2024/04/02, 확률 기초 세미나

확률 및 통계학

- 5장 이산형 확률분포 -

우 승 찬(seungchan@pel.sejong.ac.kr)

세종대학교 프로토콜공학연구실

목 차

이산형 확률분포

- 정의
  - 이산형 자료의 확률 분포
- 베르누이 시행(Bernoulli Trial)
  - 임의의 결과가 Yes or No 두 가지 중 하나의 결과가 나타나는 실험
- 종류(1/4)
  - 이산형 균일분포(Discrete Uniform Distribution)
    - 확률변수가 취할 수 있는 각 값들이 모두 같은 확률로 동일한 확률분포
      - 예시 1: P2P 네트워크에서 임의의 이웃 노드를 선택할 때, 각 선택 확률이 동일하다면, 이산형 균일분포로 나타남
      - 예시 2: 네트워크 내의 각 서버에 균등하게 부하가 분배되도록 하여 확률이 동일하다면, 이산형 균일분포로 나타남

이산형 확률분포

- 종류(2/4)
  - 이항 분포(Binomial Distribution)
    - 일련의 베르누이 시행으로부터 생성되는 확률분포
      - 예시 1: 네트워크 트래픽에서 패킷 전송의 성공 또는 실패가 이항 분포를 만족할 때, 이항 분포를 사용하여 성공 전송의 수 모델링
      - 예시 2: 네트워크 내에 들어오는 트래픽 중 악성 패킷과 정상 패킷이 이항 분포를 만족할 때, 이항 분포를 사용하여 악성 패킷의 수 모델링
  - 다항 분포(Multinomial Distribution)
    - 베르누이 시행이 아닌 각 시행에서 다양한 확률의 결과가 나타나며, 독립 시행으로 구성된 확률분포
      - 예시 1: 네트워크 내에서 여러 종류의 장애(e.g., 과부하, 장비 고장, 소프트웨어 오류 등)이 발생할 때, 다항 분포를 사용하여 특정 시간 동안 발생 가능한 각종 장애의 수가 나타날 확률을 모델링할 수 있음
      - 예시 2: 네트워크 내의 1000개의 패킷이 도착할 때, 여러 종류의 패킷(e.g., HTTP, SMTP, TCP 등)으로 분류하여 다항 분포를 통해 각 패킷 종류의 수가 나타날 확률을 모델링할 수 있음

이산형 확률분포

- 종류(3/4)
  - 초기하 분포(Hypergeometric Distribution)
    - k개의 성공과 $N-k$의 실패로 구성된 크기 N의 유한모집단에서 크기 n의 표분을 취할 때, 성공의 개수를 나타내는 확률분포
      - 예시 1: 네트워크 내의 전체 패킷 수 N에서 손실된 가능성이 있는 패킷 수 k가 존재하고 이 중 n개의 패킷을 재전송하려고 할 때, 손실된 패킷의 수를 모델링할 수 있음
  - 음이항 분포(Negative Binomial Distribution)
    - 단일 베르누이 시행에서 성공횟수 k가 정해졌을 때, 성공을 얻기까지 시행해야 하는 성공 및 실패 횟수가 따르는 확률분포
      - 예시 1: 네트워크 통신 중 패킷 손실이 발생했을 때, 정확한 수의 성공 전송을 달성하기 까지 필요한 재전송 횟수를 모델링할 수 있음
      - 예시 2: 네트워크에서 특정 서비스 및 애플리케이션에 대한 사용자 요청이 성공적으로 처리되기까지의 실패한 요청 수를 모델링하여 서비스의 가용성을 파악할 수 있음

이산형 확률분포

- 종류(4/4)
  - 기하 분포(Geometric Distribution)
    - 베르누이 시행에서 처음으로 성공할 때까지의 시행 횟수를 나타내는 확률분포
      - 예시 1: 네트워크에서의 재시도 메커니즘에서 패킷 손실 후 첫 번째의 성공적인 재전송까지의 재시도 횟수를 모델링할 수 있음
      - 예시 2: 네트워크 보안 시스템에서 첫 번째 보안 공격 시도를 감지하기까지의 실패한 감지 시도를 모델링할 수 있음
  - 포아송 분포(Poisson Distribution)
    - 단위 시간 안에서 어떠한 사건이 몇 번 발생할 것인가를 나타내는 확률분포
      - 예시 1: 네트워크 트래픽 모델링에서 특정 시간 간격에 걸쳐 발생하는 데이터 패킷의 도착 과정을 모델링하는 데 사용할 수 있음
      - 예시 2: 특정 시간 내에서 보안 이벤트 및 침입 시도의 발생을 모델링하는 데 사용될 수 있음

이산형 균일분포

- 의미
  - 이산형 확률변수가 취할 수 있는 각 값들이 모두 같은 확률로 동일한 확률분포

- 정의
  - 이산형 확률변수 X가 $x_1, x_2, \cdots, x_k$의 각 값을 취할 확률이 동일하다면, 이산형 균일분포로 정의됨
  - $f(x; k) = \frac{1}{k}, \quad x = x_1, x_2, \cdots, x_k$
    - 균일분포는 모수 k에 종속됨을 나타내기 위해 $f(x)$ 대신 $f(x; k)$ 사용

이산형 균일분포

- 예제 5.1

- 40와트, 60와트, 75와트, 100와트의 전구가 들어있는 상자에서 임의로 하나의 전구를 꺼낼 때의 확률분포

  - $S = \{40, 60, 75, 100\}$의 각 원소는 $\frac{1}{4}$ 의 발생 확률을 가짐
  - X의 확률분포는 $f(x; 4) = \frac{1}{4}, x = 40, 60, 75, 100$ 의 균일분포

- 예제 5.2

- 하나의 주사위를 던졌을 때의 확률분포

  - $S = \{1, 2, 3, 4, 5, 6\}$의 각 원소는 $\frac{1}{6}$ 의 발생 확률을 가짐
  - X의 확률분포는 $f(x; 6) = \frac{1}{6}, x = 1, 2, 3, 4, 5, 6$ 의 균일분포

이산형 균일분포

- 평균 및 분산
  - 정리
    - 평균: $\mu = \frac{1}{k}\sum_{i=1}^{k} x_i, \qquad k =$ 균일분포에 종속된 모수
    - 분산: $\sigma^2 = \frac{1}{k}\sum_{i=1}^{k}(x_i - \mu)^2$, $k =$ 균일분포에 종속된 모수
  - 증명
    - $\mu = E(X) = \sum_{i=1}^{k} x_i\, f(x;k) = \sum_{i=1}^{k} \frac{x_i}{k} = \frac{1}{k}\sum_{i=1}^{k} x_i$
    - $\sigma^2 = E[(X-\mu)^2] = \sum_{i=1}^{k}(x_i - \mu)^2\, f(x;k) = \frac{1}{k}\sum_{i=1}^{k}(x_i - \mu)^2$

이산형 균일분포

- 예제 5.3

- 하나의 주사위를 던졌을 때의 확률변수 X의 평균과 분산

  - 분산: $\sigma^2 = \frac{1}{6}\sum_{i=1}^{k}(x_i - 3.5)^2$

$$= \frac{1}{6}[(1-3.5)^2 + (2-3.5)^2 + \cdots + (6-3.5)^2] = \frac{17.5}{6} = \frac{35}{12}$$

이산형 균일분포

- 추가 예제 1: 이산형 균일분포

- P2P 네트워크에서 임의의 이웃 노드를 선택할 때, 각 선택 확률이 동일하고 이웃 노드가 10개 존재할 때의 확률분포

  - $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$의 각 원소는 $\frac{1}{10}$ 의 발생 확률을 가짐
  - X의 확률분포는 $f(x; 10) = \frac{1}{10}, x = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$의 균일분포

  - 평균: $\mu = \frac{1}{10}\sum_{i=1}^{k} x_i = \frac{1+2+3+4+5+6+7+8+9+10}{10} = 5.5$

  - 분산: $\sigma^2 = \frac{1}{10}\sum_{i=1}^{k}(x_i - 5.5)^2$
    $= \frac{1}{10}[(1-5.5)^2 + (2-5.5)^2 + \cdots + (10-5.5)^2] = \frac{99}{12} = 8.25$

이항 분포

- 의미
  - 일련의 베르누이 시행으로부터 생성되는 확률분포
  - 이항확률변수(Binomial Random Variable)의 확률분포

    - 이항확률변수
      - n번의 베르누이 시행에서의 성공 횟수 X

- 베르누이 과정(Bernoulli Process)
  - 실험은 n번의 반복 베르누이 시행으로 구성
  - 각 시행의 결과는 성공 또는 실패 두 가지 중 하나
  - P로 표시되는 성공확률은 매 시행마다 일정
  - 각 시행은 서로 독립

이항 분포

- 정의
  - 베르누이 시행의 n회 독립시행에서 성공 횟수 x를 나타내는 이항확률변수 X의 확률분포(p =성공 확률, q = 실패 확률)
  - $b(x;n,p) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n$
    - 조합(Combination): $\binom{n}{x} = nCx = \frac{n!}{x!(n-x)!}$
- 증명
  - 한번의 베르누이 시행에서 성공 확률$= p$, 실패 확률$= q = 1 - p$
  - 따라서, 한 번의 시행 결과는 성공(S) 혹은 실패(F)로 나타남
  - n번의 독립 시행에서 x번 성공하고 $n - x$번 실패하는 하나의 구체적인 순서(e.g., SSSFFFF)가 발생할 확률은 $p^x q^{n-x}$
  - n번의 시행에서 x번 성공할 수 있는 서로 다른 순서가 있고 이 경우의 수는 '같은 것이 있는 순열'이므로 $\binom{n}{x}$로 계산할 수 있음
  - 따라서, n번의 시행에서 정확히 x번 성공할 모든 경우의 확률은 $\binom{n}{x} p^x q^{n-x}$ 로 계산됨

이항 분포

- 예제 5.4

- 어떤 종류의 부품이 충격실험에서 충격을 견딜 확률이 3/4 일 때, 4개의 부품에 대한 실험에서 정확히 2개가 충격을 견딜 확률

  - 각 실험은 모두 독립이고 성공 확률 $p = \frac{3}{4}$ 이므로, 이항 분포가 성립

  - $b(x; n, p) = b\left(2; 4, \frac{3}{4}\right) = \binom{4}{2}\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)^2 = \left(\frac{4!}{2!\,2!}\right)\left(\frac{3^2}{4^4}\right) = \frac{27}{128}$

이항 분포

- 이항 정리
  - 두 개의 항의 거듭제곱에 대한 정리
  - $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$
  - 증명
    - $(a+b)^n$을 전개하면 a와 b의 서로 다른 조합을 모두 포함하는 항들의 합이 됨
      - 여기서 $a^{n-k}b^k$의 항의 계수는 $a^{n-k}b^k$형태의 항에서 발생하는 조합의 수와 같음
      - e.g., $(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$
        계수 3은 aab 에서 a를 두 번 선택하고 b를 한 번 선택하는 방법의 수 $\binom{3}{2}$ 와 같고, 이는 n개 중에서 k개를 선택하는 조합
    - 따라서, $(a+b)^n$의 전개는 $\sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$ 로 표현할 수 있음

이항 분포

- 확률분포의 조건
  - 이항 정리에서
    $(q+p)^n = \binom{n}{0}q^n + \binom{n}{1}pq^{n-1} + \binom{n}{2}p^2q^{n-2} + \cdots + \binom{n}{n}p^n$ 이 성립
    - 이는 $b(0;n,p) + b(1;n,p) + b(2;n,p) + \cdots + b(n;n,p)$ 와 같음
  - 따라서, $p+q = \sum_{x=0}^{n} b(x;n,p) = 1$ 이 성립되어 확률분포의 조건을 만족
- 이항분포 누적합
  - $B(r;n,p) = \sum_{x=0}^{r} b(x;n,p)$
    - 계산 편리성을 위해 누적이항분포표(책 부록 A.1) 참고

이항 분포

- 예제 5.5

- 빈혈환자가 회복될 확률 $p = 0.4$, 15명이 빈혈에 걸렸을 경우의 확률
  - (a) 적어도 10명이 회복될 확률
  - (b) 3명에서 8명 사이의 사람이 회복될 확률
  - (c) 정확히 5명이 회복될 확률

- $X =$ 회복된 환자의 수

- (a) $P(X \geq 10) = 1 - P(X < 10) = 1 - \sum_{x=0}^{9} b(x; 15, 0.4) = 1 - 0.9662$
  $= 0.0338$

- (b) $P(3 \leq X \leq 8) = \sum_{x=3}^{8} b(x; 15, 0.4) = \sum_{x=0}^{8} b(x; 15, 0.4) - \sum_{x=0}^{2} b(x; 15, 0.4)$
  $= 0.9050 - 0.0271 = 0.8779$

- (c) $P(X = 5) = b(5; 15, 0.4) = \binom{15}{5}\left(\frac{2}{5}\right)^{5}\left(\frac{3}{5}\right)^{10} = 0.1859$
  $= \sum_{x=0}^{5} b(x; 15, 0.4) - \sum_{x=0}^{4} b(x; 15, 0.4) = 0.4032 - 0.2173 = 0.1859$

이항 분포

- 예제 5.6

- 대형 할인점 상인은 전자장비를 제조업자로부터 납품 받음
- 해당 전자장비의 불량률은 3%
  - (a) 상인이 납품된 제품 중 20개를 임의로 선택하였을 때, 20개 중 적어도 한 대의 불량품이 있을 확률
  - (b) 한 달에 10번의 납품을 받고 납품시마다 20개의 장비를 검사할 경우, 적어도 한 대의 불량품이 포함된 납품이 3번 있을 확률

- (a) 불량품의 수 $= X$, 불량품이 하나도 없을 확률$= b(x; 20, 0.03)$
  $P(X \geq 1) = 1 - P(X = 0) = 1 - b(x; 20, 0.03) = 1 - 0.03^0(1 - 0.03)^{20}$
  $= 0.4562$

- (b) 각 납품시마다 하는 검사는 $p = 0.4562$의 베르누이 과정으로 볼 수 있음
  최소한 한 대의 불량품을 포함된 납품 건 수 $= Y$
  $P(Y = 3) = b(y; 10; 0.4562) = \binom{10}{3} 0.4562^3(1 - 0.4562)^7 = 0.1602$

이항 분포

- 평균 및 분산
  - 정리
    - $P(X=k)=\binom{n}{k}p^k q^{n-k},\ p+q=1, k=0,1,2,\dots,n$ 의 이항분포를 따르는 이항확률 변수 X의 평균
    - 평균: $\mu = np$
  - 증명
    - $\mu = E(X) = \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k} = \sum_{k=0}^{n} k\frac{n!}{k!(n-k)!}p^k q^{n-k}$

$= \sum_{k=0}^{n} k\frac{n!}{k!(n-k)!}p^k q^{n-k} = \sum_{k=1}^{n} \frac{n(n-1)!}{(k-1)!(n-k)!}p\cdot p^{k-1} q^{n-k}$

$= np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1} q^{n-1-(k-1)} = np(p+q)^{n-1} = np$

이항 분포

- 평균 및 분산
  - 정리
    - $P(X=k)=\binom{n}{k}p^k q^{n-k},\ p+q=1, k=0,1,2,\dots,n$ 의 이항분포를 따르는 이항확률 변수 X의 분산
    - 분산: $\sigma^2 = npq$
  - 증명
    - $\sigma^2 = E(X^2) - [E(X)]^2 = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k q^{n-k} - \{np\}^2$
      $= \sum_{k=0}^{n} (k^2 + k - k) \binom{n}{k} p^k q^{n-k} - \{np\}^2$
      $= \sum_{k=0}^{n} k(k+1) \binom{n}{k} p^k q^{n-k} + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} - \{np\}^2$
      $= \sum_{k=2}^{n} n(n-1) \binom{n-2}{k-2} p^k q^{n-k} + np - \{np\}^2$
      $= n(n-1)p^2 \sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2} q^{n-2-(k-2)} - \{np\}^2$
      $= n(n-1)p^2 (p+q)^{n-2} + np - \{np\}^2 = np - np^2 = np(1-p)$
      $= npq$

이항 분포

- 예제 5.7

- 어느 시골에 있는 전체 우물의 30%는 불순물이 포함되며, 이 지역의 우물 중 10개의 우물만 선정하여 검사를 수행
  - (a) 정확히 3개 우물에 불순물이 있을 확률
  - (b) 3개를 초과한 우물에 불순물이 있을 확률

  - (a) 불순물이 있는 우물의 수 $= X$,
    $P(X = 3) = b(3; 10,0.3) = \binom{10}{3}(0.3)^3(1 - 0.3)^7 = 0.2668$

  - (b) $P(X > 3) = 1 - P(X \leq 2) = 1 - \sum_{x=0}^{2} b(x; 10,0.3) = 1 - 0.6496 = 0.3504$

이항 분포

- 예제 5.8

- 빈혈환자가 회복될 확률 $p = 0.4$, 15명이 빈혈에 걸렸을 경우의 확률 평균과 분산을 구하고, 체비셰프 정리를 사용하여 구간 $\mu \pm 2\sigma$의 의미 설명

  - 체비셰프 정리
    - 어떤 k에 대하여 적어도 자료의 $1 - \frac{1}{k^2}$ 만큼의 비율이 k표준편차 이내에 존재

  - 평균: $\mu = np = 15 \cdot 0.4 = 6$
  - 분산: $\sigma^2 = npq = 15 \cdot 0.4 \cdot 0.6 = 3.6$
  - $\sigma = 1.897$

  - $\mu \pm 2\sigma = 6 \pm 2 \cdot 1.897, \;\; 2.206 \leq X \leq 9.794$
  - 자료가 이산형이므로 15명의 환자 중, 3명에서 9명 사이의 사람이 회복될 확률은 $\frac{3}{4}$ 이상

이항 분포

- 예제 5.9

- 예제 5.7에서 ’30%가 오염’ 은 추측 기반의 가설
- 10개의 우물을 검사했을 때, 6개의 우물에서 불순물이 발견되었다면, 확률적인 개념을 통한 가설의 의미

  - 가설이 맞다면, 6개 이상의 우물에서 불순물이 발견된다는 것이 있을 법한 일인지 판단해야함
  - $P(X \geq 6) = \sum_{x=0}^{10} b(x; 10,0.3) \ - \sum_{x=0}^{5} b(x; 10,0.3) = 1 - 0.9527 = 0.0473$

  - 6개의 이상에서 불순물이 발견되는 일은 드물게 나타남(4.7%)
  - 따라서, 실제 검사 시 6개의 우물이 오염된 경우 ’30%가 오염’ 이라는 가정에 의문이 생길 수 있음

이항 분포

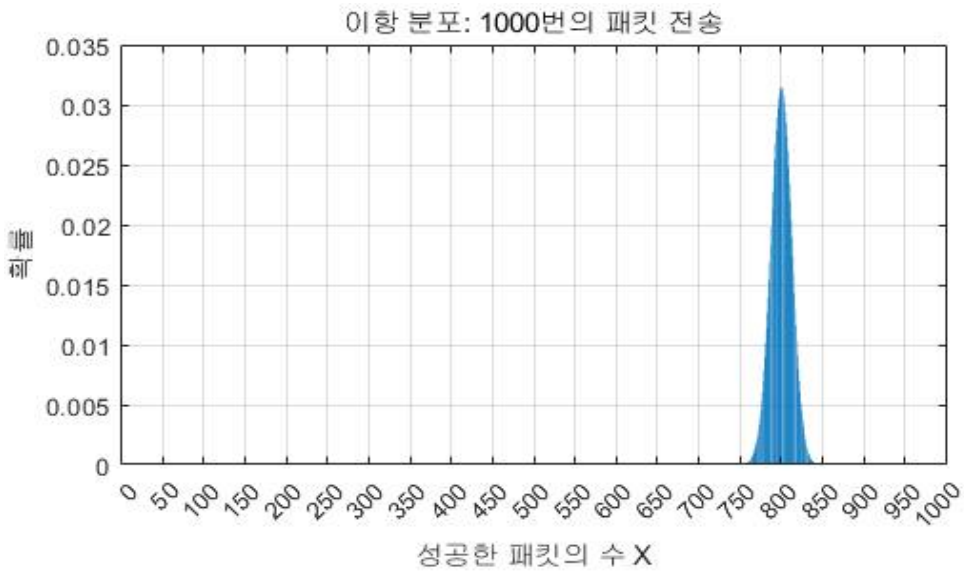
- 추가 예제 2: 이항 분포

- 네트워크 트래픽에서 패킷 전송이 성공할 확률이 80%이고 이항 분포를 만족할 경우, 1000번의 패킷을 전송할 때, 800번 초과하여 성공할 확률

  - 전송에 성공한 패킷의 수 $= X$
  - $P(X > 800) = \sum_{x=0}^{1000} b(x; 1000, 0.8) - \sum_{x=0}^{800} b(x; 1000, 0.8)$

    $= 1 - 0.5127 = 0.4873$

  - 평균: $\mu = np = 1000 \cdot 0.8 = 800$
  - 분산: $\sigma^2 = npq = 1000 \cdot 0.8 \cdot 0.2 = 160$
  - $\sigma = 12.649$



다항 분포

- 의미
  - 베르누이 시행이 아닌 각 시행에서 다양한 결과가 두 가지 이상이며 각 시행은 독립 시행으로 구성된 다항 실험(Multinomial Experiment)의 확률 분포

- 정의
  - 각 시행에서 $p_1, p_2, \dots, p_k$의 확률로 k개의 결과 $E_1, E_2, \dots, E_k$ 중 하나가 발생
  - N번의 독립시행에서 각각 $E_1, E_2, \dots, E_k$의 발생 횟수를 나타내는 확률변수 $X_1, X_2, \dots, X_k$의 확률분포
    - $f(x_1, x_2, \dots, x_k; p_1, p_2, \dots, p_k, n) = \binom{n}{x_1, x_2, \dots, x_k} \left(p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}\right)$
    - $\sum_{i=1}^{k} x_i = n$, $\sum_{i=1}^{k} p_i = 1$

다항 분포

- 예제 5.10

- 항공기 이착륙 상황에 대한 이상적인 조건을 알아보기 위해 컴퓨터 시뮬레이션 수행
- 3개의 활주로와 각 활주로가 사용될 확률은 다음과 같음
  - 활주로 1: $p_1 = \frac{2}{9}$, 활주로 2: $p_2 = \frac{1}{6}$, 활주로 3: $p_3 = \frac{11}{18}$
- 임의로 도착하는 6대의 비행기가 다음과 같이 활주로에 도착할 확률
  - 활주로 1: 2대, 활주로 2: 1대, 활주로 3, 3대

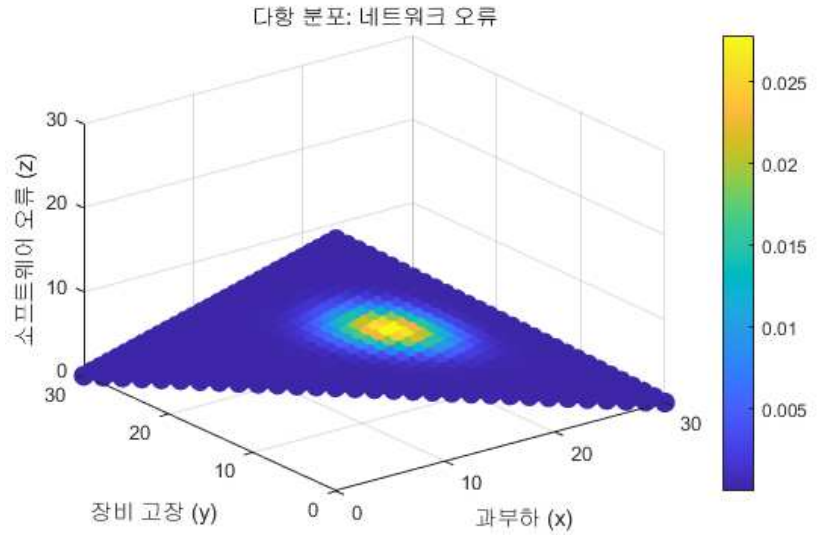
- 각 시행은 독립이고 다항 실험으로 구성되어 있어, 다항 분포를 이용함

- $f\left(2,1,3;\frac{2}{9},\frac{1}{6},\frac{11}{18},6\right) = \binom{6}{2,1,3}\left(\frac{2}{9}\right)^2\left(\frac{1}{6}\right)^1\left(\frac{11}{18}\right)^3 = \frac{6!}{2!1!3!}\cdot\frac{2^2}{9^2}\cdot\frac{1}{6}\cdot\frac{11^3}{18^3} = 0.1127$

다항 분포

- 추가 예제 3: 다항 분포

- 네트워크 내에서 여러 종류의 장애(e.g., 과부하, 장비 고장, 소프트웨어 오류 등)이 발생할 때, 해당 오류들이 발생할 확률
  - 과부하 = $\frac{1}{3}$, 장비 고장 = $\frac{2}{9}$, 소프트웨어 오류 = $\frac{4}{9}$
- 임의의 30건의 오류 중 과부하 8건, 장비 고장 8건, 소프트웨어 오류 14건이 발생할 확률
  - 각 시행은 독립이고 다항 실험으로 구성되어 있어, 다항 분포를 이용함
  - $f\left(8,8,14;\frac{1}{3},\frac{2}{9},\frac{4}{9},30\right) = \binom{30}{8,8,14}\left(\frac{1}{3}\right)^{8}\left(\frac{2}{9}\right)^{8}\left(\frac{4}{9}\right)^{14} = 0.01991$



초기하 분포

- 의미
  - 각 시행은 비복원 추출로 구성되며 초기하 실험에 대한 확률분포
    - 초기하 실험(Hypergeometric Experiment)
      - N개의 모집당 중 n개를 비복원추출할 때, k번의 성공을 확 확률에 대한 실험

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

초기하 분포

- 예제 5.11

- 어떤 분사장치는 묶음 단위로 팔림
- 생산자는 10개 중 불량이 1개 이하라면 묶음을 팔 수 있다고 판단
- 10개 중 3개를 임의로 선정하여 검사했을 때, 불량품이 하나도 없으면 그 묶음을 합격시키는 샘플링검사를 수행한다고 했을 경우, 해당 검사법의 유효성을 입증

  - 해당 묶음이 실제로는 10개 중 2개가 불량으로 불합격한다고 했을 경우, 이 샘플링검사에 의해 묶음이 합격될 확률
    - $P(X=0) = h(0;10,3,2) = \frac{\binom{2}{0}\binom{8}{3}}{\binom{10}{3}} = \frac{1 \cdot 336}{720} = 0.467$

- 불량품이 2개인 불합격 묶음을 합격시키는 경우가 약 47%만큼 발생할 수 있다는 결과가 도출되었으므로, 해당 검사법은 잘못되었다고 판단 가능

초기하 분포

- 예제 5.12

- 40개의 부품들로 구성된 한 묶음에 불량품이 3개 이상 들어 있다면, 그 묶음을 거부한다고 가정
- 한 묶음에서 임의로 5개의 부품을 취하여 불량품이 하나도 없으면 해당 묶음을 합격시키는 샘플링검사를 수행
  - 3개의 불량품이 들어 있는 묶음에서 5개를 취했을 때, 정확히 1개의 불량품이 발견될 확률

- $N = 5, N = 40, k = 3$이며, $x = 1$인 초기하분포를 따름
  - $h(1; 40, 5, 3) = \frac{\binom{3}{1}\binom{37}{4}}{\binom{40}{5}} = 0.3011$

- 불량품이 3개인 불합격 묶음을 탐지해 내는 경우가 약 30%에 불가하므로, 해당 검사법은 잘못되었다고 판단 가능

초기하 분포

- 평균 및 분산
  - 정리
    - 초기하 분포 $h(x;N,n,k)$를 따르는 초기하확률변수 X의 평균
    - 평균: $\mu = \frac{nk}{N}$
  - 증명
    - $\mu = E(X) = \sum_{x=0}^{n} x \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} = k\sum_{x=1}^{n} \frac{(k-1)!}{(k-1)!(k-x)!} \cdot \frac{\binom{N-k}{n-x}}{\binom{N}{n}}$

      $= k\sum_{x=1}^{n} \frac{\binom{k-1}{x-1}\binom{N-k}{n-x}}{\binom{N}{n}}$
    - $y = x - 1$로 치환

$$E(X) = k\sum_{y=0}^{n-1} \frac{\binom{k-1}{y}\binom{N-k}{n-1-y}}{\binom{N}{n}} = k\sum_{y=0}^{n-1} \frac{\binom{k-1}{y}\binom{N-k}{n-1-y}}{\frac{N(N-1)!}{n(n-1)!((N-1)-(n-1))!}} = k\sum_{y=0}^{n-1} \frac{\binom{k-1}{y}\binom{N-k}{n-1-y}}{\frac{N}{n}\binom{N-1}{n-1}}$$

$$= \frac{nk}{N}\sum_{y=0}^{n-1} \frac{\binom{k-1}{y}\binom{(N-1)-(k-1)}{(n-1)-y}}{\binom{N-1}{n-1}} = \frac{nk}{N}\sum_{y=0}^{n-1} h(y;N-1,n-1,k-1) = \frac{nk}{N}$$

초기하 분포

- 평균 및 분산
  - 정리
    - 초기하 분포 $h(x; N, n, k)$를 따르는 초기하확률변수 X의 분산
    - 분산: $\sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N}(1 - \frac{k}{N})$
  - 증명
    - $\sigma^2 = E(X^2) - [E(X)]^2 = E\big(X(X-1)\big) + E(X) - [E(X)]^2$

      $= \frac{n(n-1)k(k-1)}{N(N-1)} + \frac{nk}{N} - \left(\frac{nk}{N}\right)^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N}(1 - \frac{k}{N})$

초기하 분포

- 예제 5.13

- 100개의 제품 중 12개가 불량품
- 100개 중 10개를 임의로 뽑았을 경우, 3개가 불량일 확률

- $n = 10, N = 100, k = 12$이며, $x = 3$인 초기하 분포를 따름
  - $h(3; 100, 10, 12) = \frac{\binom{12}{3}\binom{88}{7}}{\binom{100}{10}} = 0.0807$

초기하 분포

- 예제 5.14

- 40개의 부품들로 구성된 한 묶음에 불량품이 3개 이상 들어 있다면, 그 묶음을 거부한다고 가정
- 한 묶음에서 임의로 5개의 부품을 취하여 불량품이 하나도 없으면 해당 묶음을 합격시키는 샘플링검사를 수행
  - 확률변수의 평균과 분산, 구간 $\mu \pm 2\sigma$의 의미를 설명

- $n = 5, N = 40, k = 3$인 초기하분포를 따름
  - $\mu = \frac{nk}{N} = \frac{15}{40} = \frac{3}{8} = 0.375$
  - $\sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N}\left(1 - \frac{k}{N}\right) = \frac{40-5}{40-1} \cdot 5 \cdot \frac{3}{40}\left(1 - \frac{3}{40}\right) = 0.3113$
    - $\sigma = 0.5580$

- $0.375 \pm 2 \cdot 0.5580 \rightarrow -0.7410$에서 1.4910까지, 5개에 포함된 불량품이 2개 미만일 확률이 적어도 $\frac{3}{4}$ 이상
  - $P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$

초기하 분포

- 초기하 분포 및 이항 분포와의 관계
  - 상대적으로 N보다 n이 작다면, 각 추출에서 확률의 변화는 크지 않음
    - 비복원 추출의 특성이 약해짐
  - 따라서, 초기하 분포를 이항 분포로 근사시킬 수 있어, 평균과 분산을 이항 분포로 표현할 수 있음
    - 평균: $\mu = np = \frac{nk}{N}$
    - 분산: $\sigma^2 = npq = n \cdot \frac{k}{N}(1 - \frac{k}{N})$
  - 초기하 분포의 분산 식을 이항 분포식과 비교하면 모집단이 유한(Finite)하기 때문에 $\frac{N-n}{N-1}$의 차이를 나타내는 수정계수 (Correction Factor) 발생
    - $\frac{n}{N} \leq 0.05$일 때 N보다 n이 상대적으로 작다고 판단하여 무시 가능

초기하 분포

- 예제 5.15

- 자동차 타이어 제조업자는 판매대리점으로 보내기 위해 선정된 5000개의 타이어 중 1000개가 약간의 결함 존재
- 구매자가 타이어 10개를 구입했을 경우, 3개가 결함을 가질 확률

  - 표본크기 $n = 10$에 대해서 $N = 5000$으로 상대적으로 크기 때문에, 이항 분포를 활용하여 근사적으로 계산이 가능함
    - $h(3; 5000, 10, 1000) \approx b(3; 10, 0.2) = \sum_{x=0}^{3} b(x; 10, 0.2) - \sum_{x=0}^{2} b(x; 10, 0.2) = 0.8791 - 0.6778 = 0.2013$

  - $h(3; 5000, 10, 1000)$의 정확한 확률값을 계산하면 0.2015 가 도출됨

초기하 분포

- 다변량 초기하 분포 (Multivariate Hypergeometric Distrubution)
  - 의미
    - 둘 이상의 초기하 확률변수를 가지는 초기하 확률분포
  - 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
      - $f(x_1, x_2, \dots, x_k; a_1, a_2, \dots, a_k, N, n) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2}\dots\binom{a_k}{x_k}}{\binom{N}{n}}$

초기하 분포

- 예제 5.16

- 생물학 실험의 대상이 되는 10명의 그룹에서 이 중 3명의 혈액형은 O형, 4명은 A형, 나머지 3명은 B형일 경우, 5명을 임의로 선택하였을 때, O형이 1명, A형이 2명, B형이 2명일 확률

  - $x_1 = 1, x_2 = 2, x_3 = 2, \quad a_1 = 3, a_2 = 4, a_3 = 3, \quad N = 10, n = 5$

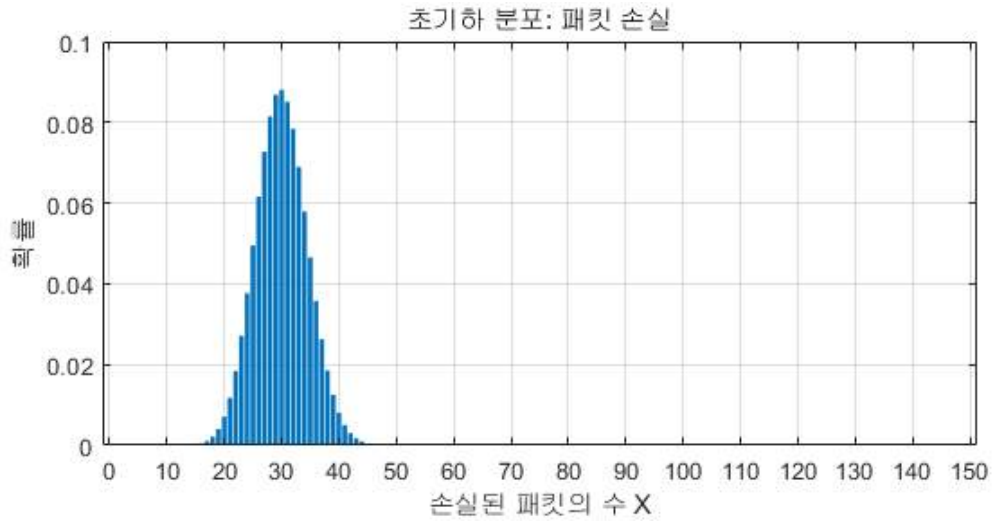
  - $f(1,2,2;3,4,3,10,5) = \frac{\binom{3}{1}\binom{4}{2}\binom{3}{2}}{\binom{10}{5}} = \frac{3}{14} = 0.2143$

초기하 분포

- 추가 예제 4: 초기하 분포

- 네트워크 내의 전체 패킷 수 1000에서 손실된 가능성이 있는 패킷 수 150개가 존재함
- 1000개 중 200개의 패킷을 재전송하려고 할 때, 손실된 패킷의 수가 90개 존재할 확률

  - $n = 150, N = 1000,\ k = 200$이며, $x = 90$인 초기하 분포를 따름
    - $h(90; 1000,200,150) = \frac{\binom{200}{90}\binom{800}{60}}{\binom{1000}{200}} = 4.74 \times 10^{-33}$



음이항 분포

- 의미
  - 단일 베르누이 시행에서 성공횟수가 일정할 때, 성공까지 시행한 횟수가 따르는 확률분포
  - k번째 성공이 일어날 때까지의 시행횟수 X는 음이항확률변수이며 해당 확률변수의 분포를 음이항 분포로 분류함

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

음이항 분포

- 예제 5.17

- NBA 결승전은 7전 4선승제로 운영되며, 결승전에서 만난 두 팀 A와 B의 한 경기에서 A팀이 B팀을 이길 확률은 0.55
  - (a) A팀이 6게임째에서 우승할 확률
  - (b) A팀이 우승할 확률
  - (c) 두 팀이 5전 3선승제인 플레이오프에서 만난 경우 A팀이 플레이오프에서 우승할 확률

- (a) $b^*(6,4,0.55) = \binom{5}{3}0.55^4(1-0.55)^{6-4} = 0.1853$

- (b) $P(\text{A팀이 NBA 우승}) = b^*(4,4,0.55) + b^*(5,4,0.55) + b^*(6,4,0.55) + b^*(7,4,0.55) = 0.0915 + 0.1647 + 0.1853 + 0.1668 = 0.6083$

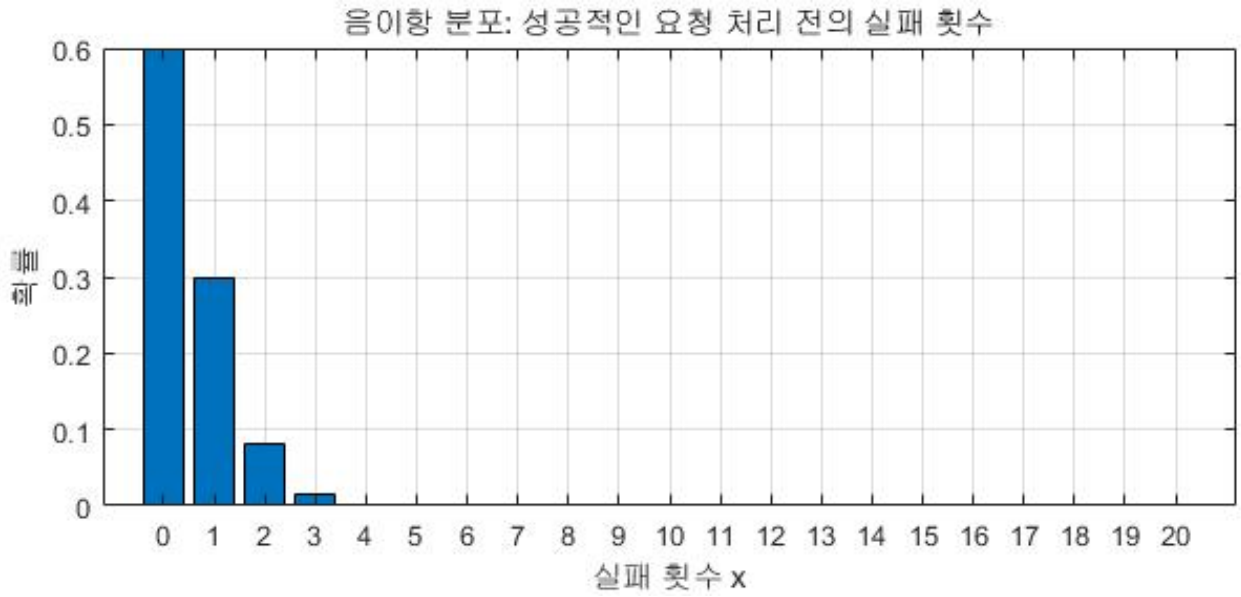
- (c) $P(\text{A팀이 플레이오프 우승}) = b^*(3,3,0.55) + b^*(4,3,0.55) + b^*(5,3,0.55) = 0.1664 + 0.2246 + 0.2021 = 0.5931$

음이항 분포

- 추가 예제 5: 음이항 분포

- 네트워크 서비스에서 성공적으로 요청을 처리할 확률은 0.95
- 해당 네트워크 서비스에서 10번의 성공 요청이 처리되기까지 3번의 실패가 발생할 확률

  - $x = 10, k = 3,\ p = 0.95$인 음이항 분포를 따름
    - $b^*(10, 3, 0.95) = \binom{10}{3} 0.95^{10-3}(1 - 0.95)^7 = 0.01647$



기하 분포

- 의미
  - 단일 베르누이 시행에서 처음으로 성공할 때까지의 시행 횟수를 나타내는 확률분포
  - 음이항 분포에서 한 번의 성공이 일어날 때까지의 시행 횟수에 따른 확률분포
- 정의
  - 독립적인 반복시행에서 성공확률이 p, 실패확률이 $q = 1 - p$일 때, 첫 번째 성공이 일어날 때까지의 시행횟수인 확률변수 X의 확률분포
    - $g(x; p) = pq^{x-1}, \quad x = 1, 2, 3, \ldots$
  - 평균 및 분산
    - $\mu = \frac{1}{p}, \quad \sigma^2 = \frac{1-p}{p^2}$

기하 분포

- 예제 5.18

- 제조공정에서 100개의 제품마다 평균적으로 한 개의 불량품이 검출됨
- 제품을 하나씩 검사할 경우 5번째에서 불량품이 발견될 확률

  - $x = 5, p = 0.01$인 기하분포를 따름
  - $g(5, 0.01) = (0.01)(1 - 0.01)^4 = 0.0096$

- 예제 5.19

- 전화통화가 폭주하는 시간대에서 한 번의 시도로 상대방과 통화할 수 있는 가능성 $p = 0.05$
- 5번째에서 통화에 성공할 확률

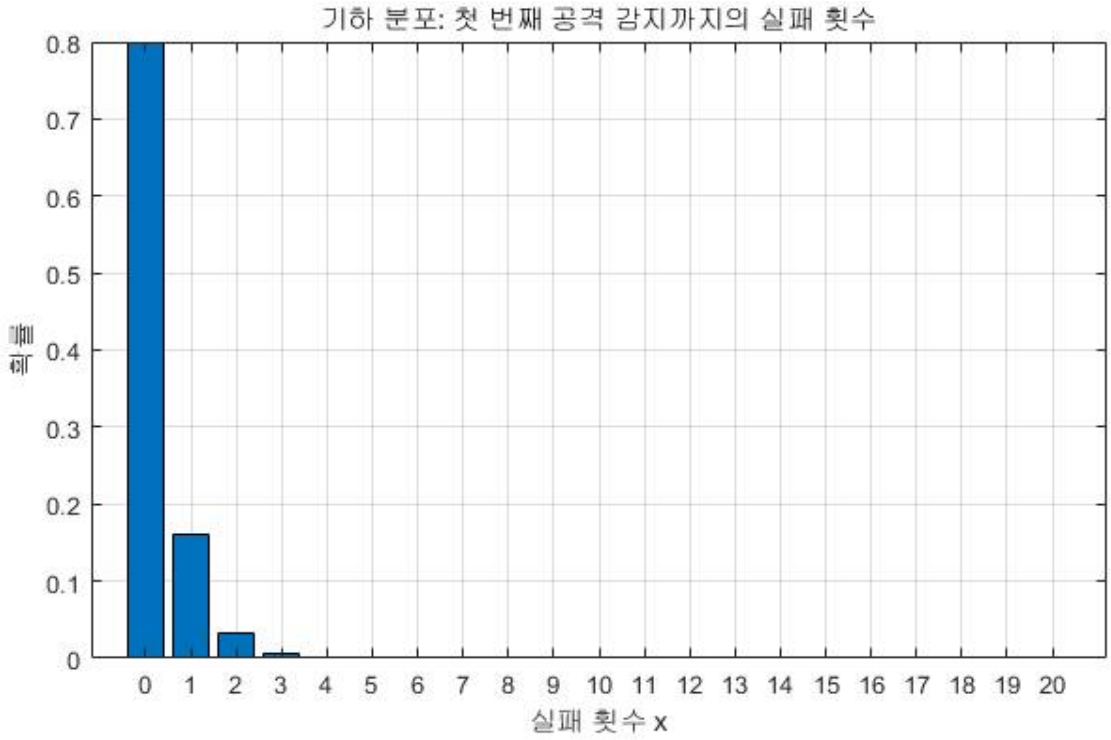
  - $x = 5, p = 0.05$인 기하분포를 따름
  - $g(5, 0.05) = (0.05)(1 - 0.05)^4 = 0.0407$

기하 분포

- 추가 예제 6: 기하 분포

- 네트워크 보안 시스템이 공격을 감지하는 확률은 0.8
- 첫 번째 공격 감지에 성공하기까지 3회의 실패가 발생하는 확률

  - $x = 4, p = 0.8$인 기하분포를 따름
    - $g(4, 0.8) = 0.8(1 - 0.8)^3 = 0.0064$



포아송 분포

- 의미
  - 단위 시간 안에서 어떠한 사건이 몇 번 발생할 것인가를 나타내는 확률분포
- 포아송 과정(Poisson Process) 성질
  - 단위 시간 간격 또는 일정 영역에서 발생하는 결과의 수는 서로 겹치지 않으며 다른 영역과 독립
    - 건망성(No Memory)
      - 미래의 사건 발생이 과거에 일어난 사건들에 의존하지 않는 성질
  - 단위 시간 간격 또는 일정 영역에서 발생하는 평균 비율은 일정
  - 매우 잛은 시간 간격이나 작은 영역에서 단 한 번의 결과가 일어날 확률은 시간간격의 길이나 영역의 크기에 비례함
  - 매우 짧은 시간 간격이나 작은 영역에서 둘 이상의 결과가 일어날 확률은 무시할 수 있음

포아송 분포

- 정의
  - 일정한 시간간격 또는 영역 t에서 발생하는 결과의 수 λ를 나타내는 포아송확률변수 X의 확률분포

    - $p(x;\lambda_t) = \frac{e^{-\lambda_t}(\lambda_t)^x}{x!}, \quad x = 0, 1, 2, \ldots$

    - λ: 단위시간 또는 단위면적에서 발생하는 결과의 평균 수
    - e: 자연 상수, $2.71828 \ldots$

- 포아송 분포 누적합
  - $P(r;\lambda_t) = \sum_{x=0}^{r} p(x;\lambda_t)$
    - 계산 편리성을 위해 누적포아송분포표(책 부록 A.2) 참고

포아송 분포

- 예제 5.20

- 실험실에서 $\frac{1}{1000}$초 동안 카운터를 통과하는 방사능 입자의 평균 수는 4일 때, $\frac{1}{100\,0}$ 초 동안 6개의 입자가 카운터를 통과할 확률

  - $x = 6,\ \ \lambda_t = 4$ 인 포아송분포를 따름
  - $p(6; 4\ ) = \frac{e^{-4}(4)^6}{6!} = 0.1042$

- 예제 5.21

- 어느 항구도시에 하루에 도착하는 유조선은 평균 10척으로 알려져 있으며, 항구에 있는 시설은 하루 최대 15척을 처리할 수 있는 경우, 항구에 온 배를 돌려보내야 할 확률

  - $\lambda_t = 10$인 포아송분포를 따름
  - $p(X > 15) = 1 - P(X \leq 15) = \sum_{x=0}^{15} p(x; 10) = 1 - 0.9513 = 0.0487$

포아송 분포

- 평균 및 분산
  - 정리
    - 포아송 분포$p(x;\lambda_t)$를 따르는 포아송확률변수 X의 평균
    - 평균: $\mu = \lambda_t$
  - 증명
    - $\mu = E(X) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda_t}(\lambda_t)^x}{x!} = \sum_{x=1}^{\infty} x \frac{e^{-\lambda_t}(\lambda_t)^x}{x!} = \lambda_t \sum_{x=1}^{\infty} \frac{e^{-\lambda_t}(\lambda_t)^{x-1}}{(x-1)!}$
    - $y = x - 1$ 로 치환
    - $E(X) = \lambda_t \sum_{y=0}^{\infty} \frac{e^{-\lambda_t}(\lambda_t)^y}{(y)!} = \lambda_t \sum_{y=0}^{\infty} p(y;\lambda_t) = \lambda_t$

포아송 분포

- 평균 및 분산
  - 정리
    - 포아송 분포$p(x;\lambda_t)$를 따르는 포아송확률변수 X의 분산
    - 분산: $\sigma^2 = \lambda_t$
  - 증명
    - $\sigma^2 = E(X^2) - [E(X)]^2 = \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda_t}(\lambda_t)^x}{x!} - {\lambda_t}^2$

$$= \sum_{x=1}^{\infty} x \frac{e^{-\lambda_t}(\lambda_t)^x}{(x-1)!} - {\lambda_t}^2$$

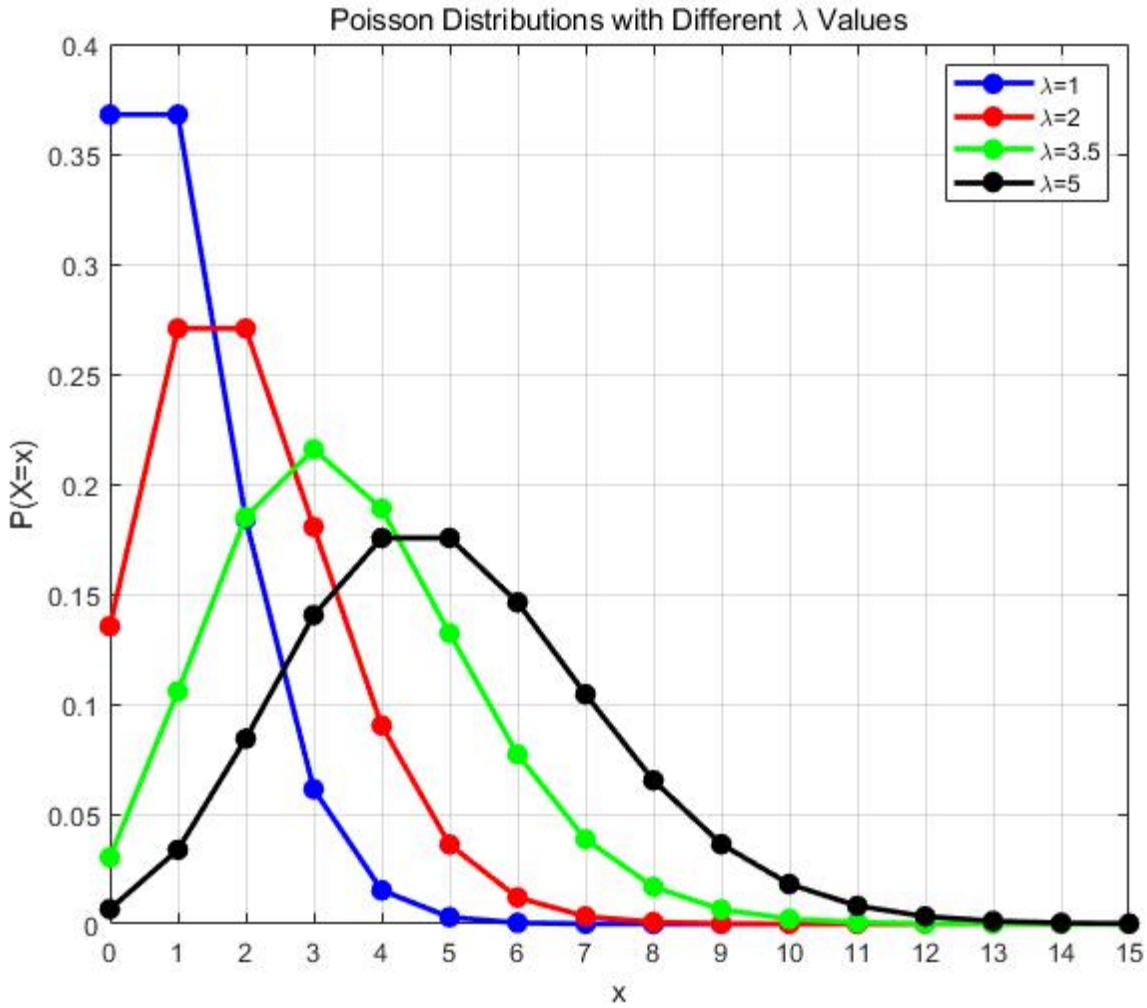
$$= \sum_{x=1}^{\infty} (x-1) \frac{e^{-\lambda_t}(\lambda_t)^x}{(x-1)!} + \sum_{x=1}^{\infty} \frac{e^{-\lambda_t}(\lambda_t)^x}{(x-1)!} - {\lambda_t}^2$$

$$= {\lambda_t}^2 \sum_{x=2}^{\infty} \frac{e^{-\lambda_t}(\lambda_t)^{x-2}}{(x-2)!} + \lambda_t \sum_{x=1}^{\infty} \frac{e^{-\lambda_t}(\lambda_t)^{x-1}}{(x-1)!} - {\lambda_t}^2$$

$$= {\lambda_t}^2 + \lambda_t - {\lambda_t}^2 = \lambda_t$$

포아송 분포

- 평균 및 분산
  - 람다 λ_t 값에 따른 포아송 분포 그래프



포아송 분포

- 포아송 분포 및 이항 분포와의 관계
  - 포아송 분포는 공간 및 시간관련 문제로 활용되지만, 이항 분포의 극단적인 형태로도 볼 수 있음
    - 이항분포에서 n을 무한으로 매우 크게 나타내고, p를 매우 작은 0에 가까운 값을 가지는 경우, 포아송 과정의 조건이 갖춰져 이항 분포를 포아송 분포로 근사시킬 수 있음
  - 정리
    - X를 $b = (x; n; p)$를 따르는 이항확률변수라 할 때,
      - $n \to \infty, p \to 0$를 만족하고 $np \xrightarrow{n\to\infty} \mu$가 상수일 경우,
      - $b(x; n, p) \xrightarrow{n\to\infty} p(x; \mu)$

포아송 분포

- 예제 5.22

- 산업현장에서는 사고가 종종 발생하며, 어느 날 사고가 일어날 확률은 0.005이고 각 사고발생은 독립적
  - (a) 400일 동안 사고가 한 건 발생할 확률
  - (b) 사고일이 많아야 3일이 될 확률

- $n = 400, p = 0.005$인 이항확률변수 X를 따르는 이항 분포를 포아송 분포로 근사, $\mu = np = 2$

- (a) $P(X = 1) \approx \frac{e^{-\mu}(\mu)^x}{x!} = \frac{e^{-2}(2)^1}{1!} = e^{-2}2^1 = 0.2707$

- (b) $P(X \leq 3) \approx \sum_{x=0}^{3} \frac{e^{-2}(2)^x}{x!} = 0.8571$

포아송 분포

- 예제 5.23

- 유리제품의 제조 공정에서 평균적으로 1000개의 제품마다 1개에 불량품이 생기는 것으로 알려짐
- 8000개의 제품 중 불량품의 개수가 7개보다 적을 확률

  - $n = 8000, p = 0.001$인 이항확률변수 X를 따르는 이항 분포를 포아송 분포로 근사, $\mu = np = 8$

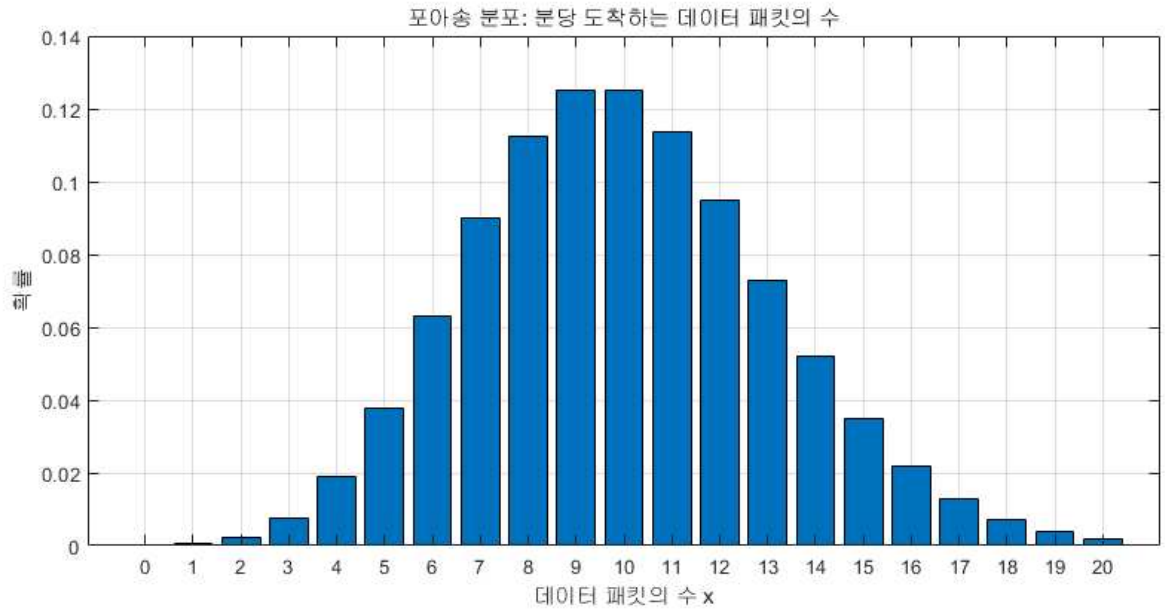
  - $P(X < 7) = \sum_{x=0}^{6} b(x; 8000; 0.001) \approx \sum_{x=0}^{6} p(x; 8) = \sum_{x=0}^{6} \frac{e^{-8}(8)^x}{x!} = 0.3134$

포아송 분포

- 추가 예제 7: 포아송 분포

- 어느 네트워크에서 평균적으로 분당 10개의 데이터 패킷이 도착함
- 특정 분에 정확히 15개의 데이터 패킷이 도착할 확률

  - $x = 15,\ \ \lambda_t = 10$ 인 포아송분포를 따름
  - $p(15; 10\ ) = \frac{e^{-10}(10)^{15}}{15!} = 0.0347$



Thanks!

우 승 찬 (seungchan@pel.sejong.ac.kr)

부록#1 MATLAB 코드

- 포아송 분포 그래프

```
lambdas = [1, 2, 3.5, 5];
colors = ['b', 'r', 'g', 'k'];

for i = 1:length(lambdas)
    lambda = lambdas(i);
    k = 0:15;
    p = poisspdf(k, lambda);

    plot(k, p, 'o-', 'Color', colors(i), 'LineWidth', 2,
'MarkerFaceColor', colors(i));
    hold on;
end

title('Poisson Distributions with Different \lambda Values');
xlabel('x');
ylabel('P(X=x)');
legend({'\lambda=1', '\lambda=2', '\lambda=3.5', '\lambda=5'},
'Location', 'best');
set(gca, 'xtick', [0:1:15]);
grid on;
```

부록#1 MATLAB 코드

- 추가 예제 2: 이항분포 그래프

```
n = 1000;
p = 0.8;

x = 0:n;
y = binopdf(x, n, p);

bar(x, y)
xlabel('성공한 패킷의 수 X')
ylabel('확률')
title('이항 분포: 1000번의 패킷 전송')
xticks(0:50:n)
grid on
```

부록#1 MATLAB 코드

- 추가 예제 3: 다항분포 그래프

```
p = [1/3, 2/9, 4/9];
n = 30;

[X, Y] = meshgrid(0:n, 0:n);
Z = n - X - Y;
X = X(Z >= 0);
Y = Y(Z >= 0);
Z = Z(Z >= 0);
probs = zeros(size(X));

for i = 1:length(X)
    outcome = [X(i), Y(i), Z(i)];
    if sum(outcome) == n
        probs(i) = mnpdf(outcome, p);
    end
end

scatter3(X, Y, Z, 100, probs, 'filled')
xlabel('과부하 (x)')
ylabel('장비 고장 (y)')
zlabel('소프트웨어 오류 (z)')
colorbar
title('다항 분포: 네트워크 오류')
grid on
```

부록#1 MATLAB 코드

- 추가 예제 4: 초기하분포 그래프

```
N = 1000;
K = 150;
n = 200;

x = 0:min(n, K);
y = hygepdf(x, N, K, n);

bar(x, y)
xlabel('손실된 패킷의 수 X')
ylabel('확률')
title('초기하 분포: 패킷 손실')
xticks(0:10:n)
grid on
```

부록#1 MATLAB 코드

- 추가 예제 5: 음이항분포 그래프

```
p = 0.95;
r = 10;

x = 0:20;

y = nbinpdf(x, r, p);

bar(x, y)
xlabel('실패 횟수 x')
ylabel('확률')
title('음이항 분포: 성공적인 요청 처리 전의 실패 횟수')
xticks(0:1:20)
grid on
```

부록#1 MATLAB 코드

- 추가 예제 6: 기하분포 그래프

```
p = 0.8;

x = 0:20;

y = geopdf(x, p);

bar(x, y)
xlabel('실패 횟수 x')
ylabel('확률')
title('기하 분포: 첫 번째 공격 감지까지의 실패 횟수')
xticks(0:1:20)
grid on
```

부록#1 MATLAB 코드

- 추가 예제 7: 포아송분포 그래프

```
lambda = 10;

x = 0:20;
y = poisspdf(x, lambda);

bar(x, y)
xlabel('데이터 패킷의 수 x')
ylabel('확률')
title('포아송 분포: 분당 도착하는 데이터 패킷의 수')
xticks(0:1:20)
grid on
```