확률통계 및 프로그래밍

5장 확률분포

: 이산확률분포(포아송분포)

포아송 분포(Poisson distribution)는 단위 시간 안에 어떤 사건이 몇 번 발생할 것인지를 표현하는 이산 확률분포이다. **즉, 일정한 단위 시간 또는 단위 공간에서 어떤 사건이 랜덤하게 발생하는 경우에 사용할 수 있는 이산형 확률분포**이다. 가령, 1시간 동안 잘못 걸려온 전화의 수, 하루 동안 어떤 사거리에서 발생하는 교통사고 건수, 책 1페이지당 오타가 발생하는 건수 등과 같이 단위 시간 혹은 단위 공간에서의 드물게 발생하는 사건의 수(확률변수 X)를 추정하는 확률분포이다.

포아송 실험의 두 가지 속성

- 어떤 구간에서 발생하는 사건은 다른 구간에서의 사건발생과 독립적이다(무관하다).
- 동일한 길이의 어떤 두 구간에서 사건이 발생 확률은 동일하다.
- 해당 구간에서 평균적으로 사건이 발생하는 수를 있을 때

이항 분포와 달리 푸아송 분포에서 모수는 한 개다. 포아송 분포의 모수 λ는 단위 시간 당 평균 발생 건수를 뜻한다. 비율(rate)이라고도 부른다.

확률변수 X가 이항분포 $B(n, p)$를 따를 때, $np = \lambda$ 로 일정할 때, n이 충분히 크고 p가 0에 가까울 때 이항분포로부터 포아송 분포(Poisson distribution)를 도출할 수 있다.

[정의] 포아송분포(Poisson distribution) $X \sim Poi(\lambda)$

일정한 단위시간에서 발생한 희소한 사건수의 확률분포

$$f(x) = \lambda^x \frac{e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots$$

$$\lim_{n\to\infty} f(x) = \lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} = \lim_{n\to\infty} \frac{n!}{x!(n-x)!} p^x (1-p)^x$$

$$\lambda = np \Rightarrow p = \frac{\lambda}{n} \qquad = \frac{1}{x!} \lim_{n\to\infty} \frac{n!}{(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{\lambda^x}{x!} \lim_{n\to\infty} \underbrace{\boxed{\frac{n(n-1)\cdots(n-x+1)}{n^x}}}_{1} \underbrace{\boxed{\left(1-\frac{\lambda}{n}\right)^n}}_{e^{-\lambda}} / \underbrace{\boxed{\left(1-\frac{\lambda}{n}\right)^x}}_{1}$$

$$= \lambda^x \frac{e^{-\lambda}}{x!}$$

 # 포아송분포 R 함수

 # 확률분포함수 (λ =기댓값, 단위시간의 발생 건수)
dpois(x, λ)

 # 누적분포함수 (q=분위수(특정한 값), lower.tail=T)
ppois(q, λ, lower.tail = T)

 # 분위수 (p=누적확률) 특정 확률 값에 해당하는 분위수 계산
qpois(p, λ, lower.tail = T)

#포아송 확률변수(n=난수의 개수)
rpois(n, λ)

누적확률이란 확률변수가 특정 값보다 같거나 작을 확률 $P(X \le x)$을 말한다.

예제 1 과거의 경험상 시내 어떤 은행에서는 매일 10시와 11시 사이에 고객이 평균 60명씩 몰려든다고 하자. 그러면 10시와 11시 사이에 1분당 2명이 도착할 확률을 구하여라.

$$f(x) = \lambda^x \frac{e^{-\lambda}}{x!}, x = 0,1,2, \ldots$$

풀이 1분당 평균도착률은 $\lambda = \frac{60}{60} = 1$이므로 $X=2$가 될 확률은

X: 1분 동안 은행에 방문하는 고객의 수

$$P(X=2) = \frac{e^{-1}1^2}{2!} = 0.5e^{-1} = 0.1839$$

```
# dpois(발생 건수, 평균)을 이용하여 포아송 분포로부터 P(X=2)를 구한다.
dpois(2, 1)

# 포아송 분포 그래프(Poisson distribution plot)
win.graph(7,6)
plot( dpois(x=c(0, 1, 2, 3, 4, 5), lambda=1),
      main = "Poisson distribution, lambda=1",
      type='h', col="Orange")
```

"h" : 수직선으로 된 "histogram" 또는 이산확률변수의 확률분포

이항분포의 포아송 분포로의 근사

$\lambda = np$ 가 일정한 모수 n, p인 이항분포에서 n→∞이면 이항분포 B(n, p)는 점근적으로 모수 λ 인 포아송 분포에 가까워진다. 일반적으로 . np 가 일정하고 n≥ 30 인 경우에 대하여 이항분포표를 이용하기보다는 근사적으로 포아송 분포를 이용하여 확률을 계산하는 것이 편리하다. 또한 . np 가 일정하고 n이 충분히 크다면 p는 상대적으로 매우 작아진다. 따라서 포아송 분포는 어떠한 사건이 발생하기 매우 어려운 상황에서 나타나는 확률모형임을 알 수 있다.

$X \sim Poi(\lambda)$

예제 2 어떤 회사에서 생산된 반도체의 불량률은 0.0001이라 한다. 그 회사의 생산라인에서 50,000개를 임의로 추출하여 2개 이하의 불량품이 나올 확률을 구하여라.

$X \sim B(50000, 0.0001)$ ⟶ $X \sim Poi(5)$

풀이 임의로 추출된 반도체에 들어 있는 불량품 개수를 X라 하면, $X \sim B(50000, 0.0001)$인 이항분포를 이룬다. 따라서 확률변수 X는 $\lambda = np = 5$인 푸아송 분포에 근사적으로 따르므로 구하고자 하는 확률의 근삿값은 $P(X \leq 2) = 0.1247$이다. ■

ppois(분위수, 평균 발생 건수)을 이용하여 포아송 분포로부터 P(X≤2)를 구한다.

```
ppois(2, 5)
```

다음은 누적분포함수 ppois( )의 결과와 같다.

```
sum( dpois(x=c(0:2), lambda = 5) )
```

pbinom(분위수, 시행횟수, 성공 확률)을 이용하여 이항분포로부터 P(X≤2)를 구하여 비교한다.

```
pbinom(2, 50000, 0.0001)
```

lambda=5인 포아송 분포에서 임의로 100개의 데이터를 추출한다.

```
rpois(100, 5)
mean(rpois(100, 5))
mean(rpois(100, 5))
mean(rpois(100, 5))
mean(rpois(100, 5))
```

확률통계 및 프로그래밍

5장 확률분포

: 이산확률분포(기하분포)

서로 독립적인 **n번의 베르누이 시행**에서 성공한 횟수를 나타내는 확률변수는 **이항분포**를 따른다.
“새로운 기술이 몇 번의 실패를 거쳐야 성공할 수 있을까? " 즉, 처음으로 성공하기 까지 몇 번 베르누이 시행을 해야 하는가에 대한 문제로 이항분포와는 다르다.

[정의] 기하분포(geometric distribution) $X \sim G(p)$

성공 확률(p)이 일정한 시행에서 첫 번째 성공이 발생할 때까지 시행한 횟수(X)의 확률분포

$$f(x) = (1-p)^{x-1} p, \quad x = 1, 2, \ldots$$

- **확률분포함수의 조건**

$$\sum_{x=1}^{\infty} f(x) = \sum_{x=1}^{\infty} (1-p)^{x-1} p = \frac{p}{1-(1-p)} = 1$$

누적분포함수

$$F(x) \equiv P(X \le x) = \sum_{y=1}^{x} (1-p)^{y-1} p \; = \frac{p[1-(1-p)^x]}{1-(1-p)} = 1-(1-p)^x$$

```
# 기하분포 R 함수

# 확률분포함수 (x=실패횟수, prob=p=성공확률)
dgeom(x, prob)

# 누적분포함수 (q=분위수, prob=p=성공확률, lower.tail=T)
pgeom(q, prob, lower.tail = T)

# 분위수 (p=누적확률)
qgeom(p, prob, lower.tail = T)

# 기하 확률변수(n=난수의 개수) 임의추출
rgeom(n, prob)
```

$X \sim G(p)$ 이면 $E(X) = \frac{1}{p}$ $Var(X) = \frac{1-p}{p^2}$

예제 1 1의 눈이 나올 때까지 주사위를 반복해서 던진다고 하자. 이때 주사위를 던진 횟수에 대한 확률함수를 구하고 평균과 표준편차를 구하여라.

풀이 주사위를 독립적으로 던지는 매 시행에서 1의 눈이 나올 가능성은 $p=1/6$이다. 따라서 1의 눈이 나올 때까지 던진 횟수를 확률변수 X라 하면 X는

$$f(x:1,\ 1/6) = \frac{1}{6}\left(\frac{5}{6}\right)^{x-1}, \quad x = 1,\ 2,\ \cdots$$

인 기하분포에 따른다. 그리고 이 경우의 확률과 표준편차는 각각 다음과 같다.

$$E(X) = \frac{1}{p} = 6,\ \sigma = \sqrt{Var(X)} = \sqrt{q/p^2} = \sqrt{30} = 5.477$$

기하분포는 x = 1 일 경우 가장 높은 확률을 가진다. 시도횟수가 많아질수록 확률은 기하급수적으로 작게 된다.

```
# 확률분포함수 (x=실패횟수, p=성공확률)
dgeom(0, 1/6) ; dgeom(1, 1/6) ; dgeom(2, 1/6)
dgeom(3, 1/6) ; dgeom(4, 1/6) ; dgeom(3, 1/6)
# 기하분포 그래프
win.graph(7,6)
plot( dgeom(x=c(0, 1, 2, 3, 4, 5), 1/6),
    main = "geometric distribution",
    type='h', col="Orange")
# pgeom(분위수, 성공확률)을 이용하여 기하분포로부터 P(X≤6)를 구한다.
pgeom(5, 1/6)
sum( dgeom(x=c(0:5), 1/6) )
#성공확률이 1/6일 때, 누적확률이 0.5가 될 때가지의 실패한 횟수이다. 이산확
률 분포이기 때문에 누적확률이 0.5 이상인 x중 최소값을 반환한다.
qgeom(0.5, 1/6)
```

성공확률이 1/6인 기하분포에서 임의로 100개의 데이터를 추출한다.(실패한 횟수)

```
rgeom(100, 1/6)
mean(rgeom(100, 1/6))  # 실패한 횟수의 평균
mean(rgeom(100, 1/6)+1)  # 처음으로 성공할 때까지 시행한 횟수의 평균
mean(rgeom(1000, 1/6)+1)
mean(rgeom(1000, 1/6)+1)
mean(rgeom(1000, 1/6)+1)
mean(rgeom(1000, 1/6)+1)

var(rgeom(1000, 1/6)+1)
```

확률통계 및 프로그래밍

5장 확률분포

: 이산확률분포(음이항분포)

$X \sim NB(r, p)$

[정의] 음이항분포(negative binomial distribution)

성공확률이 일정한 시행에서 r번 성공할 때까지 시행한 횟수 (X)의 에 대한 확률분포

$$f(x) = \binom{x-1}{r-1} p^r q^{x-r}, \quad x = r, r+1, r+2, \ldots$$

○ ○ ○ … ● ○ ○ … ○ ○ ○… ● ○ ○ … ○ ○ ○… ●

(x-1)회 시행, (r-1)회 성공 → (x-r)회 실패

x회 시행
r 번째 성공

$$P(X=x) = P(x\text{번째 시행에서 } r\text{번째 성공이 발생})$$

$$= P(x-1\text{번의 시행 중에서 } r-1\text{번 성공}, x\text{번째 시행은 성공})$$

$$= P(x-1\text{번의 시행 중에서 } r-1\text{번 성공})P(x\text{번째 시행은 성공})$$

$$= \binom{x-1}{r-1} p^{r-1} q^{x-r} \times p, \quad \underline{x = r, \ r+1, \ \cdots}$$

- **기하분포와 음이항분포의 관계**

$$X_1, X_2, \dots, X_r \overset{iid}{\sim} G(p) \;\; \Rightarrow \;\; X = X_1 + X_2 + \cdots + X_r$$

$$\mathrm{X} = \sum_{i=1}^{r} X_i \sim NB(r, p)$$

- **기댓값과 분산**

$$\mathrm{E}(X) = E(X_1 + X_2 + \cdots + X_r) \qquad E(X) = \sum_{i=1}^{r} E(X_i) = \frac{r}{p}$$

$$\mathrm{Var}(X) = Var(X_1 + X_2 + \cdots + X_r) \qquad Var(X) = \sum_{i=1}^{r} Var(X_i) = \frac{r(1-p)}{p^2}$$

음이항 분포의 R-함수

```
# 확률분포함수 (x=실패횟수, size=달성해야 할 성공 횟수, prob=성공확률)
dnbinom(x, size, prob)
# 누적분포함수 (q=분위수, lower.tail=T)
pnbinom(q, size, prob, lower.tail = T)
# 분위수 (p=누적확률)
qnbinom(p, size, prob, lower.tail = T)
# 음이항 확률변수(n=난수의 개수) 임의 추출
rnbinom(n, size, prob)
```

[예] 성공확률이 0.4인 실험에서 각각 1번, 2번, 4번의 성공할 때까지 시행한 횟수에 대한 확률분포

(1) 1번의 성공

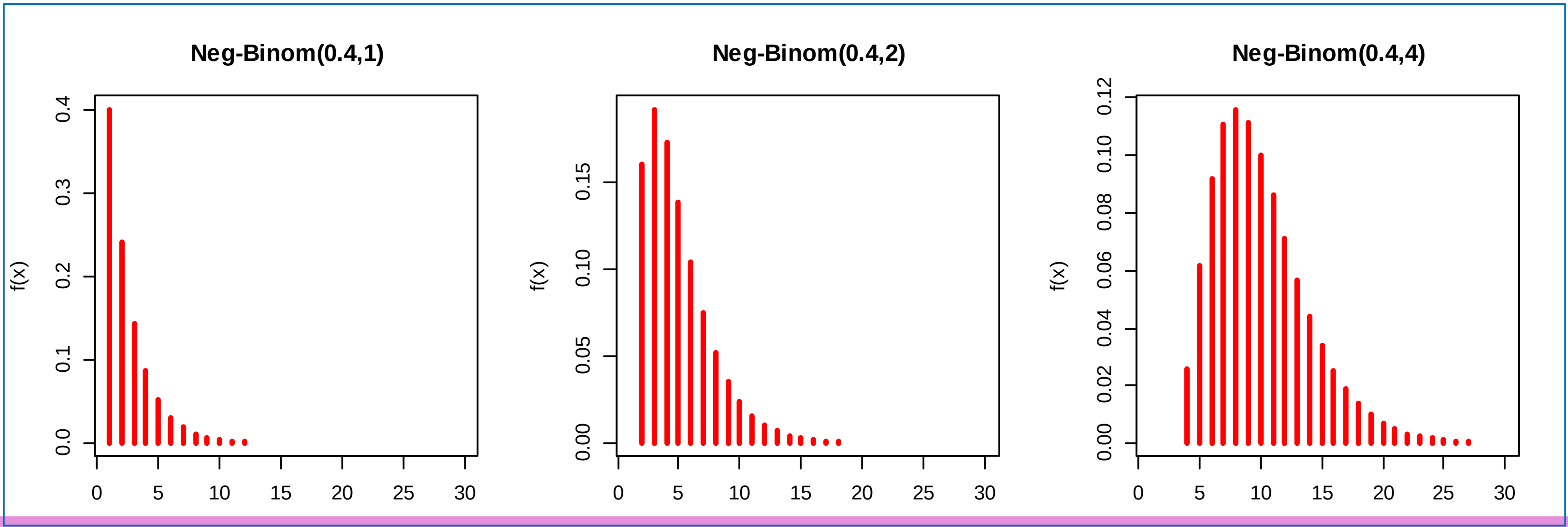
$X \sim NB(1, 0.4) = G(0.4)$ $\quad f(x) = \binom{x-1}{0}(0.4)(0.6)^{x-1}, x = 1,2,3, \ldots$

(2) 2번의 성공

$X \sim NB(2, 0.4)$ $\quad f(x) = \binom{x-1}{1}(0.4)^2(0.6)^{x-2}, x = 2,3,4, \ldots$

(3) 4번의 성공

$X \sim NB(4, 0.4)$ $\quad f(x) = \binom{x-1}{3}(0.4)^4(0.6)^{x-4}, x = 4,5,6, \ldots$



예제 4 5전 3승이면 승리하는 경기를 생각하자. A팀과 B팀의 시합에서 A팀이 1회 게임을 이길 확률은 0.52이다. A팀이 이 시합에서 승리할 때까지의 시합횟수를 X라고 할 때 X의 분포와 평균, 분산을 구하여라.

풀이 이것은 $r=3$, $p=0.52$인 음 이항분포이므로, 확률함수는

$$f(x)=\binom{x-1}{r-1}p^r(1-p)^{x-r}=\binom{x-1}{3-1}(0.52)^3(0.48)^{x-3}$$

이다. 한편 평균과 분산은

$$E(X)=\frac{r}{p}=\frac{3}{0.52}=5.77,$$

$$Var(X)=\frac{r(1-p)}{p^2}=\frac{3\cdot 0.48}{0.52^2}=5.33$$

확률분포함수 (x=실패횟수, size=달성할 성공 횟수, prob=성공확률)

dnbinom(0, 3, 0.52) # 3번째 경기에서 우승할 확률

dnbinom(1, 3, 0.52)   # 4번째 경기에서 우승할 확률

dnbinom(2, 3, 0.52)   # 5번째 경기에서 우승할 확률

dnbinom(3, 3, 0.52) ?

dnbinom(x=0:2, size =3, prob=0.52) #0~2번 실패하고, 3번 성공할 확률

누적분포함수 (q=분위수, lower.tail=T)

pnbinom(2, 3, 0.52)   # A팀이 경기에서 우승할 확률

sum (dnbinom(x=0:2, size =3, prob=0.52))

누적확률이 0.4인 누적확률변수 X(실패한 횟수)

qnbinom(0.4, size=3, prob=0.52)

음이항 확률변수(n=난수의 개수) 임의 추출

```
rnbinom(100, 3, 0.52)
mean(rnbinom(100, 3, 0.52)) ; mean(rnbinom(100, 3, 0.52)+3)
var(rnbinom(100, 3, 0.52))
```