10-1 표본분포 1

제주권역 대학 e-러닝 지원센터

실험

10,000명의 학생이 있는 J대학교 학생들의 키의 분포가 평균이 168cm이고 분산이 25cm인 정규분포 N(168,25)를 따른다고 할 때, 학생 10명을 임의로 추출하여 키를 측정한다고 하자.

확률변수 X가 '학생들의 키'를 나타낸다고 할 때, 표본으로 추출된 10명의 학생의 키는 다음과 같은 확률변수로 나타낼 수 있다.

$\{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}\}$

여기서 X_1 이 '처음 추출된 학생의 키'라고 하면 X_1 의 분포는 전체 학생의 분포와 같다.

즉, $X_1 \sim N(168, 25)$

이는 X_1이 표본을 반복해서 추출할 때마다 다른 값을 가질 수 있다는 것을 의미한다

또한 X_2, ... X_{10} 에 대해서도 각 값은 표본을 반복해서 추출할 때마다 다른 값을 가질 수 있으므로 X_1, X_2, X_3 , X_4 , X_5 , X_6 , X_7 , X_8, X_9, X_{10} 의 분포는 다음과 같이 나타낼 수 있다.

$X_i \sim N(168, 25)$, $i = 1, 2 \ldots, 10$

예 두 개의 다른 표본이 다음의 표와 같이 추출되었다면 <표본 1>에서 X_1=165이나 <표본 2>에서 X_1=180이다.

구분	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
표본 1	165	173	159	183	171	169	165	173	149	177
표본 2	180	144	163	172	185	172	166	174	170	165

위의 예에서 가능한 표본의 수는 10,000명에서 10명을 추출하는 방법인 ${}_{10000}C_{10}$ 과 같으므로 X_1의 가능한 값은 ${}_{10000}C_{10}$ 개 이다.

따라서 X_1의 분포는 ${}_{10000}C_{10}$ 개의 X_1값의 분포를 의미하며 이 값의 분포는 모집단의 분포 $X \sim N(168,25)$ 와 같다.

이 설명은 X_2, ..., X_{10} 에도 동일하게 적용될 수 있으며 X_1 , ..., X_{10} 이 서로 독립적으로 추출되었으므로 다음과 같은 분포를 가진다.

$X_i \sim NI(168,25)$, $i=1,2,$ .. 10

확률표본 (random sample)

✓확률표본이란 특정한 확률분포로부터 독립적으로 반복하여 표본을 추출하는 것으로 각각의 관찰값들은 서로 독립이며 동일한 분포를 가짐

✓확률변수 X가 특정 확률분포를 따른다고 할 때, 이 확률분포로부터 각각 독립적으로 관측된 n개의 표본임

이 표본을 ($X_1, X_2, \ldots, Xn$)이라 할 때 $X_1, X_2, \ldots, Xn$은 확률변수로 **상호독립**이며, 각각은 X와 **동일한 분포**를 갖는다.

예 1

10개의 표본을 관측할 때 평균이 168이고 분산이 25인 확률표본을 다음과 같이 나타낸다.

$$X_i \sim N(168, 25)\ ,\ i = 1, 2, .., 10$$

예 2

확률변수 X가 N(0, 1)인 분포를 따른다고 할 때, 이 확률분포로부터 n개의 관측값으로 된 확률표본을 추출한다고 한다. 각 관측값을 확률변수 $X_1, X_2, \ldots, X_n$ 으로 표현하면 $X_1, X_2, \ldots, X_n$ 은 서로 독립이며 각각의 확률분포는 다음과 같이 나타낼 수 있다.

$$X_i \sim N(0,1)\ ,\ i = 1, 2, .., n$$

통계량(statistic)

- ✓확률표본의 각 원소 $X_1, X_2, \ldots, X_n$이 확률변수이므로 이 **확률변수들의 함수로 정의된 통계량도 또한 확률변수**이다.
- ✓통계량이 확률변수라는 것은 각각의 표본에 따라 구하여지는 통계량의 값이 확률분포를 갖는다는 것을 의미한다.
- ✓일반적으로 확률표본은 $(X_1, X_2, \ldots, X_n)$과 같이 대문자 X로 표현하며, 이 확률표본의 함수인 추정량도 대문자로 표현하여 이 변수들이 확률변수임을 나타낸다.
- ✓반면에 구체적인 표본에 의하여 구한 값은 $(x_1, x_2, \ldots, x_n)$ 소문자를 사용하여 이 값들이 실제표본에 근거한 구체적인 값임을 나타낸다.

① 추정량(estimator) : 모수를 추정하는 통계량

② 추정치(estimate) : 구체적인 표본에 근거하여 구한 추정량의 값

③ 확률표본 $(X_1, X_2, \dots, Xn)$ 에 근거한 통계량

(a) $X_{(1)} = \min(X_1, X_2, \dots, Xn)$: 최소값

(b) $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$: 표본평균

(c) $X_{(n)} = \max(X_1, X_2, \dots, Xn)$: 최대값

(d) $\tilde{X}$ = median$(X_1, X_2, \dots, Xn)$: 중위수

(e) $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$: 표본분산

제주권역 대학 e-러닝 지원센터

표본분포

모집단에서 일정한 크기의 모든 가능한 표본을 추출하였을 때 그 모든 표본으로부터 계산된 통계량의 확률분포

J제약회사는 많은 종류의 신약을 개발하였다. 이 제약회사가 신약을 개발하기 위해서 1,2,3 혹은 4년의 시간이 걸렸으며 각각의 발생확률은 동등하다고 가정하자.
이때 각 개발기간의 발생확률이 동등하다는 가정을 이용하여 신약의 평균 개발기간인 모집단 평균 및 분산은 다음과 같이 구할 수 있다.

$$\mu = E(X) = \sum xf(x) = (1 \times \frac{1}{4}) + .. + (4 \times \frac{1}{4}) = 2.5$$

$$\sigma^2 = V(X) = \sum (x-\mu)^2 \cdot P_X(x)$$

$$= (1-2.5)^2 \times \frac{1}{4} + \cdots + (4-2.5)^2 \times \frac{1}{4} = 1.25$$

그림 모집단으로부터 n=2인 임의표본추출

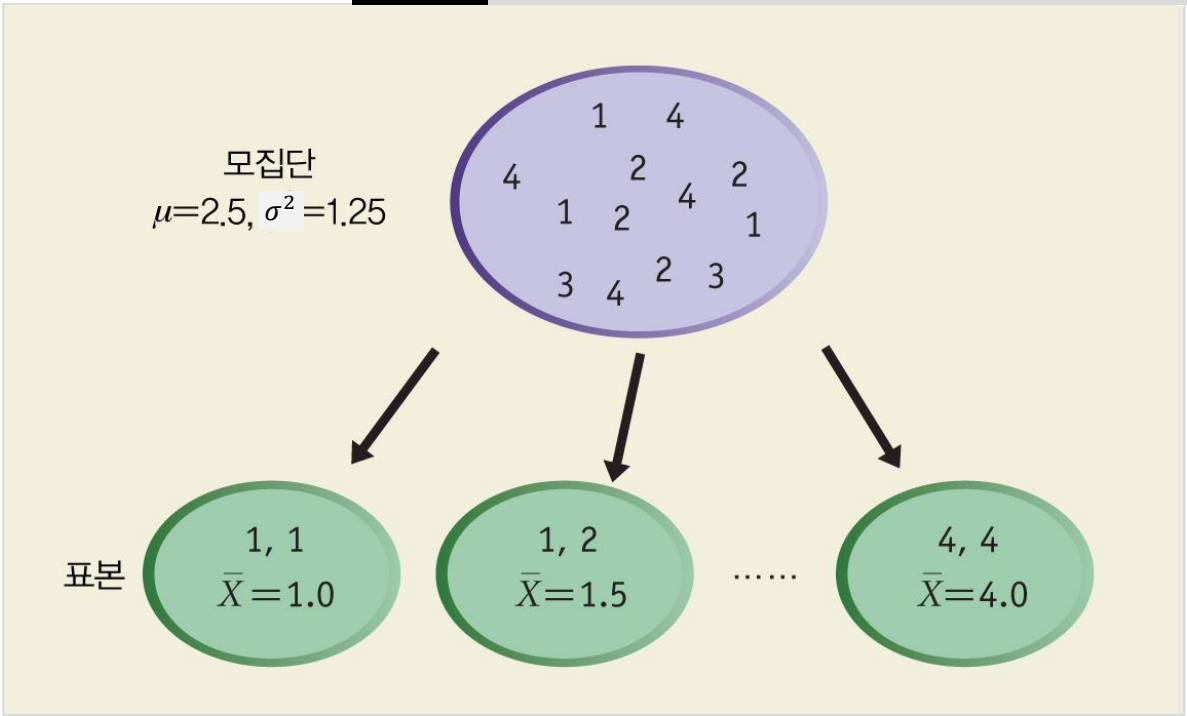


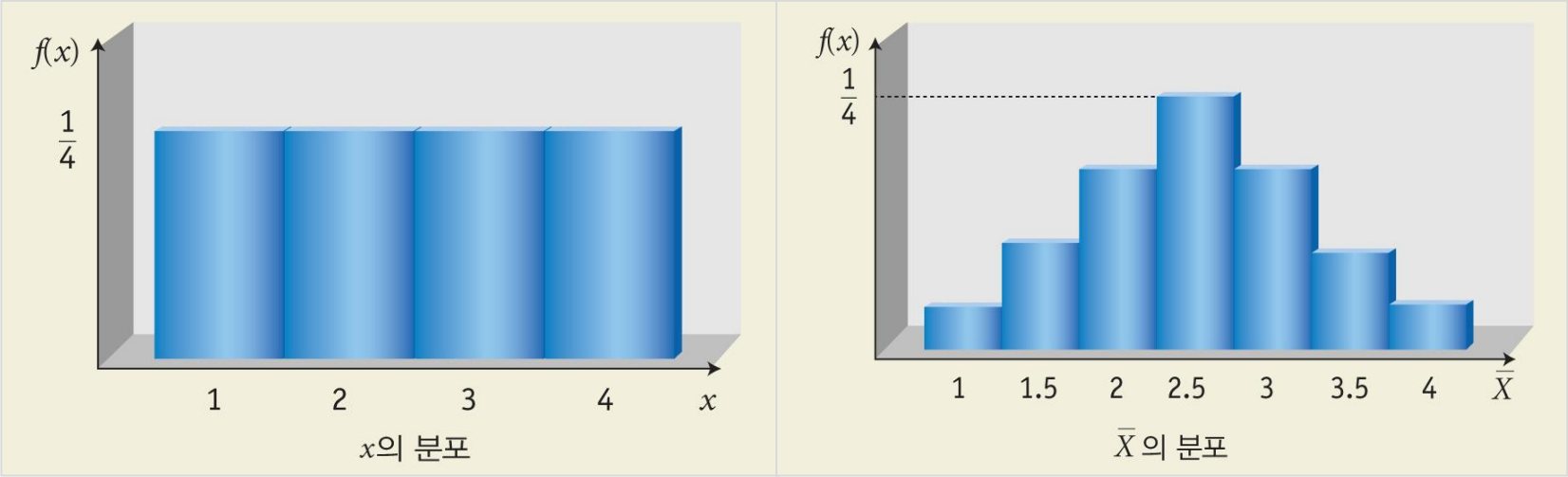
그림 표본크기 2의 모든 가능한 표본과 표본평균

표본	표본평균($\overline{X}$)	표본	표본평균($\overline{X}$)
1, 1	1.0	3, 1	2.0
1, 2	1.5	3, 2	2.5
1, 3	2.0	3, 3	3.0
1, 4	2.5	3, 4	3.5
2, 1	1.5	4, 1	2.5
2, 2	2.0	4, 2	3.0
2, 3	2.5	4, 3	3.5
2, 4	3.0	4, 4	4.0

그림 표본평균의 표본분포

$\overline{X}$	1	1.5	2	2.5	3	3.5	4
$P(\overline{X})$	1/16	2/16	3/16	4/16	3/16	2/16	1/16

그림 x와 $\bar{X}$의 분포



x의 분포 $\bar{X}$의 분포

$\bar{X}$의 평균은

$$\mu_{\bar{X}} = E(\bar{X}) = \sum \bar{X} \cdot P_X(\bar{X})$$

$$= (1.0 \times \frac{1}{16}) + (1.5 \times \frac{2}{16}) + \cdots + (4.0 \times \frac{1}{16}) = 2.5$$

이고 $\bar{X}$의 분산은

$$\sigma^2_{\bar{X}} = V(\bar{X}) = \sum (\bar{X} - \mu_{\bar{X}})^2 \cdot P_X(\bar{X})$$

$$= (1.0-2.5)^2 \times \frac{1}{16} + (1.5-2.5)^2 \times \frac{2}{16} + \cdots + (4.0-2.5)^2 \times \frac{1}{16}$$

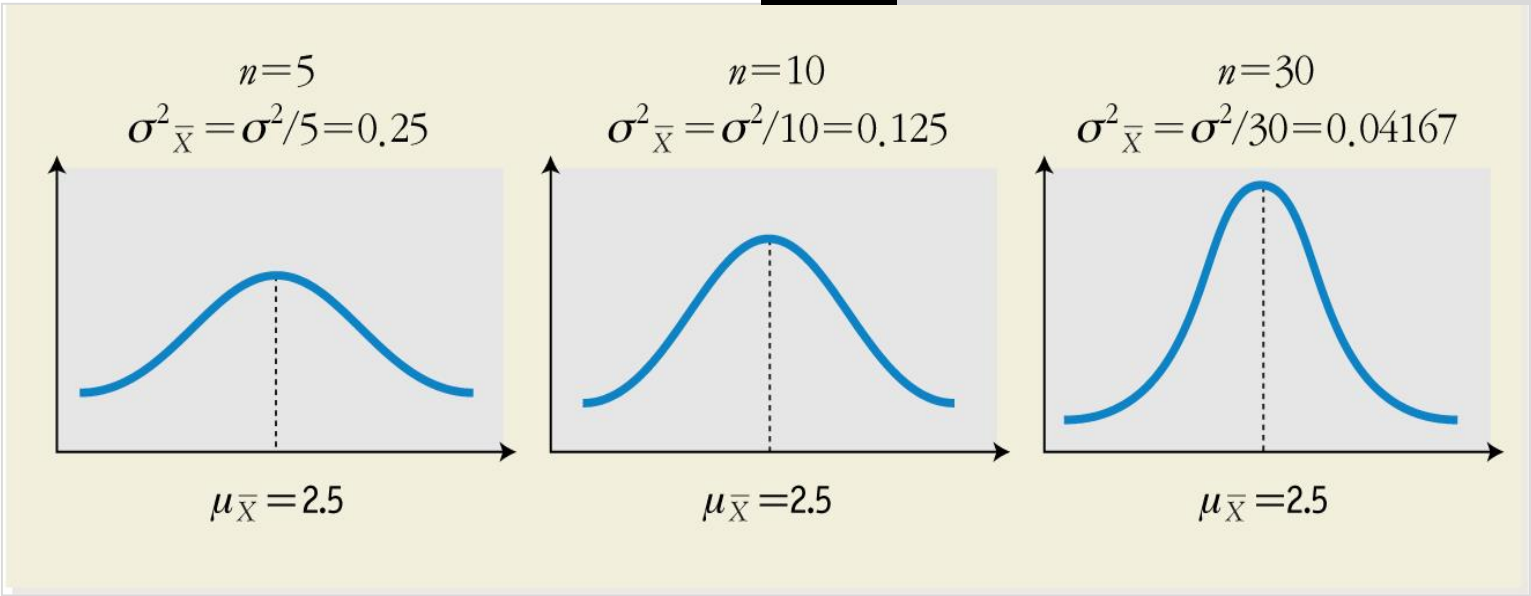
$$= 0.625$$

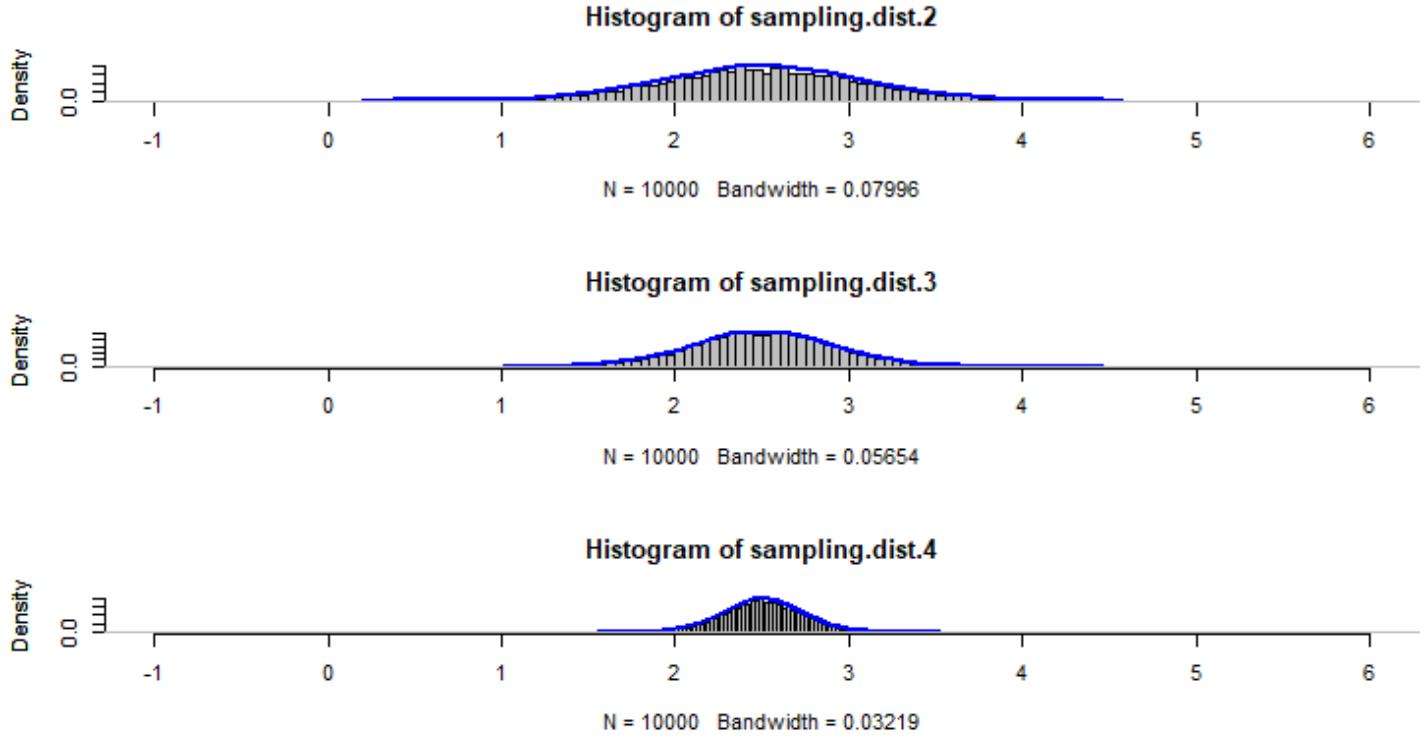
```
> sampling.dist.1<-NULL
> for(sample.count in 1:10000){
+   set.seed(sample.count)
+   sample.mean.1<-mean(rnorm(2,2.5,1.118))
+   sampling.dist.1<-c(sampling.dist.1, sample.mean.1)
+ }
> mean(sampling.dist.1)
[1] 2.502412
> var(sampling.dist.1)
[1] 0.6193981
```

계산결과를 보면 두 분포의 평균은 2.5로 같지만 x의 분산 σ_x^2 와 $\bar{X}$의 분산 $\sigma^2_{\bar{X}}$ 는 서로 같지 않으며, $\sigma^2_{\bar{X}}$는 σ_x^2 의 1/2이다.

표본평균의 표본분포

그림 N=5,10,30일 경우 $\overline{X}$의 표본분포




Histogram of sampling.dist.2
Density
0.0
-1
0
1
2
3
4
5
6
N = 10000 Bandwidth = 0.07996
Histogram of sampling.dist.3
Density
0.0
-1
0
1
2
3
4
5
6
N = 10000 Bandwidth = 0.05654
Histogram of sampling.dist.4
Density
0.0
-1
0
1
2
3
4
5
6
N = 10000 Bandwidth = 0.03219

표본평균의 표본분포 특성

평균 μ와 분산 σ^2 을 갖는 모집단으로부터 크기 n의 임의표본을 추출하였으며, 이 표본의 평균을 $\overline{X}$ 라고 하자.

1. $\mu_{\overline{X}} = \mu$

2. $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$ 혹은 $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

$\overline{X}$의 표준편차는 평균의 **표준오차(standard error)**라고도 한다.

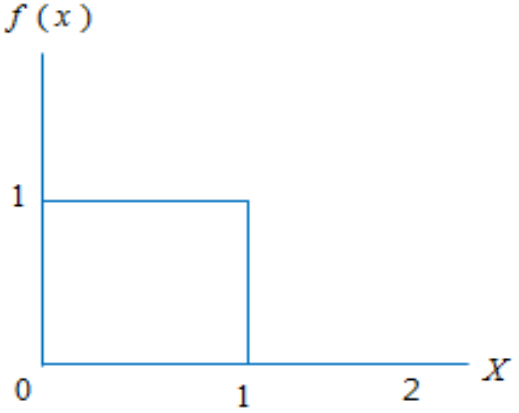
10-1-4. 중심극한정리

배경

관찰된 자료의 모집단이 정규분포에 따를 경우 관찰자료 역시 정규분포에 따르고, 관찰된 자료의 모집단이 실제로 정규분포가 아니면 관찰 자료 역시 정규분포에 따르지 않는다.
그러나 관찰된 자료의 모집단이 실제로 정규분포가 아닌 경우에도 중심극한정리에 의해 정규확률분포를 이용한 추정량의 근사확률을 구할 수 있다.

예

확률변수 X가 균등분포 U(0,1)을 따른다고 할 때 X~U(0,1)이며, X는 0과 1사이에서 균등한 분포를 갖는 연속형 확률변수이다.



이러한 확률분포로부터 크기 n=11인 확률표본을 1,000개 추출하는 실험을 해 보자.

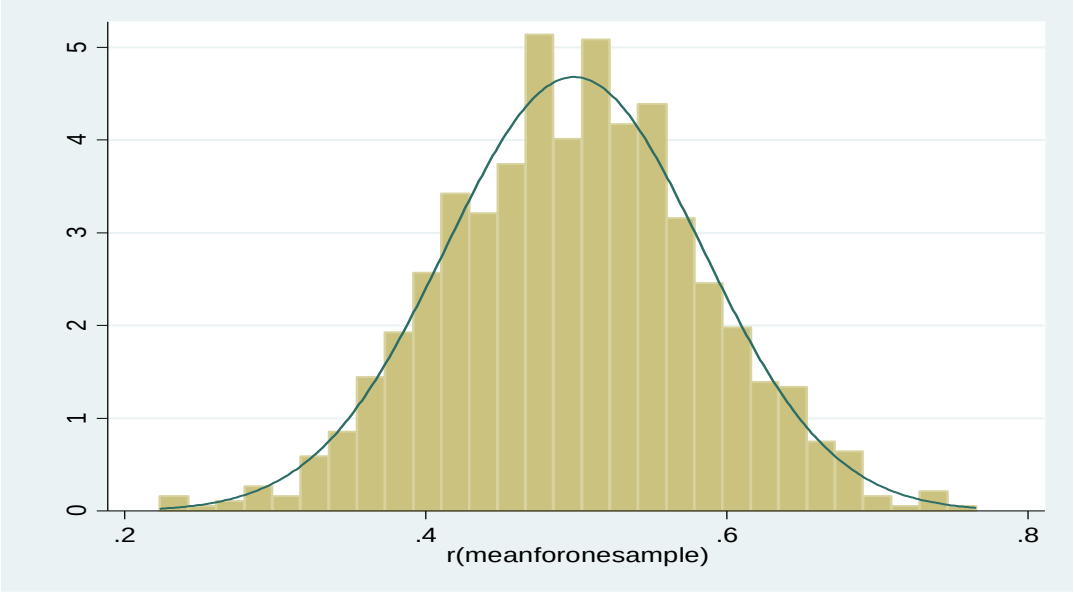
1,000개 표본 중 처음 5개의 표본이 다음과 같다고 하면,

각 표본의 평균 $\bar{x}_1, \bar{x}_2, \ldots \bar{x}_5$을 구할 수 있다.

표본	관측값											$\bar{x}$
1	.217	.786	.757	.125	.139	.919	.506	.771	.138	.516	.419	.481
2	.303	.703	.812	.650	.848	.392	.988	.469	.632	.012	.065	.534
3	.383	.547	.383	.584	.098	.676	.091	.535	.256	.163	.390	.373
4	.218	.376	.248	.606	.610	.055	.095	.311	0.86	.165	.665	.304
5	.144	.069	.485	.739	.491	.054	.953	.179	.865	.429	.648	.460

1,000개 표본으로부터 구한 표본평균 $\bar{x}_1, \bar{x}_2, .. \bar{x}_{1000}$ 을 이용하여 막대그림표를 그리면 다음과 같다.

즉, **균등분포로부터 구한 표본평균의 분포가 정규분포와 근사한 분포**를 갖는다



X~U(0,1)에서 E(X)=1/2, Var(X)=1/12이므로 n=11인 경우 표본평균 $\overline{X}$의 평균과 분산은 다음과 같다.

$$E(\overline{X}) = \frac{1}{n} nE(X) = E(X) = \frac{1}{2}$$

$$Var(\overline{X}) = \frac{1}{n^2} nVar(X) = \frac{1}{n} Var(X) = \frac{1}{11} \bullet \frac{1}{12} = \frac{1}{132} = 0.008$$

$$\sqrt{Var(\overline{X})} = \sqrt{0.008} = 0.087$$

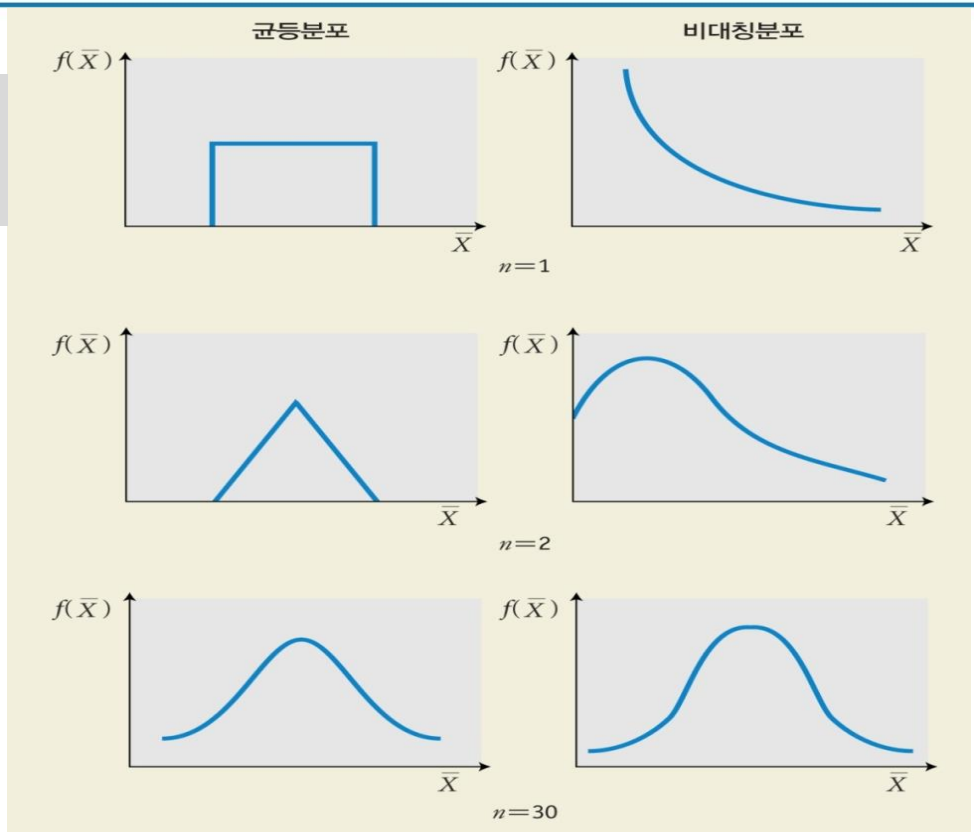
즉, 표본평균 $\overline{X}$ 는 근사적으로 $N(\frac{1}{2}, \frac{1}{132})$ 을 따른다.

1,000개 표본의 표본평균과 표준편차는 각각 0.498과 0.085로서 모집단의 이론적인 평균과 표준편차에 근접함을 알 수 있다.

Variable	Obs	Mean	Std. Dev.	Min	Max
xbar	1000	.4985796	.0852457	.2234436	.7655036

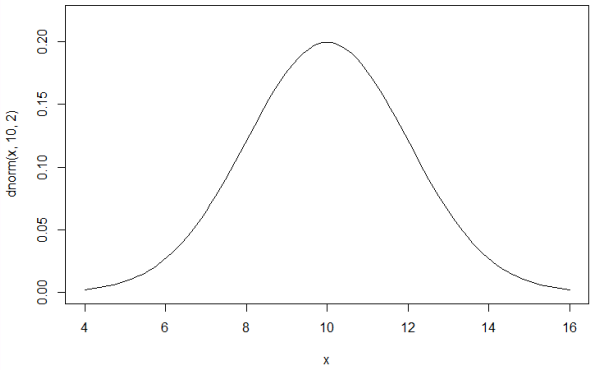
그림

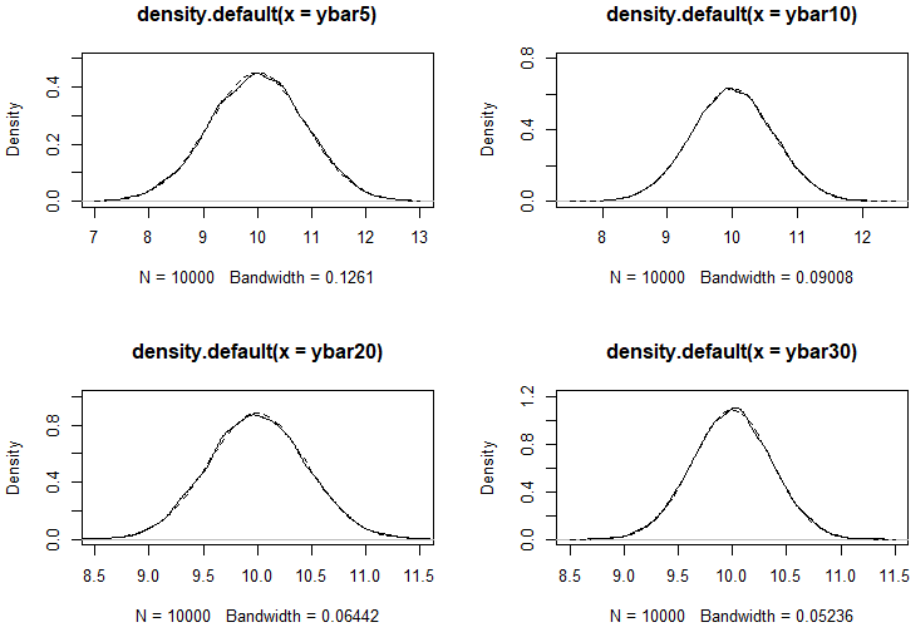
n=의 변화에 따른 $\bar{X}$의 확률 밀도함수의 모양



그림

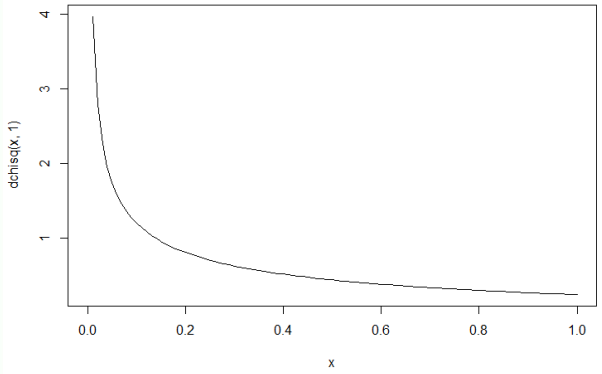
n의 변화에 따른 $\overline{X}$의 확률 밀도함수의 모양

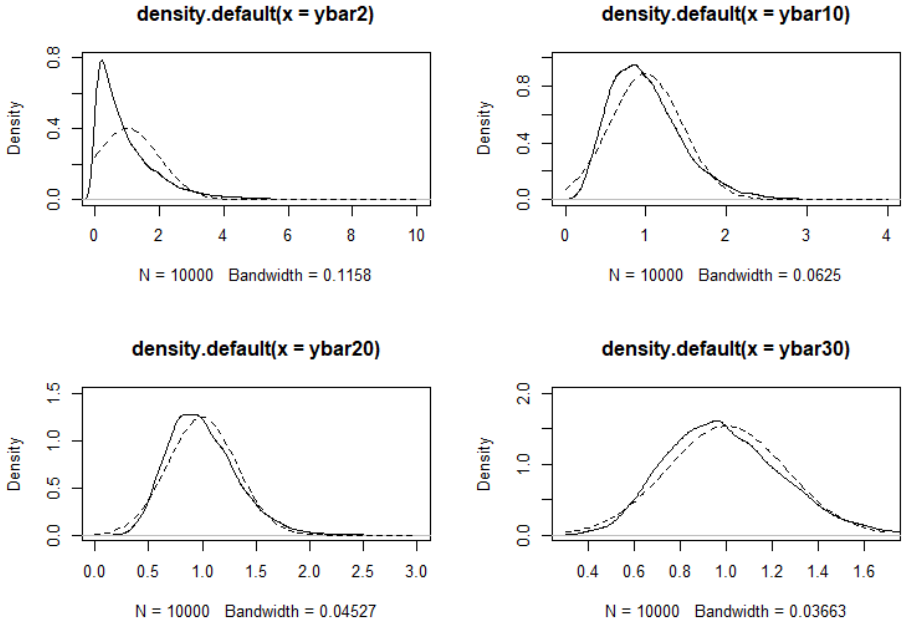




그림

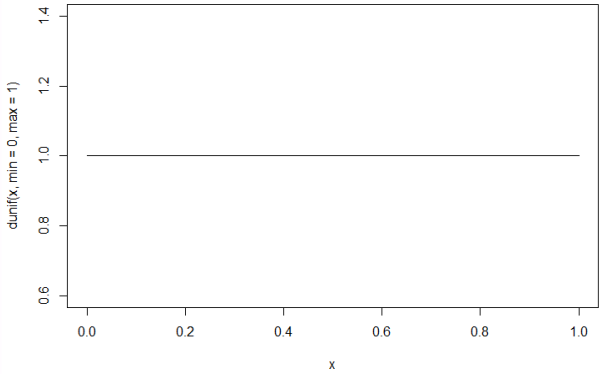
n의 변화에 따른 $\overline{X}$의 확률 밀도함수의 모양

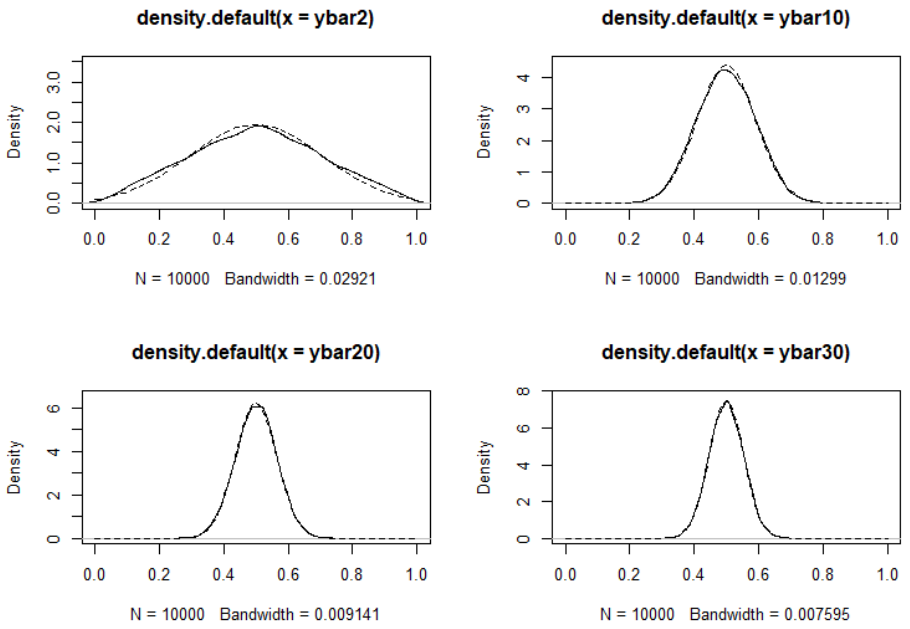




그림

n의 변화에 따른 $\overline{X}$의 확률 밀도함수의 모양





중심극한정리

평균이 μ 이고 분산이 σ^2 인 확률분포로부터 크기가 n인 확률표본 (X_1, X_2, ... , Xn)을 추출할 때, 표본평균 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 는 n이 클수록 평균이 μ 이고 분산이 σ^2/n 인 정규분포와 근사한 분포를 갖는다.

즉, $\overline{X}$ 의 분포는

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

과 같이 표현한다.

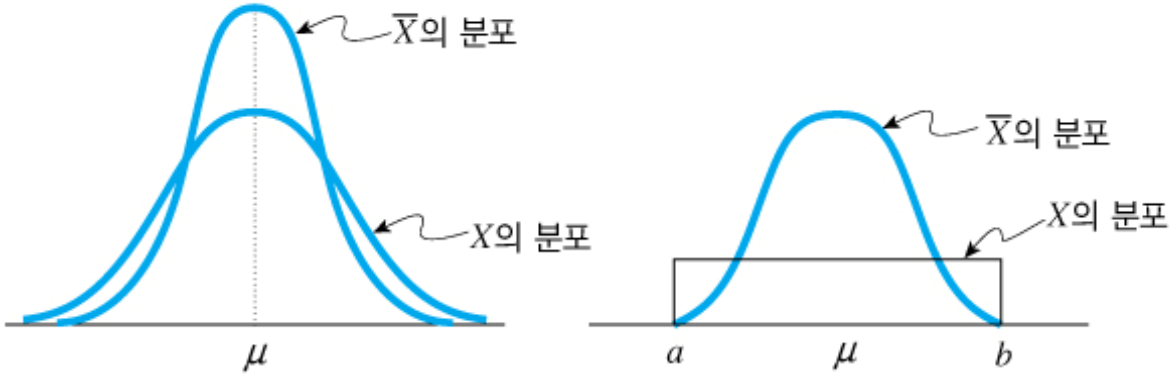
만약에 확률표본 (X_1, X_2, ... , Xn) 이 평균μ와 분산 σ^2 을 갖는 정규분포에서 추출되었다면, 표본평균 $\overline{X}$ 의 분포는 n의 크기에 관계없이 평균 μ 와 분산 σ^2/n 을 갖는 정규분포를 따른다.

즉, $\overline{X}$ 의 분포는

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

이다.

그림 X와 $\overline{X}$의 분포 모형



(a) $X \sim N(\mu, \sigma^2)$인 경우 (b) $X \sim U(a, b)$인 경우

$\overline{X}$ 의 분포는 X의 분포와 중심은 같으나 분산은 작아지며, 분포의 모양은 X의 분포와 관계없이 표본의 수 n이 커질수록 정규분포 모양을 갖는다.
위 그림은 X의 분포가 정규분포인 경우와 균등분포인 경우에 대해 그림으로 표현한 것이다.

10-1-5. 표본평균의 표준화

평균이 μ 이고 분산이 σ^2 인 확률변수의 X를 표준화하는 방법은 평균을 빼주고 표준편차로 나눈다. 따라서 평균이 μ 이고 분산이 σ^2/n 인 표본평균 $\overline{X}$ 의 표준화된 확률변수는 $\dfrac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$ 이다.