3강 통계계산(1)

한림대학교 금융정보통계학과 심송용교수

목 차

1 모의실험

모의실험(simulation)이란?

- 모의실험은 현실 세계 또는 시스템에서 발생하는 과정을 흉내(imitation) 내는 것
- 대부분 컴퓨터를 사용하며 이런 경우 컴퓨터 모의실험이라고도 함.
- 공평한 동전을 1,000번 던질 때 앞면이 몇 번 나오는지
  - 실제로 1,000번을 던져볼 수 있으나
  - 컴퓨터를 사용하여 0과 1사이의 실수를 임의로 발생하여 이 값이 ½보다 작으면 앞면이 나오는 것으로 판정
  - 위의 0과 1사이의 수를 난수, 또는 확률난수라고 함.

2 복원 및 비복원추출

복원 및 비복원 추출

- **복원추출** : 한번 관찰한 값을 다시 관찰할 수 있도록 모집단에 다시 포함시키는 경우의 추출방법.
- **비복원추출** : 한번 관찰한 값은 다시 모집단에 포함하지 않는 추출방법.
- **R-언어에서의 복원 및 비복원 추출 함수** : sample 또는 sample.int 함수

```
sample(x, size=n, replace = FALSE[TRUE], prob = NULL)
sample.int(n, size = n, replace = FALSE[TRUE], prob = NULL)
```

- x : 추출할 원소를 포함한 벡터를 주거나 자연수 값을 설정.
  - 벡터가 주어지면 주어진 벡터의 원소에서 표본을 추출하고,
  - 자연수 값을 주면 1부터 x사이의 자연수값 사이에서 표본을 추출한다.
- n : 자연수 값을 지정하며, 1부터 n사이의 자연수에서 표본을 추출.

복원 및 비복원 추출

- size : 자연수를 지정하며 추출할 표본의 크기(개수).
- replace : 논리값(T 또는 F)을 설정하며, T이면 복원추출, F이면 비복원 추출. 비복원 추출이 기본값임.
- prob : x의 원소(sample 함수) 또는 1에서 n사이의 값(sample.int 함수)이 관찰될 확률을 설정. 이 값이 설정되지 않으면 확률은 1/length(x) 및 1/n으로 등확률이 자동 설정됨.

예제 2.1. A, B, C, D가 관찰될 확률이 각각 0.1, 0.2, 0.3 및 0.4인 모집단에서 10개를 복원추출로 뽑는 모의실험을 해보자.

```
> 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
[1] "C" "C" "D" "D" "B" "D" "D" "D" "D" "C" # 결과가 매번 달라짐
```

복원 및 비복원 추출

예제 2.2. 1부터 1,000 사이의 자연수에서 비복원으로 10개의 난수를 추출하자.

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

예제 2.3. 공평한 동전을 1,000번 던져서 나오는 앞면의 수를 위의 sample 또는 sample.int 함수를 사용하여 얻어 보자.

```
> ntoss = sample(c(0,1), 1000, replace=T, prob=c(0.5, 0.5))
> # 1이 앞면
> ntoss2 = sample.int(2, 1000, replace=T) -1
> sum(ntoss) # 1(앞면)의 수

[1] 495
```

복원 및 비복원 추출

예제 2.4. 로또는 1부터 45사이의 숫자를 순서에 상관없이 6개를 맞추면 1등이다. 5등은 3개를 맞추는 경우라고 한다. 만일 당첨번호가 1,2,3,4,5,6 라고 할 때 백만 번 모의실험을 하여 로또 5등에 당첨될 확률을 얻어 보자.

```
lotto <- function(nn = 10000) { # lotto.r
luckyNo <- c(1,2,3,4,5,6) # 당첨번호
threeNo <- 0 # 번호 세 개의 맞는 회수
for (i in 1: nn) { # nn 번 모의실험
    x <- sort(sample.int(45, size=6))
    # 1에서 45사이의 난수 6개 생성하여 오름차순
    nMatch = 0
    for (j in 1:6) { # 임의의 x 값에 대해
       for(k in j:6) { # 당첨번호 LuckNo와 같은지 비교
```

복원 및 비복원 추출

```
            if (x[j] == luckyNo[k]) nMatch = nMatch +1
            # 각 번호를 당첨번호와 비교
        } # end for k
      } # end for j
      if (nMatch == 3) threeNo = threeNo + 1 # 세 개의 번호가 일치한 횟수
  } # end for i
  list(threeNo = threeNo)
} # end function
> lotto(10000)
$threeNo
[1] 241
```

10,000번 중 241번이 5등에 당첨되었다.
(결과는 매번 달라질 수 있음)

복원 및 비복원 추출

5등 당첨확률 : $\frac{{}_6C_3 \cdot {}_{39}C_3}{{}_{45}C_6} = \frac{\binom{6}{3}\binom{39}{3}}{\binom{45}{6}} = 0.0224406$

choose(6,3)*choose(39,3)/choose(45,6) = 0.0224406

choose 함수는 R-언어의 내장함수로 choose(n, x)로 사용하며

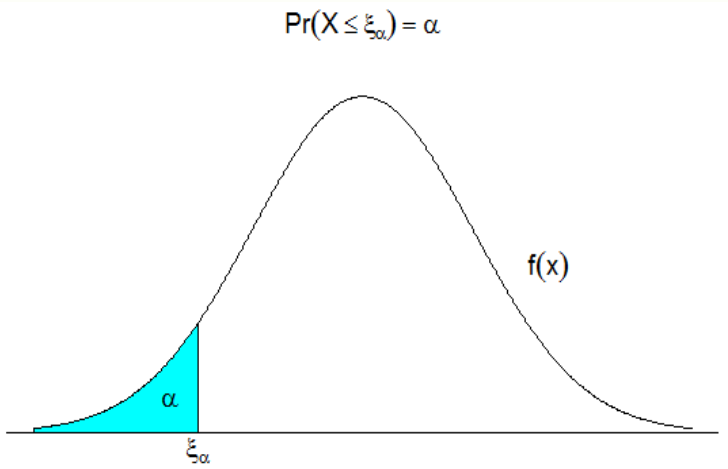
$${}_nC_x = \binom{n}{x} = \frac{n!}{(n-x)!x!}$$

을 반환하는 R 내장함수.

3 난수, 분포함수, 확률밀도함수, 분위수

난수와 분위수

- 확률난수 : 특정한 확률분포를 갖는 난수.
- 분위수 : 확률변수 X의 α **분위수** ξ_α는 Pr $[X < \xi_\alpha] = \alpha$ 가 되는 점을 말하며 백분율 개념을 추가하여 100 α % **백분위수**라고도 함.



확률밀도(질량)함수와 분포함수

- **확률질량함수** : 확률변수 X가 이산일 때 X가 특정한 값 x가 될 확률을p(x) 또는 f(x)로 표현함.
- X가 연속인 경우 X가 임의의 구간 (a,b) 사이에 포함될 확률 $\Pr[a < X < b]$이 $\int_a^b f(x)dx$ 로 표현되는 함수 f(x)를 **확률밀도함수**라고 함.
- X가 이산인 경우에도 확률질량함수를 확률밀도함수라고 함께 표현하기도 함.
- **분포함수** : $\Pr[X \leq x]$의 값을 말하며 F(x)로 표현.

R–언어의 규칙

- d + 분포이름 : x 가 매개변수로 주어졌을 때 에서의 확률밀도함수의 f(x)값 계산.
- p + 분포이름 : q가 매개변수로 주어졌을 때 q까지의 확률값(분포함수의 값) $Pr[X \leq q]=F(q)$을 계산.
- q + 분포이름 : p가 매개변수로 주어졌을 때 $Pr[X \leq x] = p$가 되는 x의 값, 즉 $F(x) = p$를 만족하는 x를 계산하며 이는 100p% 백분위수.
- r + 분포이름 : n이 매개변수로 주었을 때, 분포이름의 분포를 갖는 난수를 n개 생성.

R-언어의 분포관련규칙

- 분포이름은
  - 정규분포는 norm,
  - 이항분포는 binom 등
- 따라서
  - qnorm은 정규분포의 분위수를,
  - rnorm은 정규분포에서의 난수를,
  - dbinom 함수는 이항분포의 확률밀도함수값을,
  - pbinom은 이항분포의 누적확률을 계산한다.

4 일양분포, 정규분포

일양분포(uniform distribution)

일양분포는 구간 (a,b)의 부분집합인 임의의 구간에서 구간의 길이가 같으면 확률이 같은 분포를 말하며 균일분포라고도 함. 확률밀도함수는

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

이며 $X \sim U(a,b)$ 로 표현. R-언어의 함수는

```
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)
```

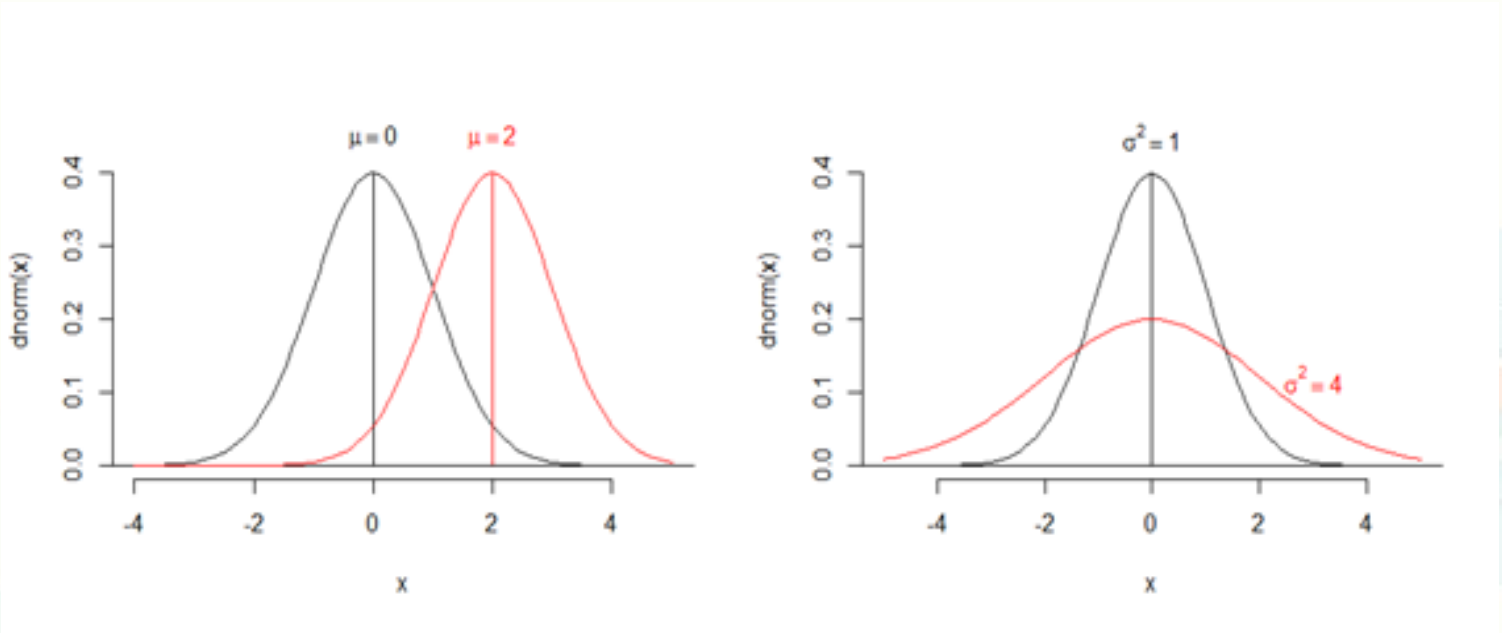
일양분포(uniform distribution)

- x, q : 각각 확률밀도함수값을 얻을 x 벡터, 누적확률을 얻을 q 벡터.
- p : 분위수를 얻을 확률값의 벡터.
- n : 발생할 난수의 개수.
- min, max : 일양분포의 범위. 즉 X ~ U의 분포를 사용하여 필요한 값을 반환한다. min과 max의 기본값은 각각 0과 1.
- lower.tail 논리값을 설정하며 TRUE이면 확률은 Pr[X ≤ x] 의 값을, 그렇지 않으면, Pr[X 〉 x]을 계산함.

```
> 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
> mean(runif(100)) # 100개의 난수의 평균
[1] 0.4954645
```

정규분포

- 가장 자주 사용하는 분포(연속, 종모양, 좌우대칭 등의 성질).
- 기댓값 μ, 분산 σ^2는 각각 중심의 위치 및 퍼짐의 정도.



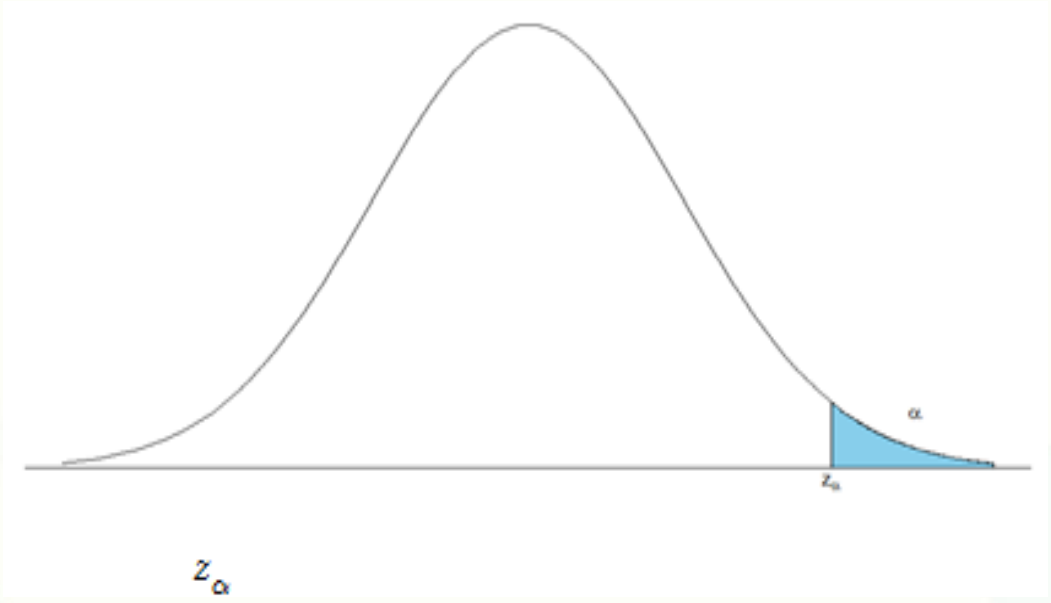
정규분포

```
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)
```

- x 또는 q : 확률밀도함수와 누적분포함수의 값을 얻고자 하는 분위수 벡터.
- p : 분위수를 얻고자 하는 확률값의 벡터.
- n : 생성할 난수의 개수.
- lower.tail : 논리값을 설정하며 TRUE이면 확률은 $Pr[X \leq x]$ 의 값을, 그렇지 않으면, $Pr[X > x]$을 계산함.
- mean : 정규분포의 모평균. 기본값은 0.
- sd : 정규분포의 표준편차. 기본값은 1.

정규분포

예제 2.6. 통계학에서 z_α는 표준정규분포에서 제 100(1−α)% 백분위수로 표시한다. 예를 들어 $z_{0.05}$ 는 제 95% 백분위수이다.



정규분포

R 언어에서 표준정규분포에서 2.5% 백분위수 $z_{0.975}$ 는

```
> 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

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

의 값을 각각 x 가 –1, 0, 1에서의 값을 계산.

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 계산.

예제 2.7. 평균이 μ , 분산이 σ^2인 정규분포로부터 독립인 n 개의 자료를 얻어서 평균을 얻으면 이 평균의 분포는 기댓값 μ , 분산 σ^2/n인 정규분포를 따른다.

정규분포

예를 들어 평균이 0, 분산이 5인 10개의 자료에서 평균을 계산하면 이 표본 평균은 기댓값 0, 분산 5/10 = 0.5인 정규분포를 따른다. 이를 확인하기 위해 10개의 정규분포 난수에서 평균을 얻는 것을 1,000번 반복하여 이 1,000개의 평균들의 평균과 분산을 계산해보자.

```
z.mean <- function(nn=10, nrep=1000) { #z.mean.r
  xbar <- rep(0, nrep)     # nrep 개의 평균을 저장할 배열
  stdev <- sqrt(5)          # 표준편차
  for (i in 1:nrep) {      # nn 개의 평균을 nrep번 계산
    xbar[i] <- mean(rnorm(nn, 0, stdev))
  } # nrep 개의 평균을 계산
  list(meanxbar= mean(xbar), varxbar=var(xbar))
} # end function
```

정규분포

```
> z.mean()
$meanxbar
[1] -0.007624802
$varxbar
[1] 0.5330471
```

결과를 보면, 평균과 분산이 각각 −0.00076, 0.5330 으로 이론적 기댓값인 0과 분산인 0.5에 가까운 값을 얻음을 알 수 있다.

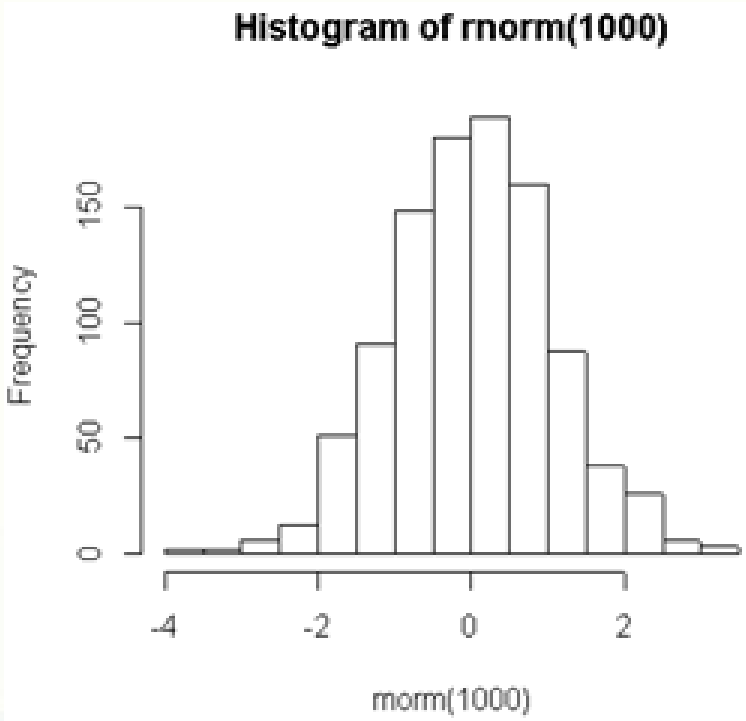
예제 2.8. 표준정규분포로부터의 난수 1,000개를 발생하여 이를 사용하여 히스토그램을 그려보자.

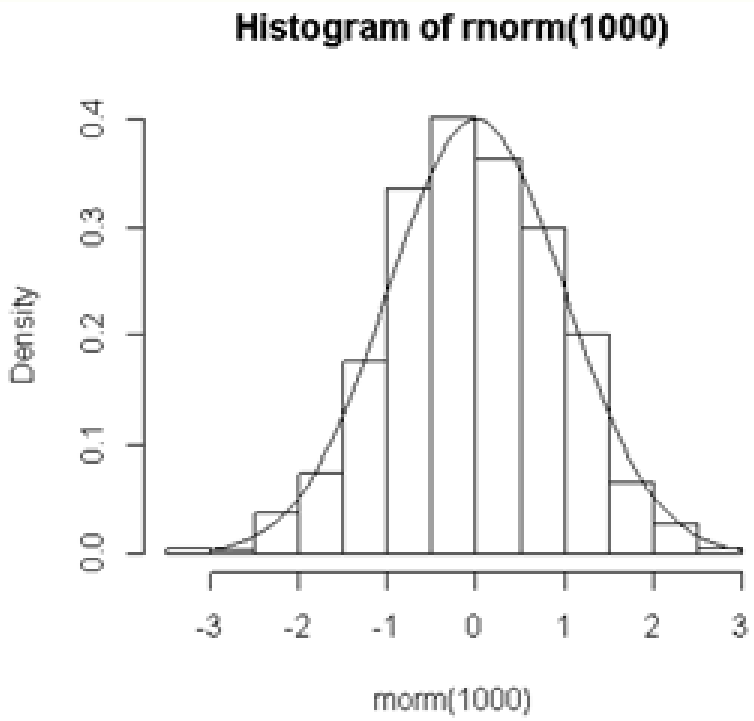
정규분포

```
z.hist <- function(nn=1000) {  # z.hist.r
 par(mfrow=c(1,2))            # 그림을 좌우로 두 개 그림
 hist(rnorm(1000))            # 난수 1000개의 히스토그램
 hist(rnorm(1000), freq=F) # 난수 1000개의 확률 히스토그램
    # freq가 F 이면 기둥의 높이는 상대도수
 lines(x<-seq(-3,3, length=100), dnorm(x))
    # lines 함수를 사용하여 –3에서 3사이의 x 값을 사용하여
    # dnorm(x)의 값(표준정규분포의 확률밀도함수 값)을 선으로
    # 추가함
} # end function

> z.hist()
```

정규분포


Histogram of rnorm(1000)
Frequency
0
50
100
150
-4
-2
0
2
morm(1000)


Histogram of rnorm(1000)
Density
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
morm(1000)

정규분포

예제 2.9. 표준정규분포로부터 난수를 10개 만들어 모평균에 대한 95% 신뢰구간을 구하면 신뢰구간은

$$\mu = \bar{x} \pm z_{0.025} \frac{\sigma}{\sqrt{n}} = \bar{x} \pm z_{0.025} \frac{1}{\sqrt{10}}$$

이다.

이 95% 신뢰구간을 1,000번 계산하여 1,000개의 신뢰구간 중 알고자 하는 모평균의 값인 0을 포함하는 경우가 몇 번이나 되는지 알아보자.

정규분포

```
z.ci <- function(alpha = 0.05, nrep = 1000) { # z.ci.r
  ndata <- 10                    # 신뢰구간을 계산할 자료의 수
  qz <-  qnorm(1-alpha/2)        #  z_{α/2}
  se <- 1/sqrt(ndata)            #  σ/√n
  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
    if ( ubound > 0 && lbound < 0) ncover = ncover + 1
    #신뢰구간에 포함되는 개수
  } # end for
```

정규분포

```
list(ncover=ncover)  # 출력
} # end function

> z.ci()
$ncover
[1] 943
```

- 이 결과는 1000개의 95% 신뢰구간 중 실제 모평균 0을 포함하는 경우는 943번임.
- 이 결과는 매번 달라지지만 950에서 크게 벗어나지 않음.

5 카이제곱분포

카이제곱분포

- $Z_1, Z_2, \cdots, Z_\nu$ 가 서로독립인 표준정규분포에서 확률변수라고 하면

$$X = Z_1^2 + Z_2^2 + \cdots + Z_\nu^2$$

의 분포를 자유도가 ν인 **카이제곱분포**라고 하며, 기호로는 $X \sim \chi_\nu^2$ 또는 $X \sim \chi^2(\nu)$로 표현.

- 만일 Z_i들의 분포가 독립인 정규분포이기는 하나 기댓값이 0이 아닌 경우 X 의 분포는 비중심 카이제곱분포.
- 자유도 ν인 카이제곱분포의 기댓값은 ν 분산은 2ν.
- 자유도가 df, 비중심모수 ncp인 카이제곱분포에서의 난수, 확률밀도함수, 누적분포함수 및 분위수는 각각 rchisq, dchisq, pchisq 및 qchisq 함수로 얻음.

카이제곱분포

```
dchisq(x, df, ncp=0, log = FALSE)
pchisq(q, df, ncp=0, lower.tail = TRUE, log.p = FALSE)
qchisq(p, df, ncp=0, lower.tail = TRUE, log.p = FALSE)
rchisq(n, df, ncp=0
```

- x, q : 확률밀도함수, 또는 확률을 계산할 값을 저장한 벡터의 이름을 설정.
- p : 100p% 백분위수를 계산할 p값의 벡터를 설정.
- n : 발생할 난수의 개수를 설정. 만일 n이 스칼라가 아닌 벡터이면 벡터의 길이만큼 난수를 발생.
- df : 자유도를 설정.
- ncp : 비중심모수를 설정. 비중심모수는 음수가 될 수 없으며 0인 경우는 중심카이제곱분포.
- lower.tail : 논리값을 설정하며 TRUE이면 확률은 $Pr[X \leq x]$ 의 값으로, 그렇지 않으면, $Pr[X > x]$을 계산함.

카이제곱분포

예제 2.10. 카이제곱분포의 모양

```
draw.chisq <- function() { # draw.chisq.r
  from <-0     # x축 하한
  to <- 20     # x축 상한
  x <- seq(from, to, length=100)  # x축의 값
  plot(x, dchisq(x, 3), type="l", col="red",  main=expression(chi[nu]^2),
     ylab=expression(f(x)) )   # 자유도 3인 카이제곱
  curve(dchisq(x, 5), from=from, to=to, lty=2, add=T, col="blue")
       # 자유도 5인 카이제곱 분포
  curve(dchisq(x, 10), from=from, to=to, lty=3, add=T, col="magenta")
      # 자유도 10인 카이제곱 분포
  abline(h=0)                    # x축 그리기
```

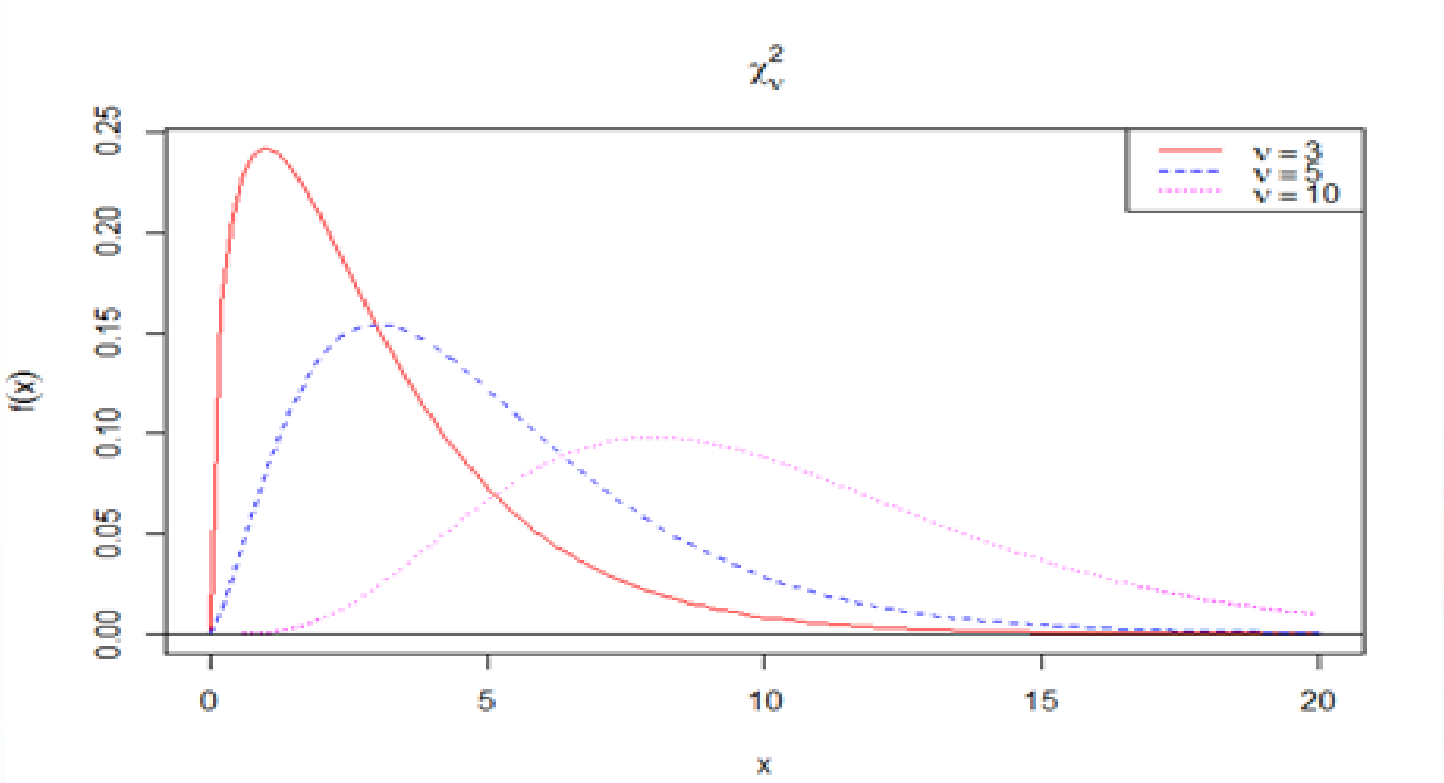
카이제곱분포

예제 2.10. 카이제곱분포의 모양

```
legend("topright", lty=1:3, col=c("red", "blue", "magenta"),
      legend=c(expression(nu == 3),expression(nu == 5),
      expression(nu == 10)) )   # 범례만들기
} # end function

>draw.chisq()
```

카이제곱분포



다음시간 안내

통계계산(2)