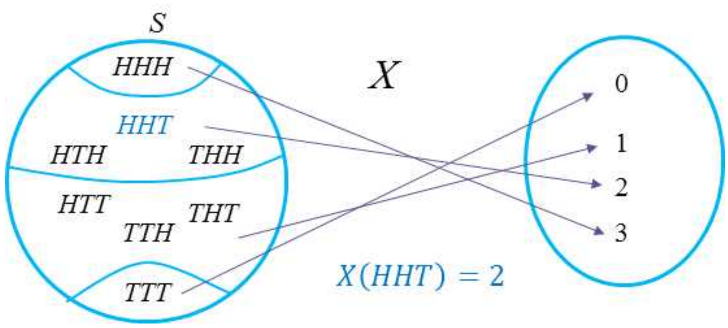
제3장 확률변수와 확률분포의 기초

3.1 확률변수

어떤 실험의 표본공간 S 위에서 정의된 하나의 실수값 함수를 확률변수(random variable)라고 한다. 즉, 어떤 실험(시행)에서의 확률변수는 모든 가능한 결과 $s \in S$에 하나의 실수 $X(s)$를 대응시키는 함수이다.

[예제1] S : 동전을 세 번 던지는 실험에서의 표본공간
확률변수 X : 동전의 앞면(H)이 나오는 횟수



[예제1] 하나의 동전을 두 번 던지는 실험에서

$$X(s) =: S\text{에서 앞면}(H)\text{의 개수}$$

라고 할 때 X의 치역을 구하여라.

```
install.packages("prob")  (설치가 되어 있지 않으면)
# "prob" 패키지 불러오기
library(prob)
# tosscoin( )을 이용하여 표본공간을 생성
S <- tosscoin(3) ; S
# 앞면의 개수를 세는 함수를 정의
Hcount <- function(x) sum(x == "H")
# 확률변수 X의 정의 ⇒ apply( ) 함수를 행별로 적용
( X <- apply(S, 1, Hcount) )
```

apply( ) 함수

- apply 함수는 행렬의 행(1) 또는 열(2) 방향으로 특정 함수를 적용한다.
- apply(행렬, 방향(1 또는 2), 함수) # 1: 행, 2: 열

확률변수를 정의하면 확률변수의 특정한 값이 발생할 가능성(확률)을 산출할 수 있다. 즉, 모든 가능한 확률변수의 값이 갖는 상대도수를 구하는 것이다. 이와 같이 확률변수가 취할 수 있는 값과 이 값들이 발생할 가능성이 상대도수에 의해 계산되면 확률분포(표)를 구성할 수 있다.

[예제2] 예제1의 실험에서 P(X=x)=**f(x)** 을 구하여라. 즉, 동전을 세 번 던지는 시행에서의 확률변수(앞면의 개수) X의 **확률분포**를 구하여라.

X	0	1	2	3	합
$P(X=x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

```
# [예제1]을 수행한 후 table( ) 함수를 이용하여 도수분포표(frequency distribution table)를 구함
( freq <- table(X) )
# length( ) 또는 nrow( )를 이용하여, 원소의 개수 또는 표본공간의 행의 수를 구하고, 확률분포표를 구함
( prob <- freq/length(X) )
( prob <- freq/nrow(S) )
# 확률을 분수로 표현하기
library(MASS)
as.fractions( prob )
```

win.graph( )

R은 다른 언어와 비교하여 그래프의 표현이 아주 쉽다는 장점이 있다.

win.graph( )는 새로운 창(팝업창)에서 그래프를 그리기 위한 함수로, 주로 그래프의 크기를 일정하게 맞추기 위해 사용한다.

두 개의 숫자(가로의 크기, 세로의 크기)를 입력하여 창의 크기를 조절한다. 예: win.graph(7, 6)

▣ plot( ) 함수

그래프를 그리는 가장 기본적인 함수로써, 첫 번째 인수는 x축, 두 번째 인수는 y축으로 받아 2차원 그래프를 그려준다.

- main : 문자열을 받아 그래프의 제목을 정한다.
- xlab, ylab : 문자열을 받아 축(x label, y label)의 이름을 정한다.
- xlim, ylim : 축이 표시되는 범위(x limit, y limit)를 조정한다.
- **type** : 그래프의 유형을 정한다. 기본값은 점으로 표시

◆ **type의 종류**

- "p" : 점
- "l" : 선
- "b" : 점과 선 ⇒ "c" : "b"에서 점 제외
- "o" : 점을 선이 통과
- **"h" : 수직선으로 된 “histogram” 또는 이산확률변수의 확률분포**
- "s" : 계단형 그래프
- "S" : 다른 계단형 그래프
- "n" : 그래프 없음

plot( )을 이용하여 제목(main), x축 이름(xlab), y축 이름(ylab), 각 점에서 축까지의 수직선(type)을 그리고, 그래프의 색상(col), y축 범위(ylim), 선의 굵기(lwd)를 지정하고 확률분포표를 그래프로 나타낸다.

```
plot(prob, main="동전 3개를 던졌을 때 앞면의 수에 대한 확률분포", xlab="앞면의 수", ylab="상대도수", type="h", col="Orange", ylim=c(0, max(prob)+0.01), lwd=4)
```

참고 선의 두께(line width) : lwd

lwd는 선의 두께를 조절하는 그래프 옵션이다. lwd=1이 디폴트 값이며, 이 숫자를 기준으로 입력한 숫자만큼 선 두께가 배수로 증가한다.

확률분포를 구성할 때 확률변수에 따라 이산확률변수(이산확률분포)와 연속확률변수(연속확률분포)로 구분한다.

[정의] 확률변수 X가 취할 수 있는 값이 유한개이거나 자연수와 같이 셀 수 있을 때, X를 **이산확률변수**(discrete random variable)라고 한다. 이산확률변수의 확률분포를 나타내는 함수를 **확률질량함수**(probability mass function)라고 한다.[1)]

$$P(X=x)=f(x)$$

확률변수 X가 취할 수 있는 값이 셀 수 없는 경우도 존재한다. 예를 들면, 지하철이 어떤 역에 도착하는 시간, 건전지의 수명시간, 대학생들의 키 등과 같이 어떤 구간 내의 모든 값을 취하는 경우가 있다. 이와 같이 시간, 길이 등과 같이 연속성을 지니는 변수를 연속변수라고 한다.

[정의] 확률변수 X가 일정한 구간의 모든 실수 값을 가질 수 있을 때, X를 **연속확률변수**(continuous random variable)라고 한다. 연속확률변수의 확률분포를 나타내는 함수 $f(x)$를 **확률밀도함수**(probability density function)라고 한다.

$$P(a \le X \le b)=\int_a^b f(x)\,dx\ ,\quad \int_{-\infty}^{\infty} f(x)\,dx=1\ ,\ f(x)\ge 0$$

[예제3] 확률변수 X의 확률밀도함수가 다음과 같다고 할 때, 상수 c를 결정하고, $P(X \ge 1)$을 구하여라.

$$f(x)=\begin{cases} c(4x-2x^2)\ , & 0<x<2 \\ 0\ , & \text{그밖의} \end{cases}$$

풀이. 1단계: 상수 c를 구하자.

$\int_{-\infty}^{\infty} f(x)\,dx=1$ 이므로

$$\int_0^2 f(x)\,dx=\int_0^2 c(4x-2x^2)\,dx=\frac{8c}{3}=1 \qquad \therefore\ c=\frac{3}{8}$$

2단계: $P(X \ge 1)$을 계산하자.

---

1) 확률분포(Probability distribution)란 확률변수의 모든 값과 그에 대응하는 확률들이 어떻게 분포하고 있는지를 말한다. 그렇다면 확률함수(Probability function)는 무엇일까요? 확률함수는 확률변수에 의해 정의된 실수를 확률(0~1사이)에 대응시키는 함수를 의미한다.

$$P(X \geq 1) = \frac{3}{8}\int_0^2 (4x - 2x^2)\,dx = \frac{1}{2}$$

R에서 정적분을 구하기 위해선 **integrate() 함수**를 사용할 수 있다.

integrate( )를 이용하여 $\int_0^2 (4x-2x^2)\,dx$ 를 구한다.

```
f1 <- function(x) {4*x - 2*x^2}
( c1 <- integrate(f1, 0, 2) )
# as.fractions( )을 이용하여 c1을 분수로 표시한 후 역수를 취함
library(MASS)
as.fractions(2.666667)
as.fractions(1/2.666667)
# integrate( )를 이용하여, 확률밀도함수(pdf)를 구함
pdf <- function(x) {(3/8)*(4*x - 2*x^2)}
# integrate( )를 이용하여, 확률를 구함
( integrate(pdf, 1, 2) )
```

[예제4] 어떤 제품의 수명시간을 확률변수 X라고 할 때, X의 확률밀도함수는

$$f(x) = \frac{100}{x^2},\ x \geq 100$$

이다. 5개의 제품 중에서 정확히 2개가 150시간 이하로 고장 나서 교체될 확률을 구하여라.

풀이. 1단계: $\int_{100}^{\infty} f(x)\,dx = \int_{100}^{\infty} \frac{100}{x^2}\,dx = 1$ (확률밀도함수)

2단계: A_i를 i 번째 제품이 150시간 이하에서 고장 나는 사건

$$P(A_i) = \int_{100}^{150} \frac{100}{x^2}\,dx = \frac{1}{3}$$

3단계: 서로 다른 5개의 제품 중에서 2개를 선택하는 경우의 수

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = {}_nC_r\,,\quad \binom{5}{2} = \frac{5!}{2!(5-2)!} = {}_5C_2 = 10$$

4단계: 나머지 3개의 제품은 정상

$$\binom{5}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^3 = \frac{80}{243}$$

```
# 확률밀도함수(pdf)를 정의하자.
pdf <- function(x) {100/x^2}
# 사건 Ai가 일어날 확률
( c <- integrate(pdf, 100, 150) )
library(MASS)
as.fractions(0.3333333)
# choose(n,r) / choose(n,r) * factorial(r)
```

서로 다른 5개의 제품 중에서 2개를 선택하는 경우의 수 $\binom{5}{2}$

```
( p <- choose(5, 2) *(1/3)^2 *(2/3)^3 )
as.fractions(p)
```