난수의 발생

- sample과 sample.int 함수
- 이산형 난수의 발생

sample과 sample.int 함수

출력

- 유한모집단에 필요한 개수의 난수를 뽑는 함수

성질

- 유한모집단이지만 복원추출(하나를 뽑고 다시 원상 복귀 후 다음 뽑기)인 경우 무한모집단과 같음

sample과 sample.int 함수

사용법

```
sample(x, size, replace = FALSE, prob = NULL)
```

```
sample.int(n, size = n, replace = FALSE, prob = NULL)
```

- sample 함수는 x에 주어진 원소들이 sample.int 함수는 1:n이 모집단임

매개변수

- x : 추출할 원소를 포함한 벡터를 주거나 자연수 값을 줄 수 있음
  벡터가 주어지면 주어진 벡터의 원소에서 표본을 추출하고, 자연수 값을 주면 1부터 x사이의 자연수에서 표본을 추출함
- n : 자연수 값을 지정하며, 1부터 n사이의 자연수에서 표본을 추출함
- size : 자연수를 지정하며 추출할 표본의 크기임
- replace : 논리값(T 또는 F)을 설정하며, T이면 복원추출, F이면 비복원 추출을 함(기본값 : False)
- prob : x의 원소 또는 1에서 n사이의 값이 관찰될 확률을 설정함
  이 값이 설정되지 않으면 확률은 1/length(x) 및 1/n으로 등확률임

이산형 난수의 발생

예제 1

표본설계 적용

어느 자치구에는 12개의 동이 있다. 조사를 위해서 이 중에서 4개의 동을 임의로 선택하려면 어떻게 하는 것이 좋은가?

12개의 동에 임의로 1부터 12까지 번호를 붙인 후 다음과 같이 난수를 발생하여 뽑힌 번호에 해당하는 동을 선택하기

```
> x <- seq(1,12)
> sample(x, size=4)
[1] 8 4 7 3
```

또는

```
> sample.int(12, size=4)
[1] 10 2 9 5
```

이산형 난수의 발생

예제 2

표본설계에 적용

뽑힌 네 개의 동에는 각각 19세 이상의 주민의 수가 12,000명, 7,000명, 5,000명 및 3,000명의 주민이 있다고 한다. 각 동에서 비례할당의 방법으로 1%의 주민을 조사대상으로 뽑으려면 어떻게 하는 것이 좋겠는지 알아보자.

뽑힌 번호에 해당하는 사람을 선택하기

```
sample.int(12000, size=120)
sample.int(7000, size=70)
sample.int(5000, size=50)
sample.int(3000, size=30)
```

이산형 난수의 발생

예제 2

- 번호를 순서대로 나오게 하게 위해 sort 함수를 사용하기

```
sort( sample.int(12000, size=120) )
sort( sample.int(7000, size=70) )
sort( sample.int(5000, size=50) )
sort( sample.int(3000, size=30) )
```

이산형 난수의 발생

예제 3

등확률이 아닌 경우

다음은 A, B, C, D가 관찰될 확률이 각각 0.1, 0.2, 0.3 및 0.4인 모집단에서 10개를 복원추출로 뽑는 경우를 모의 실험하는 과정이다.

```
> x <- c("A", "B", "C", "D")          # 모든 가능한 값을 포함한 벡터 설정
> p <- c(.1, .2, .3, .4)              # 각 값에 대응하는 확률
> sample(x, size=10, replace=T, prob=p)   # 복원 추출 10회
[1] "C" "C" "D" "D" "B" "D" "D" "D" "D" "C"
```

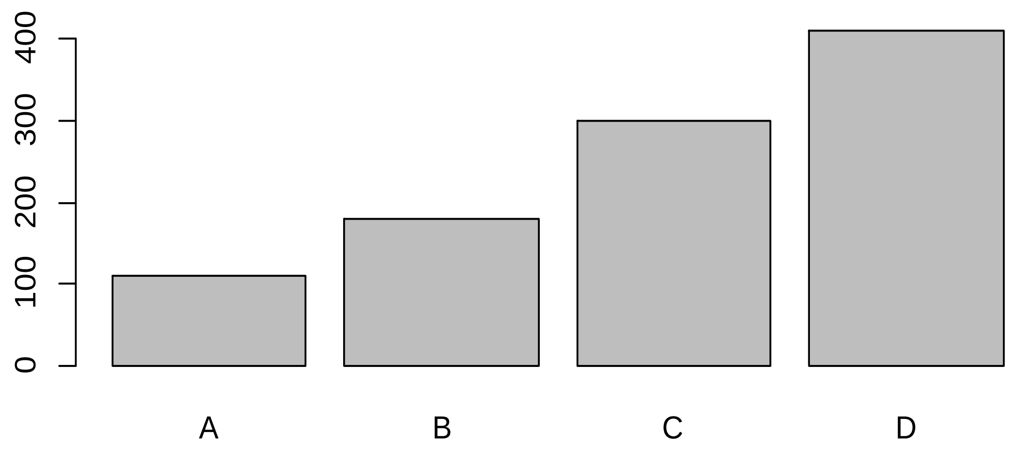
(복원 추출) 위의 보기에서 난수 1000개를 발생하여 도수분포표와 막대그래프를 그려 보자.

이산형 난수의 발생

예제 3

복원추출

```
x <- c("A", "B", "C", "D")
p <- c(.1, .2, .3, .4)
data <- sample(x, size=1000, replace=T,
               prob=p)
table(data)
data
  A   B   C   D
110 180 300 410
barplot(table(data))
```



➢ 이 경우 각각의 확률 0.1, 0.2, 0.3 및 0.4와 비슷한 비율로 A, B, C, D가 관찰됨을 확인 가능

이산형 난수의 발생

예제 4

동전 모의실험

공평한 동전을 1,000번 던져서 나오는 앞면의 수를 위의 sample 또는 sample.int 함수를 사용하여 얻어 보자.

sample 함수 사용하기

```
> sample(c(0,1), 1000, replace=T, prob=c(0.5, 0.5))
```

- R-언어로 0과 1이 될 확률이 각각 1/2인 난수를 1,000개 생성할 수 있음
- 앞면이 나온 경우를 1이라고 하면 1의 개수를 모두 합한 것이 앞면의 수이므로 이를 합하면 됨
- 위의 명령의 결과는 1000개의 0과 1

이산형 난수의 발생

예제 4

sum 함수를 사용하기

```
> sum( sample(c(0,1), 1000, replace=T, prob=c(0.5, 0.5)) )
```

- 이 경우 등확률이므로 prob는 설정하지 않아도 됨

이항분포

- 이항분포의 성질
- 이항분포 난수발생과 확률계산
- R-언어의 난수발생과 확률계산 함수 이름의 규칙
- 이항분포 난수발생과 확률계산

통계청 통계교육원

이항분포의 성질

이항분포

- 각각의 시행에서 성공확률이 성공 확률이 p이고 실패 확률이 $(1-p)$인 시행을 n번 독립적으로 시행할 때 성공의 횟수를 확률변수 X라고 하면 X의 분포를 이항분포라고 함
- $B(n,p)$로 표시하기도 함

X가 이항분포인 경우 확률밀도함수

$$\Pr[X=x] = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0,1,2,\ldots,n$$

이항분포 난수발생과 확률계산

이항분포의 확률값, 누적확률값, 분위수 및 난수의 발생

사용함수

```
dbinom(x, size, prob)
pbinom(q, size, prob, lower.tail = TRUE)
qbinom(p, size, prob, lower.tail = TRUE)
rbinom(n, size, prob)
```

매개변수

- x, q : 확률밀도함수값을 얻을 x 벡터, 누적확률을 얻을 분위수 q 벡터
- p : 분위수를 얻을 확률값의 벡터
- n : 발생할 난수의 개수
- lower.tail : 논리값을 설정하며 TRUE이면 확률은 $P[X \le x]$의 값으로 그렇지 않으면, $P[X > x]$을 계산함
- size : 이항분포 B(n, p)에서 n값을 설정함
- prob : 이항분포 B(n, p)에서 p값을 설정함

R-언어의 난수발생과 확률계산 함수 이름의 규칙

R-언어의 난수/확률밀도함수/누적확률함수/분위수 계산 함수의 이름 규칙

- R-언어는 본 과정에서 다루는 분포뿐만 아니라 다양한 분포에서 대해서 난수발생, 확률밀도함수값, 누적확률함수값, 분위수 계산을 할 수 있는 함수를 제공함

- 이들 함수가 가지고 있는 특정한 규칙
  - rname : 분포의 이름이 name인 분포에서 난수 발생하는 함수로 r이 접두어로 사용
  - dname : 분포의 이름이 name인 분포에서 확률밀도함수값을 계산하는 함수로 d를 접두어로 사용
  - pname : 분포의 이름이 name인 분포에서 누적확률함수 $F(x) = \Pr[X \le x]$를 계산
  - qname : 분포의 이름이 name인 분포에서 분위수를 계산

- 이항분포의 경우 분포이름으로 binom을 사용하고 각각 접두어에 r, d, p, q를 사용하였음을 알 수 있음

이항분포 난수발생과 확률계산

X가 이항분포 B(10, 0.2)인 경우

- $\Pr[X = 2]$

```
> dbinom(2, 10, 0.2) # B(10, 0.2)
[1] 0.3019899
```

- $\Pr[X \le 2]$

```
> pbinom(2, 10, 0.2) # B(10, 0.2)
[1] 0.6777995
```

* dbinom(0, 10, 0.2)+ dbinom(1, 10, 0.2)+dbinom(2, 10, 0.2)와 같은 값임

이항분포 난수발생과 확률계산

X가 이항분포 B(10, 0.2)인 경우

- $\Pr[X > 2]$

```
1- pbinom(2, 10, 0.2) # B(10, 0.2)
[1] 0.3222
```

```
pbinom(2, 10, 0.2, lower=F) # B(10, 0.2)
[1] 0.3222
```

- $P[X \leq a] = 0.5$가 되는 a

```
> qbinom(0.5, 10, 0.2)
[1] 2
```

이항분포 난수발생과 확률계산

예제 1

동전 모의실험

이항분포는 성공확률이 p인 실험을 독립적으로 n번 시행할 때 성공의 회수에 대한 분포이므로 앞면의 개수는 이항분포 $B(1000, 0.5)$를 따른다. 1000번의 동전 던지기에서 앞면의 개수를 모의실험해보자.

rbinom 함수 사용하기

```
> rbinom(1, 1000, 0.5)
[1] 488
```

* 488번의 앞면을 보는 실험을 난수를 통하여 대신함

- 실제 실험을 컴퓨터 등에서 대신하는 실험을 모의실험(simulation)이라 하며 이 경우 실제 동전을 1,000번 던지는 실험을 rbinom(1, 1000, 0.5) 한 명령으로 대신함
- 이 명령은 난수를 발생한 것이므로 이 명령을 실행할 때마다 488 대신 다른 숫자를 얻음(우연히 같은 값을 얻을 수도 있음)
- 동전던지기를 1,000번 할 때 앞면의 개수가 이항분포임을 알고 있다면 앞에서와 같이 앞면의 확률이 1/2인 난수를 1,000개를 얻지 않고 한 개의 이항분포 난수로 모의실험을 끝낼 수 있음을 확인함

이항분포 난수발생과 확률계산

예제 2

주사위던지기

공평한 주사위를 던지면 1의 눈금이 나올 확률이 1/6이다. 주사위를 다섯 번 던질 때 1의 눈금이 나올 회수를 확률변수 X라고 하면 X의 분포는 $B(5,1/6)$이며 X의 기댓값과 분산은 각각 $E[X] = np = 5/6 = 0.8333,\ Var(X) = npq = 25/36 = 0.6944$이다.
주사위를 다섯 번 던지는 실험을 100회 반복할 때 1의 눈금이 나온 회수의 평균과 분산이 이 값들과 얼마나 가까운지 모의실험을 해보자.

함수 작성하기

```
binom.par <- function(nrep=100, n=5, p=1/6) {
    x <- rbinom(nrep, n, p)   # B(n,p)에서 nrep 개의 난수 생성
    meanx <- mean(x)   # 이들 난수의 평균
    varx <- var(x)   # 이들 난수의 평균
    list(meanx = meanx, varx = varx)
} # end function
```

이항분포 난수발생과 확률계산

예제 2

함수 사용 결과

```
binom.par()
$meanx
[1] 0.75
$varx
[1] 0.7348485
```

- 이 값은 함수 안에서 사용한 x가 난수로써 호출할 때마다 달라지므로 매번 다른 값이 얻어지며(우연히 같을 수는 있지만 아주 희박한 확률) 그 값은 이론적인 주변에 있게 됨

초기하분포

- 초기하분포의 성질
- 난수발생과 확률계산

초기하분포의 성질

초기하분포(hypergeometric distribution)

- m개의 빨간 공과 n개의 흰 공이 들어가 있는 주머니에서 k개를 비복원으로 뽑을 때 빨간 공의 개수의 분포

사용함수

```
dhyper(x, m, n, k, log = FALSE)
phyper(q, m, n, k, lower.tail = TRUE, log.p = FALSE)
qhyper(p, m, n, k, lower.tail = TRUE, log.p = FALSE)
rhyper(nn, m, n, k)
```

초기하분포 난수발생과 확률계산

예제 1

로또 확률

로또는 45개 중에서 6개의 번호를 골라 이 중에서 당첨번호와 일치하는 개수에 따라 등수가 정해진다. 6개 중 3개 이상이 맞으면 당첨금을 받을 수 있다.
이 때 당첨번호와 일치하는 번호의 개수를 X라 하면 일치하는 번호의 개수가 x개일 확률은

$$\Pr[X=x] = \frac{\binom{6}{x}\binom{39}{6-x}}{\binom{45}{6}}, \quad x = 0,1,2,\ldots,6$$

x가 6이면(1등) 이 확률은 1/8,145,060=0.0000001227738이다. 이를 R에서 계산해보자.

초기하분포 난수발생과 확률계산

예제 1

풀이

- 로또의 경우 6개의 당첨번호(빨간 공)와 나머지 39개의 비당첨번호(흰 공)로 45개의 공이 들어가 있는 주머니에서 6개를 뽑는 경우로 m=6, n=39, k=6인 경우임

dhyper 함수로 확률 계산하기

```
> x <- 0:6
> dhyper(x, 6, 39, 6)    # 로또에서 0-6개 의 번호를 맞출 확률
[1] 4.005646e-01 4.241273e-01 1.514740e-01 2.244060e-02 1.364631e-03
[6] 2.872907e-05 1.227738e-07
```

* 위 값은 소수점사용시(반올림 적용)

0.400564, 0.4241273, 0.151474, 0.0224406, 0.00136463, 0.00002873 0.000001228

초기하분포 난수발생과 확률계산

예제 2

자료

위의 로또에서 3개 이상의 번호를 맞출 확률을 계산해보자.

phyper 함수로 확률 계산하기

```
> 1- phyper(2, 6, 39, 6)   # 로또에서 3개 이상의 번호를 맞출 확률
[1] 0.02383408
```

* 50번 중의 한 번 정도임 (5개의 번호와 특정번호 하나를 맞추는 2등은 제외)

동일한 난수 생성하기

- set.seed 함수
- 동일한 난수 생성: set.seed 함수

set.seed 함수

정의

- 한 번 발생한 난수를 그대로 발생하고자 하는 경우 동일한 난수를 얻게 설정하는 함수

사용함수

```
set.seed(seed, ... )
```

- 이 함수에서 seed에는 사용자가 임의로 정한 값(숫자)을 설정함

동일한 난수 생성: set.seed 함수

rbinom 함수의 호출

```
> set.seed(100)
> x <- rbinom(10, 100, 0.5)
> x
[1] 54 46 46 52 51 51 56 46 48 48
```

```
> x <- rbinom(10, 100, 0.5)
> x
[1] 49 51 55 52 47 50 48 55 42 49
```

```
> set.seed(100)
> x <- rbinom(10, 100, 0.5)
> x
[1] 54 46 46 52 51 51 56 46 48 48
```

- 첫 번째와 세 번째의 rbinom 함수 직전에 set.seed 함수를 사용하여 같은 값 100이 설정되었으므로 이 둘의 결과가 동일함
- 두 번째는 set.seed 함수의 호출없이 rbinom 함수가 호출되었으므로 임의의 난수가 생성됨