확률통계 및 프로그래밍

2장 확률

: 여사건, 조건부 확률

R은 수많은 내장 함수를 가지고 있지만, 사용자가 함수를 만들 수도 있다.

R에는 function 키워드를 사용하여 다음과 같이 함수를 만들 수 있다.

```
만들고자 하는 함수이름 <- function(입력변수이름) {
                         출력변수를 만드는 명령(함수구성)
                         return(출력변수이름)
                         }
```

예를 들어, 숫자 x(입력변수)를 입력하면 3배가 되도록 하는 함수는 다음과 같다.

```
threetimes <- function(x) {
              y = 3 * x
              return(y)
              }
```

이렇게 만들어진 함수를 사용하려면 **함수 이름**과 그 뒤에 괄호 안에 **입력 값**(예: 10)을 넣으면 된다.

```
threetimes(10)    # 출력값 : 30
```

```
# 입력변수가 두 개인 function 예제 (제곱 합 구하기)

squaresum <- function(x, y) {
                z <- x^2 + y^2
                return(z)
                }

squaresum(3, 4)
```

apply( )

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

```
x <- matrix(1:9, c(3,3))

apply ( x, 1, function(x) {2*x} )
```

여사건의 확률

사건 A의 여사건 A^c에 대하여

$$P(A^c) = 1 - P(A).$$

예제 11 세 개의 동전을 동시에 던질 때 뒷면이 적어도 한 개 나올 확률을 구하여라.

풀이 세 개의 동전을 동시에 던질 때 뒷면이 적어도 한 개 나오는 사건을 A라 하면 A의 여사건은 세 개 모두 앞면이 나오는 사건이므로

$$P(A^c) = \frac{1}{8}.$$

따라서 $P(A) = 1 - P(A^c) = 1 - \frac{1}{8} = \frac{7}{8}$이다.

```
# library( ) 를 이용하여 패키지 ‘prob’ 을 불러오기
library(prob)

# tosscoin( )을 이용하여 표본공간을 구함
( S <- tosscoin(3) )

# function( ) 을 이용하여 앞면의 개수를 세는 함수를 정의함
Hcount <- function(x)   sum(x == "H")

# subset( ) 을 이용하여 모두 앞면이 나오는 사건  Ac 를 구함
( Ac <- subset( S,  apply(S, 1, Hcount) == 3 )  )
```

참고.

```
Tcount <- function(x)   sum(x == “T")
( Ac <- subset( S,  apply(S, 1, Tcount) >= 1 )  )

# setdiff(  )를 이용하여 여집합을 구함
( A <- setdiff(S, Ac) )
```

```
# cat( )을 활용하여 서식과 함께 여러 객체를 출력함
cat("P(A) =1 - n(Ac)/n(S) = ", nrow(A), "/", nrow(S), "\n")
```

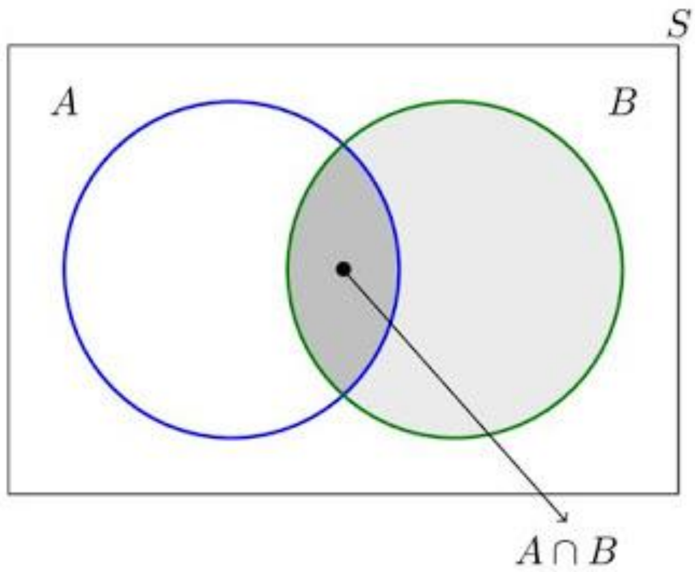
[정의] 조건부확률

어떤 사건 B가 발생했다는 조건 하에서 다른 특정 사건 A가 발생할 확률을 조건부확률(conditional probability)이라 다음과 같이 표시한다.

$$P(A\mid B)=\frac{P(A\cap B)}{P(B)}$$

$$P(A\mid B)=\frac{A\cap B\text{에 속한 원소의 개수}}{B\text{에 속한 원소의 개수}}$$

만일 사건 B가 일어난 것을 알았다면 그 실험결과는 B에 포함된 결과 중의 하나일 것이다. 그러므로 A가 일어날 확률을 계산하기 위하여, A가 발생하는 결과들을 B의 집합에서 찾아야 하므로 그러한 결과들은 집합 $A\cap B$ 의 원소이다.



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

조건부확률 $P(A \mid B)$ 은 표본공간이 S인 원래 실험으로부터 S의 부분집합인 사건 B를 새로운 표본공간으로 축소한 실험에서의 사건 A에 대한 확률을 의미한다.

예제 1 두 개의 주사위를 던지는 실험에서 두 주사위의 눈의 합을 X라고 하자. X가 홀수라고 할 때 X가 8보다 작을 확률을 구하여라.

3, 5, 7, 9, 11　　　3, 5, 7

풀이 ③ A가 $X<8$인 사건이고 ② B는 X가 홀수인 사건이라고 하면 ④ $P(A \mid B)$가 우리가 구하려는 확률이다. $A \cap B$는 X가 3, 5, 7인 사건이고 표본공간은 36가지이므로 ①

$$③\quad P(A \cap B) = \boxed{\frac{2}{36} + \frac{4}{36} + \frac{6}{36}} = \frac{1}{3},$$

$$②\quad P(B) = \boxed{\frac{2}{36} + \frac{4}{36} + \frac{6}{36}} + \frac{4}{36} + \frac{2}{36} = \frac{1}{2}$$

이다. 그러므로

$$④\quad P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{3}.$$

# library( ) 를 이용하여, 패키지 'prob' 를 불러오기
library(prob)

# rolldie( )를 이용하여, 표본공간을 구함
S <- rolldie(2)

subset( ) 을 이용하여, 두 주사위의 눈의 합이 홀수인 사건 B를 구함
(B <- subset(S, ((X1 + X2) %% 2) == 1))

subset( )을 이용하여, 두 주사위의 눈의 합이 8보다 작은 사건 A를 구함
(A <- subset(S, X1 + X2 < 8))

# intersect( )를 이용하여 교집합(사건)을 구함
AB <- intersect(A, B)

# cat( ) 을 활용하여 서식과 함께 여러 객체를 출력함
cat("P(A|B)=P(AB)/P(B)=", nrow(AB), "/", nrow(B), "\n")

rolldie( )
rolldie(times=시행횟수, makespace = FALSE)

Prob( ) ; Calculates probability and conditional probability of events.
Prob (S, event= , given= , ...)
[예제]

```
S <- rolldie(times=3, makespace=TRUE )
Prob(S, X1+X2 > 9 )
Prob(S, X1+X2 > 9, given = X1+X2+X3 > 7 )
```

```
# 또는 표본공간에 확률을 할당한 후, given 을 이용하여 조건부 확률을 구함
S <- rolldie (times=2, makespace=TRUE)
Prob(S, event=(X1 + X2<8), given=( (X1 + X2) %% 2) == 1 )
Prob(S, event=X1 + X2<8,  given= (X1 + X2) %% 2) == 1
Prob(S, X1 + X2<8, (X1 + X2) %% 2 == 1)
[1] 0.6666667
```

as.fractions( )를 이용하여, 분수로 표현함. 오류가 발생하는 경우에는 먼저 library(MASS)를 실행시킴

```
library(MASS)
as.fractions( Prob(S, X1 + X2<8, given=((X1 + X2) %% 2) == 1) )
[1] 2/3
```

R의 패키지는 분석 함수 뿐 아니라 예제 데이터도 포함하는 경우가 많다.
이 데이터를 패키지 데이터셋(Package data sets)이라고 한다.

현재 로드 된 모든 패키지가 제공하는 데이터셋의 목록을 확인하려면 data 명령을 사용한다.

```
data ( )
```

만약 특정 패키지(로드되었든 로드되지 않았든 상관없음) 가 제공하는 데이터셋 목록을 확인하고 싶다면 다음 명령을 사용한다.

```
data(package="MASS")
```

독립사건

일반적으로, 사건 B가 주어진 경우에 사건 A가 일어날 조건부확률 P(A | B)는 사건 A가 일어날 확률 P(A)와 같지 않다. 그러나 **어떤 사건 A가 일어났다는 사실이 사건 B가 일어날 확률에 영향을 미치지 않거나** 또는 반대로 **사건 B가 일어났다는 사실이 사건 A가 일어날 확률에 영향을 미치지 않는 경우**를 생각할 수 있다. 이를 수식으로 표현하면

$$P(A \mid B)=P(A), \quad P(B \mid A)=P(B)$$

이고, 이 식은 조건부 확률의 정의에 의해 다음과 같이 표현할 수 있다.

[정의] 독립사건

$$P(A|B) = P(A) = \frac{P(A \cap B)}{P(B)}$$

두 사건 A와 B가

$$P(A \cap B) = P(A) \cdot P(B)$$

를 만족할 때, A와 B는 서로 독립 (independent)이라고 한다.

예제 1 두 개의 주사위를 던지는 실험에서 두 주사위의 눈의 합을 X라고 하자. X가 홀수라고 할 때 X가 8보다 작을 확률을 구하여라.

3, 5, 7, 9, 11 3, 5, 7

풀이 ③ A가 $X<8$인 사건이고 ② B는 X가 홀수인 사건이라고 하면 ④ $P(A \mid B)$가 우리가 구하려는 확률이다. $A \cap B$는 X가 3, 5, 7인 사건이고 표본공간은 36가지이므로 ①

$$③ \quad P(A \cap B) = \boxed{\frac{2}{36} + \frac{4}{36} + \frac{6}{36}} = \frac{1}{3},$$

$$② \quad P(B) = \boxed{\frac{2}{36} + \frac{4}{36} + \frac{6}{36}} + \frac{4}{36} + \frac{2}{36} = \frac{1}{2}$$

이다. 그러므로

$$④ \quad P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{3}.$$

```
# [예제1]에서 사건 A와 B가 서로 독립인가?
# library( ) 를 이용하여, 패키지 'prob' 를 불러오기
library(prob)

S <- rolldie (times=2, makespace=TRUE)

PA <- as.fractions( Prob(S, event=(X1 + X2<8) ) ) ;  PA

PB <- as.fractions( Prob(S, event= (X1 + X2) %% 2 == 1 ) ) ;  PB

PAB <- as.fractions( Prob(S, event= X1 + X2<8 & (X1 + X2) %% 2 == 1 ) ) ; PAB

PA*PB == PAB
```

베이즈 정리란 새로운 조사나 실험을 통해 얻은 추가적인 정보가 주어질 경우 기존에 알고 있던 확률, 즉 사전확률에 이 새로운 정보를 반영하여 수정된 확률, 즉 사후확률을 계산하는 논리방법을 **베이즈 정리(Bayes' theorem)**라고 한다.

$A_1, A_2, \cdots, A_n$: 원인 B : 결과

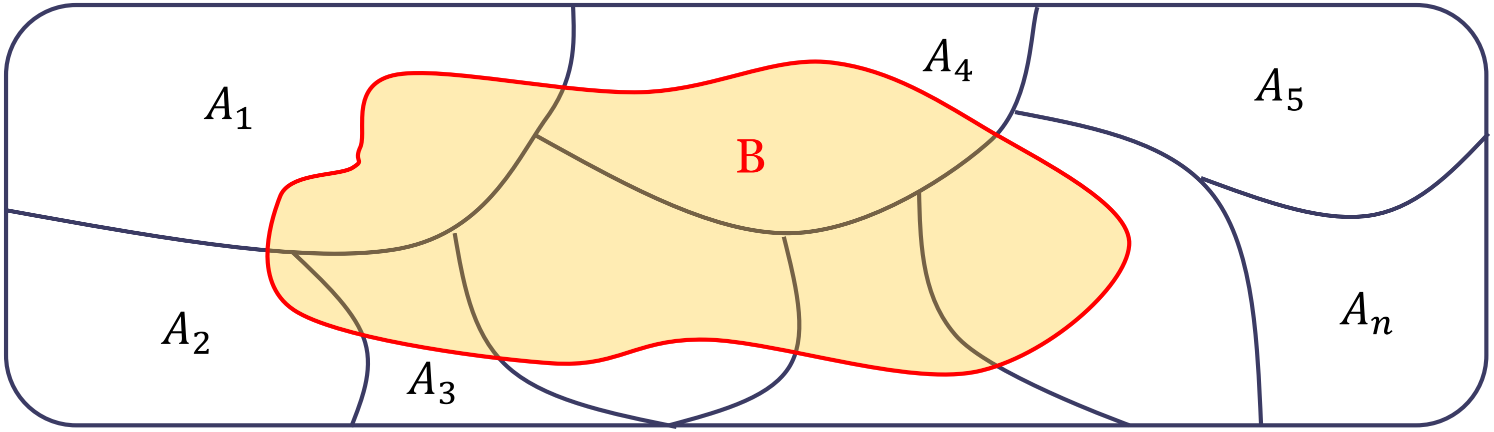
* **사전확률**(prior probability) : 관측자가 이미 알고 있는 정보, 결과가 나타나기 전에 결정되어 있는 확률로서 $P(A_1), P(A_2), \cdots, P(A_n)$ 을 의미

* **우도**(likelihood) : 이미 알고 있는 사건(원인)이 발생했다는 조건하에 결과(B)가 발생할 확률로 베이즈 정리에서 $P(B|A_1), P(B|A_2), \cdots, P(B|A_n)$을 의미함

* **사후확률**(posterior probability) : 사전확률과 우도를 통해서 알게 되는 조건부 확률, 즉 B라는 사건(결과)이 일어난 뒤에 원인별로 알아보는 확률로 베이즈 정리에서는 $P(A_k|B)$ 부분을 의미

[정리] 전확률 정리(theorem of total probability)

- 표본공간 S를 상호배반인 사상들로 분할(partition)



$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup \cdots \cup (A_n \cap B)$$

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \cdots + P(A_n \cap B)$$

$$P(A_i \cap B) = P(A_i)\, P(B|A_i)$$

$$P(B) = \sum_{k=1}^{n} P(A_k \cap B) = \sum_{k=1}^{n} P(A_k)\, P(B|A_k)$$

일반적으로, 어떤 사건 B의 발생 가능한 원인(요인)이 여러 가지인 경우에 원인 A_i 에서 발생할 확률은 조건부 확률의 정의

$$P(A_i|\, B) = \frac{P(A_i \cap B)}{P(B)}$$

와 전확률공식에 의하여 다음 성질을 얻을 수 있다.

[정리] 베이즈 정리(Bayes theorem)

$$P(A_i|B) = \frac{P(A_i)\, P(B|A_i)}{\sum_{k=1}^{n} P(A_k)\, P(B|A_k)}$$

[예제1] 어느 기업에서 세 개의 생산라인 A_1, A_2, A_3 를 이용하여 각각 전 제품의 20%, 30%, 50% 를 생산하고 각 라인에서 생산되는 제품의 불량률은 각각 1%, 2%, 3% 라고 하자. 이 기업에서 생산된 제품들 중에서 무작위로 뽑은 하나의 제품이 불량품이었다면, 이 불량품이 생산라인 A_2 에서 생산되었을 확률을 구하여라.

생산라인	A_1	A_2	A_3
생산비율	20%	30%	50%
불량률	0.01	0.02	0.03

B : 무작위로 뽑은 하나의 제품이 불량품인 사건

$$P(A_1) = 0.2, \qquad P(A_2) = 0.3, \qquad P(A_3) = 0.5$$

각 라인에서 생산된 하나의 제품이 불량품일 확률은

$$P(B|A_1) = 0.01, \qquad P(B|A_2) = 0.02, \qquad P(B|A_3) = 0.03$$

$$P(A_2|B) = \frac{P(A_2)\,P(B|A_2)}{\sum_{k=1}^{3} P(A_k)\,P(B|A_k)}$$

$$= \frac{(0.3)(0.02)}{(0.2)(0.01) + (0.3)(0.02) + (0.5)(0.03)}$$

$$= 0.26$$

```
# 사전확률을 prior 객체에 할당함
prior <- c(0.2, 0.3, 0.5)

# 조건부확률(우도)을 객체에 cond 할당함
cond <- c(0.01, 0.02, 0.03)

# sum( )을 이용하여 전확률을 구하고 tot 객체에 할당함
( tot <- sum(prior*cond) )

# 불량품일 확률을 구함
( post <- (prior*cond)/sum(tot) )

# 불량품이 라인 A2 에서 생산되었을 확률을 구함. round() 를 이용하여 소수
둘째짜리까지 나타냄
round(post[2], 2)
```

round(개체, digits= 소수점 아래 반올림 하고 싶은 위치)

```
# 사후확률 계산
prior = c(0.2, 0.4, 0.3, 0.1); cond = c(0.04, 0.02, 0.01, 0.05)
(tot = prior*cond); (stot = sum(tot))
[1] 0.008 0.008 0.003 0.005
[1] 0.024
(post = tot / stot)
[1] 0.3333333 0.3333333 0.1250000 0.2083333
# 결과 그림 ⇒ [그림 3-6]
bayes.plot(prior, post)
```