I. 표본평균의 표본분포

II. 중심극한정리

III. 표본분산의 표본분포

Ⅰ. 표본평균의 표본분포

- X~ N(μ, σ^2)이면, $\bar{X}$~N(μ, $\frac{\sigma^2}{n}$)

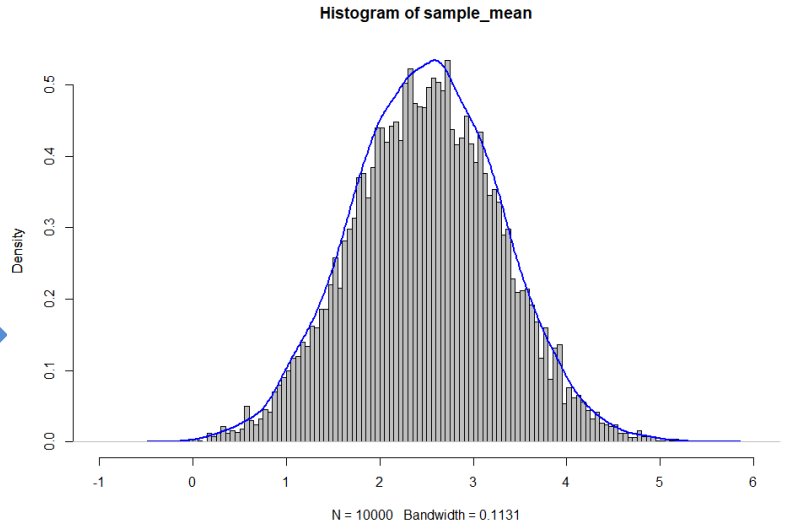
  단, n은 표본의 크기(sample size)

- (예 1) 평균이 2.5이고, 분산이 1.25인 모집단에서 표본크기가 2인 표본을 추출할 경우 표본평균의 평균은 모평균과 동일한 2.5가 되고, 분산은 1.25/2인 분포를 근사적으로 따르는데 10,000개 표본평균의 평균과 분산은 각각 2.494004 및 0.6284706로서 모집단의 이론적인 평균과 분산에 근접함을 알 수 있음

b1-ch5-1-new.R

```
# set sample size and number of samples
set.seed(1)
n <- 2
reps <- 10000
# perform random sampling
samples <- replicate(reps, rnorm(n,2.5,1.118))
# compute sample means
sample_mean <- rep(NA, reps)
for (i in 1:reps) {
  sample_mean[i] <- mean(samples[,i]) }
# check that 'sample_mean' is a vector
is.vector(sample_mean)
(mean(sample_mean)) ; (var(sample_mean)) ;(sd(sample_mean))
hist(sample_mean,  freq=F, col="grey", xlab="", xlim=c(-1, 6), breaks=10
0)
par(new=T)
plot(density(sample_mean), axes=F, main="", xlim=c(-1, 6), lwd=2, col="
blue")
```

```
> (mean(sample_mean))
[1] 2.494004
> (var(sample_mean))
[1] 0.6284706
```

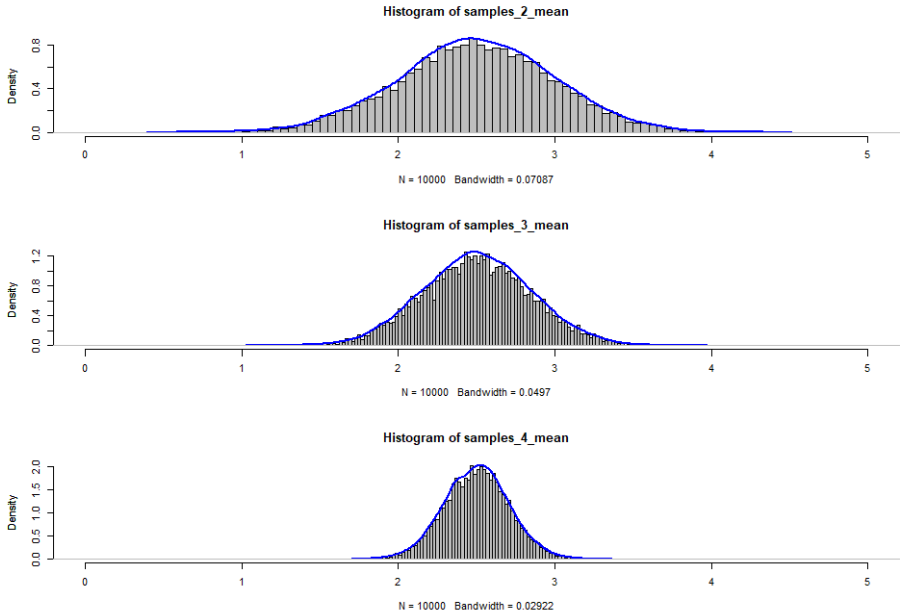


- **(예 2) 표본평균의 평균은 표본크기에 관계없이 모평균과 동일하고, 표본평균의 분산은 모분산을 표본크기로 나눈 값과 같으므로 표본크기가 커짐에 따라 표본평균의 분산은 작아짐을 보일 수 있음**

b1-ch5-2-new.R

```
set.seed(1)
n1 <- 2 ; n2 <- 5 ; n3 <- 10; n4 <- 30 ; reps <- 10000
# perform random sampling
samples_1 <- replicate(reps, rnorm(n1,2.5,1.118))
samples_1_mean <- rep(NA, reps)
for (i in 1:reps) {
  samples_1_mean[i] <- mean(samples_1[,i])}
mean(samples_1_mean) ;var(samples_1_mean)
(중략)
samples_4 <- replicate(reps, rnorm(n4,2.5,1.118)) # 30 x 1000
0 sample matrix
samples_4_mean <- rep(NA, reps)
for (i in 1:reps) {
  samples_4_mean[i] <- mean(samples_4[,i]) }
mean(samples_4_mean) ; var(samples_4_mean)
par(mfrow=c(3,1))
#hist(samples_1_mean,   freq=F, col="grey", xlab="", xlim=c(0,
5), breaks=100)
#par(new=T)
#plot(density(samples_1_mean), axes=F, main="", xlim=c(0, 5),
lwd=2, col="blue")
(이하 생략)
```

```
> mean(samples_2_mean)
[1] 2.49543

> var(samples_2_mean)
[1] 0.2497486

> mean(samples_3_mean)
[1] 2.502499

> var(samples_3_mean)
[1] 0.122446

> mean(samples_4_mean)
[1] 2.499477

> var(samples_4_mean)
[1] 0.04197086
```



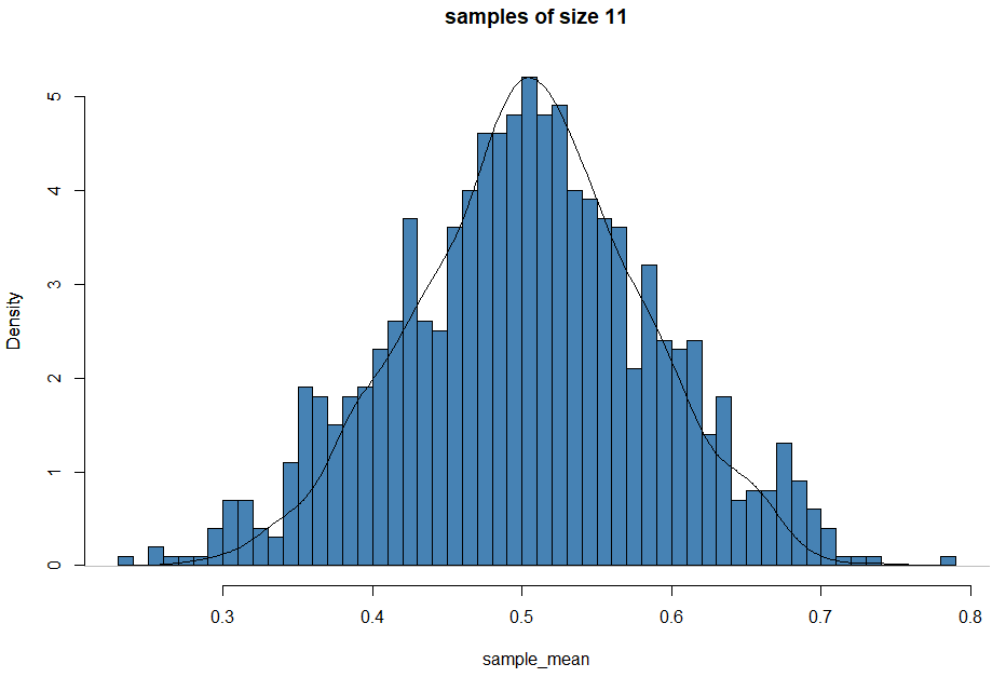
II. 중심극한정리

- X~ (μ, σ^2)이면, n이 커짐에 따라 $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, n은 표본의 크기(sample size)
- (실험 1) X~U(0,1)에서 X의 평균은 1/2, 분산은 1/12이므로 n=11인 경우 표본평균 $\bar{X}$는 근사적으로 N(1/2, 1/132)을 따르는데 1,000개 표본평균의 평균과 분산은 각각 0.5019 및 0.0078로서 모집단의 이론적인 평균과 표준편차에 근접함을 알 수 있음

b1-ch5-3-new.R

```
# set sample size and number of samples
set.seed(23456789)
n <- 11
reps <- 1000
# perform random sampling
samples <- replicate(reps, runif(n))
# compute sample means
sample_mean <- colMeans(samples)
# check that 'sample_mean' is a vector
is.vector(sample_mean)

(mean(sample_mean))
(var(sample_mean))
(sd(sample_mean))

hist(sample_mean, breaks=40, prob=T, main=paste(
"samples of size 11" ),col="steelblue")
par(new=T)
plot(density(sample_mean), xlab="", axes=F, main=""
, col="black")
```

```
> (mean(sample_mean))
[1] 0.5019778

> (var(sample_mean))
[1] 0.007837079
```

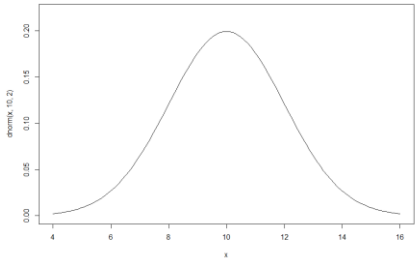


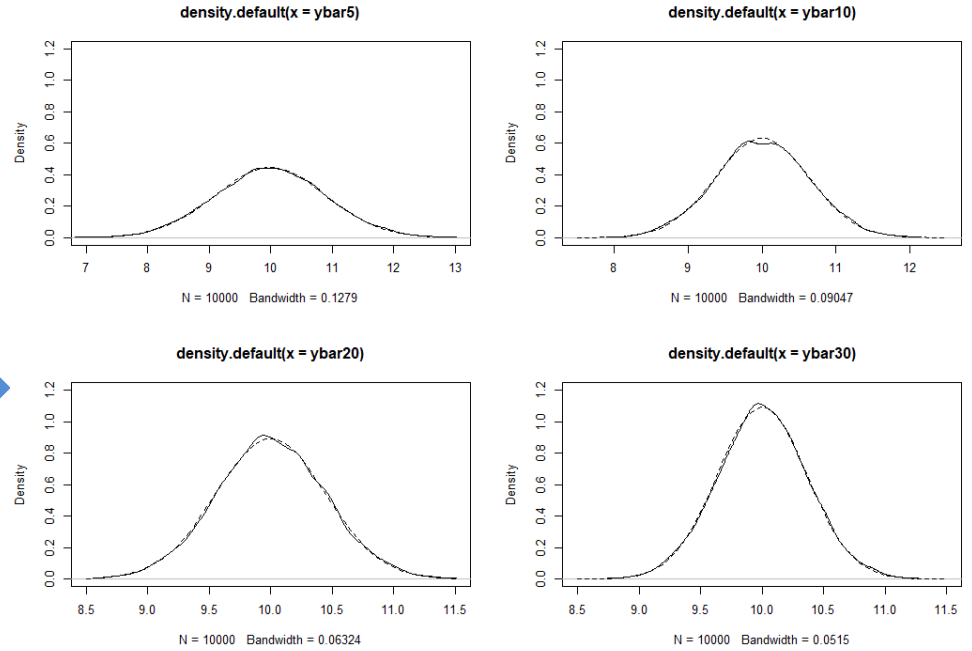
- **(실험 2) 확률변수 X가 평균이 10, 표준편차가 2인 정규분포를 따른다고 할 때 이러한 확률분포로부터 표본크기가 각각 5, 10, 20, 30인 확률표본을 10,000개 추출하는 실험을 통해 표본크기에 따라 10,000개 표본평균의 평균과 분산은 모집단의 이론적인 평균과 분산에 근접함을 알 수 있음**

b1-ch5-4-new.R

```
set.seed(123456)
curve(dnorm(x,10,2), xlim = c(4, 16), ylim = c(0.0, 0.22))
par(mfrow=c(2,2))
n1 <- 5 ;n2 <- 10;n3 <- 20;n4 <- 30 ;reps <- 10000
ybar5<-numeric(10000)
samples_1 <- replicate(reps, rnorm(n1,10,2))
ybar5 <- rep(NA, reps)
for (i in 1:reps) {
  ybar5[i] <- mean(samples_1[,i]) }
mean(ybar5);var(ybar5)
plot(density(ybar5), xlim = c(7, 13), ylim = c(0.0, 1.2))
curve(dnorm(x,10,sqrt(0.8)), add=T, lty=2)
(중략)
par(mfrow=c(1,1))
plot(density(ybar5), col="black", xlim = c(7, 13), ylim = c(0
.0, 1.2))
lines(density(ybar10), col="red", xlim = c(7, 13), ylim = c(0
.0, 1.2))
lines(density(ybar20), col="green", xlim = c(7, 13), ylim = c
(0.0, 1.2))
lines(density(ybar30), col="blue", xlim = c(7, 13), ylim = c(
0.0, 1.2))
```

구분	n=5	n=10	n=20	n=30
표본평균의 평균	10.00494	10.00259	9.998591	9.993856
표본평균의 분산	0.7818611	0.3992823	0.2039642	0.1347668



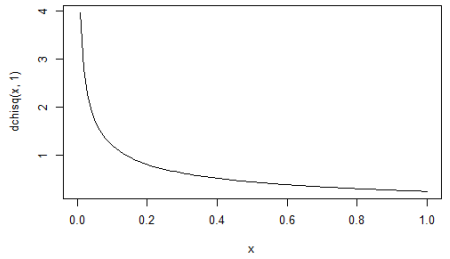


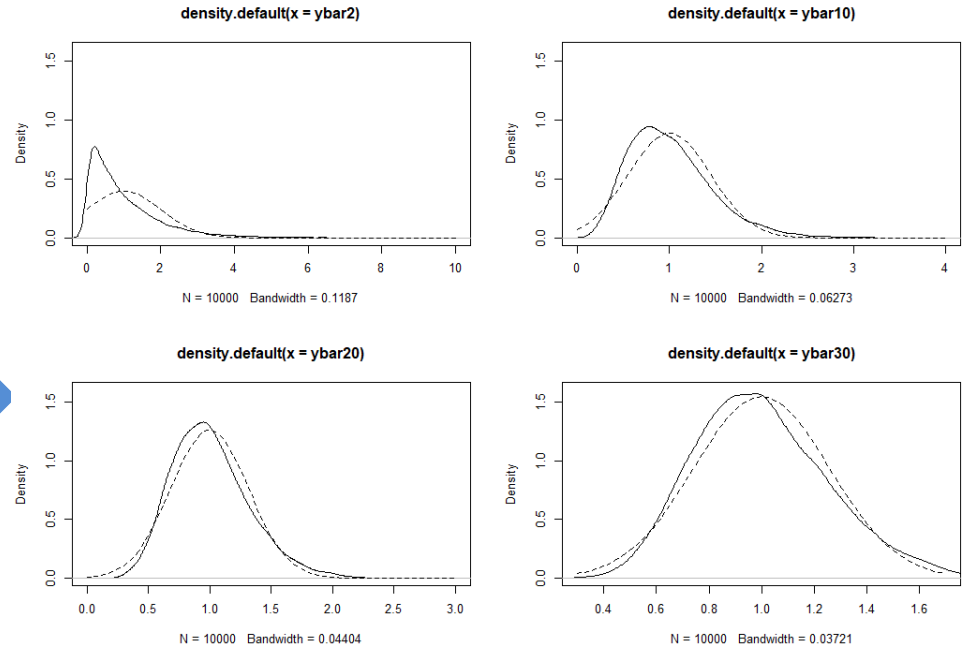
- **(실험 3) 확률변수 X가 자유도가 1인 χ^2-분포를 따른다고 할 때(따라서 평균은 1, 분산은 2됨) 이러한 확률분포로부터 표본크기가 각각 2, 10, 20, 30인 확률표본을 10,000개 추출하는 실험을 통해 표본크기에 따라 10,000개 표본평균의 평균과 분산은 모집단의 이론적인 평균과 분산에 근접함을 알 수 있음**

b1-ch5-5-new.R

```
set.seed(123456) ;curve(dchisq(x,1));par(mfrow=c(2,2))
n1 <- 2;n2 <- 10;n3 <- 20;n4 <- 30;reps <- 10000
ybar2<-numeric(10000)
samples_1 <- replicate(reps, rchisq(n1,1))
ybar2 <- rep(NA, reps)
for (i in 1:reps) {
  ybar2[i] <- mean(samples_1[,i])}
mean(ybar2);var(ybar2)
plot(density(ybar2), xlim = c(0, 10), ylim = c(0.0, 1.6))
curve(dnorm(x,1,sqrt(1.0)), add=T, lty=2)
(중략)
ybar30<-numeric(10000)
samples_4 <- replicate(reps, rchisq(n4,1)) # 30 x 10000 s
ample matrix
ybar30 <- rep(NA, reps)
for (i in 1:reps) {
  ybar30[i] <- mean(samples_4[,i])
}
mean(ybar30)
var(ybar30)
plot(density(ybar30),xlim = c(0.3, 1.7), ylim = c(0.0, 1.6))
curve(dnorm(x,1,sqrt(0.0667)), add=T, lty=2)
```

구분	n=2	n=10	n=20	n=30
표본평균의 평균	0.9873457	1.002666	0.9988942	0.9961374
표본평균의 분산	0.9841301	0.2016694	0.1021074	0.06725091



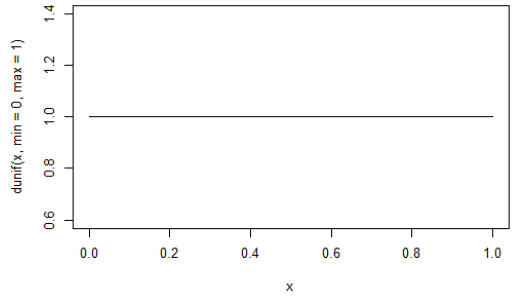


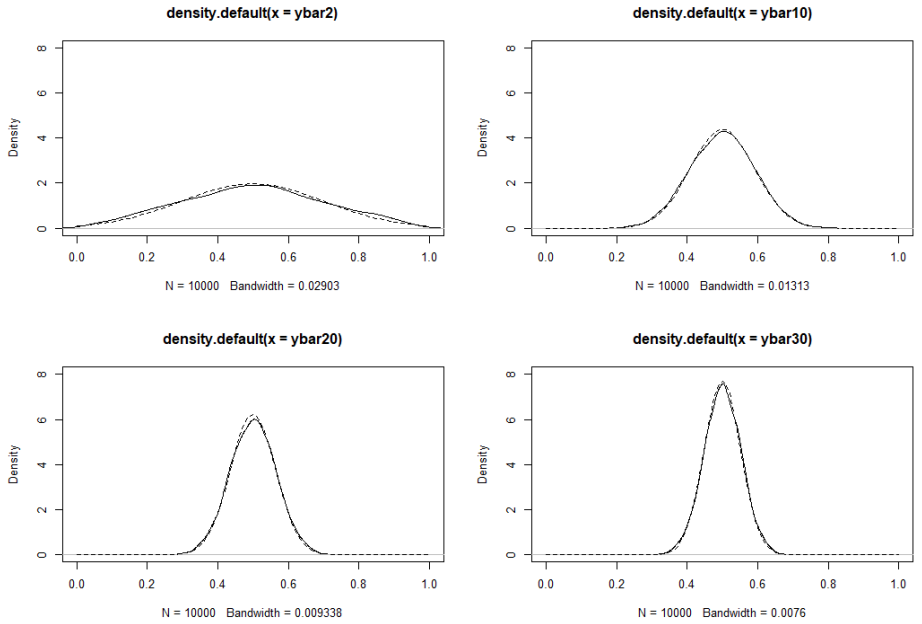
- **(실험 4) 확률변수 X가 균등분포 U(0,1)을 따른다고 할 때(평균은 0.5, 분산은 0.083333) 이러한 확률분포로부터 표본크기가 각각 2, 10, 20, 30인 확률표본을 10,000개 추출하는 실험을 통해 표본크기에 따라 10,000개 표본평균의 평균과 분산은 모집단의 이론적인 평균과 분산에 근접함을 알 수 있음**

b1-ch5-6-new.R

```
set.seed(123456);curve(dunif(x,min=0, max=1))
par(mfrow=c(2,2))
n1 <- 2;n2 <- 10;n3 <- 20;n4 <- 30;reps <- 10000
ybar2<-numeric(10000)
samples_1 <- replicate(reps, runif(n1,min=0, max=1))
ybar2 <- rep(NA, reps)
for (i in 1:reps) {
  ybar2[i] <- mean(samples_1[,i])}
mean(ybar2);var(ybar2)
plot(density(ybar2), xlim = c(0, 1), ylim = c(0,8))
curve(dnorm(x,0.5,sqrt(0.0417)), add=T, lty=2)
(중략)
ybar30<-numeric(10000)
samples_4 <- replicate(reps, runif(n4,min=0, max=1))
ybar30 <- rep(NA, reps)
for (i in 1:reps) {
  ybar30[i] <- mean(samples_4[,i])}
mean(ybar30);var(ybar30)
plot(density(ybar30),xlim = c(0, 1), ylim = c(0, 8))
curve(dnorm(x,0.5,sqrt(0.0028)), add=T, lty=2)
```

구분	n=2	n=10	n=20	n=30
표본평균의 평균	0.5013918	0.4987733	0.5003994	0.4996023
표본평균의 분산	0.04192942	0.008292283	0.004107226	0.002834952





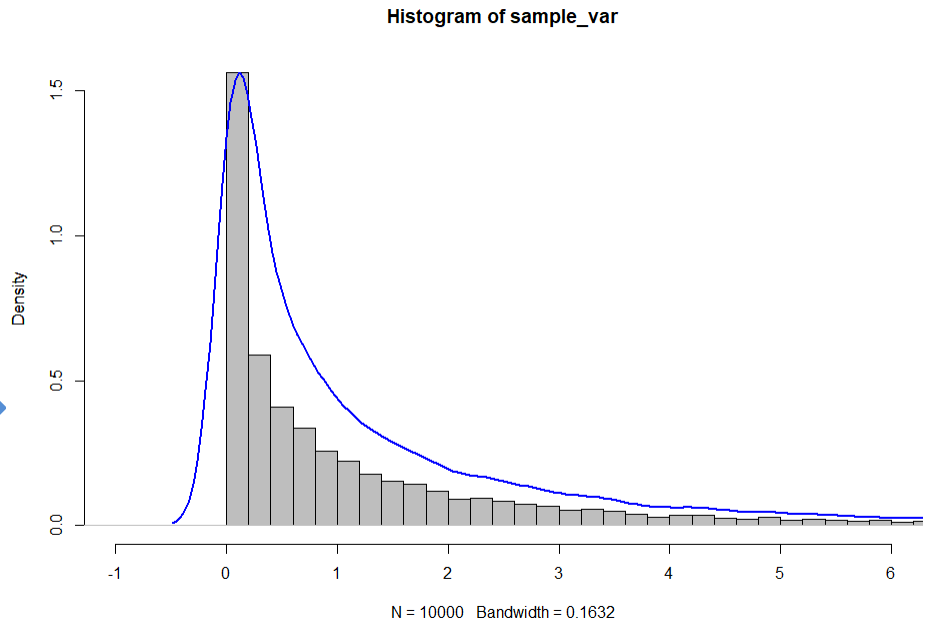
Ⅲ. 표본분산의 표본분포

- 모집단이 정규분포에 따르더라도 표본분산의 표본분포는 정규분포에 따르지 않고 표본분산의 평균은 σ^2, 분산은 $\frac{2\sigma^4}{n-1}$인 분포에 따름
- (예 1) 평균이 2.5이고, 분산이 1.25인 모집단에서 표본크기가 2인 표본을 추출할 경우 표본분산은 근사적으로 평균 1.25, 분산 3.125인 분포를 따르는데 10,000개 표본의 표본분산의 평균 1.232382, 분산 3.227298로서 모집단의 이론적인 평균과 분산에 근접함을 알 수 있음

b1-ch5-7-new.R

```
# set sample size and number of samples
set.seed(2345);n <- 2;reps <- 10000
# perform random sampling
samples <- replicate(reps, rnorm(n,2.5,1.118)) # 2 x 1000 sa
mple matrix
# compute sample variances
sample_var <- rep(NA, reps)
for (i in 1:reps) {
  sample_var[i] <- var(samples[,i])}
# check that 'sample_var' is a vector is.vector(sample_var)
mean(sample_var);var(sample_var)


hist(sample_var,  freq=F, col="grey", xlab="", xlim=c(-1, 6), b
reaks=100)
par(new=T)
plot(density(sample_var), axes=F, main="", xlim=c(-1, 6), lwd
=2, col="blue")
```

```
> mean(sample_var)
[1] 1.251813

> var(sample_var)
[1] 3.089847
```

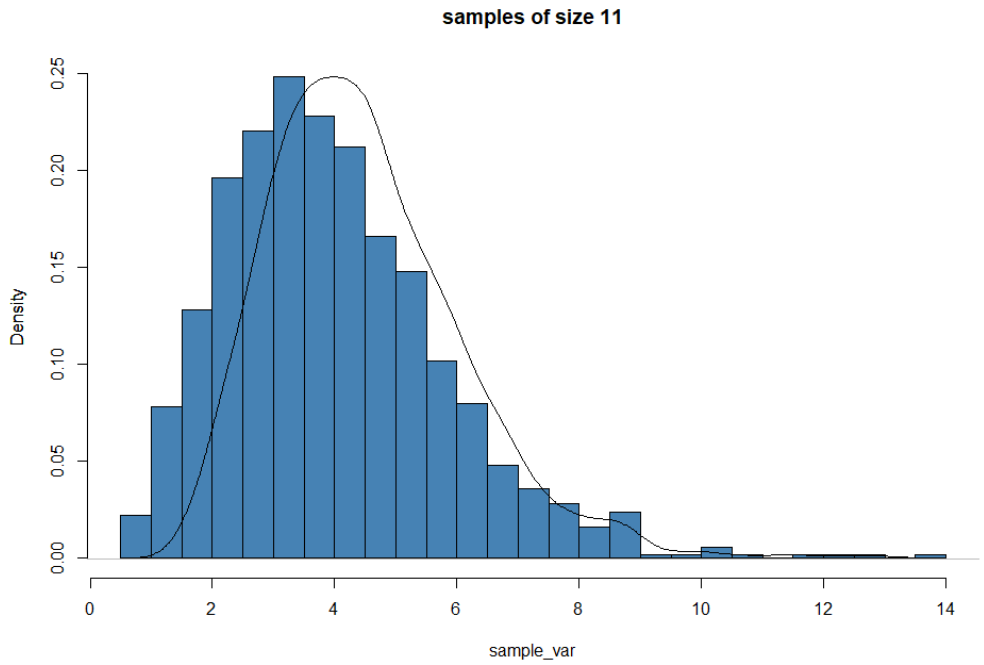


- **(예 2) 평균이 10이고, 표준편차가 2인 정규분포에 따르는 모집단에서 표본크기가 11인 1,000개 표본으로부터 구한 표본분산 s_1^2 , s_2^2 , … , s_{1000}^2은 정규분포에 따르지 않는다는 것을 확인할 수 있음**

b1-ch5-8-new.R

```
# set sample size and number of samples
set.seed(23456789);n <- 11;reps <- 1000
# perform random sampling
samples <- replicate(reps, rnorm(n,10,2)) # 11 x 10
00 sample matrix
# compute sample variances
sample_var <- rep(NA, reps)
for (i in 1:reps) {
  sample_var[i] <- var(samples[,i])
}
# check that 'sample_var' is a vector
is.vector(sample_var)


(mean(sample_var));(var(sample_var))


hist(sample_var, breaks=40, prob=T, main=paste( "s
amples of size 11" ),col="steelblue")
par(new=T)
plot(density(sample_var), xlab="", axes=F, main="",
col="black")
```

```
> (mean(sample_var))
[1] 3.986259

> (var(sample_var))
[1] 3.31678
```



- **(예 3) 표본크기를 다르게 할 경우 표본분산의 분포는 모집단의 이론적인 평균인 σ^2과 분산인 $\frac{2\sigma^4}{n-1}$에 근접함을 알 수 있음**

b1-ch5-9-new.R

```
set.seed(123456);par(mfrow=c(2,2))
n1 <- 10;n2 <- 20;n3 <- 30;n4 <- 100
reps <- 10000
vbar10<-numeric(10000)
samples_1 <- replicate(reps, rnorm(n1,10,2)) # 10 x
10000 sample matrix
vbar10 <- rep(NA, reps)
for (i in 1:reps) {
  vbar10[i] <- var(samples_1[,i])}
mean(vbar10);var(vbar10)
hist(vbar10, freq=F, xlab="", breaks=100)
par(new=T)
plot(density(vbar10), axes=F, main="", col="blue")
(중략)
vbar100<-numeric(10000)
samples_4 <- replicate(reps, rnorm(n4,10,2))
vbar100 <- rep(NA, reps)
for (i in 1:reps) {
  vbar100[i] <- var(samples_4[,i])}
mean(vbar100);var(vbar100)
hist(vbar100, freq=F, xlab="", breaks=100)
par(new=T)
plot(density(vbar100), axes=F, main="", col="blue")
```

구분		n=10	n=20	n=30	n=100
평균	표본분산	4.020385	4.002698	3.993987	3.993146
	모집단	4	4	4	4
분산	표본분산	3.60356	1.676938	1.087883	0.3217269
	모집단	3.5555	1.6842	1.1034	0.3232