# Lesson4

*Sung Won Kang*

*2016년 11월 19일*

## A. 모수와 통계량

1. 모수: 모집단의 특성을 나타내는 값
   - 모집단의 분포를 정의
2. 표본: 모집단의 일부를 추출한 값
   - 표본 추출: 임의 추출(random sample), 복원추출을 가정
   - 임의 추출: 개별 sample의 독립성(independency)을 확보하는 과정
   - 비복원추출은 개별 sample 의 확률분포가 달라지는 문제가 발생(교과서 p. 156.(4.1))

$$X_1(\text{처음 추출}), X_2(\text{다음추출})$$

$$P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2|X_1 = x_1) \neq P(X_1 = x_1)P(X_2 = x_2)$$

$$\because P(X_2 = x_1|X_1 = x_1) = 0 \qquad P(X_1 = x_1)P(X_2 = x_1) = P(X_1 = x_1)^2$$

- i.i.d sample: 동일 모집단에서 추출된 독립적 표본. 일반적인 센서스, 설문조사 sample의 가정

3. 통계량(Statistic) : 표본의 특성
   - 통계량: 공식 (예: 표본평균, 표본분산, 표본표준편차)
   - 통계치: 공식에 표본을 대입하여 도출된 값
4. 표본분포 : 표본 통계량의 확률분포
   - 표본평균의 확률분포? 모집단 평균, 표준편차 = $\mu, \sigma$ 인 경우

$$\text{평균} \qquad E(\bar{X}) = E((1/N)\sum_i X_i) = (1/N)\sum_i E(X_i) = (1/N)\cdot N \cdot \mu = \mu$$

$$\text{분산} \qquad V(\bar{X}) = E[((1/N)\sum_i X_i - \mu)^2] = E[(1/N)^2(\sum_i X_i - N\mu)^2]$$

$$= (1/N^2)E[(\sum(X_i - \mu))^2] = (1/N^2)\cdot N\sigma^2 = \sigma^2/N$$

$$\text{표준편차} \qquad \sqrt{V(\bar{X})} = \sigma/\sqrt{N}$$

$$X_i \sim (\mu, \sigma) \to \bar{X} \sim (\mu, \sigma/\sqrt{N})$$

5. 대표본이론

- 대수의 법칙 (law of large numbers) : 표본의 수가 크면 표본집단의 모집단의 평균에 수렴한다.

```r
set.seed(9)
t=10
p=0.1
x=0:10
n=1000
b.2.mean=rep(NA,n)
b.4.mean=rep(NA,n)
b.32.mean=rep(NA,n)

for (i in 1:n) {
  b.2.mean[i] =mean(rbinom(2,size=t,prob=p))
  b.4.mean[i] =mean(rbinom(4,size=t,prob=p))
  b.32.mean[i] =mean(rbinom(32,size=t,prob=p))
}
options(digits=4)
c(mean(b.2.mean),sd(b.2.mean))
```

```
## [1] 1.0090 0.6763
```

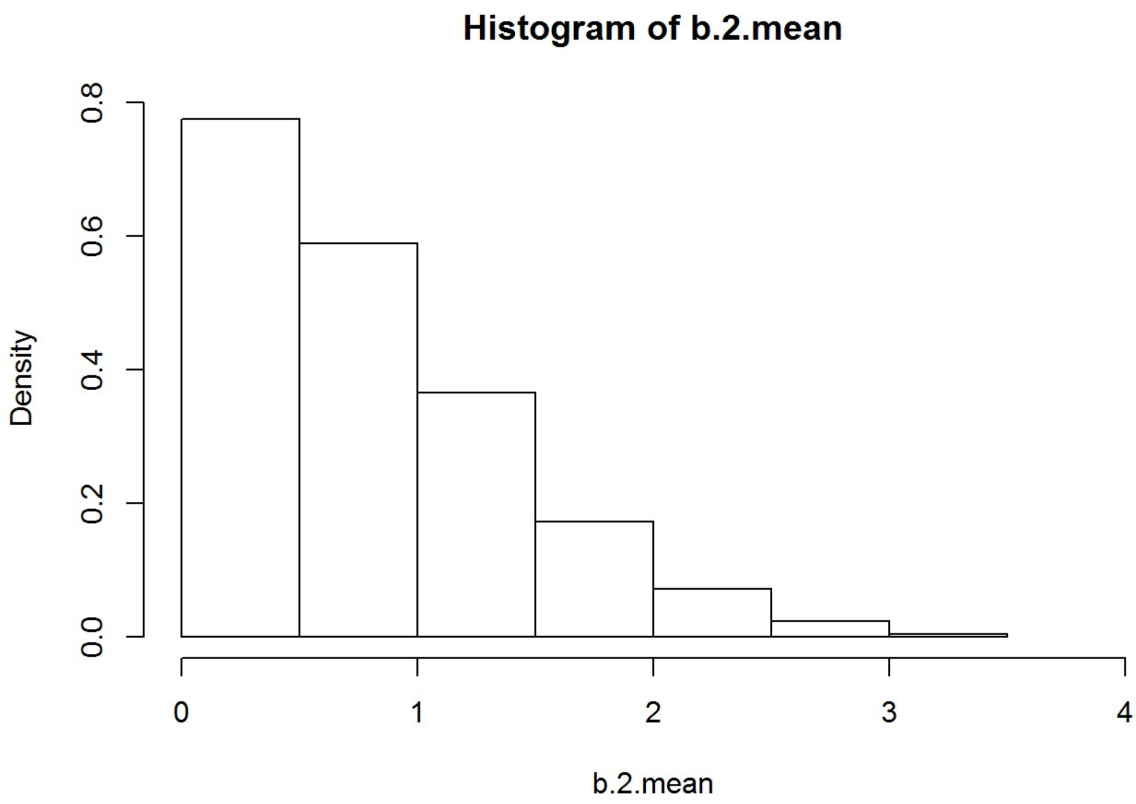
```r
c(mean(b.4.mean),sd(b.4.mean))
```

```
## [1] 1.006 0.481
```

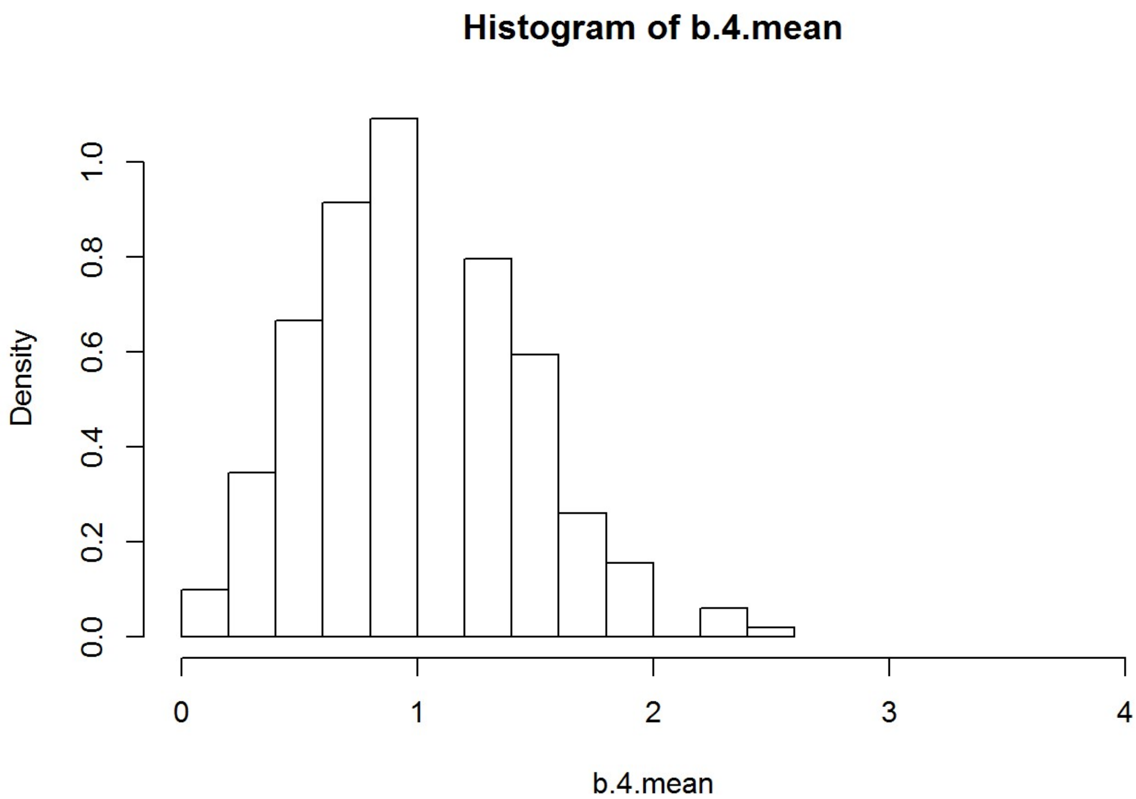
```r
c(mean(b.32.mean),sd(b.32.mean))
```

```
## [1] 0.9989 0.1624
```

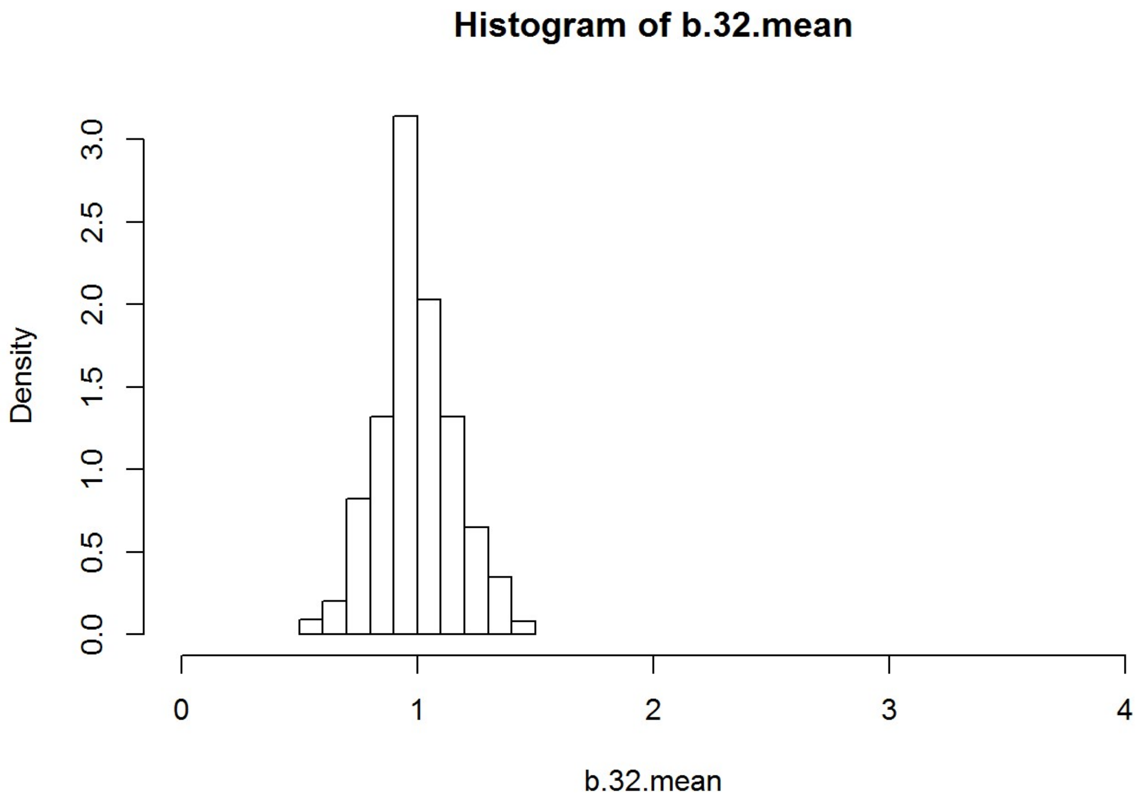
```r
#LLN
hist(b.2.mean, prob=T,xlim=(c(0,4)))
```




```
hist(b.4.mean, prob=T,xlim=(c(0,4)))
```



```
hist(b.32.mean, prob=T,xlim=(c(0,4)))
```



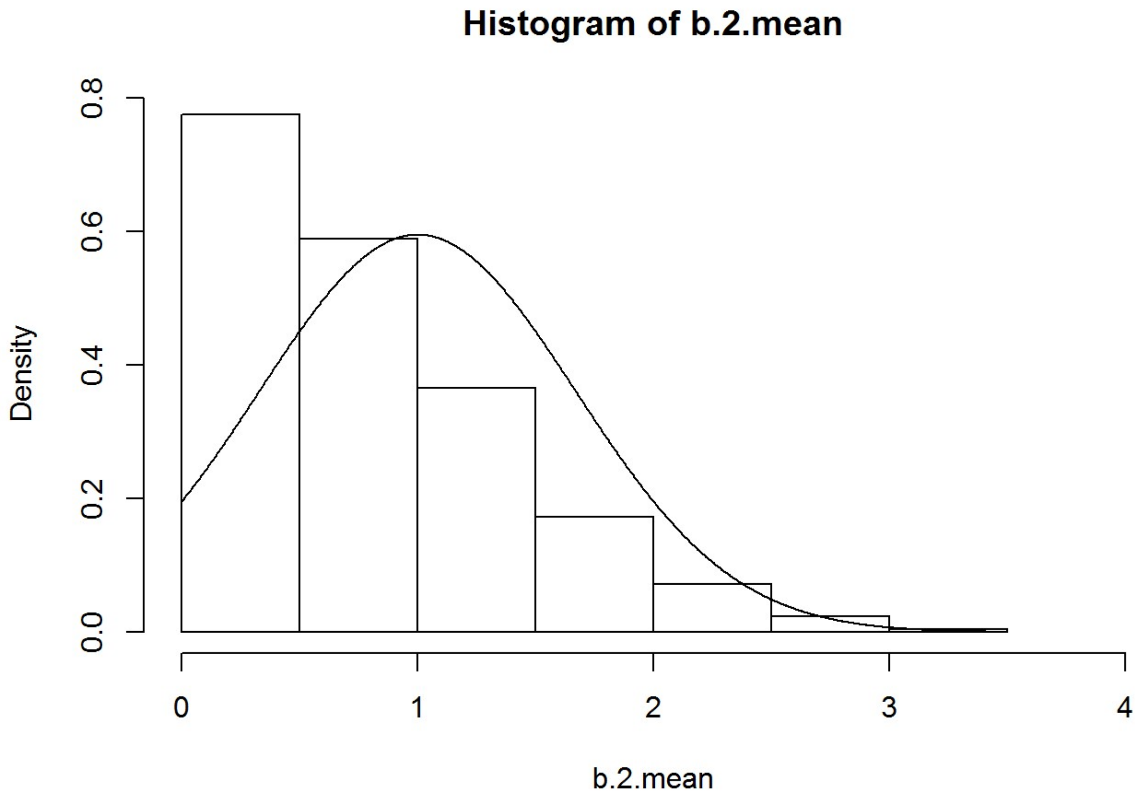

- 중심극한정리(Central Limit Theorem) : i.i.d 표본 표본평균의 확률분포는 표본의 수가 많으면 정규분포로 수렴한다. (모집단과 관계 없음)

$$X_i \sim N(\mu, \sigma) \rightarrow \bar{X} \sim N(\mu, \sigma/\sqrt{N}) \quad \text{정규분포의 특성}$$

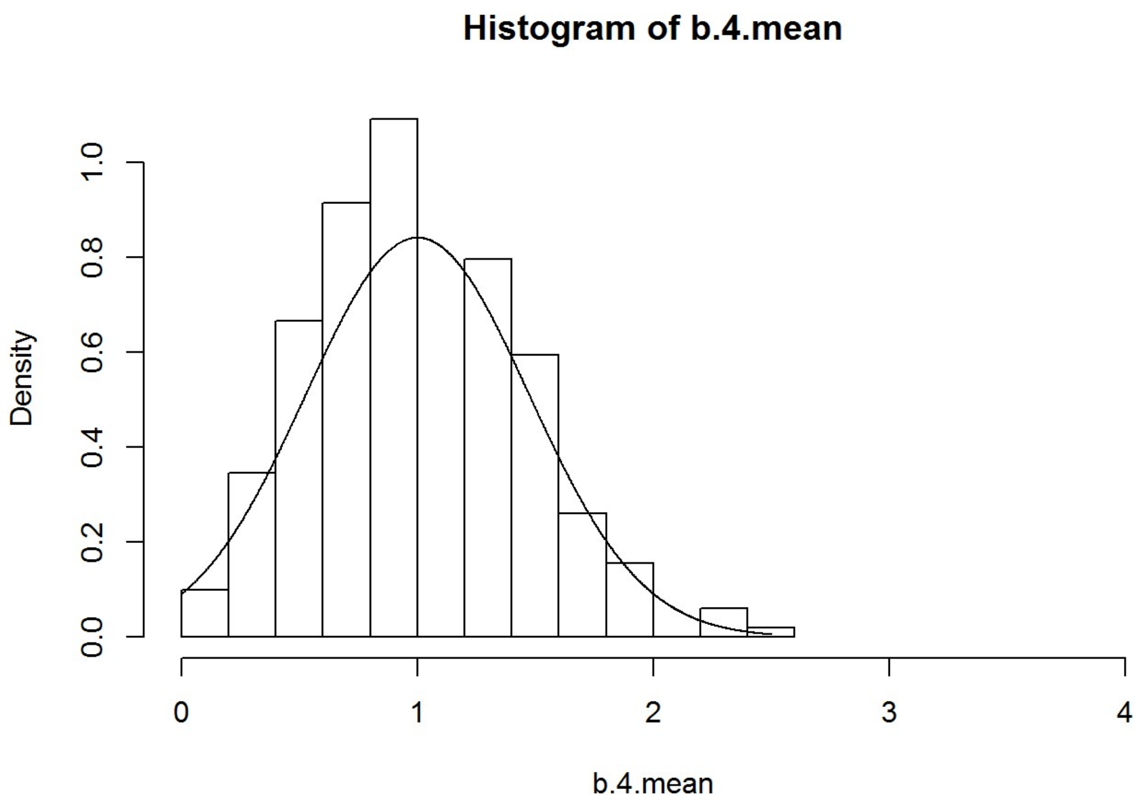
$$X_i \sim (\mu, \sigma) \rightarrow \bar{X} \sim N(\mu, \sigma/\sqrt{N}) \quad \text{모든분포}$$

$$X_i \sim (\mu, \sigma) \rightarrow \sqrt{N}(\bar{X} - \mu)/\sigma \sim N(0, 1) \quad \text{모든분포}$$

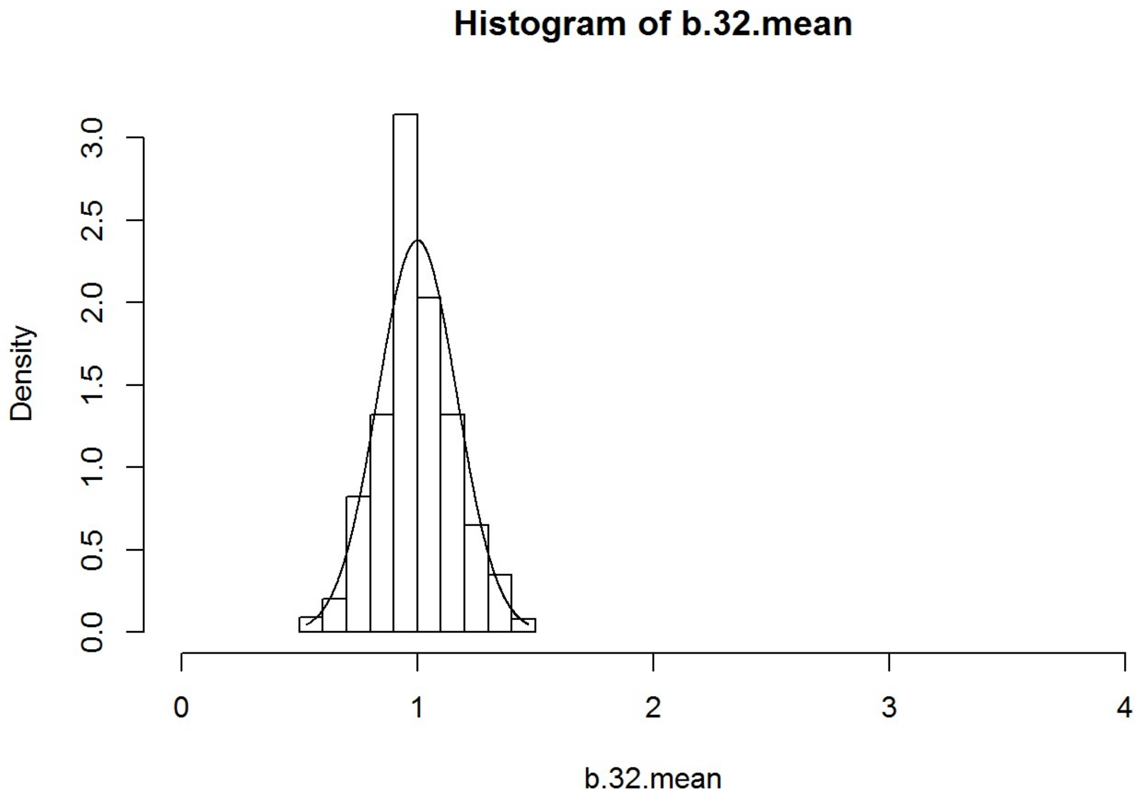
```
#CLT
hist(b.2.mean, prob=T,xlim=(c(0,4)))
x1=seq(min(b.2.mean),max(b.2.mean),length=1000)
y1=dnorm(x=x1,mean=1,sd=sqrt(t*p*(1-p))/sqrt(2))
lines(x1,y1)
```




```
hist(b.4.mean, prob=T,xlim=(c(0,4)))
x2=seq(min(b.4.mean),max(b.4.mean),length=1000)
y2=dnorm(x=x2,mean=1,sd=sqrt(t*p*(1-p))/sqrt(4))
lines(x2,y2)
```




```
hist(b.32.mean, prob=T,xlim=(c(0,4)))
x3=seq(min(b.32.mean),max(b.32.mean),length=1000)
y3=dnorm(x=x3,mean=1,sd=sqrt(t*p*(1-p))/sqrt(32))
lines(x3,y3)
```




## B. 다양한 (표본)분포

1.$\chi^2$ 분포

$$X = \sum_i Z_i^2 \qquad Z_i \sim N(0,1) \qquad i.i.d$$

모수: 자유도 $k$

$$Z_i{}^2 \sim \chi^2(1) \qquad X \sim \chi^2(k)$$

$$E(X) = k \qquad V(X) = 2k$$

표본분산 $S^2 = \sum_i (X_i - \bar{X})^2/(n-1) \qquad X_i \sim N(\mu, \sigma)$

$$(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$$

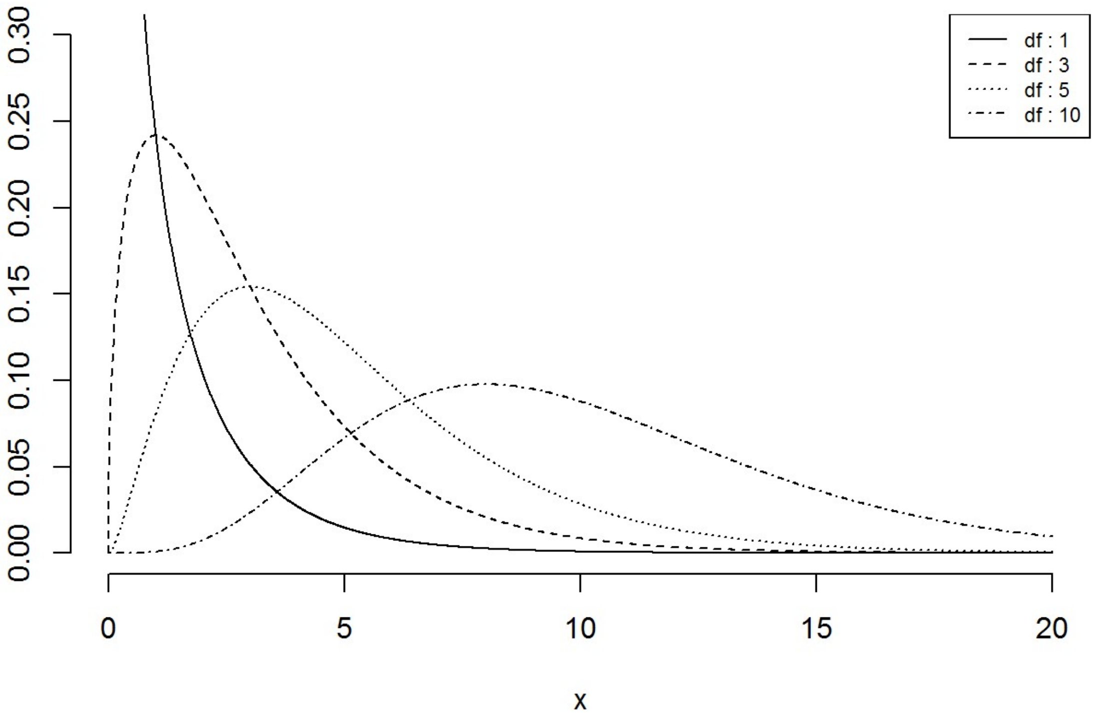
$$E(S^2) = \frac{\sigma^2}{n-1} E((n-1)S^2/\sigma^2) = \frac{\sigma^2}{n-1} \times (n-1) = \sigma^2$$

$$V(S^2) = V[\frac{\sigma^2}{n-1}(n-1)S^2/\sigma^2] = \frac{(\sigma^2)^2}{(n-1)^2} \times V((n-1)S^2/\sigma^2)$$
$$= \frac{(\sigma^2)^2}{(n-1)^2} \times 2(n-1) = \frac{2\sigma^4}{n-1}$$

교과서 p.170 그림 4-10.

```
df <- c(1, 3, 5, 10)
x <- seq(0, 20, by=0.01)
chi2.1 <- dchisq(x, df[1])
chi2.3 <- dchisq(x, df[2])
chi2.5 <- dchisq(x, df[3])
chi2.10 <- dchisq(x, df[4])
plot(x, type="n", xlim=c(0, 20), ylim=c(0, 0.3), main="", xlab="x", ylab
="", axes=F)

axis(1)
axis(2)
lines(x, chi2.1, lty=1)
lines(x, chi2.3, lty=2)
lines(x, chi2.5, lty=3)
lines(x, chi2.10, lty=4)
legend("topright", paste("df :", df), lty=1:4, cex=0.7)
```




2. t 분포

- $Z \sim N(0,1), V \sim \chi^2(k)$ Z, V 상호 독립

$$T = \frac{Z}{\sqrt{V/k}} \sim t(k)$$

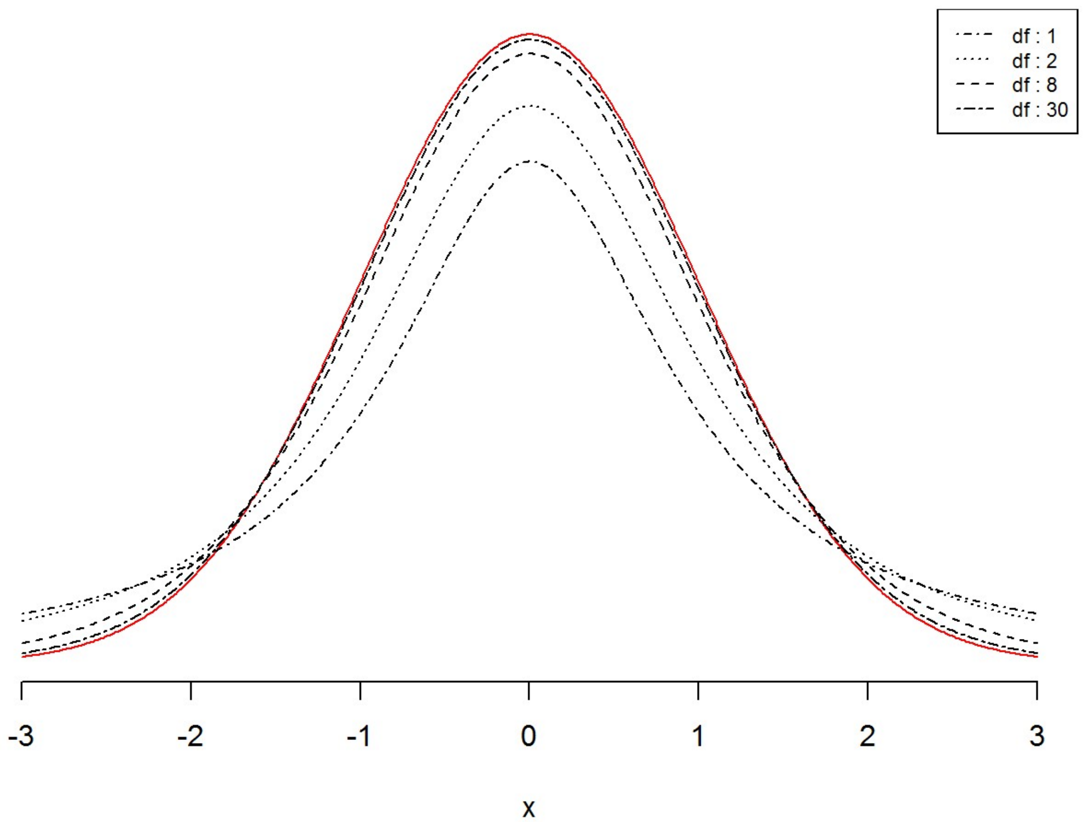
- $\bar{X} \sim N(\mu, \sigma^2/N), (N-1)S^2/\sigma^2 \sim \chi^2(N-1)$ , $\bar{X}, (N-1)S^2/\sigma^2$ 상호독립

$$T = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{N})}{\sqrt{(N-1)S^2/\sigma^2(N-1)}} = \frac{\bar{X}-\mu}{S/\sqrt{N}} \sim t(N-1)$$

교과서 p.172 그림 4-11

```
df <- c(1, 2, 8, 30)
x <- seq(-3, 3, by=0.01)
y <- dnorm(x)
t.1 <- dt(x, df=df[1])
t.2 <- dt(x, df=df[2])
t.8 <- dt(x, df=df[3])
t.30 <- dt(x, df=df[4])

par(mar=c(4,2,2,2))
plot(x, y, type="l", lty=1, axes=F, xlab="x", ylab="", col="red")
axis(1)
lines(x, t.1, lty=4)
lines(x, t.2, lty=3)
lines(x, t.8, lty=2)
lines(x, t.30, lty=6)
legend("topright", paste("df :", df), lty=c(4, 3, 2, 6), cex=0.7)
```




- 평균과 분산
  - 자유도가 1인 t 분포는 Caushy 분포라는 다른 이름이 있는데, 이 분포의 평균과 분산은 정할 수 없다.(undetermiend)
  - 자유도가 2 이상인 t 분포의 평균과 분산은 다음과 같다.

$$T = \frac{Z}{\sqrt{V/k}} \sim t(k) \qquad (k \geq 2)$$
$$E(T) = 0$$
$$V(T) = \frac{k}{k-2}$$

3. F 분포.

- $V_1 \sim \chi^2(k_1), V_2 \sim \chi^2(k_2)$ V1 V2 상호 독립

$$F = \frac{V_1/k_1}{V_2/k_2} \sim F(k_1, k_2)$$

- 평균과 분산

$$F \sim F(m, n)$$
$$E(F) = \frac{m}{m-2} \quad (m \geq 3)$$
$$V(F) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)} \quad (m > 5)$$

- 분산의 비율과 F 분포

$$X_i \sim N(0, \sigma_1^2), \quad i = 1...n$$
$$Y_j \sim N(0, \sigma_2^2) \quad j = 1...m$$
$$X_i, Y_j \quad \text{independent}$$

$$S_1^2 = \frac{\sum_i^m (X_i - \bar{X})^2}{n-1}, \quad (n-1)S_1^2/\sigma_1^2 \sim \chi^2(n-1)$$
$$S_2^2 = \frac{\sum_j^m (Y_j - \bar{Y})^2}{m-1} \quad (m-1)S_2^2/\sigma_2^2 \sim \chi^2(m-1)$$
$$F = \frac{(n-1)S_1^2/\sigma_1^2(n-1)}{(m-1)S_2^2/\sigma_2^2(m-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} \sim F(n-1, m-1)$$

그래서?

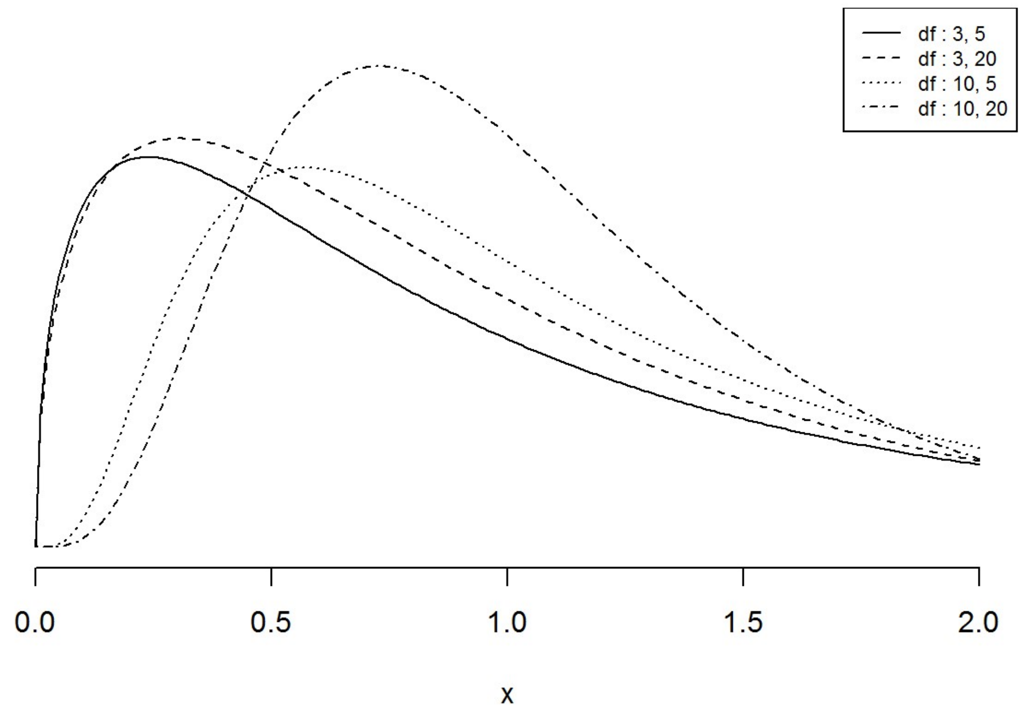
$$E(\frac{S_1^2}{S_2^2} \times \frac{\sigma_2^2}{\sigma_1^2}) = \frac{n-1}{n-3}$$
$$\frac{\sigma_2^2}{\sigma_1^2} \sim \frac{S_2^2}{S_1^2} \frac{n-1}{n-3}$$

```
df1 <- c(3, 10)
df2 <- c(5, 20)
x <- seq(0, 2, by=0.01)

f3.5 <- df(x, df1[1], df2[1])
f3.20 <- df(x, df1[1], df2[2])
f10.5 <- df(x, df1[2], df2[1])
f10.20 <- df(x, df1[2], df2[2])

plot(x, f3.5, type="l", ylim=c(0, 0.9), axes=F, xlab="x", ylab="")
axis(1)
lines(x, f3.20, lty=2)
lines(x, f10.5, lty=3)
lines(x, f10.20, lty=4)

legend("topright", paste("df :", c("3, 5", "3, 20", "10, 5", "10, 20")), lty
=1:4, cex=0.7)
```




- F test.

$$X_i \sim N(\mu_1, \sigma^2), \quad i = 1 ... n$$
$$Y_j \sim N(0, \sigma^2) \quad j = 1 ... m$$
$$X_i, Y_j \quad \text{independent}$$

(평균이 0 이 아닌 경우)

$$S_1^2 = \frac{\sum_i^m (X_i - \bar{X})^2}{n-1}, \quad (n-1)S_1^2/\sigma^2 \sim \chi^2(n-1, \mu^2) \qquad \text{noncentral chi}$$

$$S_2^2 = \frac{\sum_j^m (Y_j - \bar{Y})^2}{m-1}, \quad (m-1)S_2^2/\sigma^2 \sim \chi^2(m-1)$$

non central F 분포

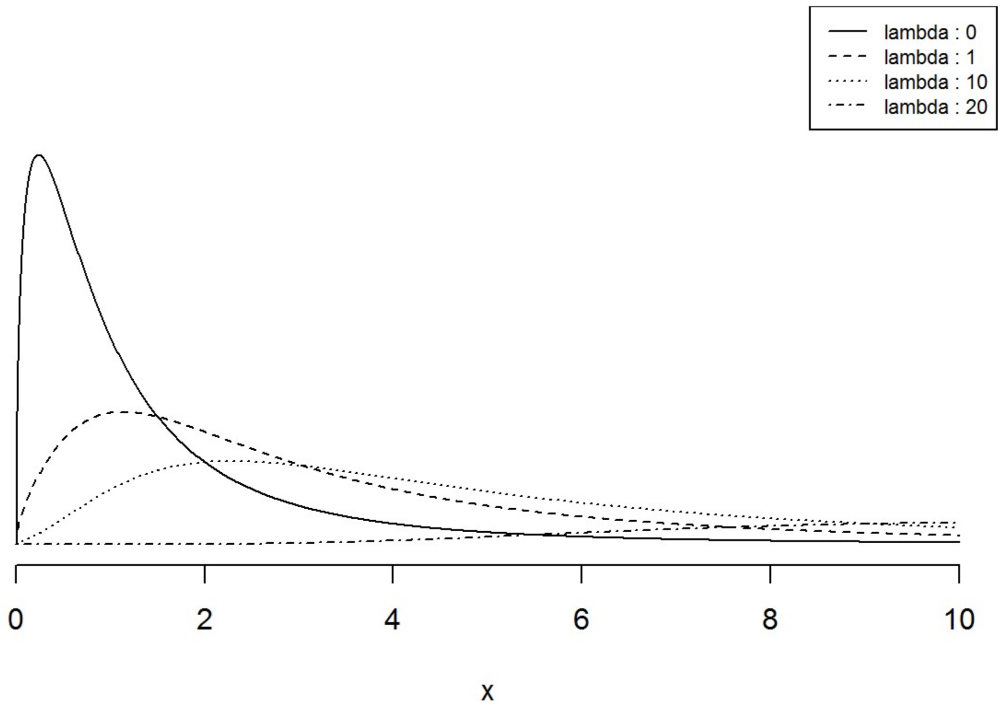
$$F = \frac{(n-1)S_1^2/\sigma^2(n-1)}{(m-1)S_2^2/\sigma^2(m-1)} = \frac{S_1^2/\sigma^2}{S_2^2/\sigma^2} = \frac{S_1^2}{S_2^2} \sim F(n-1, m-1, \lambda)$$

```
df1 =3
df2 =5
x <- seq(0, 10, by=0.01)

f3.5 <- df(x, df1, df2)
f3.20 <- df(x, df1, df2,5)
f10.5 <- df(x, df1, df2,10)
f10.20 <- df(x, df1, df2,50)

plot(x, f3.5, type="l", ylim=c(0, 0.9), axes=F, xlab="x", ylab="")
axis(1)
lines(x, f3.20, lty=2)
lines(x, f10.5, lty=3)
lines(x, f10.20, lty=4)

legend("topright", paste("lambda :", c("0", "1", "10", "20")), lty=1:4, cex=
0.7)
```




- F test 회귀분석

$$\text{Model 1:} \quad y = [X, 1]\beta + \epsilon, \quad X_[n \times (k-1)], \quad \epsilon \sim N(0, \sigma^2)$$

$$\hat{\beta} = (X'X)^{-1}X'y, \quad E(\hat{\beta}) = \beta$$
$$\hat{y} = X\hat{\beta}$$

$$\text{Model 2:} \quad y = [1]\beta_1 + \epsilon_1, \quad \epsilon \sim N(0, \sigma^2)$$

$$\hat{\beta}_1 = (1'1)^{-1}1'y = \bar{y} \quad E(\hat{\beta}_1) = \beta_1$$
$$\hat{y}_1 = [1]\hat{\beta}_1 = [1]\bar{y}$$

회귀분석 F test 검정통계량

$$F = \frac{\sum_i (\hat{y}_i - \hat{y}_1)/(k-1)}{\sum_i (y_i - \hat{y}_1)/(n-1)}$$

만약 $\{\beta_2, \beta_3, \ldots \beta_k\} = \{0, 0, \ldots, 0\}, i.e. \beta = \beta_1$

$$\hat{\beta} \sim \hat{\beta}_1 \rightarrow \hat{y} \sim \hat{y}_1$$
$$E(\hat{y} - \hat{y}_1) = E(\hat{y} - [1]\bar{y}) = XE(\hat{\beta}) - XE(\hat{\beta}_1) = X(\beta - \beta_1) = 0$$
$$F \sim F(k-1, n-1, 0)$$

만약 $\{\beta_2, \beta_3, \ldots \beta_k\} \neq \{0, 0, \ldots, 0\} i.e. \beta \neq \beta_1$

$$\hat{\beta} <> \hat{\beta}_1 \rightarrow \hat{y} <> \hat{y}_1$$

$$E(\hat{y} - \hat{y}_1) = E(\hat{y} - [1]\bar{y}) \neq 0$$

$$F \sim F(k-1, n-1, \lambda), \quad \lambda \neq 0 \quad \text{non central F}$$

그래서 central F 분포일 경우보다는 큰 값이 나올 확률이 높아진다. 실제로 높은 값의 F 통계치가 나오면 $\{\beta_2, \beta_3, \ldots \beta_k\} = \{0, 0, \ldots, 0\}$ 일 가정을 폐기한다.

## C. 다음단원 준비(R function)

```
options(digits=4)
var.p2 <- function(x, na.rm=FALSE) {
  if(na.rm == TRUE){
    x <- x[!is.na(x)]
  }
  n <- length(x)
  m <- mean(x, na.rm=na.rm)
  num <- sum( (x - m)^2, na.rm=na.rm )
  denom <- n
  var <- num / denom
  return( var )
}

radius <- c(234, 234, 234, 233, 233, 233, NA, 231, 232, 231)
weight=c(146.3,146.4,144.1,146.7,145.2,144.1,143.3, 147.3,146.7,147.3)

var.p2(radius)
```

```
## [1] NA
```

```
var.p2(radius, na.rm=TRUE)
```

```
## [1] 1.284
```

```
var.p2(weight)
```

```
## [1] 1.908
```