확률통계 및 프로그래밍

5장 확률분포

: 이산확률분포(이항분포)

확률변수 X에 대한 확률함수 f(x) 따라 확률분포가 결정된다. 확률변수 X에 따라 이산확률분포로는 이항분포, 초기하분포, 푸아송 분포, 기하분포, 음이항분포 등이 있으며, 연속확률분포에는 정규분포를 비롯하여 균등분포, 지수분포, 코시 분포, 감마 분포, 베타 분포 등이 있다.

베르누이 분포와 이항분포

두 가지 실험결과만을 가지는 통계실험을 생각할 때 앞면이나 불량품 또는 성공의 가능성을 $p(0 \le p \le 1)$라 하고, 확률변수 X를 실험결과가 '성공'이면 $X=1$, '실패'이면 $X=0$이라 할 때, 성공과 실패 확률은 각각

$$\mathrm{P}(X=0)=1-p, \quad \mathrm{P}(X=1)=p$$

어떤 시행을 독립적으로 반복할 때, 발생할 수 있는 결과가 오직 두 개 뿐인 시행을 베르누이 시행 이라고 한다.

[정의] 베르누이분포(Bernoulli distribution) $X \sim B(1,p)$

성공 확률이 p인 1회의 시행에 대한 성공 횟수의 확률분포

- **pdf** $f(x)=\begin{cases} p, & x=1 \\ 1-p, & x=0 \end{cases} \Leftrightarrow f(x)=p^{x}(1-p)^{1-x},\ x=0,1$
- **기댓값** $E(X)=\sum_{x=0}^{1} xf(x)=0\times(1-p)+1\times p=p$
- **분산** $E(X^2)=\sum_{x=0}^{1} x^2 f(x)=0^2\times(1-p)+1^2\times p=p$

$$\Rightarrow Var(X)=E(X^2)-E(X)^2=p-p^2=p(1-p)$$

❖ 성공확률이 p 인 베르누이분포의 평균과 분산은 p에 의하여 결정되는 것을 알 수 있다. 이와 같이 확률분포를 결정짓는 상수를 확률분포의 **모수**parameter라 한다.

[예제1] 앞면이 나올 가능성이 $p = 1/3$ 인 찌그러진 동전을 반복해서 3번 던질 때 확률변수 X를 앞면이 나온 횟수 라 하자. 이때 확률변수 X의 확률분포를 구하여라.

베르누이 실험을 독립적으로 3번 반복하여 성공한 횟수 X의 확률분포

표본공간	(TTT)	$(HTT)\,(THT)\,(TTH)$	$(HHT)\,(THH)\,(HTH)$	(HHH)
X 값	0	1	2	3
각 표본점의 확률	$\left(\frac{2}{3}\right)^3$	$\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)$	$\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^2$	$\left(\frac{1}{3}\right)^3$
$P(X=x)$	$\left(\frac{2}{3}\right)^3$	$3\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)$	$3\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^2$	$\left(\frac{1}{3}\right)^3$

확률변수 X의 확률함수 : $f(x) = \binom{3}{x}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{3-x} \quad (x = 0,1,2,3)$

names( )을 사용하여 데이터 프레임(표본공간) 변수명(열이름) 변경

 # 열 이름(변수 명)을 출력하기

names(테이블 또는 표본공간)

모든 열의 이름(변수명) 바꾸기. 단, 이 경우 입력하지 않는 컬럼은 NA로 표시된다.

names(테이블 또는 표본공간) <- c("변경하고자 하는 변수명1", "변경하고자 하는 변수명2", ...)

특정한 열만 선택하여 변경하기

names(테이블) [names(테이블) == "원래 변수명"] <- c("변경하고자 하는 변수명")

변수명을 다시 출력해 보자.

names(테이블)

rename( )을 사용해서 컬럼 이름 바꾸기

rename(테이블 이름, "바꿀 이름" = "원래 이름")

R에서는 많은 종류의 확률분포에 대한 함수가 내장되어 있다.

이항분포(Binomial distribution) 는 어떤 통계적 실험에서 n번을 실행했을 때 성공 횟수가 x번 일어나는 확률분포이다

R에서 확률질량함수 P(X=x) 값을 구하기 위해서는 dbinom 함수를 사용한다.

dbinom(x, size, prob)

dbinom 함수의 인자는 x, size, prob으로 구성되어 있고 x는 n번 시행에서 **성공 횟수**이며 단일 숫자뿐만 아니라 숫자벡터를 입력해도 된다. size는 전체 **시행횟수**, prob는 단일 **성공확률**이다.

dbinom(성공횟수, 시행횟수, 성공확률)

```
# dbinom(성공 횟수, 시행 횟수, 성공 확률)을 이용하여 이항확률 값을 구한다.
( S <- dbinom(0, 3, 1/3) )     # 성공횟수가 0인 경우
( S <- dbinom(1, 3, 1/3) )
( S <- dbinom(2, 3, 1/3) )
( S <- dbinom(3, 3, 1/3) )
( S <- dbinom(0:3, 3, 1/3) )
( S <- dbinom( c(0,1,2,3), 3, 1/3) )
# as.fractions( )를 이용하여 소수를 분수로 표현함
library(MASS)
( S <- as.fractions( c(0.29629630, 0.44444444, 0.22222222, 0.03703704 ) ) )
# 열 이름(변수 명)을 출력하기
names(S)
# names( )를 이용하여 변수명을 변경(첨가)하여 이항확률분포표의 형태로 나타내자.
( names(S) <- c("0", "1", "2", "3") )
S
names(S)
```

[정의] 이항분포(binomial distribution) $X \sim B(n, p)$

성공 확률이 일정한 n 회의 시행에서 나오는 성공 횟수의 확률분포

- **pdf** $$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \; x = 0, 1, 2, \ldots, n$$

$$X_1, X_2, \cdots, X_n \overset{iid}{\sim} B(1, p) \;\Rightarrow\; X \equiv \sum_{i=1}^{n} X_i \sim B(n, p)$$

- **기댓값** $$E(X) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} p = np$$

- **분산** $$Var(X) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

[예제2] 한 개의 공정한 주사위를 4번 던질 때, 1의 눈이 나타날 확률분포를 구하여라.

X : 1의 눈이 나온 횟수(성공횟수), 전체 시행횟수: 4, 1회 시행에서의 성공확률: 1/6

확률변수 X의 확률함수 : $f(x) = \binom{4}{x}\left(\frac{1}{6}\right)^{x}\left(\frac{5}{6}\right)^{4-x} \quad (x = 0,1,2,3,4)$

이항누적확률분포(Binomial cumulative distribution) 구하기

누적확률이란 확률변수가 특정 값보다 같거나 작을 확률 P(X≤x)을 말한다.

R에서는 pbinom( )을 이용하여 이항누적확률을 쉽게 구할 수 있다.

pbinom(x, size, prob, lower.tail)

x는 P(X≤x)에서 x값, size는 총 시행횟수, prob는 단일시행 성공확률이다. 그리고 lower.tail이 TRUE이면 P(X≤x)이고(default), False이면 P(X>x)이다.

확률분포의 평균과 분산을 구하기 위한 패키지 **"distrEx"**를 설치하자.

```
install.packages("distrEx")
library(distrEx)
```

[예제3] 한 개의 공정한 동전을 5번 던질 때, 앞면이 나온 횟수를 확률변수 X라고 하자
이때 확률변수 X의 평균, 분산 그리고 $P(X \leq 2)$ 를 구하여라.
X : 1의 눈이 나온 횟수(성공횟수), 전체 시행횟수: 5, 1회 시행에서의 성공확률: 1/2

확률변수 X의 확률함수 : $f(x) = \binom{5}{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{5-x} \quad (x = 0,1,2,3,4,5)$

$$P(X \leq 2) = f(0) + f(1) + f(2)$$

dbinom(성공횟수, 시행횟수, 성공확률)을 이용하여 이항분포를 구하자.
```
( S <- dbinom(0:5, 5, 1/2) )
( S <- dbinom( c(0,1,2,3,4,5), 5, 1/2) )
```
names( )를 이용하여 변수명을 변경(첨가)하여 이항확률분포표의 형태로 나타내자.
```
( names(S) <- c("0", "1", "2", "3", "4", "5") )
S
```

pbinom( ) 함수를 이용하여, 이항누적확률을 구하기
```
library(MASS)
as.fractions( pbinom(2, 5, 0.5) )
```

Binom(시행횟수, 성공확률)으로 이항분포를 정의한 후, **E( )**와 **var( )** 함수를 이용하여 평균과 분산을 구한다.
```
library(distrEx)
BD <- Binom(5, 0.5)
as.fractions( E(BD) )
as.fractions( var(BD) )
```