DSL Seminar: 통입+통방 (2)

Kyung-han Kim

Data Science Lab

January, 2023

Contents

- 확률변수 (Random Variable)
- 확률분포 (Probability distribution)
  - 이산형 확률분포 (Discrete random variable): 이항분포
  - 기타 이산형 확률분포들
- 공분산 (Covariance), 상관계수 (Correlation coefficient)

Review

- 통계학은 **모집단**에서 **표본**을 추출하고,
  그 표본에서 얻은 **추정량**으로 **모수**를 추정하는 학문이다.
  - 모수? (Parameter)
- 평균과 분산
  - 모평균, 모분산의 정의
- 확률
- 조건부 확률과 베이즈 정리

확률변수 (Random Variable)

- 우리의 주변에서는 다양한 사건 (Event) 들이 발생합니다.
- 그 사건을 수학적/통계학적으로 분석 가능한 형태로 변환해 주는 것이 확률변수라고 생각할 수 있습니다.
- 보다 정확하게는, 확률변수를 아래와 같이 받아들일 수 있습니다: 확률변수는 각각의 사건에 숫자를 대응시키는 **함수**다.
- 같은 사건이더라도, 숫자를 어떻게 대응시키냐에 따라 다른 확률변수가 될 수 있다.
- Ex] 주사위 3개를 던지는 시행의 결과, $(5, 2, 1)$ 이 나왔다고 하자.
  - X_1 (홀수의 개수): 2
  - X_2 (세 눈의 총합): 8
  - X_3 (가장 작은 값): 1

확률분포 (Probability Distribution)

- 결과가 우연에 의해 결정되는 사건이라면 이론적으로 어떤 것이든지 확률변수로 표현할 수 있습니다.
- 우리는 이러한 확률변수를 설정한 후, 각각의 확률변수 값이 발생할 확률을 계산합니다.
- 이러한 과정을 거쳐 실제 상황을 수학적/통계학적으로 표현할 수 있고, 이를 바탕으로 추가적인 분석이 가능해집니다.
- 이때 확률변수 값 (혹은 구간) 에 일정한 확률이 대응되는데, 이 관계를 **확률분포**라고 합니다.
  - 추후 이산확률변수와 연속확률변수를 다룰 텐데, 두 경우에서 확률을 구하는 방식이 조금 다릅니다.
- 특히 몇몇 확률분포는 실생활에서 자주 발생하거나, 이론적인 분석이 용이해 이름이 붙어 있는 경우가 있습니다.
  - 이항분포, 정규분포, 포아송분포, 지수분포, 균일분포, 감마분포, 카이제곱분포, t분포, F분포 등

모수 (Parameter)

- 확률분포에는 확률분포의 성질을 결정하는 값이 반드시 있습니다.
  그 값을 모수(parameter)라고 합니다. : 통계학의 최종 목표!
- 확률분포는 옷의 종류, 모수는 옷의 치수로 비유할 수 있을 것 같습니다.
  - Ex] 정규분포: 청바지
    평균이 0이고 분산이 1인 정규분포: 허리가 30인치, 다리가 100cm인 청바지
  - 평균과 분산을 정해 주면 해당 정규분포는 유일하게 결정됩니다.
  - 다른 이름이 붙은 확률분포들도 다 각각의 모수를 갖고 있고,
    그에 맞게 모수 값이 정해지면 해당 확률분포는 유일하게 결정됩니다.
  - 하지만 같은 사이즈의 옷이 여러 개 존재할 수 있듯이,
    똑같이 생긴(Identical) 확률분포가 여러 개 존재할 수도 있습니다.
- 통계학에서는 주로 "확률변수 X가 OOO분포를 따른다" 라고 표현하고,
  기호로는 $X \sim N(0, 1^2)$처럼 적습니다.
- 확률분포의 종류를 알고 있고, 그 분포의 모수를 모두 안다면
  해당 확률변수의 모든 확률을 정확히 구할 수 있습니다!
- iid (Independent and Identically Distributed) assumption

확률변수의 구분

- 확률변수에는 크게 **이산확률변수**와 **연속확률변수**의 두 종류가 있습니다.
- 이산 확률변수: 확률변수가 가질 수 있는 값이 유한하거나, countable
  - 이항분포 (Binomial distribution)
  - 포아송분포 (Poisson distribution)
  - 음이항분포 (Negative Binomial distribution)
- 연속 확률변수: 확률변수가 가질 수 있는 값이 연속적/무한함
  - 정규분포 (Normal/Gaussian distribution)
  - 카이제곱분포 (Chi-squared(χ^2) distribution)
  - 감마분포 (Gamma distribution)
  - 지수분포 (Exponential distribution)
  - t분포 (t distribution)
  - F분포 (F distribution)
  - 균등(균일)분포 (Uniform distribution): 이산도 가능

이산/연속 구분은 왜 하는가?

- 두 종류의 확률변수에서 **확률이 정의되는 방식**이 다릅니다.
- 이산확률변수 (Discrete random variable)
  - 확률질량함수(Probability Mass Function, pmf)를 사용함
  - pmf는 확률변수가 특정 **값**일 때의 확률을 구할 수 있게 해 주는 함수.
  - 평균: $\sum_{i=1}^{n} x_i P(X = x_i)$
  - 분산: $\sum_{i=1}^{n} (x_i - \mu)^2 P(X = x_i)$
- 연속확률변수 (Continuous random variable)
  - 확률밀도함수(Probability Density Function, pdf)를 사용함
  - pdf는 확률변수가 특정 **구간** 내의 값을 가질 때의 확률을 알려준다.
  - 평균: $\int_a^b x f(x) dx$
  - 분산: $\int_a^b (x - \mu)^2 f(x) dx$
  - 확률변수 X가 연속형 확률변수일 때, $P(X = a) = 0$ 이다. (Why?)
- Riemann-Stieltjes Integration?

이산 확률변수 (Discrete random variable)

- 확률변수가 가질 수 있는 값이 유한 (finite) 하거나, countable한 경우
- Ex] 동전을 10번 던졌을 때 앞면이 나오는 **횟수**
  Ex] 양궁 선수가 활을 한 발 쐈을 때 얻을 수 있는 점수
- X가 가질 수 있는 값이 제한적이고,
  특정 확률변수 값마다 그 값이 나올 확률을 구할 수 있습니다.
- 확률질량함수 (pmf) 를 이용하면 해당 확률을 바로 구할 수 있습니다.
  - Ex] $f(x) = P(X = x) =_{10} C_x p^x (1-p)^{n-x}$ (binomial distribution)
- 평균과 분산을 직접 구해 봅시다!

이항 분포 (Binomial Distribution)

- 통계학입문에서부터 배우는, 가장 대표적인 이산 확률분포입니다.
- 어떤 시행의 결과를 성공/실패의 두 가지로 분류할 수 있고,
  각각의 시행이 서로 독립이라면
  해당 확률변수는 이항분포로 설명할 수 있습니다.
  - 결과가 두 가지만 존재하는 것과, 두 가지로 분류할 수 있는 것은 다릅니다!
  - Ex] 주사위 눈은 6개니까 이항분포로 설명이 불가능하다..? (No!)
- 모수(Parameter)는 총 2개입니다.
  - 1) 시행 횟수 (n)
  - 2) 성공 확률 (p)
- 확률변수 X가 시행 횟수가 n이고 성공확률이 p인 이항분포를 따른다면,
  이를 $X \sim B(n, p)$로 표기합니다.
- pmf: $P(X = x) =_n C_x p^x (1-p)^{n-x}$
- $E[X] = np$, $V[X] = np(1-p)$ (증명?)
- 이항 분포는 서로 독립인 베르누이 분포 n개의 합으로 정의됩니다.

포아송 분포 (Poisson Distribution)

- 포아송 분포는 어떤 사건이 **아주 드물게 발생**할 때 사용할 수 있다고 알려져 있습니다.
- 즉, 포아송 분포는 이항 분포의 특수한 케이스라고 볼 수 있고, 이항분포로부터 포아송분포를 유도할 수 있습니다.
- 모수는 1개입니다. (λ)
- pmf:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- $E[X] = \lambda, V[X] = \lambda$ (증명?)

공분산과 상관계수

- 확률변수 간에 **선형적** 상관관계가 있는지 알고 싶을 때 사용합니다.
- 공분산 (Covariance): $COV[X, Y] = E[XY] - E[X]E[Y]$
  - 만약 X와 Y 사이에 양의 상관관계가 있다면,
    X가 커질 때 Y가 커진다고 기대할 수 있으므로
    X와 Y의 부호가 서로 일치하는 경우가 많을 것입니다. (centralized data)
  - X와 Y의 부호가 일치하면 둘의 곱인 XY는 양수가 되므로,
    이 때 $COV[X, Y] > 0$이 될 것이라고 기대할 수 있습니다.
  - 반대로 X와 Y 사이에 음의 상관관계가 있다면,
    $COV[X, Y] < 0$이 될 것임을 기대할 수 있습니다.
  - 즉, 공분산은 절대적인 수치보다도 **부호**가 중요합니다.
- 상관계수 (Correlation coefficient):

$$Corr[X, Y] = \frac{COV[X, Y]}{\sqrt{V[X]V[Y]}}$$

  - 공분산을 $\sqrt{V[X]V[Y]}$로 나눠 값이 -1부터 1 사이에 위치하게 scaling한 값
  - Perfect linear correlation: $Y = aX$