확률 및 통계학

- 5장 이산형 확률분포(2) -

명 세인(sein@pel.smuc.ac.kr)

상명대학교 프로토콜공학연구실

목 차

- 이산형 확률분포
- 초기하 분포
- 음이항 분포
- 기하분포
- 포아송 분포

이산형 확률분포

- 이산형 자료의 확률 분포 분류
  - 초기하분포(Hypergeometric Distribution)
    - 각 시행이 2 가지의 결과만 나오며, 비 복원 시행인 확률분포
  - 음이항분포(Negative Binomial Distribution)
    - 베르누이 시행에서, 특정 성공횟수가 발생하기까지의 **시행 횟수에 대한** 확률분포
  - 기하분포(Geometric Distribution)
    - 베르누이 시행에서, 1번째 성공까지 **시행횟수**의 확률분포
  - 포아송분포(Poisson Distribution)
    - 특정 범위 안에서 발생하는 사건의 확률분포
      - 범위는 **시간 또는 영역**이 될 수 있음

초기하분포

- 의미
  - 각 시행이 비복원 추출(Sampling Without Replacement)이며, 시행 결과가 두 가지인 확률분포

- 정의
  - k개의 성공과 $N-k$개의 실패로 구성된 크기 N인 유한모집단에서 크기 n인 확률표본을 취할 때, 성공의 개수를 나타내는 초기하 확률변수 X의 확률분포
    - $h(x;N,n,k)=\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \max\{0,n-(N-k)\}\le x\le \min\{n,k\}$

초기하분포

- 초기하분포의 평균과 분산
  - 초기하분포$h(x;\ N,n,k)$의 평균과 분산
    - 평균 $\mu = \frac{nk}{N}$
      - $E[X] = \sum_{x=0}^{n} x \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} = k\sum_{x=1}^{n} \frac{(k-1)!}{(x-1)!(k-x)!} \cdot \frac{\binom{N-k}{n-x}}{\binom{N}{n}} = k\sum_{x=1}^{n} \frac{\binom{k-1}{x-1}\binom{N-k}{n-x}}{\binom{N}{n}}$
      - $y = x - 1$ 이라 하면,
        - $E(X) = k\sum_{y=0}^{n-1} \frac{\binom{k-1}{y}\binom{N-k}{n-1-y}}{\binom{N}{n}}$
        - $\binom{N-k}{n-1-y} = \binom{(N-1)-(k-1)}{n-1-y}$, $\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{N}{n}\binom{N-1}{n-1}$ 을 이용하여
        - $E[X] = \frac{nk}{N}\sum_{y=0}^{n-1} \frac{\binom{k-1}{y}\binom{(N-1)-(k-1)}{n-1-y}}{\binom{N-1}{n-1}} = \frac{nk}{N} \to np$
    - 분산 $\sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N}\left(1 - \frac{k}{N}\right) \to \frac{N-n}{N-1} \cdot npq$

초기하분포

- 예제 5.12
  - 40개의 부품으로 구성된 한 로트에 불량품이 3개 이상 들어 있으면 일반적으로 그 로트를 거부할 때, 한 로트에서 임의로 5개의 부품을 취하여 하나의 불량품이라도 발견되면 그 로트를 거부하려고 한다.
    - 3개의 불량품이 들어있는 로트에서 5개를 취했을 때 정확히 1개의 불량품이 발견될 확률
      - $N = 40, n = 5, k = 3, x = 1 \rightarrow h(x; N, n, k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$
      - $h(1; 40,5,3) = \frac{\binom{3}{1}\binom{37}{4}}{\binom{40}{5}} = 0.3011$

초기하분포

- 예제 5.13
  - 예제 3.4에서 100개의 제품 중 12개가 불량품이었다. 100개 중 10개를 임의로 뽑았을 때 3개가 불량일 확률
    - $n = 10, N = 100, k = 12, x = 3$
    - $h(3; 100,10,12) = \frac{\binom{12}{3}\binom{88}{7}}{\binom{100}{10}} = 0.0807$

초기하분포

- 예제 5.14
  - 예제 5.12에서 확률변수의 평균과 분산을 구하고, 체비셰프 정리를 이용하여 구간 $\mu \pm 2\sigma$ 의 의미를 설명
    - 예제 5.12에서 $N = 40, n = 5,\ k = 3$
    - $\mu = \frac{5\times 3}{40} = \frac{3}{8} = 0.3750$
    - $\sigma^2 = \frac{40-5}{40-1} \times 5 \times \frac{3}{40}\left(1 - \frac{3}{40}\right) = 0.3113$
      - $\sigma\ = 0.5580$
    - $0.375 \pm (2)(0.5580) \rightarrow -0.7410$에서 1.4910까지, 5개에 포함된 불량품이 2개 미만일 확률이 적어도 $\frac{3}{4}$ 이상
      - $P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$

초기하분포

- 초기하분포와 이항분포
  - 상대적으로 N보다 n이 작다면, 각 추출에서 확률의 변화는 크지 않음(비복원 추출의 특징이 약해짐)
  - 따라서, 초기하분포를 이항분포(독립시행)로 근사 시킬 수 있으므로, 평균과 분산 또한 표현 가능
    - $\mu = np = \frac{nk}{N}$
    - $\sigma^2 = npq = n \cdot \frac{k}{N}(1 - \frac{k}{N})$
  - 분산을 위 증명과 비교하면, 모집단이 유한(Finite)하기 때문에 **수정계수(Correction Factor)**로 나타나는 $\frac{N-n}{N-1}$의 차이가 발생
    - n이 N보다 상대적으로 작다면 무시 가능

초기하분포

- 예제 5.15
  - 자동차 타이어 제조업자는 판매대리점으로 보내기 위해 선적된 5000개의 타이어 중 1000개가 약간의 결함을 가지고 있다고 한다. 어떤 사람이 타이어 10개를 구입했을 때 3개가 결함을 가질 확률
    - 표본크기$n = 10$에 대해 $N = 5000$이 상대적으로 크기 때문에 이항분포를 이용하여 근사적으로 계산 가능
    - $N = 5000, n = 10, p = 0.2$ 이므로
      - $h(3; 5000,10,1000) \approx b(3; 10,0.2)$
        $= \sum_{x=0}^{3} b(x; 10,0.2) - \sum_{x=0}^{2} b(x; 10,0.2) = 0.8791 - 0.6778 = 0.2013$
      - $h(3; 5000,10,1000)$의 확률값을 계산하면 0.2015

초기하분포

- 다변량 초기하분포 (Multivariate Hypergeometric Distribution)
  - 의미
    - 둘 이상의 초기하 확률변수에 대한 초기하확률분포
  - 정의
    - N개의 유한모집단이 k개의 집합 $A_1, A_2, \dots, A_k$로 분할되고, 각각의 집합은 $a_1, a_2, \dots, a_k$개의 원소를 가질 때, n개의 확률표본 중 $A_1, A_2, \dots, A_k$의 원소의 수를 나타내는 확률변수 $X_1, X_2, \dots, X_k$의 확률분포
      - $f(x_1, x_2, \dots, x_k; a_1, a_2, \dots, a_k, N, n) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$

초기하분포

- 예제 5.16
  - 생물학 실험의 대상이 되는 10명의 그룹이 있고, 이중 3명의 혈액형이 O형, 4명은 A형, 나머지 3명은 B형일 때, 5명을 임의로 선택하는 경우 O형이 1명, A형이 2명, B형이 2명일 확률
    - $x_1 = 1, x_2 = 2, x_3 = 2, a_1 = 3, a_2 = 4, a_3 = 3, N = 10, n = 5$
    - $f(1, 2, 2; 3, 4, 3, 10, 5) = \frac{\binom{3}{1}\binom{4}{2}\binom{3}{2}}{\binom{10}{5}} = \frac{3}{14} = 0.2143$

음이항분포

- 의미
  - 베르누이 시행을 고정된 수의 성공이 특정횟수의 시행에서 발생할 확률을 음이항실험(Negative Binomial)으로 분류
  - k번째 성공이 일어날 때까지의 시행횟수 X를 음이항확률변수라 하고, 이 확률변수의 분포를 음이항분포로 분류

- 정의
  - 독립적인 반복시행에서 성공확률이 p, 실패확률이 q 일 때, k번째 성공이 일어날 때 까지의 시행 횟수인 확률변수 X의 확률분포
    - $b^*(x;k,p) = \binom{x-1}{k-1} p^k q^{x-k}, \quad x = k, k+1, k+2, \ldots$

음이항분포

- 예제 5.17
  - NBA 결승전을 7전 4선승제이며, 결승전에서 만난 두 팀 A와 B의 한 경기에서 A팀이 B팀을 이길 확률이 0.55일 때
    a. A팀이 6게임째에서 우승할 확률은
       - $b^*(6;4,0.55) = \binom{5}{3}0.55^4(1-0.55)^{6-4} = 0.1853$

    b. A팀이 우승할 확률은
       - $P(A\text{팀이 }NBA\text{우승}) = \sum_{x=4}^{7} b^*(x;4,0.55) = 0.6083$

    c. 두 팀이 5전 3선승제인 플레이오프에서 만난 경우 A팀이 플레이오프에서 우승할 확률은
       - $P(A\text{팀이 플레이오프 우승}) = \sum_{x=3}^{5} b^*(x;3,0.55) = 0.5931$

기하분포

- 의미
  - 음이항분포에서 한번의 성공이 일어날 때 까지 시행횟수에 대한 확률분포
  - 음이항분포 식으로 표현할 경우 기하수열(등비수열)이 나타남

- 정의
  - 독립적인 반복시행에서 성공확률이 p, 실패확률이 $q = 1 - p$ 일 때, 한번의 성공이 일어날 때 까지의 시행횟수인 확률변수 X의 확률분포$g(x; p)$와 평균과 분산
    - $g(x; p) = pq^{x-1},\ x = 1, 2, 3, \ldots$
    - $\mu = \frac{1}{p},\ \sigma^2 = \frac{1-p}{p^2}$

기하분포

- 예제 5.18
  - 제조공정에서 100개의 제품마다 평균적으로 한 개의 불량품이 들어있고, 하나씩 검사할 때 4번째 검사까지 양품, 5번째에서 불량품이 발견될 확률은?
    - $x = 5, p = 0.01$ 인 기하분포이므로
    - $g(5; 0.01) = (0.01)(0.99)^4 = 0.0096$

기하분포

- 예제 5.18
  - 전화통화가 폭주하는 시간대에 한 번의 시도로 상대방과 통화할 수 있는 가능성 $p = 0.05$일 때, 5 번째 시도에서 상대방과 통화할 확률
    - $x = 5, p = 0.05$ 인 기하분포이므로
    - $g(5; 0.05) = (0.05)(0.95)^4 = 0.0407$

포아송분포

- 의미
  - 일정한 시간, 공간 내에서 발생하는 사건의 횟수에 따른 확률분포
  - 포아송과정(Poisson Process)의 특징
    - 단위시간간격 또는 일정영역에서 발생하는 결과의 수는 서로 겹치지 않고, 다른 영역과 독립
      - 건망성(No Memory)
    - 매우 짧은 시간이나 영역에서 한 번의 결과가 일어날 확률은 시간간격의 길이나 영역의 크기에 비례하며, 대상 영역 외부의 확률은 관련이 없음(영역의 크기가 같다면 확률이 같음)
    - 매우 짧은 시간간격이나 영역에서 둘 이상의 결과가 일어날 확률은 무시 가능

포아송분포

- 정의
  - 일정한 시간간격 또는 영역 t에서 발생하는 결과의 수(λ_t)를 나타내는 포아송 확률변수 X의 확률분포
    - $p(x;\lambda_t) = \frac{e^{-\lambda_t}(\lambda_t)^x}{x!}, x = 0, 1, 2, \ldots$
    - λ: 단위시간 또는 단위면적에서 발생하는 결과의 평균 수
    - e: 2.71828 ...
    - 부록 A.2에 0.1과 18사이의 몇 가지 λ_t값에 대한 누적 합 $P(r;\lambda_t) = \sum_{x=0}^{r} P(x;\lambda_t)$의 값이 정리됨

포아송분포

- 예제 5.20
  - 실험실에서 $\frac{1}{1000}$초 동안 카운터를 통과하는 방사능 입자의 평균 수는 4 일 때, $\frac{1}{1000}$초 동안 6개의 입자가 카운터를 통과할 확률
    - $p(6;4) = \frac{e^{-4}(4)^6}{6!} = \sum_{x=0}^{6} P(x;4) - \sum_{x=0}^{5} P(x;4)$
      $= 0.8893 - 0.7851 = 0.1042$

포아송분포

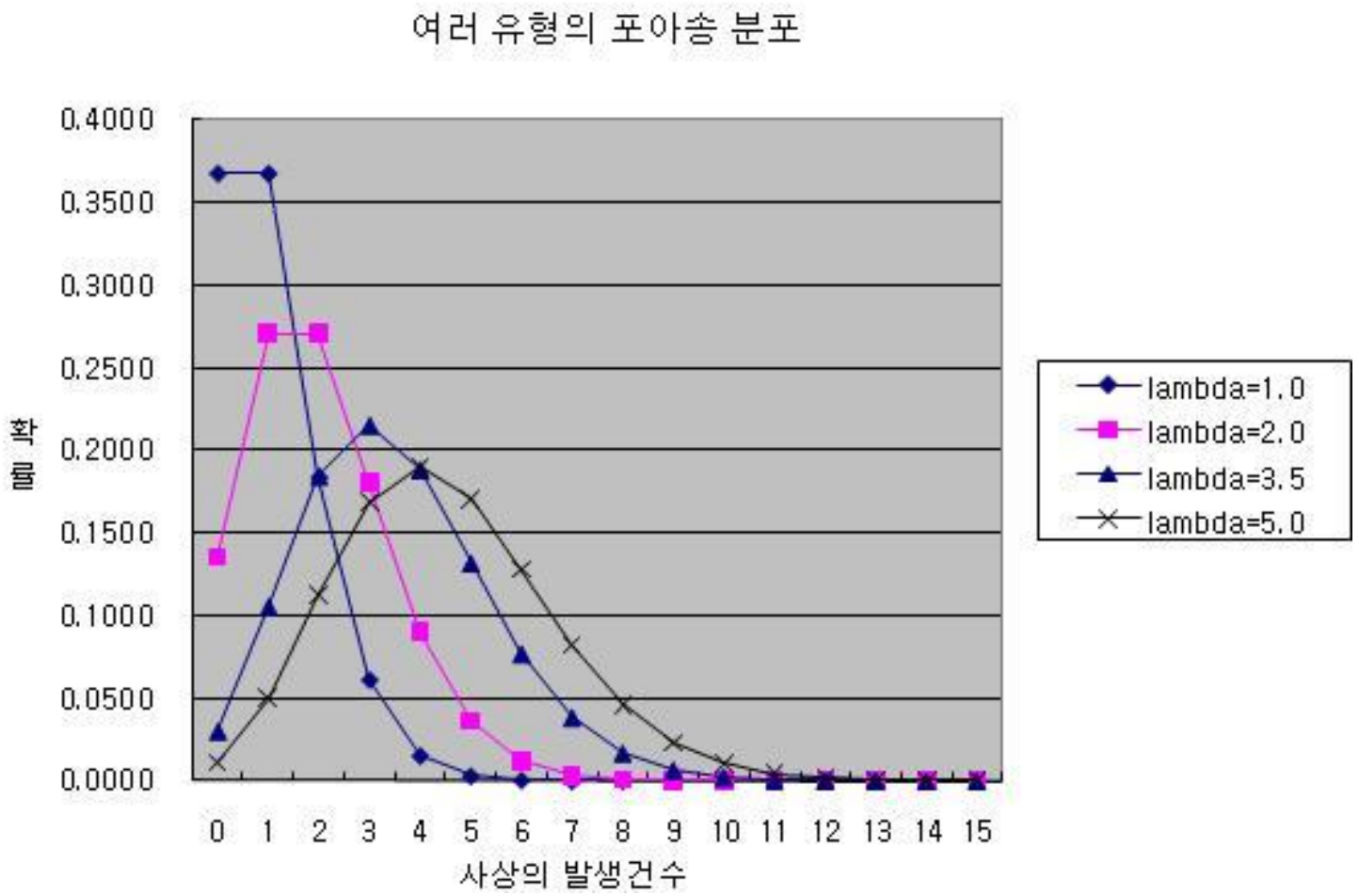
- 예제 5.21
  - 항구도시에 하루에 도착하는 유조선은 평균 10척으로 알려져 있으며, 항구에 있는 시설은 하루 최대 15척을 처리할 수 있을 때, 항구에 온 배를 돌려보내야 할 확률
    - $P(X > 15) = 1 - P(X \leq 15) = 1 - \sum_{x=0}^{15} p(x; 10)$
      $1 - 0.9513 = 0.0487$

포아송분포

- 포아송분포의 평균과 분산
  - $p(x;\lambda_t)$를 따르는 포아송 분포의 평균과 분산
    - 평균 $\mu = E(X) = \lambda_t$
      - $E(X) = \sum_{x=0}^{\infty} x \cdot \frac{e^{-\mu}\mu^x}{x!} = \sum_{x=1}^{\infty} x \cdot \frac{e^{-\mu}\mu^x}{x!} = \mu \sum_{x=1}^{\infty} x \cdot \frac{e^{-\mu}\mu^{x-1}}{(x-1)!}$
      - $y = x - 1$이라 하면
        - $E(X) = \sum_{y=0}^{\infty} \frac{e^{-\mu}\mu^y}{y!} = \mu$ 이므로, $\sum_{y=0}^{\infty} \frac{e^{-\mu}\mu^y}{y!} = \sum_{y=0}^{\infty} p(y;\mu) = 1$
    - 분산 $\sigma^2 = E[X^2] - (E[X])^2 = \lambda_t$
      - $E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1)\frac{e^{-\mu}\mu^x}{x!} = \sum_{x=2}^{\infty} x(x-1)\frac{e^{-\mu}\mu^x}{x!}$
        $= \mu^2 \sum_{x=2}^{\infty} \frac{e^{-\mu}\mu^{x-2}}{(x-2)!}$
      - $y = x - 2$라 하면
        - $E[X(X-1)] = \mu^2 \sum_{x=0}^{\infty} \frac{e^{-\mu}\mu^y}{y!} = \mu^2$
        - 따라서, $\sigma^2 = E[X^2] - (E[X])^2 = E[X(X-1)] + E(X) - (E[X])^2 = \mu = \lambda_t$

포아송분포

- 포아송분포의 평균과 분산
  - 람다에 따른 포아송 분포 그래프



포아송분포

- 포아송분포와 이항분포의 관계
  - 포아송분포는 공간 또는 시간관련 문제로 활용하지만 이항분포의 극단적인 형태로도 볼 수 있음
    - 이항분포에서 n이 매우 크고 p가 매우 작은 경우, 포아송 과정의 조건이 갖춰짐
  - 정리
    - X를 $b = (x; n, p)$를 따르는 이항확률변수라 할 때,
      - $n \to \infty, p \to 0, np \xrightarrow{n\to\infty} \mu$가 상수이면, $b(x; n, p) \xrightarrow{n\to\infty} p(x; \mu)$

포아송분포

- 예제 5.22
  - 산업현장에서 사고가 발생할 확률은 0.005이며, 각 사고발생은 독립적일 때, 확률변수X는 $n = 400$이고 $p = 0.005$인 이항확률 변수로 생각하면 $np = 2$인 포아송근사를 이용

    a. 400일 동안 사고가 한 건 발생할 확률
       - $p(X = 1) = e^{-2}2^1 = 0.2707$

    b. 사고일이 많아야 3일이 될 확률
       - $P(X \leq 3) = \sum_{x=0}^{3} \frac{e^{-2}2^x}{x!} = 0.8571$

포아송분포

- 예제 5.23
  - 유리제품 제조 시 기포가 생겨 시장에 출하할 수 없는 제품이 평균적으로 1000개의 제품 중 1개 이상의 불량품이 생길 때, 8000개의 제품 중 기포가 있는 제품이 7개보다 적을 확률
    - 원래는 $n = 8000$이고 $p = 0.001$인 이항실험이지만, p가 0에 근사하고 n이 매우 크므로 $\mu = (8000)(0.001) = 8$인 포아송 분포로 근사
      - $P(X < 7) = \sum_{x=0}^{6} b(x; 8000, 0.001) \approx p(x; 8) = 0.3134$

감사합니다!