DSL Seminar: 통입+통방 (3)

Kyung-han Kim

Data Science Lab

January, 2023

목차

- 연속형 확률변수
- 정규분포와 정규분포의 표준화
- 이항분포와 정규분포의 관계
- 다른 연속형 확률분포
  - 지수분포
  - 균일분포
  - t분포
  - F분포
  - 카이제곱 (χ^2) 분포

Review

- 확률변수
- 확률분포
- 이산형 확률변수, 확률분포
- 이항분포
- 포아송분포

연속형 확률변수 (Continuous Random Variable)

- 확률변수가 가질 수 있는 값이 무한한 (infinite) 확률변수
- Ex] 우리나라 성인 남성의 **몸무게**
- Ex] 건전지의 **수명**
- 연속형 확률변수 X 는 가질 수 있는 값이 무수히 많고,
  확률은 구간 별로 정의됩니다.
- 확률밀도함수 (Probability Density Function, pdf) 를 이용해 연속형 확률변수에서 원하는 구간의 확률을 구할 수 있습니다.
- Ex] $f(x) = \lambda e^{-\lambda x}$ (지수분포의 pdf)
- 모든 확률의 합은 1이므로, 확률변수 X 가 구간 (a, b) 에서 정의될 때 $\int_a^b f(x)dx = 1$ 입니다.
  - 이산확률변수: $\sum_{\text{all } i} P(X = x_i) = 1$
- cf] 확률밀도함수 자체의 함숫값은 큰 의미는 없습니다.

정규분포 (Normal Distribution)

- 통계학입문 수업에서 가장 중점적으로 다루는 연속형 확률분포입니다.
- 우리가 일반적으로 생각할 수 있는 예시들의 상당수가 정규분포를 따릅니다.
  - Why?
- 모수는 총 2개입니다.
  - 1) 평균 (μ) (범위 제한 없음)
  - 2) 분산 (σ^2) or 표준편차 (σ) (분산과 표준편차는 0보다 커야 함)
  - μ를 Location parameter, σ^2를 Scale parameter라고 하기도 합니다.
- 확률변수 X가 평균이 μ이고 분산이 σ^2인 정규분포를 따른다면, 이를 $X \sim N(\mu, \sigma^2)$로 표기합니다.
- pdf: $f(x) = f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$
- $E[X] = \mu$, $V[X] = \sigma^2$ (증명 생략)
- 정규분포를 가우시안 분포 (Gaussian Distribution)이라고도 합니다.

정규분포의 성질 및 특징

1. 확률밀도함수가 종 모양 (Bell-shaped) 이다.
2. 평균을 중심으로 좌우 대칭이다.
3. 분산이 커질수록 확률밀도함수가 넓게 퍼지고, 작아질수록 평균 주위에 몰린다.
   - 어느 쪽이 더 '고른' 확률분포일까?
4. $\int_{-\infty}^{\infty} f(x)dx = 1$

표준정규분포 (Standard Normal distribution)

- 정규분포 중에서 평균이 0이고 분산이 1인 분포를 표준정규분포 (Z)라고 합니다. ($N(0,1) = Z$)
- 모든 정규분포는 간단한 변수 변환 (Variable Transformation)을 통해 표준정규분포로 바꿀 수 있으며, 이를 표준화 (Standardization)라고 합니다.

$$Z = \frac{X - \mu}{\sigma}$$

- 표준화를 통해 정규분포를 따르는 확률변수의 구간 별 확률을 편리하게 구할 수 있습니다. (표준정규분포표 활용)
- $P(Z < 1.96) = 0.975$
- $P(Z < 1.65) = 0.95$

이항분포의 정규 근사

- 이항분포가 특정 조건을 만족시킬 경우 정규분포로 근사시킬 수 있습니다.
- $X \sim B(n, p)$일 때 $np > 5, n(1-p) > 5$이면
  $X \sim N(np, np(1-p))$로 근사시킬 수 있습니다.
  - 참고: $np < 5$이면 포아송분포에 더 가까워집니다.
- 평균과 분산이 그대로 유지된다고 생각하면 편합니다.
- 증명은 생략합니다. (MGF, Taylor expansion 필요)

지수분포 (Exponential Distribution)

- 지수분포는 포아송분포와 밀접한 관련이 있습니다.
- 포아송분포는 발생할 확률이 희박한 사건을 표현할 때 사용할 수 있습니다.
- 지수분포는 포아송분포를 따르는 어떤 사건이 발생할 때까지 걸리는 시간이 따르는 확률분포가 됩니다.
- 모수는 1개입니다. (λ: 발생률)
- $X \sim \text{Exp}[\lambda]$ 일 때 pdf: $f(x) = \lambda e^{-\lambda x}$
- $E[X] = \frac{1}{\lambda}$, $V[X] = \frac{1}{\lambda^2}$
- 포아송분포의 모수가 λ로 표기되므로, 헷갈리지 않기 위해 pdf를 $f(x) = \frac{1}{\theta} e^{-\frac{1}{\theta}x}$ 로 쓰는 경우도 있습니다.
  - 이 경우 $E[X] = \theta$, $V[X] = \theta^2$ 입니다. (θ: 평균 발생 시간)

지수분포와 포아송분포의 관계

- $X \sim \text{Pois}[\lambda]$ 라면, X 는 어떠한 사건이 단위 시간 당 λ 번 발생함을 의미합니다.
- 그렇다면 t 시간 동안에는 해당 사건이 λt 번 발생할 것입니다.
- 이러한 포아송분포의 pmf는 $\frac{(\lambda t)^x e^{-\lambda t}}{x!}$ 일 것입니다.
- 여기에서 t시간 동안 사건이 한 번도 발생하지 않을 확률은 $e^{-\lambda t}$ 입니다.
- 그런데 t시간 동안 사건이 한 번도 발생하지 않았다는 것은, 사건이 처음 발생할 때까지 시간이 t 보다 더 오래 걸린다는 뜻입니다.
- 따라서 이 확률을 지수분포를 이용해 구해 보자면, $\int_t^\infty \lambda e^{-\lambda x} dx = [-e^{-\lambda x}]_t^\infty = e^{-\lambda t}$ 가 됩니다.
- 즉, 지수분포와 포아송분포는 같은 상황을 다른 관점으로 해석하는 것으로 볼 수 있습니다.

기타 연속형 확률분포

- 균일분포: $f(x) = \frac{1}{b-a}, x \in [a, b]$
  - $E[X] = \frac{a+b}{2}, V[X] = \frac{(b-a)^2}{12}$
- t분포: 정규분포에서 꼬리가 약간 더 두꺼운 분포
  - 자세한 내용은 표본평균의 분포를 다룬 뒤에 설명하겠습니다.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{이면, } \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t[n-1].$$

- χ^2 분포: 표준정규분포를 따르는 확률변수들의 제곱의 합이 따르는 분포
  - $Z_1, Z_2, \cdots, Z_n \sim Z$(iid) 일 때 $\sum_{i=1}^{n} Z_i^2 \sim \chi^2[n]$
  - $E[X] = n, V[X] = 2n$
- F 분포: 두 χ^2 분포 확률변수를 자유도를 나눈 것의 비율이 따르는 분포

$$\frac{\chi_1^2/df_1}{\chi_2^2/df_2} \sim F[df_1, df_2]$$