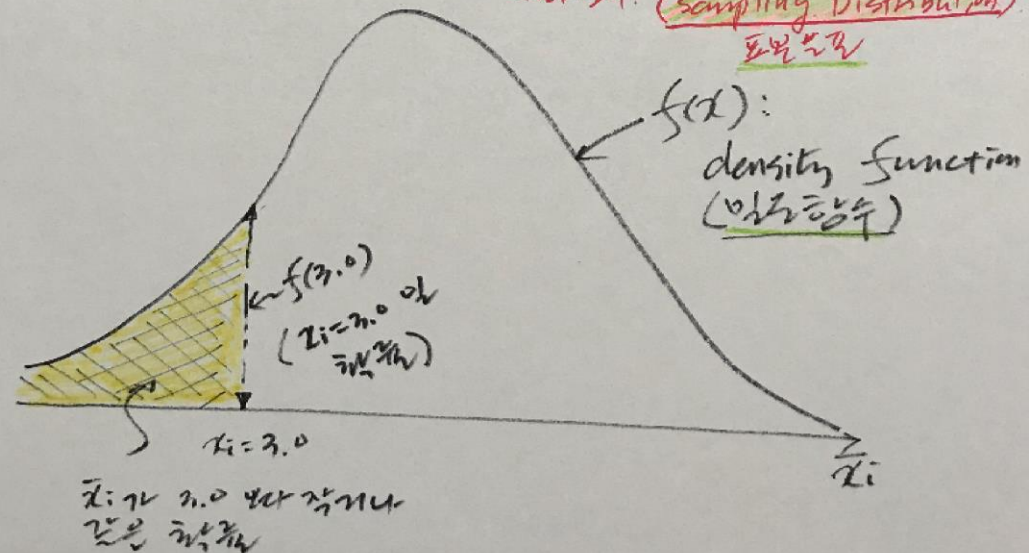


\bar{x} 는 각 표본마다
값이 다르므로 다르다.
(sampling variability)
표본 변동성

여기서 \bar{x} 는 분포(확률분포)를
나타낸다. (sampling Distribution)
표본 분포



조사 및 통계분석

IV. 표본분포/중심극한정리

1. 모수 추정의 오차
2. 표본분포
3. 표본변동의 예
4. 확률밀도함수
5. 표준정규분포
6. 정규분포
7. 누적정규분포
8. 정규분포의 표준화
9. 중심극한정리

모수 추정의 오차

Errors in Parameter Estimation

표본 통계량(a statistic)를 이용하여 모수(the parameter)를 추리하는 과정을 **모수 추정(parameter estimation)**이라 한다.

이 때 모집단을 완벽하게 대표할 수 있는 표본을 선택하지 못함으로써 발생하는 오차를 **표본오차(sampling errors)**라 부르며, 편향(bias)와 우연(chance)에 의해 발생한다.

- 1) **편향(Bias)**: 표본 통계량이 일관성있게 한쪽 방향에 치우치는 것.
- 2) **변동(variability)**: 동일한 규모의 표본을 여러 번 추출한다고 했을 때, 각 표본에서 산출된 표본 통계량들이 동일하지 않고 우연히 각기 다른 값을 갖는다. 이를 **표본 변동(sampling variability)**이라 부른다.

우연에 의한 표본오차는 표본의 크기를 증가시킴으로써 감소시킬 수 있으며, 편향에 의한 오차는 단순 무작위 표본추출방법으로 해결할 수 있다.

표본분포

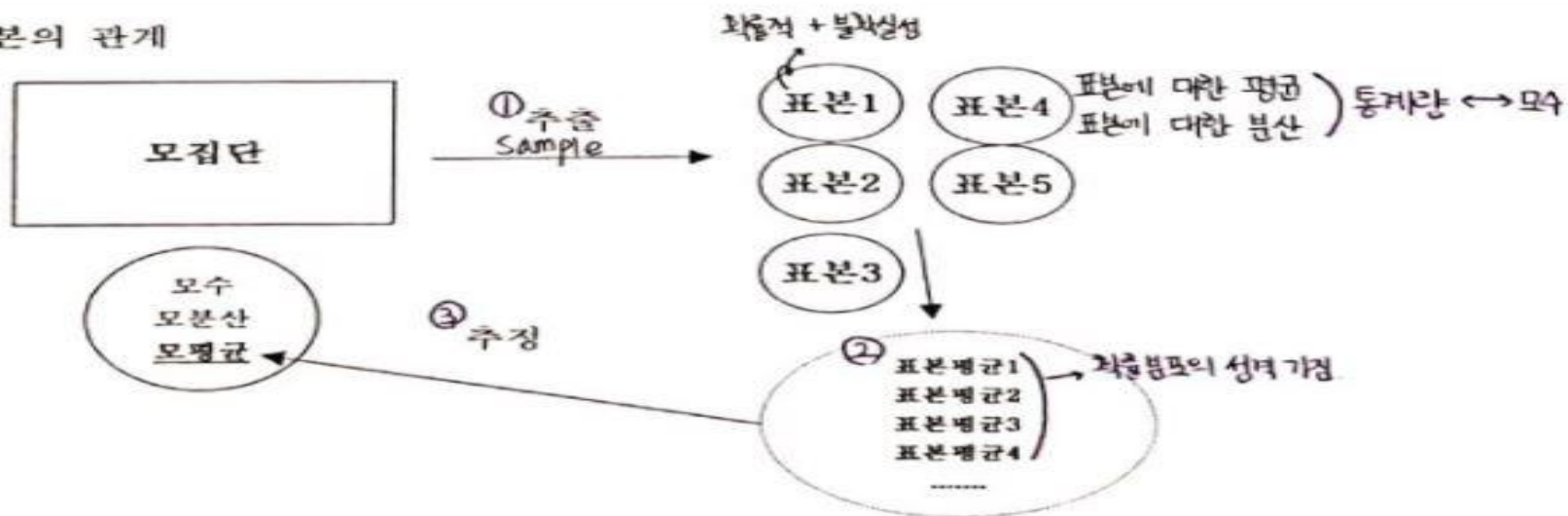
Sample Distribution

표본 통계량은 모수와 반드시 일치한다고 볼 수 없다. 또한 모수를 알수 없으므로(모수를 알면 표본조사가 필요 없음), 표본 통계량과 모수가 어느 정도 차이가 나는지 알 수 없다. 그러나 **표본분포(sample distribution)**를 통해 표본통계량이 어느 정도의 오차를 갖는가는 확실히 알 수 없으나, 오차가 어떠한 형태로 발생할 것인가는 알 수 있다.

즉, 같은 모집단에서 일정한 크기의 표본 몇 조를 취하여 비율, 평균치 등의 통계량을 계산해 보면, 표본에 따라 그 값은 크든 작든 어떤 차가 생기기 마련이다. 이 분포를 표본분포라고 한다. 이 표본분포를 통해 표본통계량의 대표값과 분산값을 알 수 있으며, 이를 이용해 모수의 값을 추론할 수 있다.

표본분포(sample distribution) 표본에서 도출되는 통계량에 대한 확률분포이다. 결국 표본분포는 모수를 추정하기 위해 사용할 수 있는 표본 통계량의 확률분포이다.

-모집단과 표본의 관계



* 모평균을 추정한다.

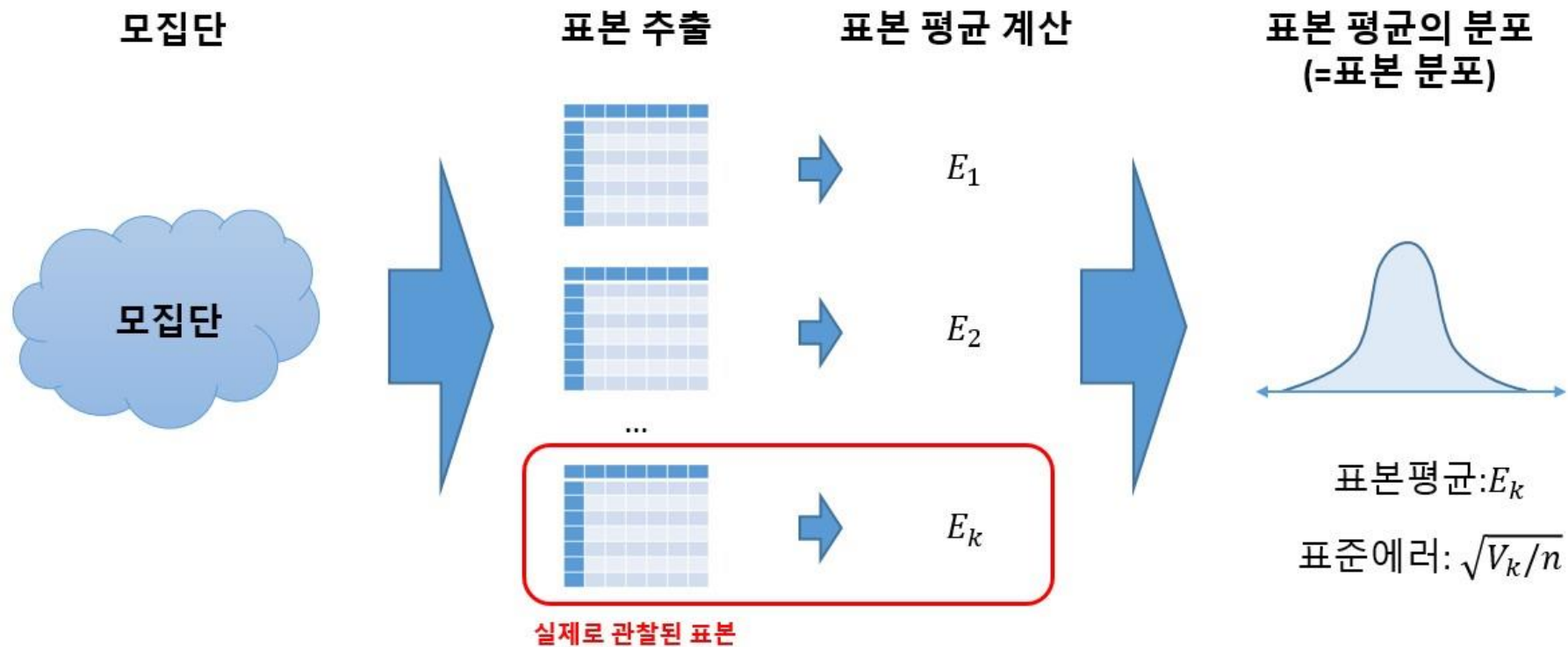
二 표본평균의 확률분포로 모평균을 추정.

통계량 $\left\{ \begin{array}{l} \text{표본평균} \\ \text{표본분산} \end{array} \right\}$ 모수 $\left\{ \begin{array}{l} \text{모평균} \\ \text{모분산} \end{array} \right\}$
 = 확률변수, 확률분포가짐

- 모수(모평균, 모분산)를 추정하기 위해 모집단으로 부터 여러 표본을 추출함
- > 각각의 표본에서 평균과 분산을 구할 수 있음. (통계량)
- > 미지의 모집단에서 추출되었기 때문에, 표본은 확률적 성격을 가지고 있음
- > 그러한 표본에서 구한 평균과 분산은 확률변수이며 확률분포를 가지게 됨.
- > 이 표본분포는 표본에서 계산되는 통계량과 모수 사이의 관계를 규명해주기 때문에 모수에 대한 추정과 검정을 가능하게 함.

표본분포

Sample Distribution



확률분포와 확률밀도함수

probability density function

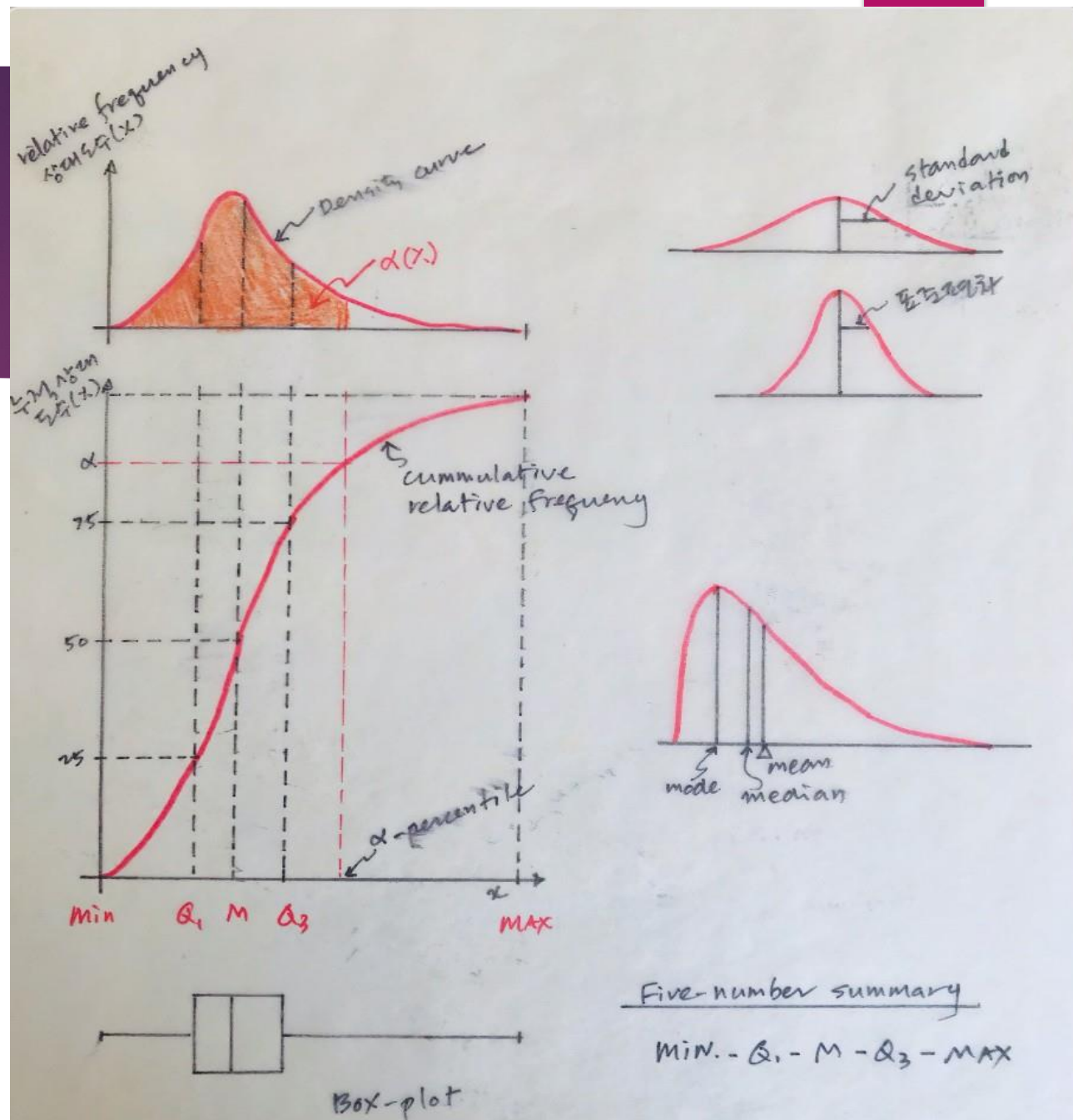
확률분포(probability distribution)는 개개의 변수값에 대한 발생 확률을 나타내는 것이다. 확률분포는 각각의 확률변수(random variable)의 값에 대응하는 발생 확률을 나타내는 함수의 형태로 볼 수 있기 때문에 이를 **확률분포함수(probability distribution function)**라고 부른다.

이 분포는 이론 상으로 연속적으로 정의되지만, 실질적으로는 표본으로부터 실험적으로 얻어진 표본 통계량에 의해 정의되며, 이산화(discreted) 구간 내에서 발생할 빈도(frequency)를 **히스토그램(histogram)**으로 표현하게 된다.

확률밀도함수(probability density function)란 주어진 변량이 정해진 구간 안에 존재할 확률(즉, 상대빈도(relative frequency))을 나타내는 함수이다.

확률밀도함수에서 y -축은 각 확률변수값에 대한 확률을 나타내므로 함수값의 총합은 1이 된다.

확률분포함수 (distribution function)
 확률밀도함수 (density function) 빈
 도 (frequency)
 상대빈도 (relative frequency)
 표준편차 (standard deviation)
 상자그림 (box-plot)
 다섯 숫자 요약 (five numbers summary)



표준정규분포

standard normal distribution

정규분포(normal distribution)는 평균값을 중심으로 좌우대칭인 종 모양을 이루는 확률분포이다. 이는 대표적인 연속확률분포(continuous probability distribution)이며, 가우스분포(Gaussian distribution)라고도 불리운다.

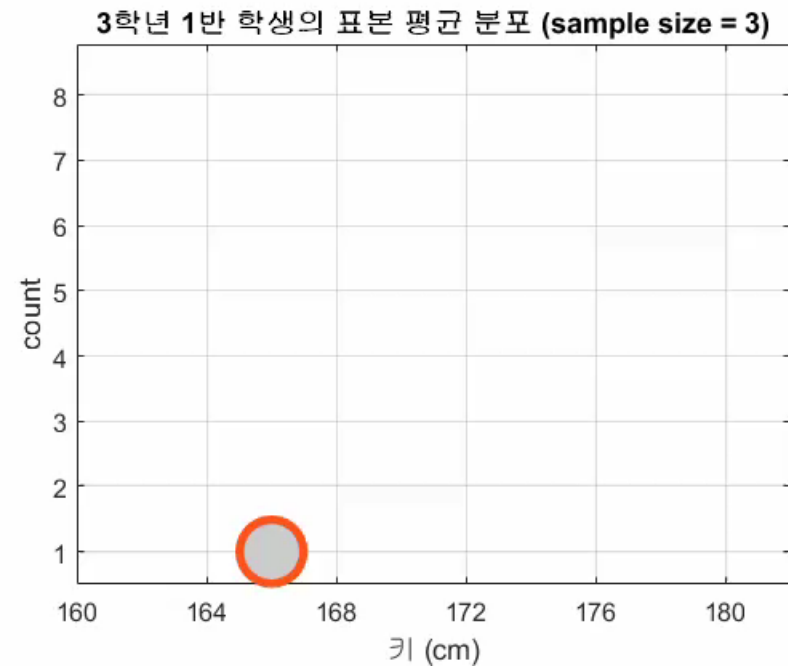
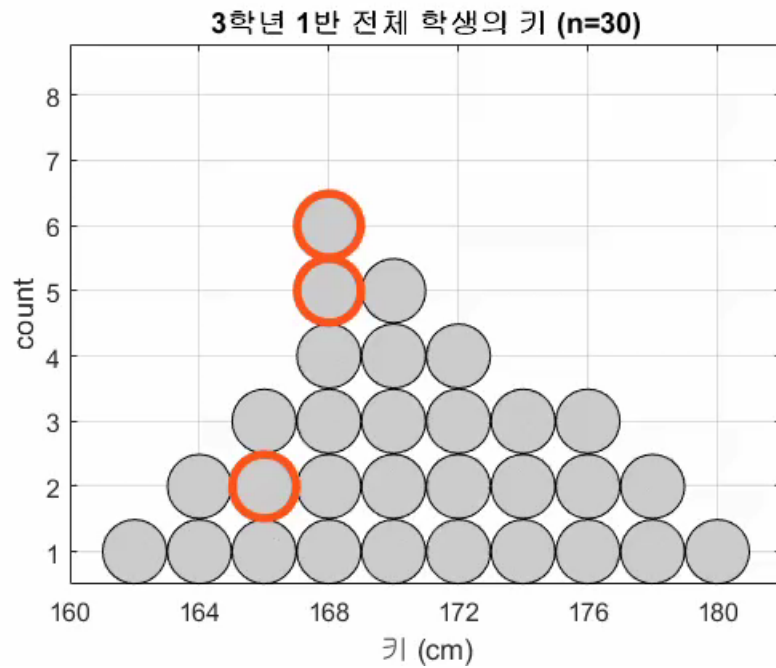
정규분포는 통계학에서 뿐 아니라 자연과학 및 사회과학에서 어떤 확률변수값에 대한 분포를 기술하는데 매우 유용하게 쓰이는데, 이는 중심극한정리에 기인한다.

중심극한정리(central limit theorem)는 임의의 확률분포를 갖는 모집단으로부터 독립적으로 동일한 규모의 표본을 추출하는 작업을 반복했을 때, 각각의 표본 통계량의 확률분포, 즉 표본분포는 정규분포에 접근한다는 것이다.

정규분포는 평균 μ 와 분산 σ^2 값으로 정의된다. 이 두 값을 평균 0, 표준편차 1로 만드는 새로운 확률변수(표준점수)로 변환시킴으로 모든 형태의 정규분포를 표준화할 수 있다. 이를 표준정규분포(standard normal distribution)라고 한다.

표준정규분포

standard normal distribution



표준정규분포

standard normal distribution

노트 참조하기

표준정규분포

standard normal distribution

문제 . 확률변수 X 가 정규분포 $N(30, 4^2)$ 을 따를 때 ,
다음 확률을 구하시오.

1) $P(26 \leq X \leq 32)$

2) $P(32 \leq X \leq 38)$

표준정규분포

standard normal distribution

확률변수 X 가 정규분포 $N(30, 4^2)$ 을 따를 때, 다음 확률을 구하여라.

(1) $P(26 \leq X \leq 32)$

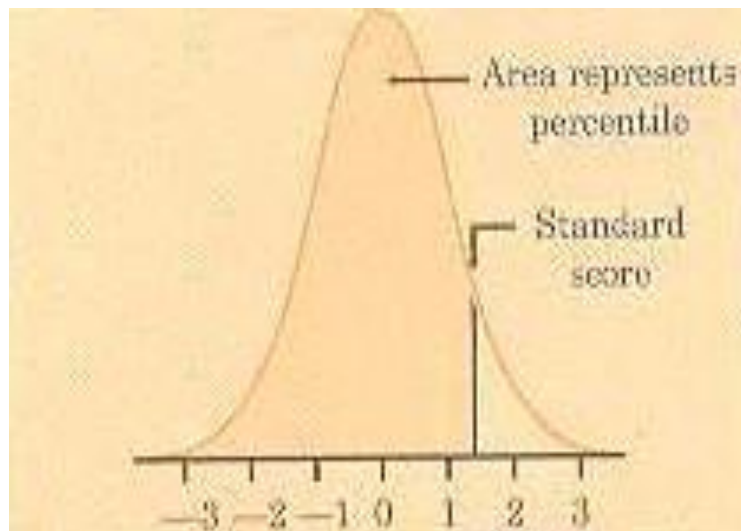
(2) $P(32 \leq X \leq 38)$

(3) $P(X \leq 34)$



$$\begin{aligned} (1) \underline{P(26 \leq X \leq 32)} &= \underline{P(-1.0 \leq Z \leq 0.5)} \\ &= \underline{P(-1.0 \leq Z \leq 0)} + \underline{P(0 \leq Z \leq 0.5)} \\ &= \underline{P(0 \leq Z \leq 1.0)} \quad 0.1915 \\ &\quad 0.3413 \\ &= \begin{array}{r} 0.3413 \\ + 0.1915 \\ \hline 0.5328 \end{array} \end{aligned}$$

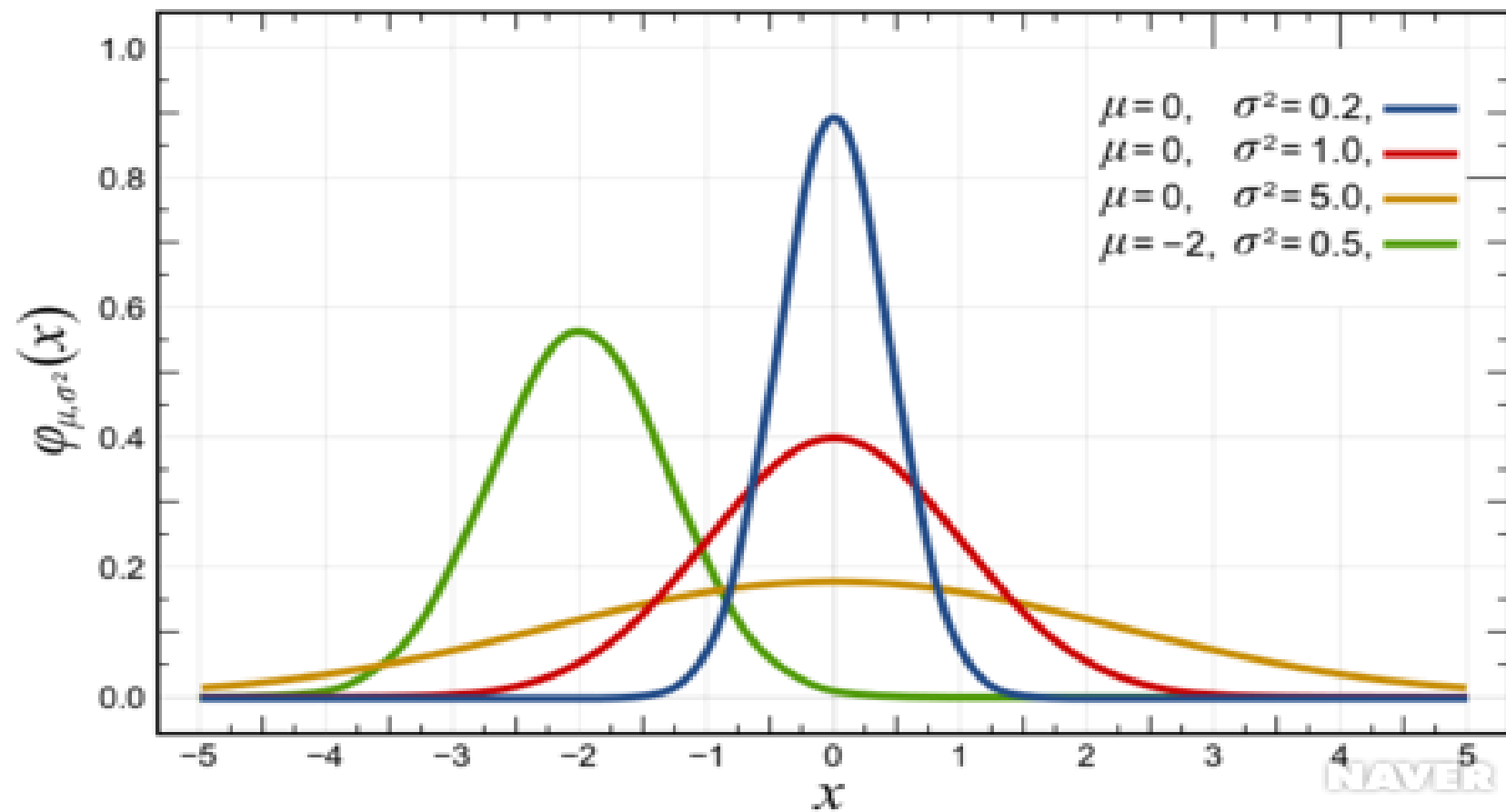
표준정규분포표



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936

정규분포

Normal Distribution



중심극한정리

Central Limit Theorem

중심극한정리(central limit theorem)란? “모집단의 확률분포형태가 어떠하든지 관계없이, 이 모집단에서 무작위로 추출된 표본의 통계량의 분포, 즉, 표본분포는 표본의 크기가 커질수록, 정규분포에 근사한(approximate) 형태를 갖는다”.

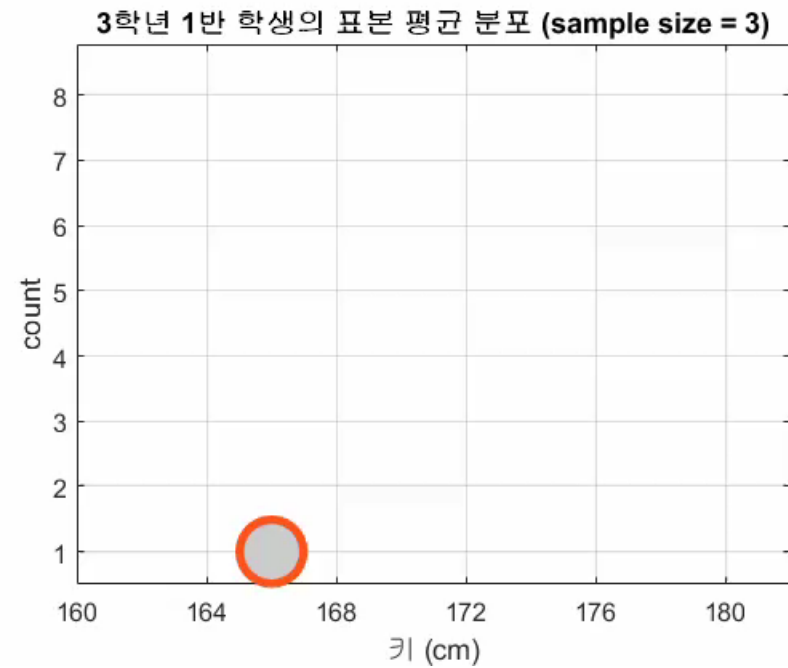
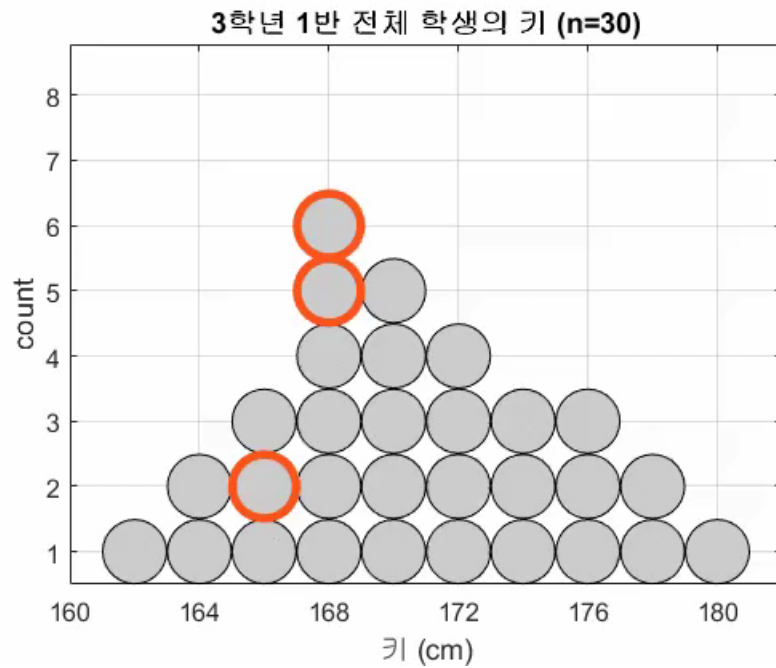
이 정리에 따르면 1)표본의 개수(n)가 충분하다면, 그리고 2)표본이 무작위로 추출된 것이라면, 3) 모집단의 분포가 어떠하든지 관계없이 표본분포는 정규분포로 가정할 수 있다.

따라서 모수를 모르는 상황에서도 표본분포(즉 표본통계량의 확률분포)가 정규분포를 따르므로 이를 근거로 모수를 추론할 수 있다(신뢰구간추정 또는 가설검정 등).

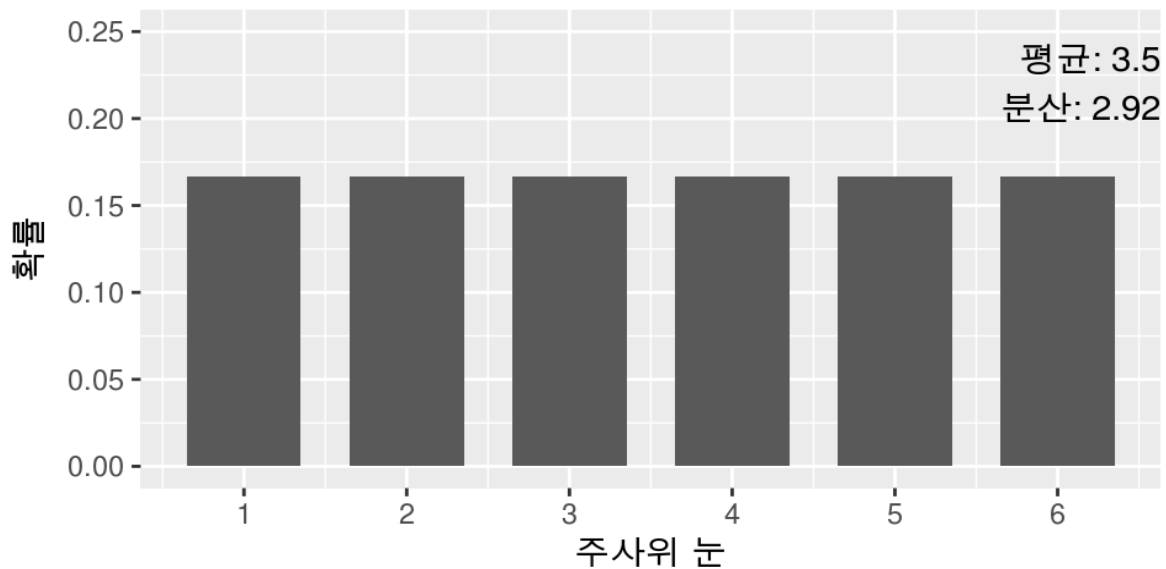
대부분의 통계 모형들은 자료의 분포가 정규분포라는 가정을 필요로 한다. 따라서 모집단의 확률분포를 알지 못하더라도, 중심극한정리에 의해 표본분포가 정규성을 갖는다고 가정할 수 있으므로 표본분포를 다양한 통계 검정에 이용할 수 있다.

표준정규분포

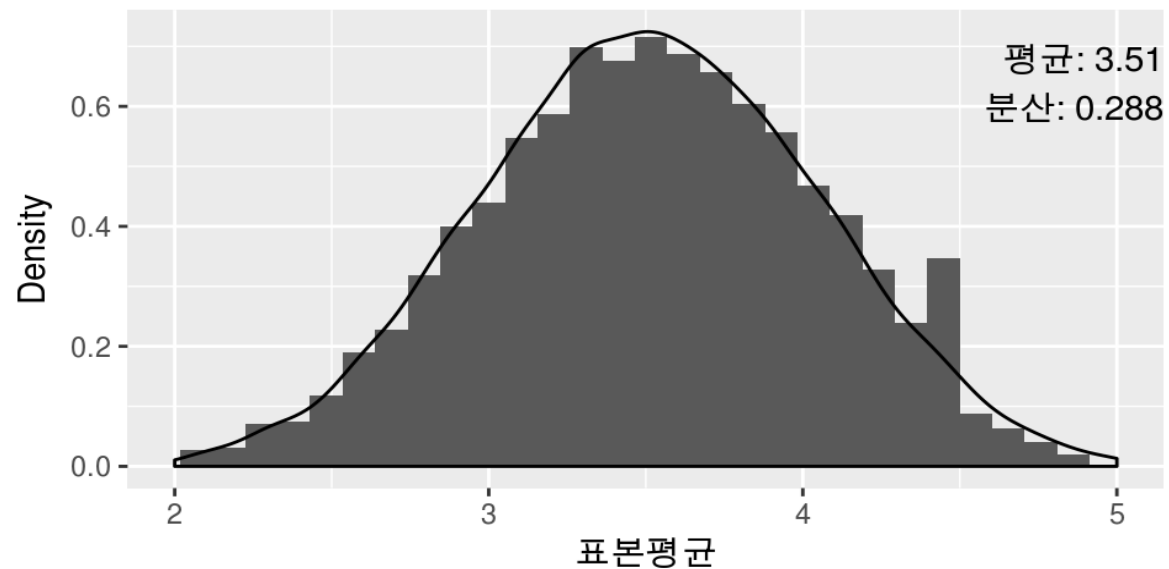
standard normal distribution



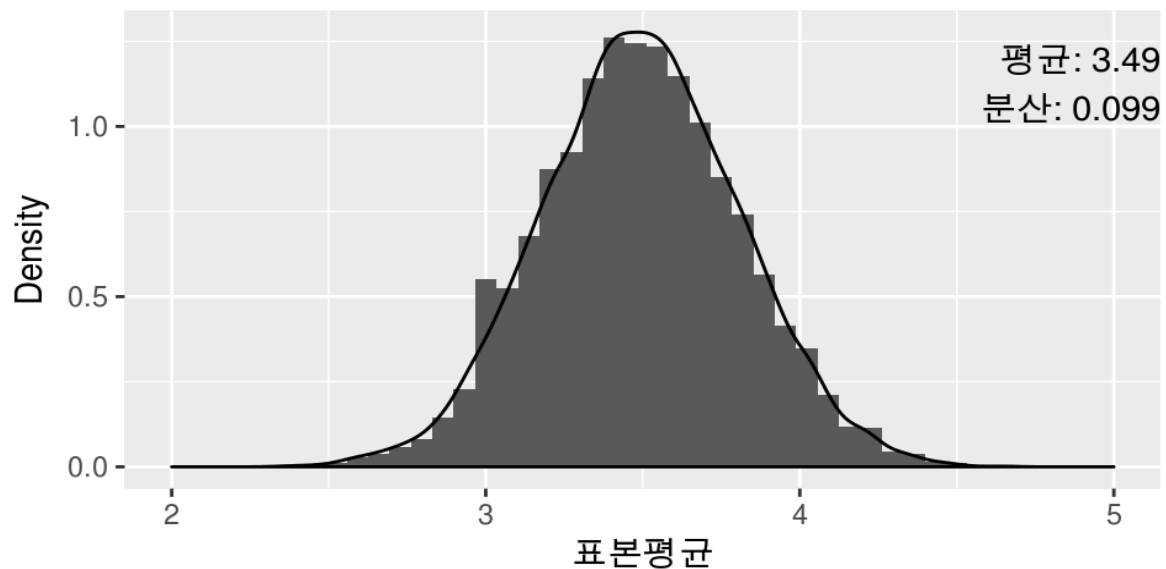
주사위 던지기의 확률분포



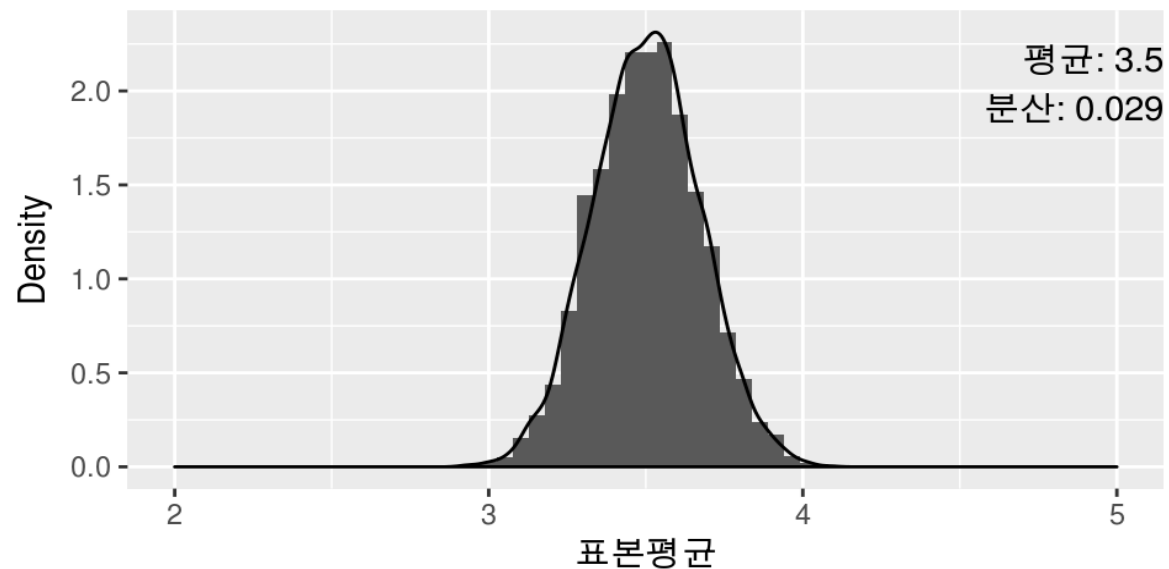
10번 던진 평균들의 분포

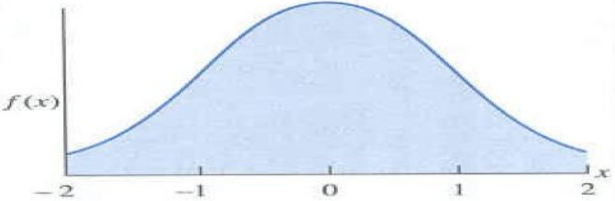
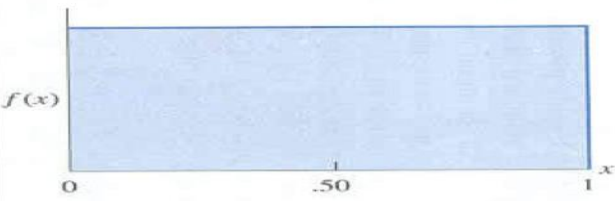
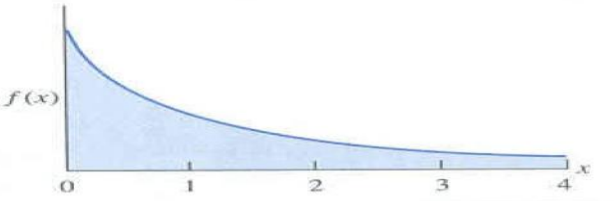
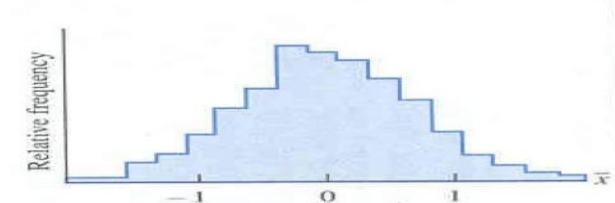
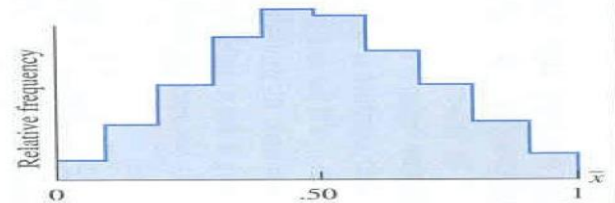
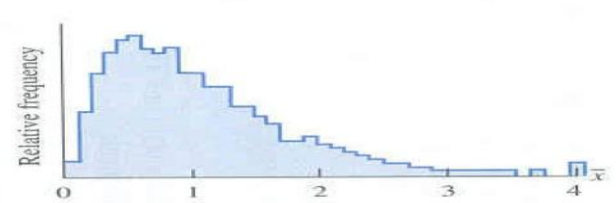
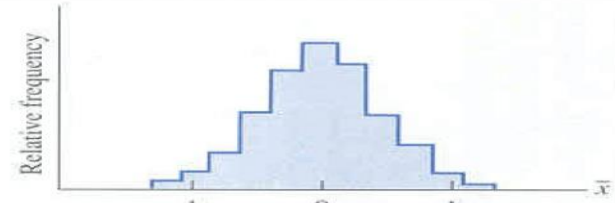
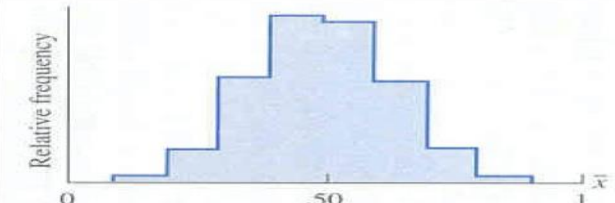
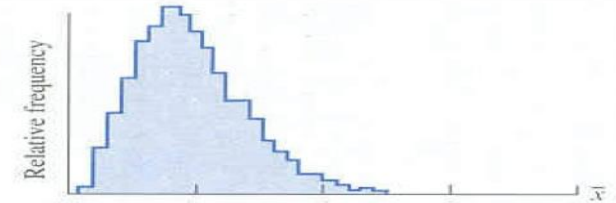
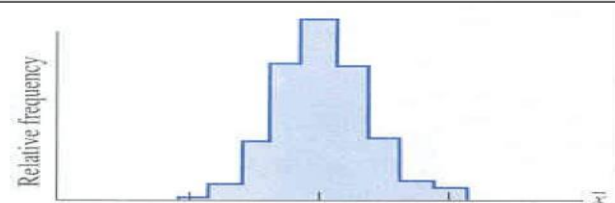
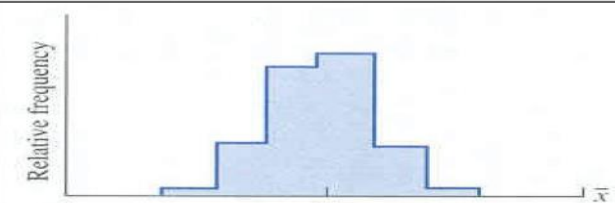
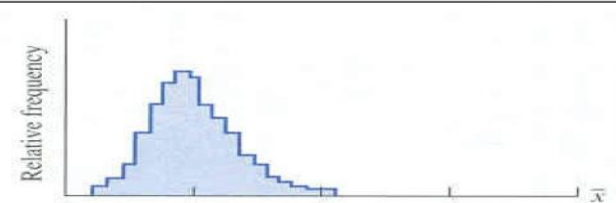
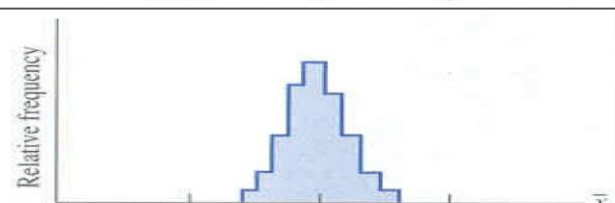
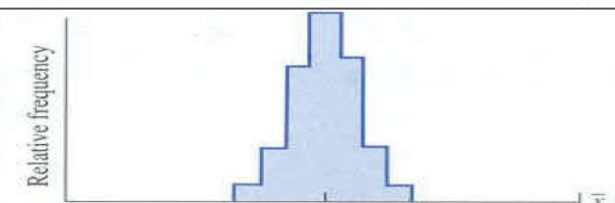
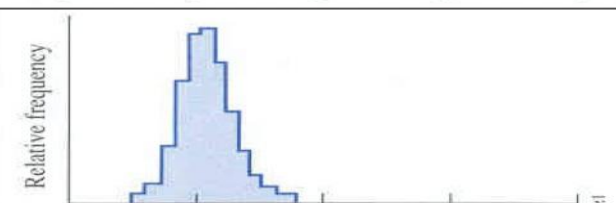


30번 던진 평균들의 분포



100번 던진 평균들의 분포



Simple size n	(a) Normal distribution $\mu = 0, \sigma = 1$	(b) Uniform distribution $\mu = .5, \sigma = .29$	(a) Negative exponential $\mu = 1, \sigma = 1$
Population $n = 1$			
$n = 2$			
$n = 5$			
$n = 10$			
$n = 25$			

확률변수의 표준화

standard score(z-score)

정규분포는 $N(\mu, \sigma^2)$ 로 표현할 수 있다. 여기서 μ 는 평균, σ 는 표준편차를 나타낸다. **표준정규분포(Standard Normal Distribution)**는 모든 정규분포를 하나로 표준화한 것으로 $N(0, 1)$ 로 표현한다. 결국 표준정규분포는 평균을 0으로, 표준편차를 1인 새로운 확률변수로 변환시킨 확률분포인 것이다. 이 새로운 변수를 **표준점수 (Z-score)**라 부른다. 이 값은 $z = (x - \mu) / \sigma$ 로 산출된다.

예제) 어느 해 한국, 미국, 일본의 대졸 신입 사원의 월급은 평균이 각각 80만원, 2000달러, 18만엔이고, 표준편차가 각각 10만원, 300달러, 2만 5천엔인 정규분포를 따른다고 한다. 위 3개국에서 임의로 한 명씩 뽑힌 대졸 신입사원 A, B, C의 월급이 각각 94만원, 2250달러, 21만엔이라 할 때, 누가 상대적으로 월급을 가장 많이 받는가?

한국 신입사원이 받는 월급의 표준점수 = $(94 - 80) / 10 = 1.40$, 따라서 91.9 percentiles에 해당.
미국 신입사원이 받는 월급의 표준점수 = $(2250 - 2000) / 300 = 0.83$, 79.6 percentiles에 해당.
일본 신입사원이 받는 월급의 표준점수 = $(21 - 18) / 2.5 = 1.20$, 89.5 percentiles에 해당.