

9-2 표본분포 1



9-2-1. 확률표본과 통계량

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 X 가 '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_1 이 표본을 반복해서 추출할 때마다 다른 값을 가질 수 있다는 것을 의미한다

또한 $X_2, \dots, X_{10}$ 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.

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

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

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_1 의 가능한 값은 $_{10000}C_{10}$ 개 이다.

따라서 X_1 의 분포는 $_{10000}C_{10}$ 개의 X_1 값의 분포를 의미하며 이 값의 분포는 모집단의 분포 $X \sim N(168, 25)$ 와 같다.

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

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

확률표본 (random sample)

- ✓ 확률표본이란 특정한 확률분포로부터 독립적으로 반복하여 표본을 추출하는 것으로 각각의 관찰값들은 서로 독립이며 동일한 분포를 가짐
- ✓ 확률변수 X 가 특정 확률분포를 따른다고 할 때, 이 확률분포로부터 각각 독립적으로 관측된 n 개의 표본임

이 표본을 $(X_1, X_2, \dots, X_n)$ 이라 할 때 $X_1, X_2, \dots, X_n$ 은 확률변수로 **상호독립**이며, 각각은 X 와 **동일한 분포**를 갖는다.

예 1

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

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

예 2

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

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

통계량(statistic)

- ✓ 확률표본의 각 원소 $X_1, X_2, \dots, X_n$ 이 확률변수이므로 이 확률변수들의 함수로 정의된 통계량도 또한 확률변수이다.
- ✓ 통계량이 확률변수라는 것은 각각의 표본에 따라 구하여지는 통계량의 값이 확률분포를 갖는다는 것을 의미한다.
- ✓ 일반적으로 확률표본은 $(X_1, X_2, \dots, X_n)$ 과 같이 대문자 X 로 표현하며, 이 확률표본의 함수인 추정량도 대문자로 표현하여 이 변수들이 확률변수임을 나타낸다.
- ✓ 반면에 구체적인 표본에 의하여 구한 값은 $(x_1, x_2, \dots, x_n)$ 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 $(X_1, X_2, \dots, X_n)$ 에 근거한 통계량

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

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

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

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

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

9-2-2. 표본분포

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.

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

$$\sigma^2 = Var(X) = \sum (x - \mu)^2 P_x(x) = (1 - 2.5)^2 \times \frac{1}{4} + \dots + (4 - 2.5)^2 \times \frac{1}{4} = 1.25$$

그림

모집단으로부터 $n=2$ 인 임의표본추출

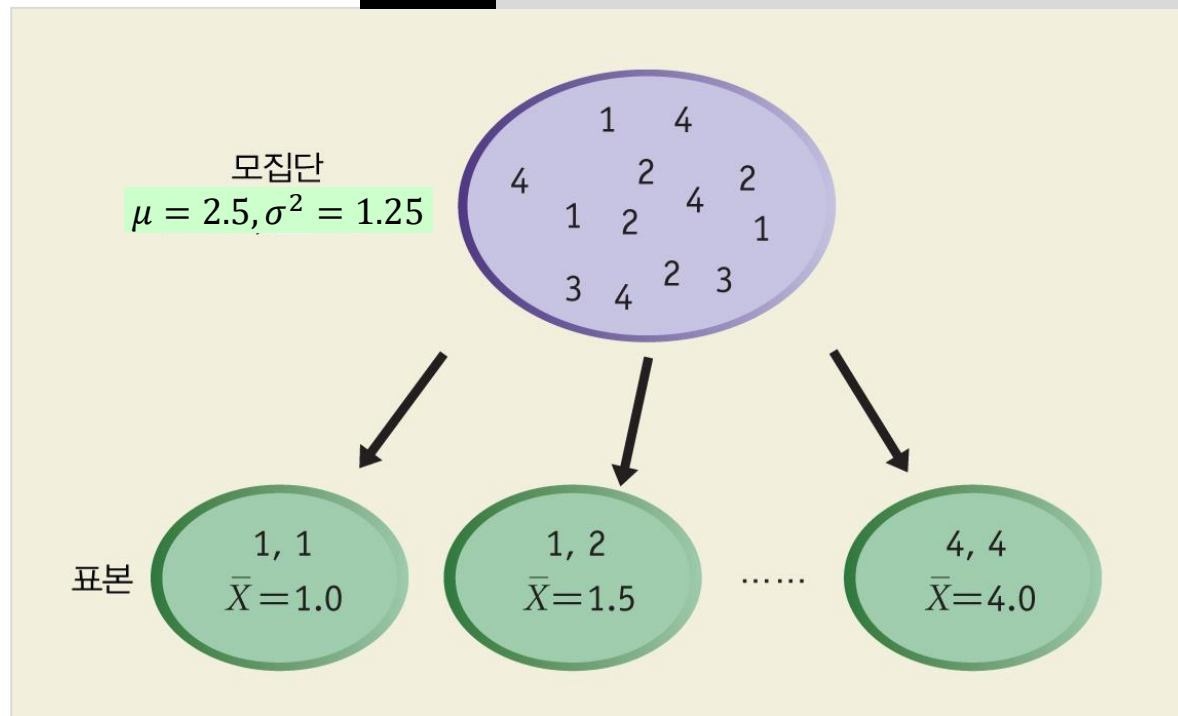


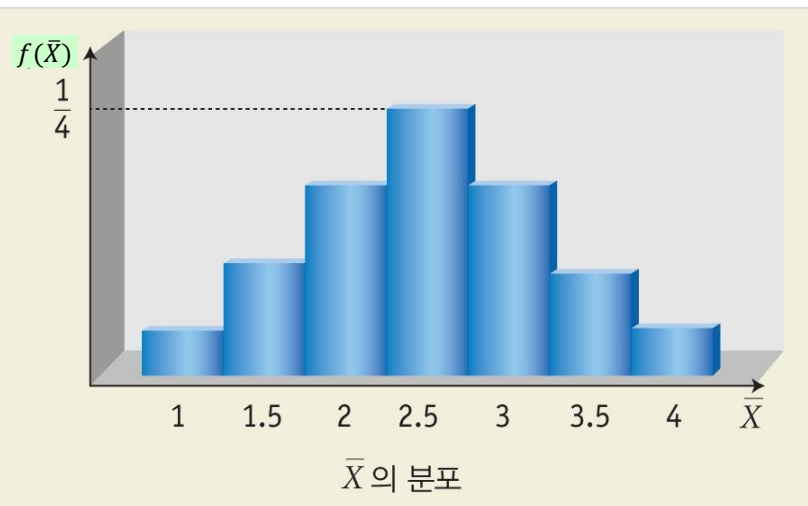
그림 표본크기 2의 모든 가능한 표본과 표본평균

표본	표본평균($\bar{X}$)	표본	표본평균($\bar{X}$)
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 확률분포

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

그림 x 와 $\bar{X}$ 의 분포



- $\bar{X}$ 의 평균과 분산은 각각 다음과 같음

· 평균 : $\mu_{\bar{X}} = E(\bar{X}) = \sum \bar{X}P(\bar{X}) = \left(1.0 \times \frac{1}{16}\right) + \left(1.5 \times \frac{2}{16}\right) + \dots + \left(4.0 \times \frac{1}{16}\right) = 2.5$

· 분산 : $\sigma_{\bar{X}}^2 = Var(\bar{X}) = \sum (\bar{X} - \mu_{\bar{X}})^2 P(\bar{X}) = (1.0 - 2.5)^2 \times \frac{1}{16} + (1.5 - 2.5)^2 \times \frac{2}{16} + (4.0 - 2.5)^2 \times \frac{1}{16} = 0.625$

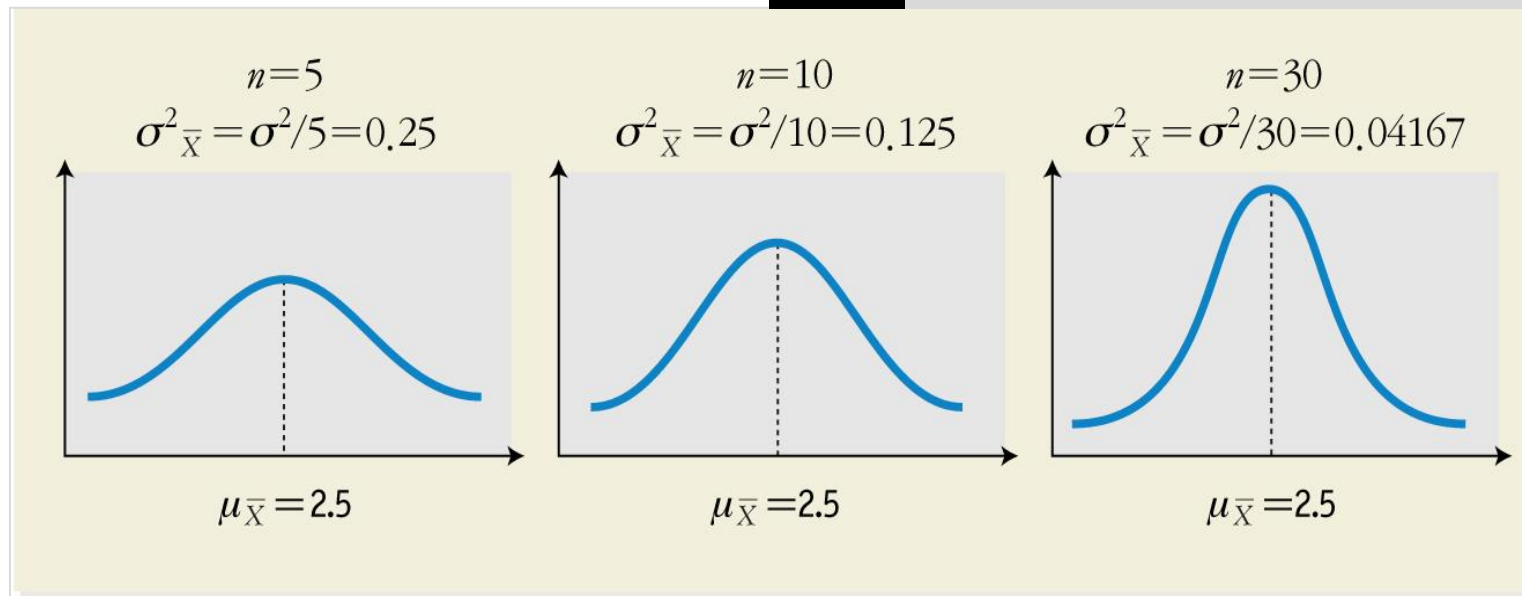
계산결과를 보면 두 분포의 평균은 2.5로 같지만 x 의 분산 σ_X^2 과 $\bar{X}$ 의 분산 $\sigma_{\bar{X}}^2$ 은 서로 같지 않으며, $\sigma_{\bar{X}}^2$ 는 σ_X^2 의 1/2이다.

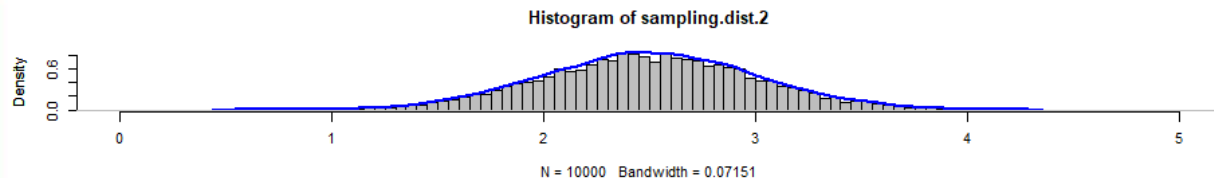
```
> sampling.dist.1<-NULL
> for(sample.count in 1:10000){
+   set.seed(sample.count)
+   sample.mean.1<-mean(rnorm(2,2.5,1.118))
+   sampling.dist.1<-c(sampling.dist.1, sample.mean.1)
+ }
> mean(sampling.dist.1)
[1] 2.502412
> var(sampling.dist.1)
[1] 0.6193981
```

표본평균의 표본분포

그림

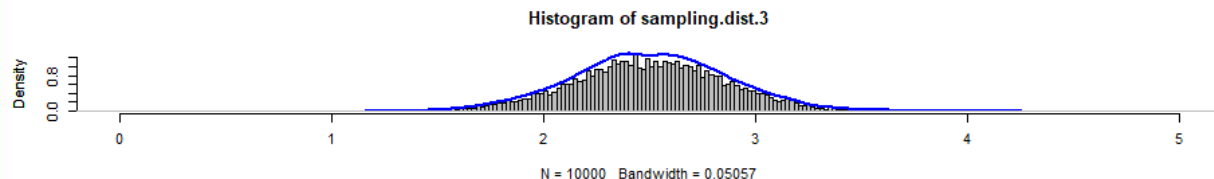
$n=5, 10, 30$ 일 경우 $\bar{X}$ 의 표본분포





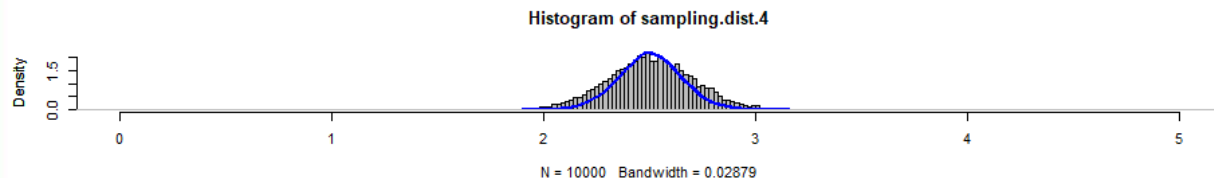
```
> mean(sampling.dist.2)
[1] 2.498498
```

```
> var(sampling.dist.2)
[1] 0.2513623
```



```
> mean(sampling.dist.3)
[1] 2.500606
```

```
> var(sampling.dist.3)
[1] 0.1256997
```



```
> mean(sampling.dist.4)
[1] 2.502484
```

```
> var(sampling.dist.4)
[1] 0.0420068
```

결론 : 표본평균의 평균은 모집단의 평균과 같고, 표본평균의 분산은 모집단의 분산을 n 으로 나눈 것과 같다.



표본평균의 표본분포 특성

평균 μ 와 분산 σ^2 을 갖는 모집단으로부터 크기 n 의 임의표본을 추출하였으며, 이 표본의 평균을 $\bar{X}$ 라고 하자.

$$\mu_{\bar{X}} = \mu$$
$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \quad \text{또는} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

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

