중간과제물 모범답안

1. 다음 관계를 증명하고, R 프로그램의 난수를 이용하여 히스토그램을 그리고 이를 바탕으로 관계가 성립함을 보이시오.

(1) $X \sim Bin(n,p)$의 확률표본일 때, n 이 커지면서 $\frac{(X-np)^2}{np(1-p)} \sim \chi^2(1)$

(증명)

$X \sim B(n,p)$에서 n 이 충분히 커지게 되면 중심극한정리에 따라서

$$E(X) = np, \qquad Var(X) = np(1-p)$$

인 정규분포에 근사된다.

X를 표준정규분포 Z로 표준화하게 되면,

$$Z = \frac{X-np}{\sqrt{np(1-p)}}$$

로 변환할 수 있다.

표준화된 Z는 $N(0,1)$ 을 따르며, 카이제곱 분포의 정의에 따라 표준정규분포의 확률변수의 제곱은 $\chi^2(1)$ 를 따르게 되므로,

$$Z^2 = \left(\frac{X-np}{\sqrt{np(1-p)}}\right)^2 = \frac{(X-np)^2}{np(1-p)}$$

이 된다. 따라서,

$$Z^2 = \frac{(X-np)^2}{np(1-p)} \sim \chi^2(1)$$

(R 프로그램)

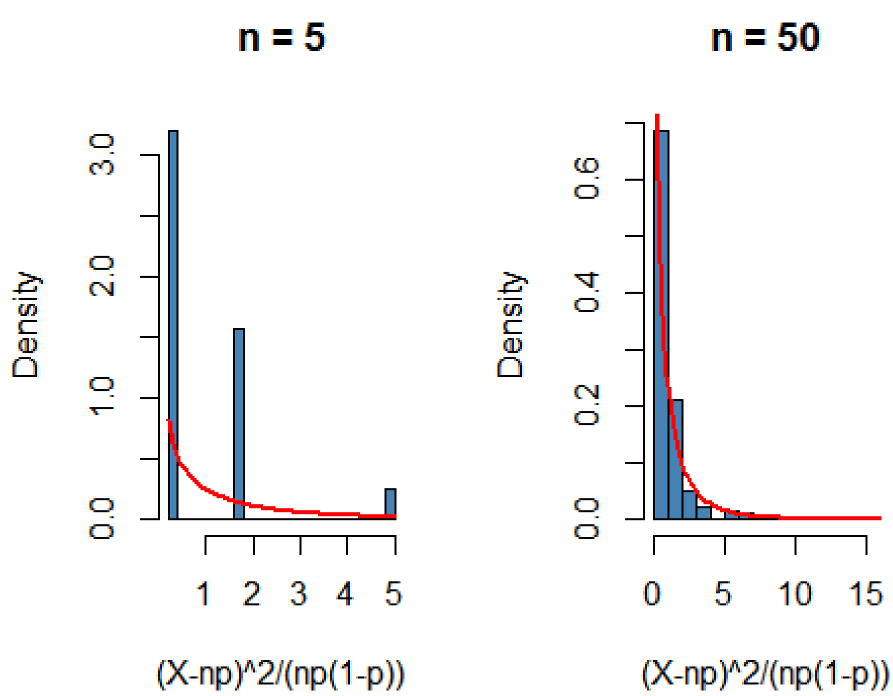
n=5 에서 n=50 으로 커지는 경우를 살펴보기 위해서 두 가지에 대한 히스토그램을 만들어 볼 수 있다.

```
n_values <- c(5, 50)
p <- 0.5
par(mfrow=c(1,2))

for (n in n_values) {
  X <- rbinom(1000, n, p)
  New_X1_1 <- (X - n*p)^2 / (n*p*(1-p))

  hist(New_X1_1, breaks=20, probability=TRUE,
       main=paste("n =", n),
       xlab="(X-np)^2/(np(1-p))", col="steelblue")

  curve(dchisq(x, df=1), add=TRUE, col="red", lwd=2)
}
```



그래프를 살펴보면 n=50 일 때 자유도가 1 인 카이제곱분포와 유사해진다.

```r
n_values <- c(5, 50)
p <- 0.5
results1 <- data.frame()

for (n in n_values) {
  X <- rbinom(1000, n, p)
  New_X1_1 <- (X - n*p)^2 / (n*p*(1-p))

  mean1_1 <- mean(New_X1_1)
  var1_1 <- var(New_X1_1)
  results1 <- rbind(results1, data.frame(n=n, Mean=mean1_1, Variance=var1_1))
}

print(results1)

##    n    Mean Variance
## 1  5 1.11040 1.843095
## 2 50 1.06144 2.259079
```

자유도가 1 인 카이제곱분포의 평균은 1 이며, 분산은 2 이다. 시뮬레이션으로 얻어진 평균과 분산을 살펴보면 n 이 늘어날수록 이론적인 평균과 분산에 수렴함을 확인할 수 있다.

따라서 n 이 늘어날수록 자유도가 1 인 카이제곱분포에 수렴함을 알 수 있다.

(2) $X_1, \cdots, X_n \sim Uniform(0,1)$의 확률표본이고 $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$를 순서통계량이라고 할 때, n이 커지면서 $n(1 - X_{(n)}) \sim Exp(1)$

(증명)

$X_{(1)}, \cdots, X_{(n)}$은 각각 $U(0,1)$ 에서 독립적으로 추출되며, $X_{(n)}$은 그 중 최대값이다. $X_{(n)}$의 누적분포함수는 모든 변수들이 x 이하의 값을 가질 확률이다.

따라서,

$$F_{X_{(n)}} = P(X_{(n)} \leq x) = P(X_{(1)} \leq x, X_{(2)} \leq x, \cdots, X_{(n)} \leq x)$$
$$= P(X_{(1)} \leq x) \times P(X_{(2)} \leq x) \times \cdots \times P(X_{(n)} \leq x) = x^n$$

이다(모든 시행이 독립이므로).

한편, $Y = n(1 - X_{(n)})$ 라고 하면,

$$P(Y \leq y) = P(n(1 - X_{(n)}) \leq y) = P\left(X_{(n)} \geq 1 - \frac{y}{n}\right) = 1 - P\left(X_{(n)} \leq 1 - \frac{y}{n}\right)$$

$P(X_{(n)} \leq x) = x^n$ 이므로,

$$P(Y \leq y) = 1 - P\left(X_{(n)} \leq 1 - \frac{y}{n}\right) = 1 - \left(1 - \frac{y}{n}\right)^n$$

이다.

충분히 n 이 커지는 경우, 이항전개에 의해서

$$(1 - \frac{y}{n})^n \approx e^{-y}$$

이며, 따라서

$$P(Y \leq y) = 1 - \left(1 - \frac{y}{n}\right)^n \approx 1 - e^{-y}$$

위 식의 우변은 Exp(1)의 cdf 와 동일하다.

결론적으로 n 이 충분히 큰 경우 $Y = n(1 - X_{(n)})$는 Exp(1)를 따르게 된다.

(R 프로그램)

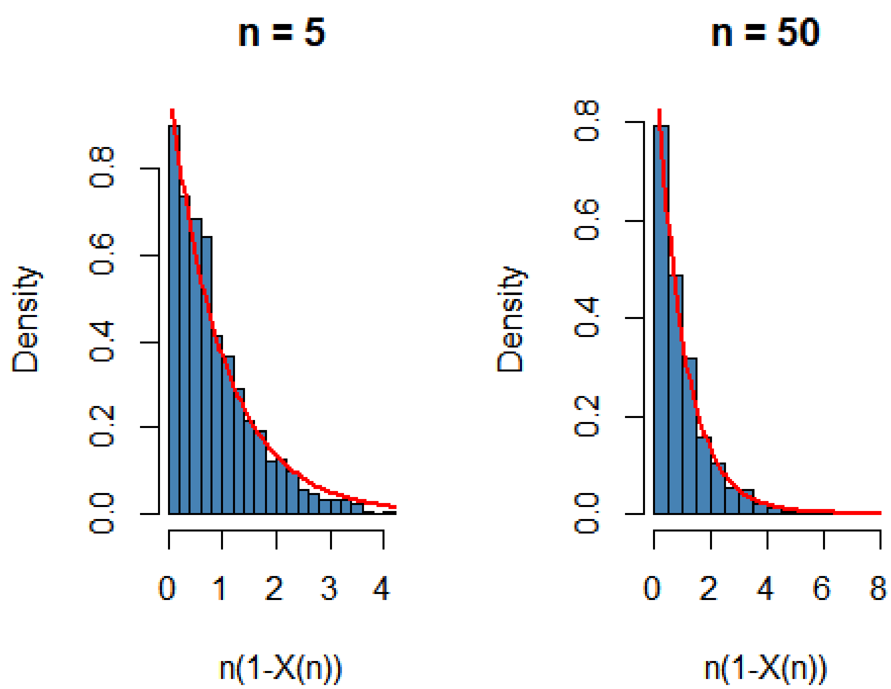
n=5 에서 n=50 으로 커지는 경우를 살펴보기 위해서 두 가지에 대한 히스토그램을 만들어 볼 수 있다.

```
n_values <- c(5, 50)
par(mfrow=c(1,2))

for (n in n_values) {
  X_matrix <- matrix(runif(n * 1000), ncol = n)
  max_X <- apply(X_matrix, 1, max)
  New_X1_2 <- n * (1 - max_X)

  hist(New_X1_2, breaks=20, probability=TRUE,
       main=paste("n =", n),
       xlab="n(1-X(n))", col="steelblue")

  curve(dexp(x, rate=1), add=TRUE, col="red", lwd=2)
}
```



그래프를 살펴보면 n=50 일 때 exp(1)과 유사해진다.

```
n_values <- c(5, 50)
results2 <- data.frame()
```

```r
for (n in n_values) {
  X_matrix <- matrix(runif(n * 1000), ncol = n)
  max_X <- apply(X_matrix, 1, max)
  New_X1_2 <- n * (1 - max_X)

  mean1_2 <- mean(New_X1_2)
  var1_2 <- var(New_X1_2)
  results2 <- rbind(results2, data.frame(n=n, Mean=mean1_2, Variance=var1_2))
}

print(results2)
```

```
##    n      Mean  Variance
## 1  5 0.8474153 0.4996276
## 2 50 0.9725498 0.9665548
```

Exp(1)의 평균은 1 이며, 분산 또한 1 이다. 시뮬레이션으로 얻어진 평균과 분산을 살펴보면 n 이 늘어날수록 이론적인 평균과 분산에 수렴함을 확인할 수 있다.

따라서 n 이 늘어날수록 exp(1)에 수렴함을 알 수 있다.

2. $X_i \sim N(\mu, \sigma^2)$의 확률표본일 때 다음 관계 중 (1)-(3)을 증명하고, (1), (2), (4)에 대해 R 프로그램의 난수(n=30 을 가정)를 이용하여 히스토그램을 그리고 이를 바탕으로 관계가 성립함을 보이시오.

(1) $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

(증명)

표본평균 $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ 이다.

$$E(\bar{X}) = E\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) = \frac{1}{n}(E(X_1) + E(X_2) + \cdots + E(X_n)) = \frac{1}{n} \times n\mu = \mu$$

$$Var(\bar{X}) = Var\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) = \frac{1}{n^2}(Var(X_1) + Var(X_2) + \cdots + Var(X_n))$$

$$= \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}$$

따라서

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

(R 프로그램)

여기서는 X_i의 평균을 0, 분산을 1 로 두고 구하였다.

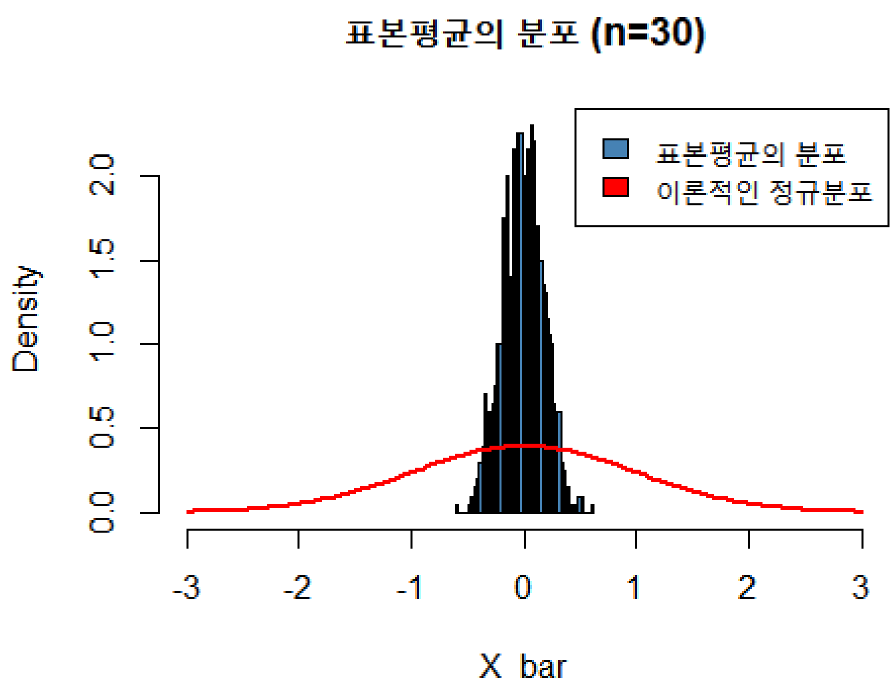
```r
mu <- 0
sigma <- 1
n <- 30

X_bar <- replicate(1000, mean(rnorm(n, mu, sigma)))

hist(X_bar, breaks=50, probability=TRUE,
     main="표본평균의 분포 (n=30)",
     xlim = c(-3,3),
     col="steelblue")

curve(dnorm(x, mean=mu, sd=sigma), add=TRUE, col="red", lwd=2)
```

```r
legend("topright", legend=c("표본평균의 분포", "이론적인 정규분포"),
       fill=c("steelblue", "red"))
```



n = 30 의 경우, 평균은 0 과 같지만 분산은 더 작아져 $\bar{X}$가 평균 주위로 몰리게 되는 것을 볼 수 있다. 직접 구해보면,

```r
mean(X_bar);mu
## [1] -0.01026516
## [1] 0
var(X_bar);sigma/n
## [1] 0.03276775
## [1] 0.03333333
```

평균은 -0.010, 분산은 0.033 으로, 계산된 평균 0, 분산 0.033 과 유사하다.

(2) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

(증명)

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2$$

위 식의 양변을 모두 σ^2으로 나누면,

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} + \frac{n(\bar{X}-\mu)^2}{\sigma^2}$$

이 된다.

한편, 표본분산 $S^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i-\bar{X})^2$ 이며, 이를 위 식에 적용하면,

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} + \frac{n(\bar{X}-\mu)^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{(n-1)} \times \frac{(n-1)}{\sigma^2} + \frac{(\sqrt{n})^2(\bar{X}-\mu)^2}{\sigma^2}$$
$$= \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

위 식을 $V_1 = V_2 + V_3$로 나타낼 수 있다. 이 때 1) 표준정규분포의 제곱의 분포는 카이제곱을 따른다는 성질과 2) 카이제곱분포의 가법성에 따라, $V_1 \sim \chi^2(n)$ 이며, $V_3 \sim \chi^2(1)$이므로,

$$V_2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

이다.

(R 프로그램)

여기서는 X_i의 평균을 0, 분산을 1 로 두고 구하였다.

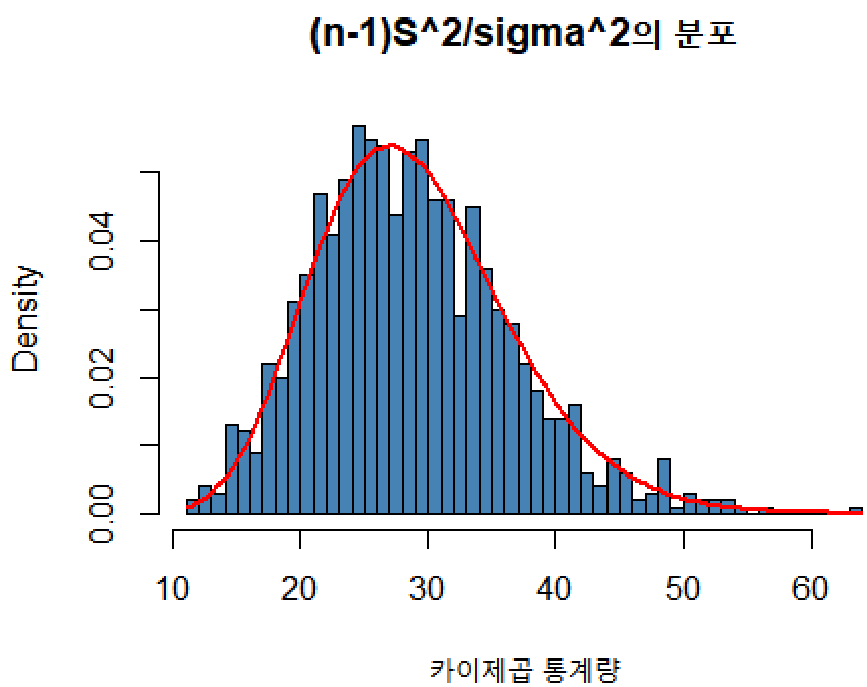
```
mu <- 0
sigma <- 1
n <- 30

chisq_values <- replicate(1000, {
```

```r
  sample2_2 <- rnorm(n, mu, sigma)
  sample_var2_2 <- var(sample2_2)
  (n-1) * sample_var2_2 / sigma^2
})

hist(chisq_values, breaks=50, probability=TRUE, main="(n-1)S^2/sigma^2 의 분포
",
     xlab="카이제곱 통계량", col="steelblue")

curve(dchisq(x, df=n-1), add=TRUE, col="red", lwd=2)
```



n = 30 의 경우 $\frac{(n-1)S^2}{\sigma^2}$는 자유도가 29 인 카이제곱분포와 유사하게 되는 것을 알 수 있다. 평균과 분산을 구해보면,

```r
mean(chisq_values);(n-1)

## [1] 28.65685

## [1] 29

var(chisq_values);(2*(n-1))

## [1] 59.68951

## [1] 58
```

이론상 평균은 29, 분산은 58 이며, 시뮬레이션으로 구한 평균 28.967, 표준편차 59.690 으로 유사해짐을 알 수 있다.

(3) $\bar{X}$와 S^2은 서로 독립

(증명)

표본분산 $S^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2$는 $(X_1 - \bar{X}, \cdots, X_n - \bar{X})$ 의 함수이므로, 이 변수들이 $\bar{X}$와 독립임을 보이면 된다.

$Y = (X_1 - \bar{X}, \cdots, X_n - \bar{X})$라 할 때, $\bar{X}$와 Y의 결합적률생성함수를 구하여 전개하면 다음과 같다.

$$M_{\bar{X},Y}\left(s, (t_1, \cdots, t_n)\right)$$
$$= E[\exp\{s\bar{X} + \ t_1(X_1 - \bar{X}) + \cdots + t_n(X_n - \bar{X})\}]$$
$$= E\left[\exp\left\{\left(\frac{s}{n} + \ (t_1 - \bar{t})\right)X_1 + \cdots + \left(\frac{s}{n} + \ (t_n - \bar{t})\right)X_n\right\}\right]$$
$$= \prod_{i=1}^{n} M_{X_i}\left(\frac{s}{n} + (t_i - \bar{t})\right)$$
$$= \exp\left[\sum_{i=1}^{n}\{\mu(\frac{s}{n} + (t_i - \bar{t})) \ + \frac{1}{2}\sigma^2\left(\frac{s}{n} + (t_i - \bar{t})\right)^2\}\right]$$
$$= \exp\,(\mu s \ + \frac{1}{2}\frac{\sigma^2}{n}s^2)\exp\,(\frac{1}{2}\sigma^2\sum_{i=1}^{n}(t_i - \bar{t})^2)$$

이로부터 $\bar{X}$와 Y의 결합적률생성함수는 각각의 주변적률생성함수의 곱인 s 함수와 t 함수의 곱으로 나타내어진다. 따라서 $\bar{X}$와 Y는 독립이므로, 결론적으로 $\bar{X}$와 S^2는 서로 독립이다.

(4) $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

(R 프로그램)

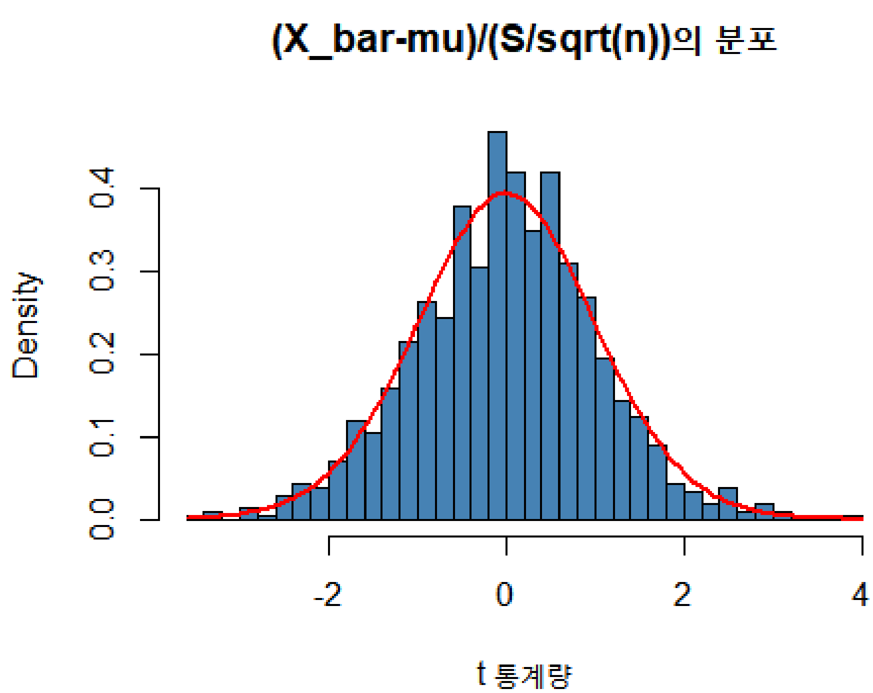
여기서는 X_i의 평균을 0, 분산을 1 로 두고 구하였다.

```
mu <- 0
sigma <- 1
n <- 30

t_values <- replicate(1000, {
  sample2_4 <- rnorm(n, mu, sigma)
  X_bar2_4 <- mean(sample2_4)
  sample_var2_4 <- var(sample2_4)
  (X_bar2_4 - mu) / (sqrt(sample_var2_4)/sqrt(n))
})

hist(t_values, breaks=50, probability=TRUE, main="(X_bar-mu)/(S/sqrt(n))의 분포",
     xlab="t 통계량", col="steelblue")

curve(dt(x, df=n-1), add=TRUE, col="red", lwd=2)
```



n = 30 의 경우 $\frac{\bar{X}-\mu}{S/\sqrt{n}}$는 자유도가 29 인 t 분포와 유사하게 되는 것을 알 수 있다. 평균과 분산을 구해보면,

```
mean(t_values);0
## [1] -0.004487855
## [1] 0
var(t_values);(n-1)/(n-1-2)
## [1] 1.087133
## [1] 1.074074
```

이론상 평균은 0, 분산은 29/27 이며, 시뮬레이션으로 구한 평균 -0.004, 표준편차 1.087 으로 유사해짐을 알 수 있다.

3. $X_i \sim Exp(2)$의 확률표본일 때 다음 물음에 답하시오.

(1) $\bar{X}$가 $E(X_1)$으로 확률적으로 수렴함을 증명하고, n=10, 100, 1000, 10000 일 때의 값과 $E(X_1)$의 값을 비교하시오.

(증명)

$X_i \sim Exp(2)$ 이며, 지수분포에서 평균은 $\frac{1}{\lambda}$, 분산은 $\frac{1}{\lambda^2}$ 이므로, $E(X_i) = \frac{1}{2}, Var(X_i) = \frac{1}{4}$ 이다.

표본의 평균 및 분산은 $E(\bar{X}) = \mu = \frac{1}{2},\ Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{1}{4n}$ 이다.

체비셰프 부등식을 이용하면, 임의의 상수 $\varepsilon > 0$에 대해

$$P(|\overline{X_n} - \mu| < \varepsilon) = P((\overline{X_n} - \mu)^2 < \varepsilon^2) \geq 1 - \frac{Var(\overline{X_n})}{\varepsilon^2} = 1 - \frac{1}{4n\varepsilon^2}$$

위 식은 n 이 ∞로 갈수록 우변이 1 으로 수렴함을 알 수 있다($\lim_{n\to\infty} P(|\overline{X_n} - \mu| < \varepsilon) = 1$).

따라서, $\overline{X_n} \xrightarrow{p} \frac{1}{2}$ 이다.

(값의 비교)

R 을 통해서 n=10, 100, 1000, 10000 일때의 값을 알 수 있다.

```
sample_means3_1 <- numeric(4)
index <- 1

for (nn in c(10,100,1000,10000)){
  XBar = rep(NA, 10000)
  for (i in 1:10000) XBar[i] = mean(rexp(nn, 2))
  sample_means3_1[index] <- mean(XBar)
  index <- index + 1
}

print(data.frame('n' = c(10,100,1000,10000), 'mean' = sample_means3_1))
```

```
##       n      mean
## 1    10 0.4982141
## 2   100 0.5003333
```

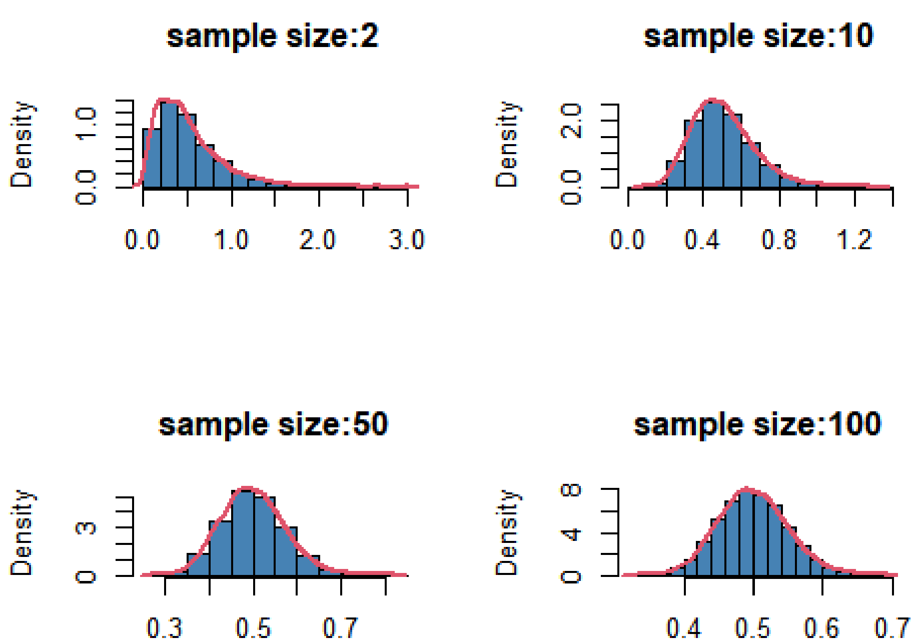
```
## 3  1000 0.4999665
## 4 10000 0.5001021
```

결과를 살펴보면 n=10 일 때 평균은 0.498, n=100 이상일 때 평균은 0.500 (소수 4 번째 자리에서 반올림)임을 알 수 있고, 이는 Exp(2)의 평균인 1/2=0.5 와 유사함을 알 수 있다.

(2) $\bar{X}$의 분포를 n = 2, 10, 50, 100 의 히스토그램을 그리고 중심극한정리로 그 의미를 정리하시오.

(히스토그램)

```
par(mfrow=c(2,2))
for (nn in c(2,10,50,100)){
  XBar = rep(NA, 10000)
  for (i in 1:10000) XBar[i] = mean(rexp(nn, 2))
  hist(XBar, main=paste0("sample size:", nn), xlab=" ", freq = FALSE, col="st
eelblue")
  lines(density(XBar), col=2, lwd=2)
}
```



(중심극한정리로 의미 정리)

중심극한정리란 표본을 동일한 분포에서 독립적으로 추출할 때, 표본수가 증가하게 되면 모집단의 분포와 상관없이 표본평균의 분포가 정규분포에 수렴한다는 정리이다(단, 모집단의 평균과 분산이 존재할 때).

위 히스토그램을 살펴보면 n 수가 적을 때는 왼쪽으로 치우친 분포를 보이지만, n 수가 증가함에 따라 그래프가 정규분포에 근사해짐을 시각적으로 확인할 수 있다.