R컴퓨팅

10강

R 데이터 탐색 I

정보통계학과 장영재 교수

1. 모의실험
2. 복원 및 비복원 추출
3. 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

1 모의실험

1 모의실험

- 컴퓨터 프로그램 등을 사용하여 실제 실험을 대신한 가상의 실험을 하는 것을 모의실험(simulation)이라 함

실제 실험을 하는 것과 같은 상황을 만들고 이를 실행

- 공평한 동전던지기 실험을 예로 들면 동전의 앞면과 뒷면
또는 좀더 간단히 0과 1이 나올 확률이 각각 1/2인 프로그램을
만들고 이 프로그램을 1,000번, 백만 번, 또는 1억 번 반복 실행

1 모의실험

- 이와 같은 모의실험은 통계학에서 특정한 분포를 갖는 확률변수를 컴퓨터 프로그램 등으로 생성하여 통계량의 특성 등을 규명하기 위해 사용하며, 확률변수의 값을 컴퓨터 등을 통해 생성한 것을 확률난수(pseudo random number) 또는 줄여서 난수라고 부름
  (모의실험에 관해서는 9장에서 다시 자세히 살펴봄)

2 복원 및 비복원 추출

2 복원 및 비복원 추출

- 복원추출이란 한번 관찰한 값을 다시 관찰할 수 있도록 모집단에 다시 포함시키는 추출방법이며 비복원 추출은 한번 관찰한 값은 다시 모집단에 포함하지 않는 추출방법

- R에서는 모집단의 모든 가능한 값을 포함한 벡터 x의 원소의 값을 복원 또는 비복원으로 임의로 추출하는 함수로 sample 함수와 sample.int 함수가 제공

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

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

2 복원 및 비복원 추출

- 보기 10-1: A, B, C, D가 관찰될 확률이 각각 0.1, 0.2, 0.3 및 0.4인 모집단에서 10개를 복원추출로 뽑는 경우를 모의 실험하는 과정

```
> 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"
```

2 복원 및 비복원 추출

- 보기 10-2: 1부터 1000 사이의 자연수에서 비복원으로 10개의 난수를 얻기 위한 명령문

```
> sample.int(1000, size=10)
 [1]  29 735 623 346 694 699 360 646 148 459
```

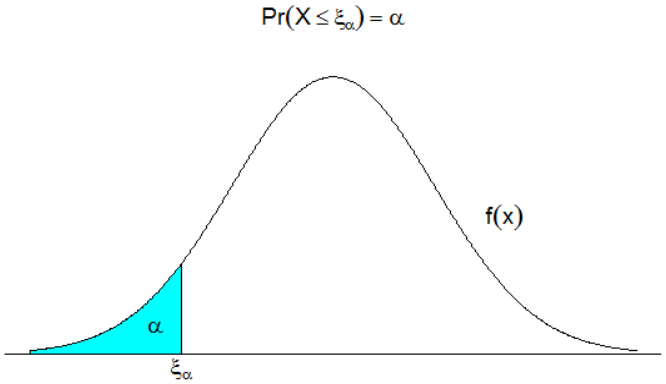
- 보기 10-3: 공평한 동전을 1000번 던져서 나오는 앞면의 수를 위의 sample 또는 sample.int 함수를 사용하여 구하기

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

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

- 0<α<1인 α에 대해 확률변수 X의 α분위수 ε_α는 P(X< ε_α)=α가 되는 점을 말하며 백분율 개념을 추가하여 100α% 백분위수 (percentile)라고 부름



- R에서 통계적 분포 관련 함수들의 이름은 규칙을 가지고 있음

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

- 함수 이름의 첫 글자는 난수발생의 경우 r, 확률밀도함수는 d, 누적분포함수는 p, 분위수는 q를 첫 문자로 사용

- 두 번째 글자부터는 분포이름의 약자를 사용

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

1 일양분포(uniform distribution)

- 일양분포( )는 구간 의 부분집합인 임의의 구간에서 구간의 길이가 같으면 확률이 같은 분포를 의미

$$f(x) = \frac{1}{b-a}, \qquad a < x < b$$

- 일양분포의 확률밀도함수, 누적분포함수, 분위수 및 난수는 각각 dunif, punif, qunif 및 runif로 얻을 수 있음

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

1 일양분포(uniform distribution)

```
dunif(x, min=0, max=1, ...)
punif(q, min=0, max=1, lower.tail = TRUE, ...)
qunif(p, min=0, max=1, lower.tail = TRUE, ...)
runif(n, min=0, max=1)
```

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

1 일양분포(uniform distribution)

- x, q: 각각 확률밀도함수값을 얻을 x 벡터 및 누적확률을 얻을 q 벡터를 설정
- p: 분위수를 얻을 확률값의 벡터
- n: 발생할 난수의 개수를 설정
- min, max: 일양분포의 범위로서 기본값은 각각 0과 1이다.
- lower.tail 논리값을 설정하며 TRUE이면 확률은 $P[X \leq x]$의 값으로 그렇지 않으면, $P[X > x]$을 계산

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

1 일양분포(uniform distribution)

› 보기 10-4: $U(0,1)$은 0부터 1사이에 균일한 분포 확률을 가지므로 $f(x)=1,\ 0<x<1$임

```
〉dunif(1)     # U(0,1)의 f(1)의 값
[1] 1
〉punif(0.5)          # U(0,1)에서  0.5보다 같거나 작을 확률
[1] 0.5
```

100개의 난수를 생성하여 평균을 계산해보면,

```
〉mean(runif(100))
[1] 0.4954645
```

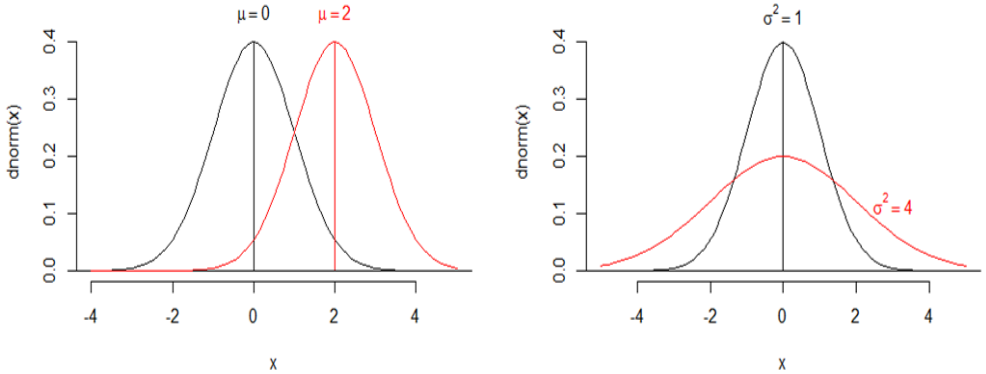
로 0.5에 근접한 값을 얻을 수 있음

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도 함수 및 분위수

2 정규분포

› 정규분포는 통계학에서 가장 자주 사용되는 분포 중의 하나이며 확률변수 가 정규분포를 따르면 평균 에 대해 좌우 대칭이며 분산이 클수록 더 많이 흩어진 분포를 갖게 됨

※평균과 분산에 따른 정규분포의 모양



3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

2 정규분포

- 모든 정규분포는 평균 에 대해서 좌우 대칭이며 분산이 같을 때 평균이 커지거나 작아지면 분포의 모양은 변하지 않고 분포만 오른쪽 또는 왼쪽으로 평행 이동하게 됨

- 평균이 같을 때 분산이 커지면 중심의 위치는 변하지 않으나 흩어짐의 정도가 커져 좌우로 더 많이 흩어진 분포를 보이게 됨

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

2 정규분포

- 기댓값 , 분산 인 정규분포에서의 난수, 확률밀도함수, 누적분포함수 및 분위수는 각각 rnorm, dnorm, pnorm 및 qnorm 함수로 얻을 수 있음

```
dnorm(x, mean = 0, sd = 1, ...)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE ,...)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, ...)
rnorm(n, mean = 0, sd = 1)
```

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

2 정규분포

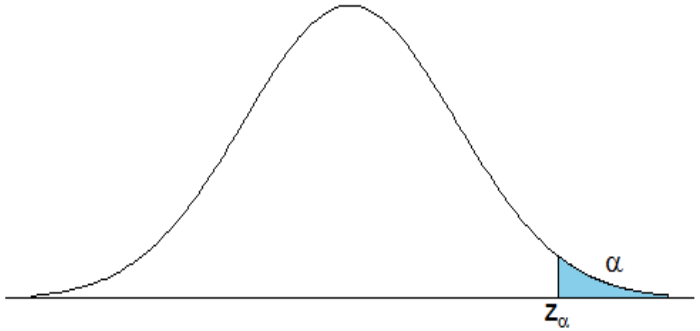
- mean에는 기댓값 , sd에는 표준편차 를 지정하며 생략된 경우 mean과 sd는 각각 0과 1인 표준 정규분포를 사용

- x, q: 확률밀도함수와 누적분포함수의 값을 얻고자 하는 벡터를 설정
- p: 분위수를 얻고자 하는 확률값의 벡터를 설정
- n: 생성할 난수의 개수를 지정
- lower.tail 논리값을 가지며 TRUE이면(기본값), 확률이 $Pr[X \leq x]$로 계산되며 FALSE이면 $Pr[X > x]$ 로 계산

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

2 정규분포

- 보기 10-5: 통계학에서 z_α를 표준정규분포의 $100(1-\alpha)\%$ 백분위수로 표시. 예를 들어 $z_0.05$는 95% 백분위수



〈그림 6-3〉 의 위치 및 해당 영역

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

2 정규분포

표준정규분포에서 2.5% 백분위수 z_0.975는

```
〉qnorm(0.025)
[1] –1.959964
```

로 잘 알려진 값 -1.96임을 알 수 있음(반올림 적용)

```
〉dnorm(c(–1, 0, 1))
[1] 0.2419707 0.3989423 0.2419707
```

은 표준정규분포의 확률밀도함수

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

의 값을 각각 -1, 0, 1에서의 값을 계산한 것임

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

2 정규분포

```
>pnorm(c(-2.54,-1.96,0, 1.96, 2.54))
[1] 0.005542623 0.024997895 0.500000000 0.975002105
0.994457377
```

은 Z가 표준정규분포일 때

$$\Pr[Z \leq -2.54], \Pr[Z \leq -1.96], \Pr[Z \leq 0],$$
$$\Pr[Z \leq 1.96], \Pr[Z \leq 2.54]$$

의 확률을 계산한 것임

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

2 정규분포

- 보기 10-6: 표준정규분포로부터 난수를 n개 만들어 모평균에 대한 95% 신뢰구간을 구하면 신뢰구간은 (α=0.05)

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \bar{x} \pm z_{\alpha/2}\frac{1}{\sqrt{n}}$$

- n=10 인 경우의 95% 신뢰구간을 1000번 계산하여 1000개의 신뢰구간 중 모평균 0을 포함하는 경우를 구해보기

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

2 정규분포

```
> z.ci <- function(alpha = 0.05, nrep = 1000) {
+ ndata <- 10          # 신뢰구간을 계산할 자료의 수
+ qz <- qnorm(1-alpha/2)
+ se <- 1/sqrt(ndata)
+ ncover <- 0        # 신뢰구간이 0을 포함하는 회수
+ for (i in 1:nrep) {            # nrep 번 (기본값 1000번) 반복
+ x <- rnorm(ndata)      # ndata 개(기본값 10개)의 난수 생성
+ meanx <- mean(x)         # ndata 개의 평균
+ ubound <- meanx + qz*se      # 신뢰상한
+ lbound <- meanx - qz*se      # 신뢰하한
```

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

2 정규분포

```
+ if ( ubound > 0 & lbound < 0) ncover = ncover + 1
+       # 신뢰구간에 포함되는 회수를 계산
+ }         # end for
+ list(ncover=ncover)          # 출력
+ }             # end function
> z.ci()
$ncover
[1] 943
```

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

3 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 하며 B(n,p)로 표시

- X가 이항분포이면 확률밀도함수는

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

- 이항분포의 확률값, 누적 확률값, 분위수 및 난수의 발생은 다음과 같은 함수로 얻을 수 있음

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

3 이항분포

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

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

3 이항분포

- 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 $P[X \leq x]$의 값으로 그렇지 않으면, $P[X > x]$을 계산
- size: 이항분포 B(n, p)에서 n값을 설정
- prob: 이항분포 B(n, p)에서 p값을 설정

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

3 이항분포

› 보기 10-7: X가 이항분포 B(10, 0.2)이면 Pr[X=2]는

```
〉 dbinom(2, 10, 0.2)
[1] 0.3019899
```

로 얻으며 Pr[X≤2]는

```
〉 pbinom(2, 10, 0.2)
[1] 0.6777995
```

로 얻을 수 있다. 이 값은

```
〉dbinom(0, 10, 0.2)+ dbinom(1, 10, 0.2)+dbinom(2, 10, 0.2)
```

와 같은 값임을 확인할 수 있다.

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

3 이항분포

› Pr[X>2]는

```
> 1- pbinom(2, 10, 0.2)
```

또는

```
> pbinom(2, 10, 0.2, lower=F)
```

로 0.3222 임을 얻을 수 있다. $P[X \leq \alpha]=0.5$가 되는 α는

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

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

3 이항분포

› 보기 10-8: 주사위를 다섯 번 던지는 실험을 100회 반복할 때 1의 눈금이 나온 회수의 평균과 분산이 이 값들과 얼마나 가까운지 알아보기

```
> 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 분산
```

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

3 이항분포

```
+ list(meanx = meanx, varx = varx)
+ }                          # end function
> binom.par()
$meanx
[1] 0.75
$varx
[1] 0.7348485
```

를 얻는다.

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

4 초기하분포 (Hypergeometric distribution)

- 주머니에 m개의 흰색 공과 n개의 검은 공이 있을 때 임의로 k개를 비복원 추출로 꺼내는 경우 흰색 공의 개수가 x개가 될 확률은

$$\Pr[X=x] = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} \qquad \max(0, n-k) \le x \le \min(m, k)$$

이고 이 때 확률변수 X의 분포를 초기하분포라고 함

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

4 초기하분포 (Hypergeometric distribution)

› 초기하분포의 확률값, 누적확률, 백분위수, 난수 발생은

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

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

4 초기하분포 (Hypergeometric distribution)

- x, q: 분포함수를 계산하거나 누적확률을 계산할 벡터
- p: 100p% 백분위를 계산할 확률값 p를 저장한 벡터의 이름
- m: 주머니 속의 흰색공의 개수
- n: 주머니 속의 검은색 공의 개수
- k: 주머니에서 비복원 추출하는 공의 수
- p: 분위수를 계산할 확률값의 벡터
- nn: 난수를 발생할 개수
- lower.tail: 논리값으로 기본값인 TRUE이면 확률이 $P[X \leq x]$로 계산되며 FALSE이면 $P[X > x]$이 구해짐

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

4 초기하분포 (Hypergeometric distribution)

› 보기 10-9: 주머니 속에 빨간공 5개 파란 공 6개가 있을 때, 이 주머니에서 3개의 공을 비복원으로 추출할 때 빨간 공의 수가 2일 확률 구하기

```
〉 dhyper(2, 5, 6, 3)
[1] 0.3636364
```

이며 이 값은

```
〉choose(5,2)*choose(6,1)/choose(11,3)
```

로 얻어도 같은 결과를 얻을 수 있음

3 특정 분포에서의 확률난수, 누적분포함수, 확률밀도함수 및 분위수

4 초기하분포 (Hypergeometric distribution)

빨간 공의 개수가 2 이상일 확률은 1-Pr[X≤1]로도 얻을 수 있으므로

```
〉 1-phyper(1,5,6,3)
[1] 0.4242424
```

이며 이 값은 직접 빨간 공의 수가 2 또는 3일 확률을

```
〉choose(5,2)*choose(6,1)/choose(11,3) +
choose(5,3)*choose(6,0)/choose(11,3)
```

또는

```
〉dhyper(2,5,6,3) + dhyper(3,5,6,3)
```

로도 얻을 수 있음

R컴퓨팅