_n$은 교환가능하다면,

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

를 만족시키는 모수 θ와 밀도함수 $f(x_i \mid \theta)$, 그리고 θ의 밀도함수 $\pi(\theta)$가 존재하는가?

Question

- 이 예제를 일반화시키면, $\theta \sim \pi(\theta)$이고 $X_1, \ldots, X_n$가 θ가 주어졌을 때 조건부 독립이며 동일한 분포 $f(x \mid \theta)$를 따르면, $X_1, \ldots, X_n$은 교환가능하다.
- 위의 역도 성립하는가? 즉, $X_1, \ldots, X_n$은 교환가능하다면,

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

를 만족시키는 모수 θ와 밀도함수 $f(x_i \mid \theta)$, 그리고 θ의 밀도함수 $\pi(\theta)$가 존재하는가?

De Finetti 정리

Let $X_1, \ldots, X_n$ be exchangeable random variables and $(X_1, \ldots, X_n) \sim f(x_1, \ldots, x_n)$. Then there exist a random variable $\theta \sim \pi(\theta)$ and a pdf $f(x \mid \theta)$ such that

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

- The last equality implies

$$f(x_1, \ldots, x_n) = \int \prod_{i=1}^{n} f(x_i \mid \theta)\pi(\theta) d\theta.$$

De Finetti 정리의 의미

- De Finetti 정리는, 교환가능성이 있는 자료에 대하여
  1. 조건부 iid가 성립하는 자료의 조건부 분포 $f(x_i \mid \theta)$와
  2. 모수 θ의 사전분포 $\pi(\theta)$

  가 반드시 존재한다는 것을 암시한다.