빅데이터의 이해와 활용

Understanding and Using Big Data

9

추론통계학

학습 내용

01 정규분포

02 표본 분포

03 중심극한정리

04 통계적 추측

학습 목표

- 정규분포의 개념을 이해하고 표준정규분포와 비교하여 설명할 수 있다.
- 다양한 표본 분포의 종류를 파악하고 표본 분포에서 활용되는 용어를 이해하고 설명할 수 있다.
- 중심극한정리의 개념을 이해하고 설명할 수 있다
- 점추정과 구간 추정의 개념과 용어를 이해하고 설명할 수 있다.

생각
해보기

선거철이 되면 언론, 선관위 등에서 정당 지지율, 후보 지지율, 당선율 등을 대한 조사를 합니다. 이럴 때 많이 듣는 용어가 표본, 신뢰도 등인데 어떤 의미일까요?

• 표본의 크기

표본의 크기			
구분		조사완료 사례수(명)	목표할당 사례수(명)
전체		500	500
성별	남	252	247
	여	248	253
연령대별	18~29세	87	89
	30대	73	82
	40대	92	96
	50대	105	100
	60세 이상	143	133
지역별	1권역	173	173
	2권역	327	327

• 여론조사 결과

가중값 산출 및 적용방법		
기본가중	산출방법	지역별, 성별, 연령별 가중치 부여(2020년 02월 행정안전부 주민등록 인구 기준)
	적용방법	셀가중
추가가중	산출방법	
	적용방법	
표본오차		95% 신뢰수준에 ±4.4%

〈출처: 중앙선거여론조사심의위원회(https://bit.ly/2TUFCnj)〉

01

정규분포

1) 정규분포 (Normal Distribution)
2) 표준정규분포 (Standard Normal Distribution)
3) R을 활용한 정규분포 계산
4) R 분포함수 의미

1) 정규분포(Normal Distribution)

● 정규분포 특징

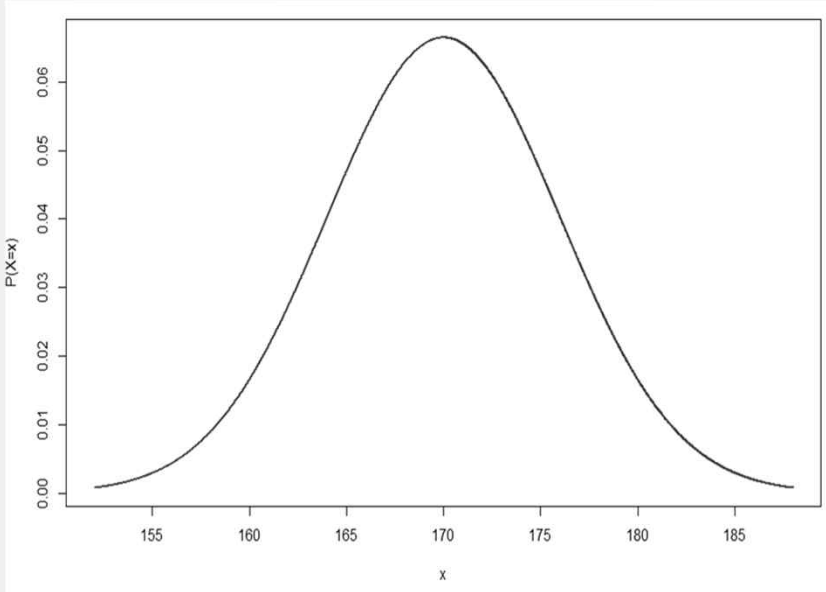
01 종모양의 형태를 가짐

- 양 끝이 아주 느린 속도로 감소하지만, 축에 닿지 않고 -∞와 ∞까지 계속됨

02 평균을 중심으로 좌우대칭임

03 평균 주변에 많이 몰려 있으며 양 끝으로 갈수록 줄어듦

어떤 학생 집단의 키의 평균 170cm, 분산이 6cm, 확률 변수 X가 30일 때의 그래프



1) 정규분포(Normal Distribution)

● 정규분포 특징

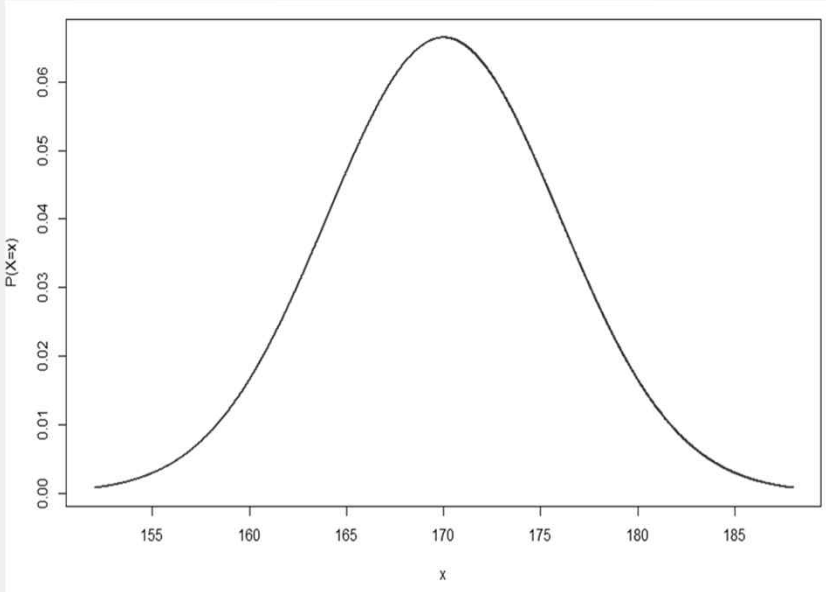
04 평균과 표준편차로 분포의 모양을 결정함

- 정규분포의 모수는 평균 μ와 표준편차 σ(분산 σ^2)로, $N(\mu, \sigma^2)$으로 나타냄

정규분포의 확률밀도함수

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \ -\infty \leq x \leq \infty$$

어떤 학생 집단의 키의 평균 170cm, 분산이 6cm, 확률 변수 X가 30일 때의 그래프



1) 정규분포(Normal Distribution)

성공 확률이 0.5이고 시행 횟수 n이 아주 큰 이항분포가
어떤 함수와 비슷해지는 것을 발견함

- 프랑스 태생의 수학자 드무아브르(1667~1754)

01 이 함수의 모양은 앞선 그림에서 n이 30인 경우와 많이 닮았음

02 좌우가 대칭인 종모양(확률분포의 확률값이 x축에 가까이 다가가나 확률이 0이 되지 않는)의 형태와 유사함

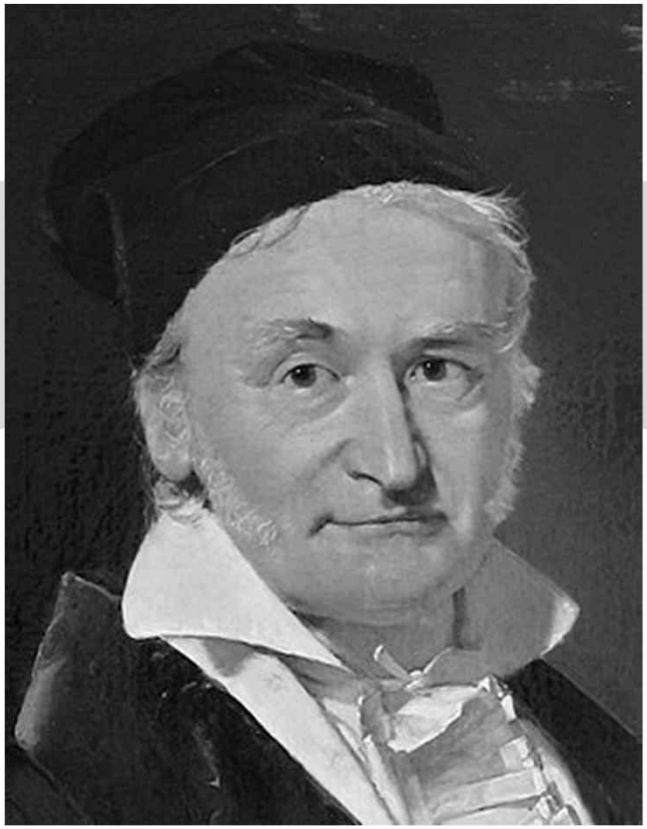
03 n이 충분히 크다면 이산형이 아닌 연속형처럼 다루는 것이 가능함

1) 정규분포(Normal Distribution)

이항분포가 아닌 다른 분포(카이제곱, 감마 등)에서도
n(시도)의 크기가 크면 좌우가 대칭이면서
종모양을 갖는 정규분포와 닮아가는 것을 밝힘

- 라플라스(1749~1827)

1) 정규분포(Normal Distribution)



관측오차가(모집단의 평균 - 관측치)도
정규분포를 따른다는 점을 발견함

- 가우스(1777~1855)

2) 표준정규분포(Standard Normal Distribution)

표준정규분포

평균이 0이고 표준편차가 1인 정규분포($N(0, 1^2)$)

→ 대문자 Z로 표시함

- 모든 정규분포는 표준정규분포로 변환할 수 있음
- 확률변수 X가 평균 μ와 표준편차 σ인 정규분포를 따른다고 할 때

$$Z = \frac{X - \mu}{\sigma}, \qquad Z \sim N(0, 1^2)$$

2) 표준정규분포(Standard Normal Distribution)

평균이 0이고 표준편차가 1인 표준정규분포로 각 값을 계산해 표로 만들고, 다음의 과정을 통해 그 값을 구함

01 임의의 정규분포를 표준정규분포로 변환함

02 구하고자 하는 값을 미리 계산된 표준정규분포의 분포표를 통해 구함

03 구한 값을 원래의 정규분포로 변환함

2) 표준정규분포(Standard Normal Distribution)

Q 어느 대학교 남학생들 키의 평균은 170cm, 표준편차는 6cm입니다.
이 대학교에서 남학생의 키가 182cm 이상일 확률은 얼마인가?
(남학생의 키는 정규분포를 따르고 있는 것으로 가정함)

풀이 : 정규분포 계산법

- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \ -\infty \le x \le \infty$
- $P(X \ge 182) = 1 - P(X \le 182) = 1 - \int_{-\infty}^{182} \frac{1}{6\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-170}{6}\right)^2} dt$

➡ 표준정규분포를 활용하여 쉽게 계산할 수 있음

2) 표준정규분포(Standard Normal Distribution)

Q 어느 대학교 남학생들 키의 평균은 170cm, 표준편차는 6cm입니다.
이 대학교에서 남학생의 키가 182cm 이상일 확률은 얼마인가?
(남학생의 키는 정규분포를 따르고 있는 것으로 가정함)

풀이 : 표준정규분포 계산법

- 표준화 변환을 통한 표준정규분포 계산

① $z = \frac{x-\mu}{\sigma} = \frac{182-170}{6} = \frac{12}{6} = 2$

② $P(Z \geq 2) = 1 - P(Z \leq 2)$

③ 표준정규분포표에서 Z=2인 값은 0.977이므로

∴ 어떤 남학생의 키가 182보다 클 확률은 0.023임

2) 표준정규분포(Standard Normal Distribution)

- 표준정규분포표

z	0.00	0.01	0.02	0.02	0.04	0.05	0.06	0.07	0.08	0.09
0.1	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.2	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.3	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
…										
1.6	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954
1.7	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.962	0.962	0.963
1.8	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971
1.9	0.971	0.972	0.973	0.974	0.974	0.974	0.975	0.976	0.976	0.977
2.0	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981	0.982

3) R을 활용한 정규분포 계산

#0. 평균이 170이 분산 6인 자료, 다른 표현으로는 확률변수 X가 N(170, 6)을 따르는 자료 작성

```
options(digits = 3)
```
▶ # 소수점 3자리

```
mu <- 170
```
▶ # 평균

```
sigma <- 6
```
▶ # 분산

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

3) R을 활용한 정규분포 계산

#0. 평균이 170이 분산 6인 자료, 다른 표현으로는 확률변수 X가 N(170, 6)을 따르는 자료 작성

```
ll <- mu - 3*sigma
```

▶ # 3배의 표준편차 3시그마 범위

```
ul <- mu + 3*sigma
```

```
print(ul)
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

3) R을 활용한 정규분포 계산

#1. 평균이 170이 분산 6인 확률분포 그래프 작성 : dnrom()

```
x <- seq(ll, ul, by = 0.01)
```

▶ # 확률변수 X 생성

```
nd <- dnorm(x, mean = mu, sd = sigma)
```

▶ # 평균 mu 분산이 sigma를 따르는 확률 변수 x가
발생할 확률

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

3) R을 활용한 정규분포 계산

#1. 평균이 170이 분산 6인 확률분포 그래프 작성 : dnrom()

```
plot(x, nd,type = 'l', xlab = 'x', ylab = "P(X=x)",
     lwd = 2, col = 'red')
```

▶ # 확률변수 x가 평균 mu와 분산 sigma의 조건에서 발생할 확률을 표시한 그래프

```
sum(nd)
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

3) R을 활용한 정규분포 계산

#2. 평균이 170이 분산 6인 확률분포 함수값 계산 : pnorm()

```
pnorm(mu, mean = mu, sd = sigma)
```

- # 확률 변수 X가 평균 mu, 분산 sigma를 따를 때 X가 평균보다 작을 확률

```
pnorm(ll, mean = mu, sd = sigma)
```

- # 확률 변수 X가 평균 mu, 분산 sigma를 따를 때 X가 152보다 작을 확률

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

3) R을 활용한 정규분포 계산

#2. 평균이 170이 분산 6인 확률분포 함수값 계산 : pnorm()

```
pnorm(ul, mean = mu, sd = sigma)
```

▶ # 확률 변수 X가 평균 mu, 분산 sigma를 따를 때
X가 188보다 작을 확률

```
pnorm(180, mean = mu, sd = sigma) -
 pnorm(175, mean = mu, sd = sigma)
```

▶ # 확률 변수 X가 평균 mu, 분산 sigma를 따를 때
X가 152보다 크고 188보다 작을 확률

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

3) R을 활용한 정규분포 계산

#3. 평균이 170이 분산 6인 확률분포 분위수 확인 : qnorm()

```
qnorm(0.25, mean = mu, sd = sigma)
```

```
qnorm(0.5, mean = mu, sd = sigma)
```

```
qnorm(0.75, mean = mu, sd = sigma)
```

```
qnorm(c(0.25, 0.5, 0.75), mean = mu, sd = sigma)
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

3) R을 활용한 정규분포 계산

#4. 평균이 170 분산이 6인 모집단에서 표본 추출 : rnorm

```
par(family = "AppleGothic")
```

▶ # 맥북일 경우 한글 깨짐 방지를 위해 필요

```
options(digits = 5)
```

```
set.seed(5)
```

```
smp <- rnorm(400, mean = mu, sd = sigma)
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

3) R을 활용한 정규분포 계산

#4. 평균이 170 분산이 6인 모집단에서 표본 추출 : rnorm

```
hist(smp, probability = T,
     main = "N(170, 6^2)으로 부터 추출한 표본의 분포(n=400)",
     xlab = "", ylab = "", col = "white", border = "black")
```

```
lines(x, nd, lty = 2, col = 'red')
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

4) R 분포함수 의미

역할	접두사(Prefix)	첫 번째 전달 인자	
확률(질량/밀도) 함수	d	x	확률을 구할 값, P(X=x)
분포함수	p	x	확률을 구할 값, P(X≤x)
분위수 함수	q	q	알고 싶은 분위수P(X≤x) = q
난수 생성	r	n	생성할 난수의 개수

02

표본 분포

1) 모수와 통계량(Parameter & Statistics)
2) 표본 분포(Sampling Distribution)
3) R을 활용한 표본 분포 실습
4) 이산형 확률분포(Discrete Distribution)
5) 연속형 확률분포(Continuous Distribution)

1) 모수와 통계량(Parameter & Statistics)

● 모수(Parameter)

모수

모집단의 특성을 나타내는 값

예 대한민국 유권자의 무당층 비율

- 모집단 : 대한민국 유권자 전체
- 모수 : 지지하는 정당이 없는 유권자의 비율

모수는 알지 못하나 존재하는 값으로
우리가 알고자 하는 대상이 됨

1) 모수와 통계량(Parameter & Statistics)

● 통계량(Statistics)

통계량

표본의 특성을 나타내는 값

↳ 수집된 표본에 따라 그 값이 달라짐

통계량의 종류	평균, 표준편차, 중앙값, 비율 등이 있음

예 대한민국 유권자의 A 정당에 대한 지지율 조사
: “앞선 여론조사에서 표본 2,529명으로 부터 무당층은 19.5%로 조사되었습니다.”

2) 표본 분포(Sampling Distribution)

표본조사를 실시하면 조사를 위해 표본을 한 번 추출해서
표본의 특성을 구하고, 모집단에 대해 추측함

모집단의 크기가 N이고 표본의 크기가 n일 때
표본을 비복원으로 추출하는 경우의 수

→ 수는 $\binom{N}{n}$ 가지로 모집단의 크기와 표본의 크기에 따라 결정됨

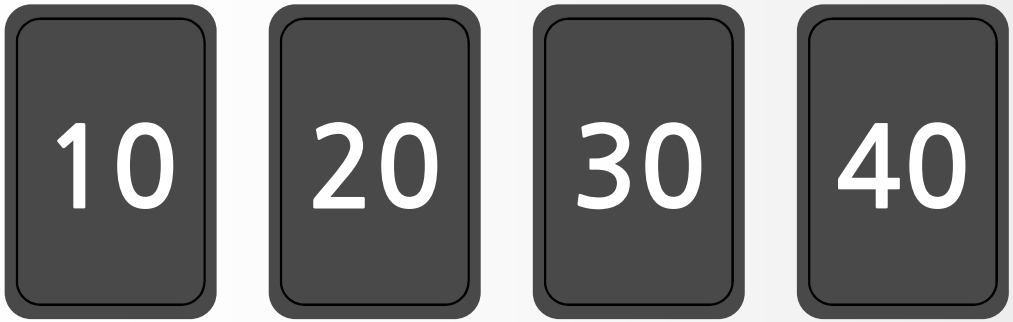
표본 분포

표본의 크기가 n으로 정해졌을 때 추출될 수 있는
모든 표본으로 부터 구한 통계량으로 구성된 확률 분포

2) 표본 분포(Sampling Distribution)

- 사례

다음과 같이 4장의 카드가 있을 때 2장의 카드를 뽑아
4장의 카드의 평균을 맞추는 게임이 있음



- 네 장의 카드에 쓰인 숫자들의 평균(μ)은 25,
  분산(σ^2)은 125임
- 게임에 참가하는 사람은 두 장의 카드를 뽑아 카드에
  쓰여져 있는 숫자들로 평균을 맞추고자 함

2) 표본 분포(Sampling Distribution)

- 사례

01 참가자들이 4장 중 2장의 카드를 비복원추출로 뽑을 수 있는 경우의 수

$$\binom{4}{2} = \frac{4!}{2! \cdot 2!} = 6\text{가지}$$

➡ 이 과정은 모집단이 모르는 숫자 4가지로 구성되어 있고, 이로부터 2개를 표본으로 뽑아 관찰하는 과정임

2) 표본 분포(Sampling Distribution)

- 사례

02 여섯 가지 경우별로 평균(표본 평균)을 구해보면 다음과 같음

구분	경우1		경우2		경우2		경우4		경우5		경우6	
추출된 개별표본	10	20	10	30	10	40	20	30	20	40	30	40
표본 평균 ($\bar{x}$)	15		20		25		25		30		35	

2) 표본 분포(Sampling Distribution)

- 사례

02 여섯 가지 경우별로 평균(표본 평균)을 구해보면 다음과 같음

- 추출된 표본 평균으로부터 모집단의 평균을 추측할 때
  - '경우 1'과 같이 표본 평균($\bar{x}$=15)이 모집단 평균(μ=25)과 차이가 있을 때도 있음
  - '경우 3'과 '경우 4'와 같이 표본 평균이 모집단 평균과 일치할 때도 있음

2) 표본 분포(Sampling Distribution)

- 사례

02 여섯 가지 경우별로 평균(표본 평균)을 구해보면 다음과 같음

- 표본의 크기 n인 표본으로부터 구하는 표본 평균 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$는 추출된 확률표본에 따라 값이 달라짐
  - 추출된 확률표본에 따라 값이 결정되는 표본 평균은 표본 평균 $\bar{x}$의 분포로부터 확률 추출된 확률변수임

2) 표본 분포(Sampling Distribution)

- 사례

표본 평균 $\bar{x}$의 분포

- 4장의 카드에서 표본으로 2장의 카드를 뽑아서 구한 표본 평균 $\bar{x}$의 분포

1. 표본으로 추출될 6가지의 경우 추출될 확률이 1/6으로 동일함
2. 각 표본으로부터 구할 수 있는 표본 평균 $\bar{x}$는 15, 20, 25, 30, 35의 5가지임
3. 표본 평균이 25가 될 확률은 '경우 3' 혹은 '경우 4'가 선택될 경우로 확률은 $\frac{1}{6}+\frac{1}{6}=\frac{2}{6}$
4. ③의 경우가 아닌 다른 표본 평균이 나타날 확률은 $\frac{1}{6}$로 모두 동일함

2) 표본 분포(Sampling Distribution)

- 사례

표본 평균 분포의 확률분포와 그 특성을 다음의 표를 통해 확인할 수 있음

$\bar{X} = \bar{x}$	$p(\bar{X} = \bar{x}) = p(\bar{x})$	① $\bar{x} \cdot p(\bar{x})$	② $\bar{x}^2 \cdot p(\bar{x})$
15	$\frac{1}{6}$	$15 \cdot \frac{1}{6} = \frac{15}{6}$	$15^2 \cdot \frac{1}{6} = \frac{225}{6}$
20	$\frac{1}{6}$	$20 \cdot \frac{1}{6} = \frac{20}{6}$	$20^2 \cdot \frac{1}{6} = \frac{400}{6}$
25	$\frac{2}{6}$	$25 \cdot \frac{2}{6} = \frac{50}{6}$	$25^2 \cdot \frac{2}{6} = \frac{1250}{6}$
30	$\frac{1}{6}$	$30 \cdot \frac{1}{6} = \frac{30}{6}$	$30^2 \cdot \frac{1}{6} = \frac{900}{6}$
35	$\frac{1}{6}$	$35 \cdot \frac{1}{6} = \frac{35}{6}$	$35^2 \cdot \frac{1}{6} = \frac{1225}{6}$
합	1	$E(\bar{X}) = \sum_{\bar{X}} \bar{x} \cdot p(\bar{x}) = \frac{150}{6} = 25$	$E(\bar{X}^2) = \sum_{\bar{X}} \bar{x}^2 \cdot p(\bar{x}) = \frac{4000}{6}$

2) 표본 분포(Sampling Distribution)

- 사례
  - 기댓값 구하기(E(X))

열의 합은 표본 평균 $\bar{x}$ 분포의 기댓값임

그 값은 25로 모집단의 평균과 동일함

2) 표본 분포(Sampling Distribution)

- 사례
  - ▶ 분산 구하는 방법 1

열의 합은 $\overline{x}^2$의 기댓값으로 이 값에서 기댓값의 제곱을 빼면
표본 평균 $\overline{x}$ 분포의 분산을 구할 수 있음

$$Var(X) = E(X^2) - [E(X)]^2 = \sum_{\text{모든}x} x^2 P(X = x) - [E(X)]^2$$

값을 대입하여 계산하면…

$$Var(\overline{X}) = \frac{4000}{6} - 25^2 = \frac{4000}{6} - 625 = \frac{4000 - 3750}{6} = \frac{250}{6} = 41.66$$

2) 표본 분포(Sampling Distribution)

- 사례
  - ▶ 분산 구하는 방법 2

N을 모집단의 수, n을 표본의 수, 모집단의 분산을 σ^2이라 할 때, 아래의 공식을 활용하여 계산함(σ^2(모분산)은 125)

$$\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}$$

$$\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}=\frac{4-2}{4-1}\cdot\frac{125}{2}=\frac{125}{3}=\frac{250}{6}=41.66$$

2) 표본 분포(Sampling Distribution)

- **표본 평균 $\bar{x}$ 분포의 기댓값과 표준편차(분산)의 의미**
  - ▸ **기댓값**

✓ 표본조사에서는 여러 번에 걸쳐 동일한 크기의 표본을 추출하는 것이 아님

↳ **단 한 번 추출한 표본을 통해 모집단의 특성을 유추함**

✓ 추출된 표본으로부터 구한 표본 평균은 표본 평균 $\bar{x}$ 의 분포에서 확률추출한 것으로 간주함

↳ **4장의 카드에서 2장을 뽑는 예제에서 2장의 카드를 확률 추출하여 (1, 2)가 나온 것은 표본 평균 $\bar{x}$ 분포에서 1.5인 값을 확률 추출한 것과 동일한 의미임**

2) 표본 분포(Sampling Distribution)

- **표본 평균 $\bar{x}$ 분포의 기댓값과 표준편차(분산)의 의미**
  - ▶ **기댓값**

표본 평균의 기댓값이 모집단의 평균과 같다는 성질

표본을 추출하기에 앞서 추출된 표본으로부터 구한
표본 평균이 모집단의 평균과 같을 것으로 기대할 수 있음

2) 표본 분포(Sampling Distribution)

- **표본 평균 $\bar{x}$ 분포의 기댓값과 표준편차(분산)의 의미**
  - ▶ **표준편차(분산)**

> 각 표본 평균들이 기댓값(모집단 평균)에 대해
> 얼마나 흩어져 있는지를 나타냄

이 값이 작을 경우 표본을 통해 관찰한 표본 평균이
모집단의 평균과 차이가 날 확률이 작을 것으로 간주함

표준편차($\frac{\sigma}{\sqrt{n}}$)를 반으로 줄이기 위해서는
표본의 수를 4배로 늘려야 함

3) R을 활용한 표본 분포 실습

표본 평균 x의 분포

```
m10 <- rep(NA, 1000)
```
▶ # 표본 크기가 10인 표본을 1000번 추출

```
m40 <- rep(NA, 1000)
```
▶ # 표본 크기가 40인 표본을 1000번 추출

```
set.seed(9)
```
▶ # 동일한 난수 생성을 위해 난수 생성 값 고정

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

3) R을 활용한 표본 분포 실습

표본 평균 x의 분포

```
for(i in 1:1000){
m10[i] <- mean(rnorm(10, mean = 170, sd = 6))
```

▶ # 학생 키의 평균이 170, 표준편차가 6인 학생 집단에서 10명의 학생을 한 표본으로 하여 이 표본의 평균을 1000번 생성

```
m40[i] <- mean(rnorm(40, mean = 170, sd = 6))
```

▶ # 학생 키의 평균이 170, 표준편차가 6인 학생 집단에서 40명의 학생을 한 표본으로 하여 이 표본의 평균을 1000번 생성

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

4) R을 활용한 표본 분포 계산

```
options(digits = 4)
```
▶ # 자릿수 설정, 출력 숫자의 자릿수를 4자리로

```
c(mean(m10), sd(m10))
```
▶ # 10명의 학생들을 한 표본으로 한 표본 평균 분포의 평균과 편차

```
c(mean(m40), sd(m40))
```
▶ # 40명의 학생들을 한 표본으로 한 표본 평균 분포의 평균과 편차

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

4) R을 활용한 표본 분포 계산

```
hist(m10, xlim = c(160, 180), main = "", xlab = "x", ylab = "",
col = 'cyan', border = "blue")
```

▶ # 10명 학생들을 한 표본으로 한 표본 평균들의 분포 그래프

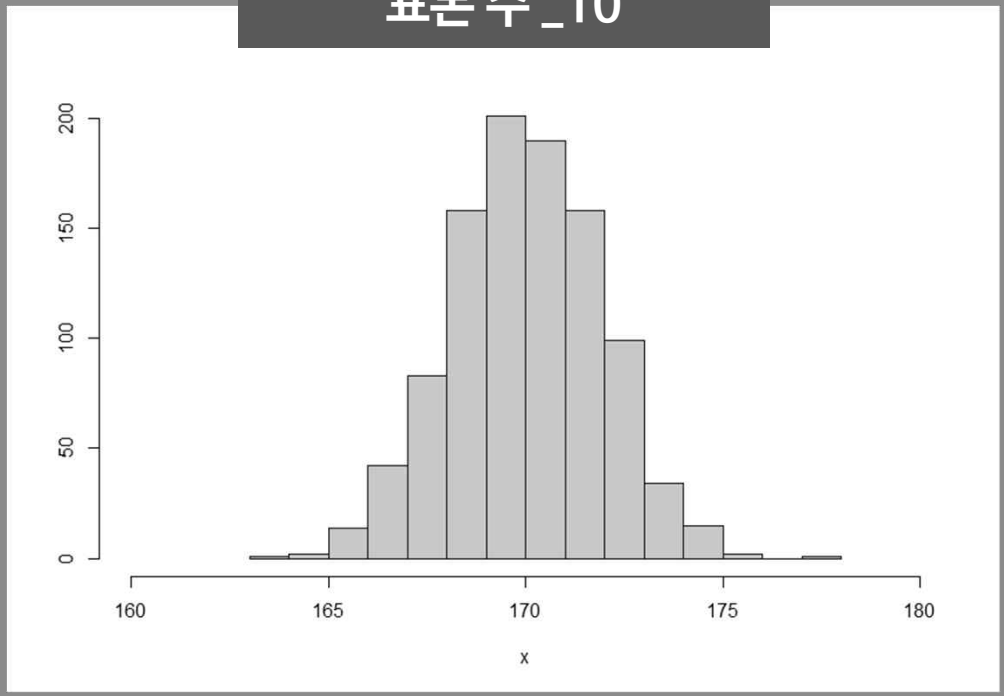
```
hist(m40, xlim = c(160, 180), main = "", xlab = "x", ylab = "",
col = 'cyan', border = "blue")
```

▶ # 40명 학생들을 한 표본으로 한 표본 평균들의 분포 그래프

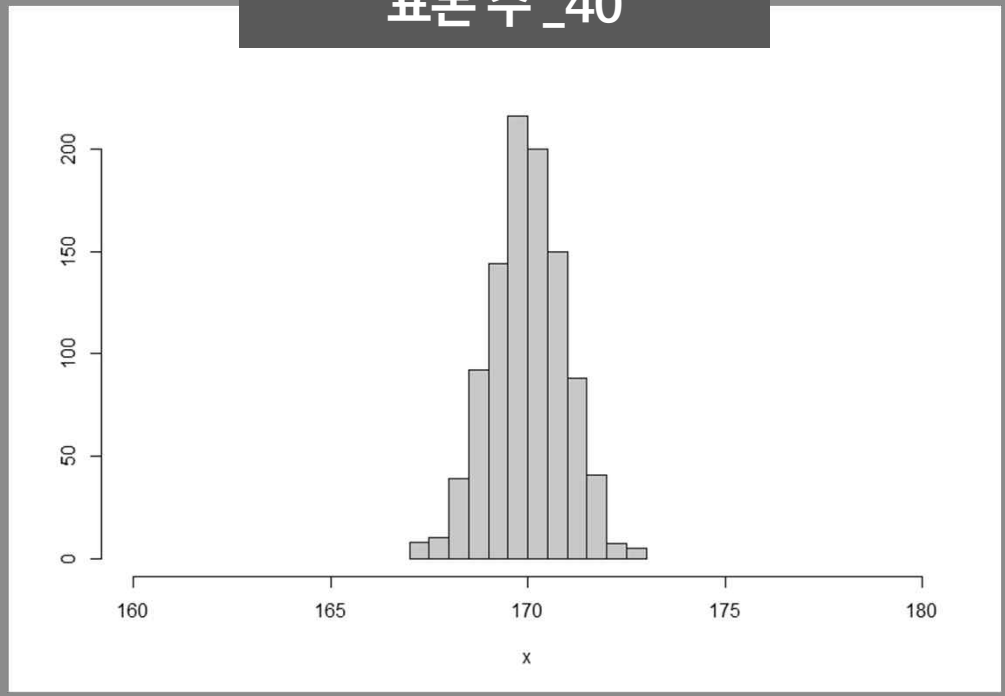
〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

▶ 결과

표본 수 _10



표본 수 _40



표본이 많을수록 모집단 평균에 집중되어 있는 것을 볼 수 있음

5) 이산형 확률분포(Discrete Distribution)

확률분포	설명
균일분포 (Uniform)	• 확률변수 X는 1부터 N까지 일정한 크기인 1/N의 확률을 갖고 있는 분포 • 주사위 실험에서 윗면에 나타날 수를 확률 변수로 간주하는 경우
베르누이분포 (Bernoulli)	성공(x=1), 실패(x=0) 두 결과 중 오직 하나로 나타나는 베르누이 시행의 결과에 대한 확률분포
이항분포 (Binomial)	베르누이 시행을 독립적으로 n번 반복했을 때 나타나는 결과에서 성공의 횟수에 대한 확률분포

5) 이산형 확률분포(Discrete Distribution)

확률분포	설명
음이항분포 (Negative Binomial)	성공확률이 p, 실패 확률이 q = 1-q인 연속적인 시행에서 x번째 성공 전까지 실패횟수
기하분포 (Geometric)	성공의 확률이 p인 베르누이 시행을 독립적으로 반복하여 실시할 때 첫 성공이 나타날 때까지의 실행 횟수에 대한 확률분포
초기하 분포 (Hypergeometric)	N개의 원소로 구성된 모집단에서 비복원 추출 방법에 의해 뽑혀진 표본에서의 확률분포
포아송 분포 (Poisson)	특정시간 또는 구간에 어떤 사건이 적은 수로 발생하는 경우 그 사건의 발생횟수를 측정하는 확률분포

6) 연속형 확률분포(Continuous Distribution)

확률분포	설명
균일분포(Uniform)	특정 구간내의 값들이 나타날 가능성이 균등한 확률분포
정규분포(Normal)	평균근처에서 확률변수 값이 발생할 확률이 높고,평균에서 멀어질 수록 그 확률이 감소하는 종 모양의 분포
지수분포 (Exponential)	어떤 사건이 포아송 분포에 의하여 발생할 때, 지정된 시점으로부터 이 사건이 일어날 때까지 걸린 시간을 측정하는 확률 분포
감마분포 (Gamma)	지수 분포가 사건이 한 번 일어날 때까지의 시간에 관한 분포임에 반하여, 감마분포는 이러한 사건이 x회 일어날 때까지 걸리는 시간의 분포

6) 연속형 확률분포(Continuous Distribution)

확률분포	설명
T분포(T)	정규분포를 따르는 집단의 평균에 대한 가설 검정 또는 정규분포를 따르는 두 집단의 평균 차이 검정에 사용되는 분포
카이제곱 분포 (Chi-Square)	정규분포를 따르는 변수의 분산에 대한 신뢰구간을 구할 때 사용되는 분포
F분포(F)	정규분포를 따르는 두 집단의 분산에 대한 가설 검정을 할 때 사용되는 분포
와이블 분포 (Weibull)	일부의 고장이 부품 전체의 파손, 기능정지 등을 야기하는 사상에 적용되는 확률 분포

03

중심극한정리

1) 중심극한정리(Central Limit Theorem)

1) 중심극한정리(Central Limit Theorem)

표본 평균 $\bar{x}$ 분포는 어떤 분포를 따를까?

- 앞서 표본 평균 $\bar{x}$ 분포의 중요한 특성인 기댓값과 분산(표준편차)에 대해 알아봤음
- 그렇다면 표본 평균 $\bar{x}$ 분포는 어떤 모양이 될지 알아보고자 함

1) 중심극한정리(Central Limit Theorem)

이 과정은 상급과정에서 수리적으로 복잡한 계산을 통해
증명하지만, 우리는 R을 통해 그래프를 그려가면서
어떤 분포와 닮아가는지 확인해 보고자 함

아주 특수한 경우

모집단이 정규분포를
따를 때 이로부터 추출한
표본들의 표본 평균 $\bar{x}$
분포가 어떤 분포를 따를지
살펴봄

일반적인 경우

모집단의 분포가
임의의 분포일 때
어떤 분포를 따를지 살펴봄

1) 중심극한정리(Central Limit Theorem)

- 모집단이 정규분포일 때

모집단이 정규분포일 때 이로부터 추출된 표본들의 표본 평균의 분포는 어떤 모양을 따를지 살펴봄

서로 다른 두 정규분포에서 4개의 표본으로부터 평균을 구하는 것을 1,000번 실시함

→ 표본 평균의 분포가 어떤 형태를 따르는지 확인해 봄

1) 중심극한정리(Central Limit Theorem)

● 모집단이 정규분포일 때

▶ Step 1 : 준비과정

```
1 set.seed(9)
2 n <- 1000
3 r.1.mean <- rep(NA, n)
4 r.2.mean <- rep(NA, n)
```

1줄	난수생성의 초깃값을 9로 고정함
2줄	표본추출 횟수 1,000을 변수 n에 저장함
3, 4줄	모집단별로 표본 평균이 저장될 두 변수 r.1.mean과 r.2.mean을 결측값(NA)으로 초기화함

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

- 모집단이 정규분포일 때
  - ▶ Step 2

두 정규분포 $N(3, 1^2)$과 $N(170, 6^2)$으로부터
표본 크기가 4인 표본을 1,000번 추출하고,
각 추출마다 평균을 저장함

```
5 for (i in 1:n ) {
6 r.1.mean[i] <- mean( rnorm(4, mean=3, sd=1) )
7 r.2.mean[i] <- mean( rnorm(4, mean=170, sd=6) )
8 }
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

● 모집단이 정규분포일 때

▶ Step 2

5, 8줄	• 1:1000으로 생성되는 벡터의 원소 수만큼 반복문을 만듦 • 이 반복문으로 인해 6, 7번째 줄을 1,000번 반복함
6줄	$N(3, 1^2)$ 으로부터 4개의 표본을 추출하고, 그 평균을 r.1.mean의 i번째 원소에 저장함
7줄	$N(170, 6^2)$ 으로부터 4개의 표본을 추출하고, 그 평균을 r.2.mean의 i번째 원소에 저장함

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

- 모집단이 정규분포일 때
  - ▶ Step 2

표본 평균들의 분포에서 평균과 표준편차를 구함

```
10 options(digits=4)
11 c(mean(r.1.mean), sd(r.1.mean))
12 c(mean(r.2.mean), sd(r.2.mean))
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

- 모집단이 정규분포일 때
  - Step 2

10줄	출력물의 자릿수를 4로 함
11, 12줄	각 정규분포로부터 추출된 표본 크기가 4인 표본 평균 분포의 평균과 표준편차를 출력함

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

● 모집단이 정규분포일 때

▶ Step 3

```
> c(mean(r.1.mean), sd(r.1.mean))
[1] 3.0214 0.5096
> c(mean(r.2.mean), sd(r.2.mean))
[1] 170.032   2.835
```

표준정규분포로부터 추출한 표본 평균의 분포

- 평균이 모집단 평균에 가까움
- 표준편차는 모집단 정규분포의 표준편차를 표본 크기의 제곱근으로 나눈 값($\frac{\sigma}{\sqrt{4}}$, 모집단 표준편차의 반)과 비슷함

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

- 모집단이 정규분포일 때
  - ▶ Step 4

01 표본 평균의 분포에 대한 히스토그램을 그림

02 그 위에 각 표본 평균의 분포가 따를 것으로 생각되는 분포의 확률도표를 그림

1) 중심극한정리(Central Limit Theorem)

- 모집단이 정규분포일 때
  - Step 4

모집단이 정규분포일 때 이로부터 추출한 표본 평균의 분포는 또 다른 정규분포를 따르는 것으로 알려져 있음

정규분포로부터 추출된 경우 알려진 표본 평균의 분포

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

1) 중심극한정리(Central Limit Theorem)

- 모집단이 정규분포일 때
  - Step 4

```
14 hist(r.1.mean, prob=TRUE, xlab="표본 평균",
   ylab="밀도", main="", col="orange", border="red")

15 x1 <- seq(min(r.1.mean), max(r.1.mean),
   length=1000)

16 y1 <- dnorm(x=x1, mean=3, sd=(1/sqrt(4)))

17 lines(x1, y1, lty=2, lwd=2, col="blue")
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

- 모집단이 정규분포일 때
  - ▶ Step 4

```
18 hist(r.2.mean, prob=TRUE, xlab="표본 평균",
   ylab="밀도", main="", col="orange", border="red")

19 x2 <- seq(min(r.2.mean), max(r.2.mean),
   length=1000)

20 y2 <- dnorm( x=x2, mean=170, sd=(6/sqrt(4)) )

21 lines(x2, y2, lty=2, lwd=2, col="blue")
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

● 모집단이 정규분포일 때

▶ Step 4

14~17줄

19~22줄

- $N(3, 1^2)$으로부터 표본 크기를 4로 하는 표본 평균의 분포에서 평균은 모집단의 평균인 3이고, 표준편차는 $\frac{1}{\sqrt{4}}$ 임
- 표본 평균의 히스토그램과 평균이 3이고 표준편차가 $\frac{1}{\sqrt{4}}$ 인 정규분포와 비교해 보고자 함

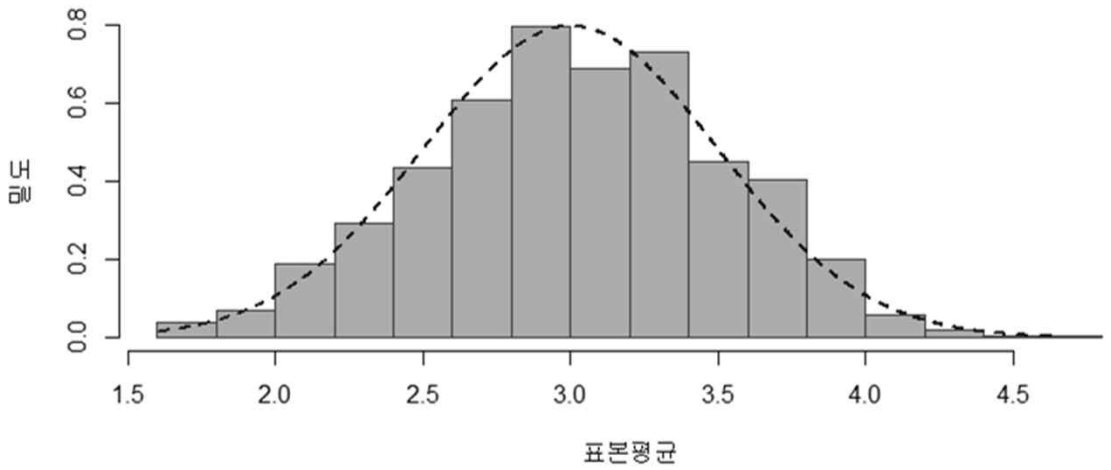
1) 중심극한정리(Central Limit Theorem)

- 모집단이 정규분포일 때
  - Step 4

14~17줄

19~22줄

- hist() 함수에서 prob으로 TRUE을 전달하면, 빈도에 대한 히스토그램이 아닌 상대빈도에 대한 히스토그램을 작성함



1) 중심극한정리(Central Limit Theorem)

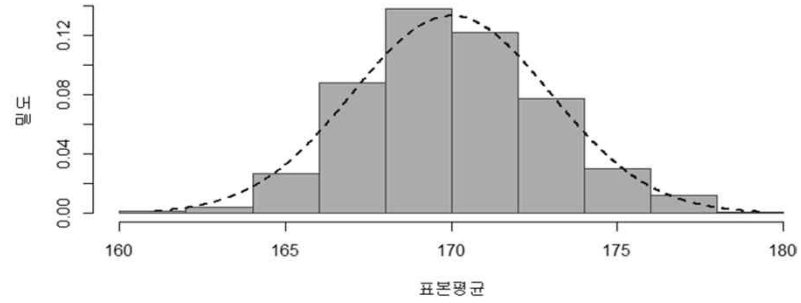
● 모집단이 정규분포일 때

▶ Step 4

14~17줄

19~22줄

- $N(170, 6^2)$으로부터 표본 크기를 4로 하는 표본 평균의 분포에서 평균은 모집단의 평균인 170이고, 표준편차는 $\frac{6}{\sqrt{4}}$ 임
- 표본 평균의 히스토그램과 평균이 170이고 표준편차가 $\frac{6}{\sqrt{4}}$ 인 정규분포와 비교해 봄



1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때

모집단이 임의의 분포일 때 이로부터 추출된 표본들의
표본 평균의 분포는 어떤 모양을 따를지 살펴보고자 함

표본의 크기가 각각 2, 4, 10이고 시행횟수가 10,
성공 확률이 0.1인 이항분포를 가정함

→ 각각이 표본 평균의 분포가 어떤 형태가 되는지
알아보고자 함

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - ▶ Step 1 : 자료준비

```
7  set.seed(9)
8  t <- 10
9  p <- 0.1
10 x <- 0:10
11 n <- 1000
12 b.2.mean <- rep(NA, n)
13 b.4.mean <- rep(NA, n)
14 b.32.mean <- rep(NA, n)
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - ▶ Step 1 : 자료준비

7~10줄	난수생성의 초기값을 9로, 시행 횟수 10을 변수 t에, 성공 확률 0.1을 변수 p에 저장하고 시행횟수가 10인 이항분포로부터 관찰 가능한 값을 변수 x에 저장함
11줄	표본을 추출할 횟수 1,000을 변수 n에 저장함
12~14줄	표본 크기에 따라 1,000번의 표본추출에서 관찰된 표본 평균이 저장될 변수 b.2.mean, b.4.mean과 b.32.mean에 대해 각각 1,000개의 NA 값을 갖는 벡터로 준비함

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - Step 2

표본 크기별로 1000번의 표본추출로 표본 평균을 구함

```
10 for(i in 1:n) {
11 b.2.mean[i] <- mean( rbinom(2, size=t, prob=p) )
12 b.4.mean[i] <- mean( rbinom(4, size=t, prob=p) )
13 b.32.mean[i] <- mean( rbinom(32, size=t,
   prob=p) )
14 }
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - ▶ Step 2

16, 20줄	17~19줄을 1000번 반복(1000번의 표본추출)하는 반복문임
17줄	$B(10, 0.1)$로부터 2개의 표본을 추출하고, 그 평균을 b.2.mean의 i번째 원소에 저장함
18줄	$B(10, 0.1)$로부터 4개의 표본을 추출하고, 그 평균을 b.4.mean의 i번째 원소에 저장함
19줄	$B(10, 0.1)$로부터 32개의 표본을 추출하고, 그 평균을 b.32.mean의 i번째 원소에 저장함

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - Step 3

표본 평균들의 분포에서 평균과 표준편차를 구함

```
22 options(digits=4)
23 c(mean(b.2.mean), sd(b.2.mean))
24 c(mean(b.4.mean), sd(b.4.mean))
25 c(mean(b.32.mean), sd(b.32.mean))
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - Step 3

22줄	• 출력물의 자릿수를 4로 함
23~25줄	

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - Step 3

22줄

23~25줄

- $B(10, 0.1)$로부터 1,000번 추출된 표본 크기가 2, 4, 32인 표본 평균 분포의 평균과 표준편차를 출력함
- 출력물에서 이항분포로부터 추출한 표본 평균의 분포는 그 평균이 이항분포의 평균과 비교해 봄

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - Step 3

22줄

23~25줄

- 표준편차는 모집단 이항분포의 표준편차를 표본 크기의 제곱근으로 나눈 값들과 비교해 봄

$$\frac{0.9487}{\sqrt{2}} \approx 0.6708\,, \frac{0.9487}{\sqrt{4}} \approx 0.4743\,, \frac{0.9487}{\sqrt{32}} \approx 0.1677$$

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - Step 3

```
> c(mean(b.2.mean), sd(b.2.mean))
[1] 1.0090 0.6763
> c(mean(b.4.mean), sd(b.4.mean))
[1] 1.006 0.481
> c(mean(b.32.mean), sd(b.32.mean))
[1] 0.9989 0.1624
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - ▶ Step 4

01 각 표본 평균 분포의 히스토그램을 그림

02 그 위에 각 표본 평균의 분포가 따를 것으로 알려진 정규분포의 확률도표를 작성함

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - Step 4

앞서 모집단이 정규분포일 경우와 마찬가지로
정규분포를 따를 것으로 생각해 봄

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

n=2일 때

$$N\left(1, \left(\frac{0.9473}{\sqrt{2}} \approx 0.6708\right)^2\right)$$

n=4일 때

$$N\left(1, \left(\frac{0.9473}{\sqrt{4}} \approx 0.4743\right)^2\right)$$

n=32일 때

$$N\left(1, \left(\frac{0.9473}{\sqrt{32}} \approx 0.1677\right)^2\right)$$

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - Step 4

```
27 hist(b.2.mean, prob=T, xlim=c(0, 4), main="표본
   크기 : 2", ylab="", xlab="", col="orange",
   border="red")

28 x1 <- seq(min(b.2.mean), max(b.2.mean),
   length=1000)

29 y1 <- dnorm( x=x1, mean=1, sd=sqrt(0.9)/sqrt(2) )

30 lines(x1, y1, lty=2, lwd=2, col="blue")
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - Step 4

```
31 hist(b.4.mean, prob=T, xlim=c(0, 4), ylim=c(0, 1.2),
   main="표본 크기 : 4", ylab="", xlab="",
   col="orange", border="red")

32 x2 <- seq(min(b.4.mean), max(b.4.mean),
   length=1000)

33 y2 <- dnorm( x=x2, mean=1, sd=sqrt(0.9)/sqrt(8) )

34 lines(x2, y2, lty=2, lwd=2, col="blue")
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - Step 4

```
35 hist(b.32.mean, prob=T, xlim=c(0, 4), main="표본
   크기 : 32", ylab="", xlab="", col="orange",
   border="red")

36 x3 <- seq(min(b.32.mean), max(b.32.mean),
   length=1000)

37 y3 <- dnorm( x=x3, mean=1,
   sd=sqrt(0.9)/sqrt(32) )

38 lines(x3, y3, lty=2, lwd=2, col="blue")
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture3.R〉

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - ▶ Step 4

27~30줄

32~35줄

37~40줄

- $B(10, 0.1)$로부터 표본 크기를 2로 하는 표본 평균의 분포의 평균은 모집단의 평균인 1이고,
  표준편차는 $\frac{0.9473}{\sqrt{2}} \approx 0.6708$ 을 가짐

➡ 평균이 1이고 표준편차가 약 0.6708인 정규분포와 비교해 봄

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - Step 4

27~30줄

32~35줄

37~40줄

- $B(10, 0.1)$로부터 표본 크기를 4로 하는 표본 평균의 분포의 평균은 모집단의 평균인 1이고,
  표준편차는 $\frac{0.9473}{\sqrt{4}} \approx 0.4743$ 을 가짐

➡ 평균이 1이고 표준편차가 약 0.4743인 정규분포와 비교해 봄

1) 중심극한정리(Central Limit Theorem)

- 모집단이 임의의 분포일 때
  - ▶ Step 4

27~30줄

32~35줄

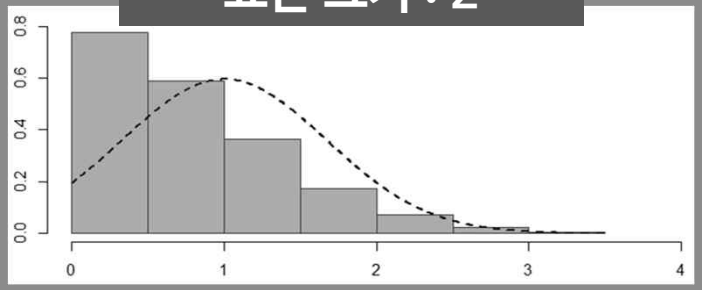
37~40줄

- $B(10, 0.1)$로부터 표본 크기를 32로 하는 표본 평균의 분포의 평균은 모집단의 평균인 1이고,
  표준편차는 $\frac{0.9473}{\sqrt{32}} \approx 0.1677$ 을 가짐

➡ 평균이 1이고 표준편차가 약 0.1677인 정규분포와 비교해 봄

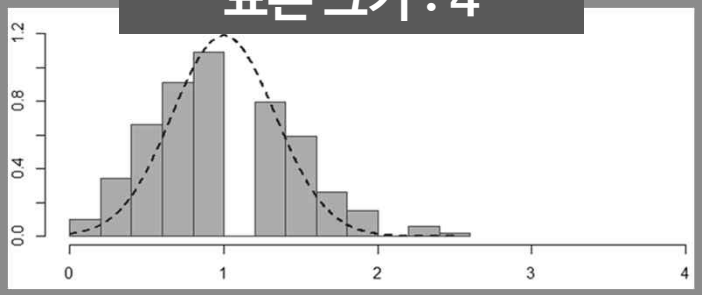
▶ 결과

표본 크기 : 2



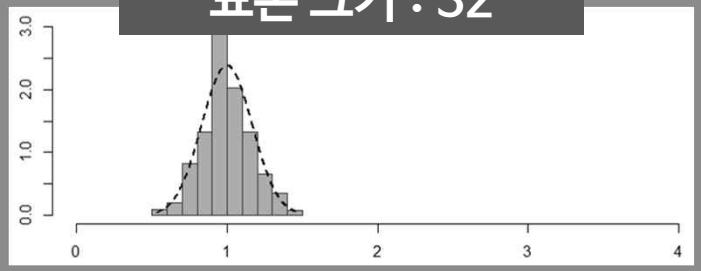
- 여전히 표본 평균의 분포는 오른쪽으로 늘어져 있음

표본 크기 : 4



- n=2일 때 보다는 비교적 좌우대칭으로 보이지만, 중간중간 빈 구간이 보임

표본 크기 : 32



- n이 증가할 수록 좌우대칭을 보이고, 점점 정규분포와 닮아감

1) 중심극한정리(Central Limit Theorem)

- 중심극한정리

표본의 개수가 증가할수록 표본 평균의 분포가
정규분포와 닮아감을 확인해 보았음

이와 같은 성질을 수리적으로 밝혀낸 것이
중심극한정리임

- 모집단의 분포와 상관없이 평균과 표준편차가 μ와 σ로 존재하는 모집단에서 추출할 때 표본의 크기 n이 충분히 크면, 표본 평균의 분포가 근사적으로 정규분포를 따름

중심극한정리는 모집단의 분포에 대한 사전 지식 없이도
표본 평균의 분포를 알 수 있게 하여 통계학에서 유용함

1) 중심극한정리(Central Limit Theorem)

● 중심극한정리

모집단의 분포와 상관없이 모집단의 평균 μ 와 표준편차 σ가 존재할 때
표본 크기 n이 충분히 클 경우

- 표본 평균의 분포는 다음과 같이 근사적으로 정규분포를 따름

$$\bar{X} \cong N(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2)$$

표본 평균의 분포가 정규분포를 따르므로 다음 같이 표준화하여 사용할 수 있음

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

04

통계적 추측

1) 추론통계학
2) 점 추정(Point Estimation)
3) 구간 추정(Interval Estimation)

1) 추론통계학

표본 분포

모집단의 특성을 나타내는 모수(모집단 평균, 모집단 분산)의 값을 알고 있는 상태에서 표본으로 부터 추출된 통계량(표본 평균, 표본 분산, 표본 비율)의 분포

→ 통계량의 값과 모수의 값의 차이, 즉 표본 오차에 대한 확률적 분석을 가능하게 해주는 것으로 통계적 추측의 기초

1) 추론통계학

추론통계학

표본을 추출하여 구한 표본 평균이나 표본 비율의 값을 근거로 모집단 평균이나 모집단 비율의 값이 얼마가 될 것이라고 추측하는 것

→ **모집단의 특성을 알고 표본의 특성에 대한 분석을 하는 기술통계학과는 그 과정이 반대임**

1) 추론통계학

● 방법론

추정

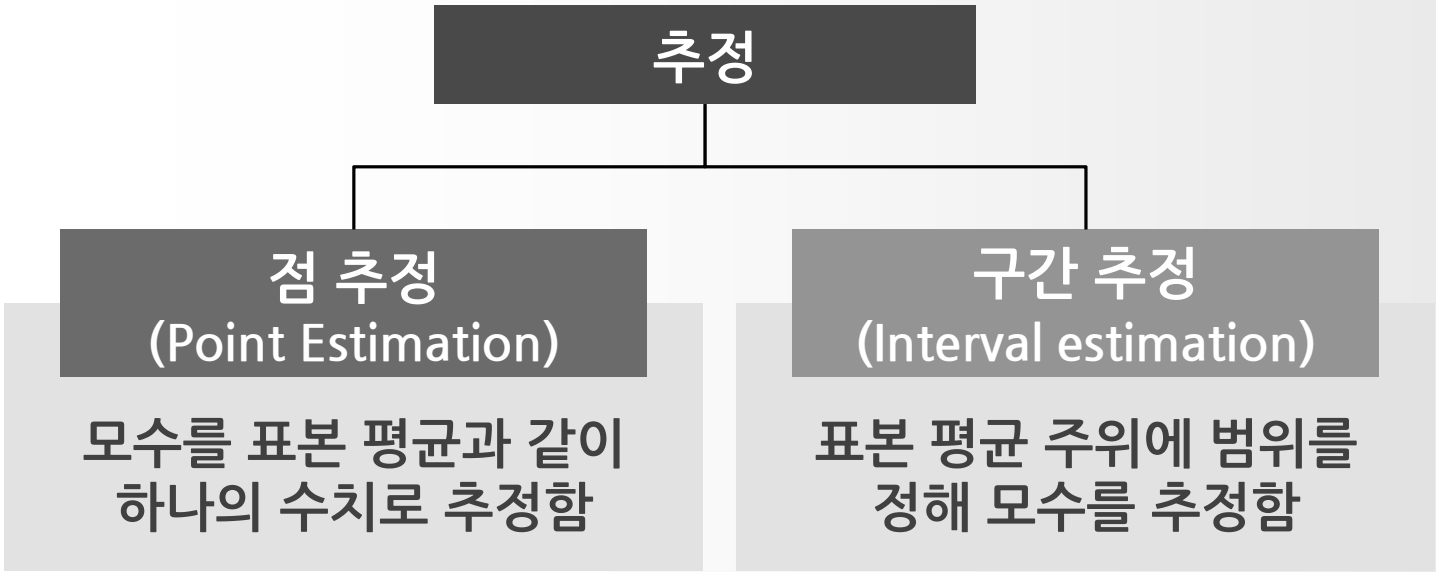
표본으로부터 통계량의 값

예 표본 평균이나 표본 비율의 값을 구하여 그 값을 근거로 모수의 값

↳ 즉, 모집단 평균이나 모집단 비율의 값이 얼마나 될 것이라고 추정하는 방법

1) 추론통계학

- 방법론


추정
점 추정
(Point Estimation)
모수를 표본 평균과 같이
하나의 수치로 추정함
구간 추정
(Interval estimation)
표본 평균 주위에 범위를
정해 모수를 추정함

1) 추론통계학

● 방법론

가설검정

모수에 대해서 어떤 값을 가정하고 표본 정보를 이용하여
그 가정이 합당한가 합당하지 않은가를 결정하는 방법

2) 점 추정(Point Estimation)

- **추정량**

추정량

알고자 하는 모수(θ)를 추측하기 위해 표본으로부터
관찰된 값으로 계산되는 표본의 통계량

→ $\hat{\theta}$으로 표기함

추정치

표본으로부터 관측된 자료를 통해 계산된
추정량의 결과(값)

2) 점 추정(Point Estimation)

- 추정량
  - 모수와 추정량

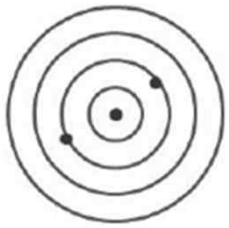
모수(θ)	구분	추정량($\hat{\theta}$)
μ	평균	$\bar{X}$
σ^2	분산	s^2
P	비율	$\hat{P}$

2) 점 추정(Point Estimation)

- 추정량

좋은 추정량은 다음 중 어떤 추정량일까?

→ 동심원은 과녁을 나타내어 중심이 모수임

	큰 표준오차	작은 표준오차
편이된 추정량 (불편 추정량이 아닌 경우)		
불편 추정량		

2) 점 추정(Point Estimation)

● 불편성과 불편추정량(Unbiased Estimator)

불편성

추정량이 갖춰야 할 가장 기본적인 성질로 한쪽으로 치우쳐지지 않음

불편 추정량

'치우쳐지지 않음'은 추정량의 기대값이 모수와 같음을 나타내며 이런 성질을 만족하는 추정량

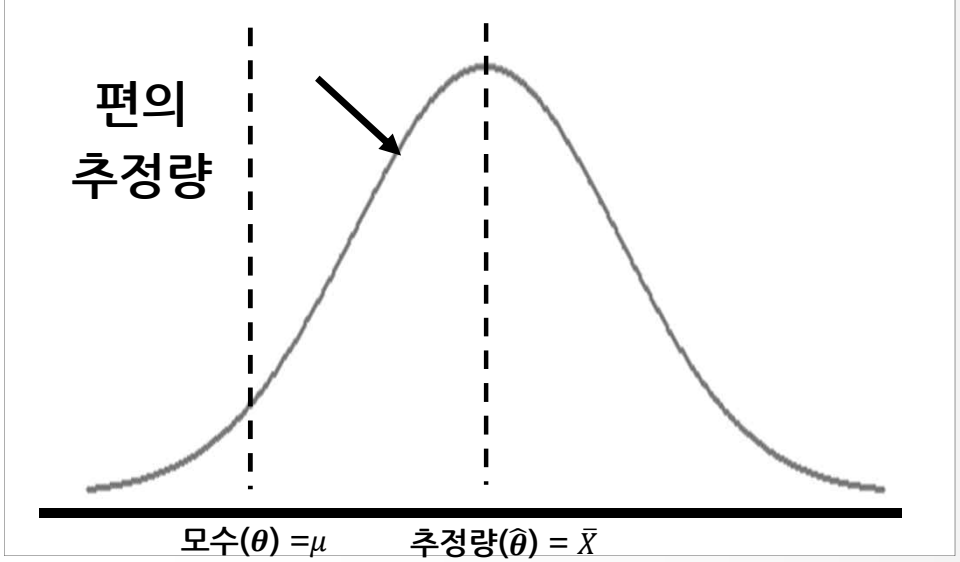
→ 모수 θ에 대한 추정량 $\hat{\theta}$이 다음을 만족할 때 $\hat{\theta}$ 은 θ에 대한 불편추정량이라고 함

$$E(\hat{\theta}) = \theta$$

2) 점 추정(Point Estimation)

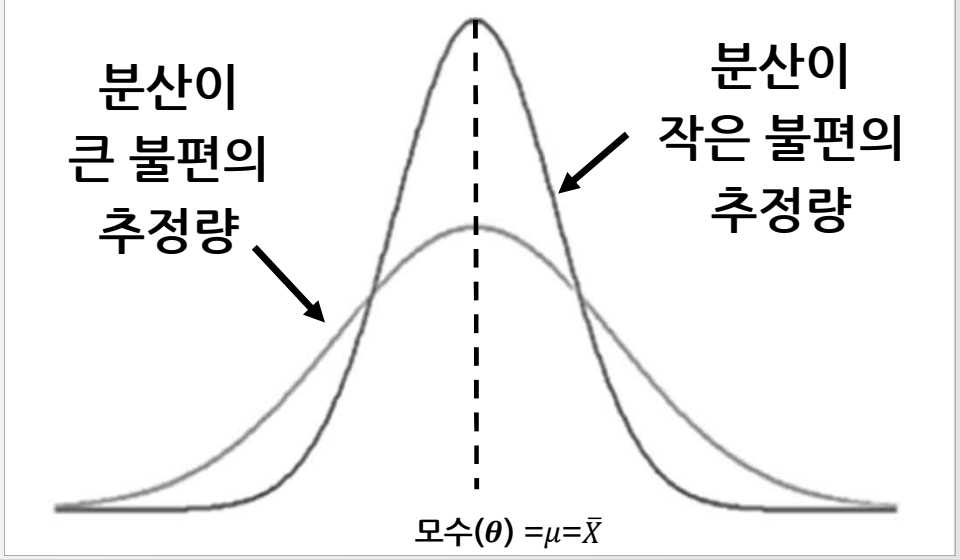
- 불편성과 불편추정량(Unbiased Estimator)

편의 추정량의 기대값



모수(θ) ≠ 추정량($\hat{\theta}$)

실제 모수의 불편의 추정량의 기대값



모수(θ) = 추정량($\hat{\theta}$)

2) 점 추정(Point Estimation)

● 불편성과 불편추정량(Unbiased Estimator)

적합한 추정량(Goodfit Estimator)이란?

- 모수에 대한 불편추정량이어야 함
- 여러 불편추정량 중 분산이 가장 작은 불편추정량이어야 함
- 분산이 가장 작은 추정량을 최소분산 불편추정량(Minimum Variance Unbiased Estimator)이라고 함

3) 구간 추정(Interval Estimation)

구간 추정

모수의 참값이 존재할 것으로 추정되는 구간을
표본으로부터 구하여 추정하는 방법

신뢰구간

- 구간 추정을 위해 표본으로부터 구한 하한과 상한을 각각 $\widehat{\theta_L}$, $\widehat{\theta_U}$이라 할 때, 0 < α < 1인 α에 대해

$$P\left(\widehat{\theta_L} < \alpha < \widehat{\theta_U}\right) = 1 - \alpha$$

를 만족하는 구간 $\widehat{\theta_L}$, $\widehat{\theta_U}$ 을 "모수 θ에 대한 100(1 - α)% 신뢰구간"이라 부르고, (1 - α)를 신뢰수준이라고 함

3) 구간 추정(Interval Estimation)

- 모집단의 분산을 알 때 모평균의 구간 추정

01 $\overline{X}$에서 기대값인 μ를 빼고 표준편차인 $\frac{\sigma}{\sqrt{n}}$로 나눈 표준화 변환을 사용함

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

02 표준정규분포표를 이용하여 하한과 상한을 구한 후 원래의 정규분포로 돌아와 구할 것임

3) 구간 추정(Interval Estimation)

● 모집단의 분산을 모를 때 모평균의 구간 추정

모집단이 미지의 평균과 분산을 갖는 정규분포를 따를 때

- 이로부터 추출된 n개의 확률표본 $X_1, X_2, \cdots, X_n$의 표본 평균을 $\overline{X}$, 표본분산을 S^2(표준편차 S)이라 하면, 다음의 통계량 T는 자유도가 (n-1)인 t분포를 따름

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

정리
하기

추론통계학(Statistics Inference)

✓ 정규분포

- 종모양의 형태를 가짐
- 양 끝이 아주 느린 속도로 감소하지만,
  축에 닿지 않고 $-\infty$와 ∞까지 계속됨
- 평균을 중심으로 좌우 대칭임
- 평균 주변에 많이 몰려 있으며 양 끝으로
  갈수록 감소함
- 평균과 표준편차로 분포의 모양이 결정됨
- 정규분포의 모수는 평균 μ와
  표준편차 σ(분산 σ^2)로, $N(\mu, \sigma^2)$으로 나타냄

정리
하기

추론통계학(Statistics Inference)

✓ **표준정규분포**

- **평균이 0이고 표준편차가 1인 정규분포($N(0, 1^2)$)를 표준정규분포라 하고 대문자 Z로 표시함**
- **모든 정규분포는 표준정규분포로 변환할 수 있음**
  - **확률변수 X가 평균 μ와 표준편차 σ인 정규분포를 따른다고 할 때,**

$$Z = \frac{X - \mu}{\sigma}, \qquad Z \sim N(0, 1^2)$$

정리
하기

추론통계학(Statistics Inference)

- ✓ **표본 분포**
  - **모집단의 크기가 N이고 표본의 크기가 n일 때 표본을 비복원으로 추출하는 경우의 수는 $\binom{N}{n}$가지로 모집단의 크기와 표본의 크기에 따라 결정됨**
  - **표본 분포는 표본의 크기가 n으로 정해졌을 때 추출 될 수 있는 모든 표본으로 부터 구한 통계량으로 구성됨**

정리
하기

추론통계학(Statistics Inference)

✓ **중심극한정리(Central Limit Theorem)**

- **동일한 확률 분포를 가진 독립 확률 변수 n개의 평균의 분포는 n이 적당히(n 〉 30)크다면 정규분포에 가까워 진다는 정리**

cf 대수의 법칙은 큰 모집단에서 무작위로 뽑은 표본의 평균이 전체 모집단의 평균과 가까울 가능성이 높다는 통계와 확률 분야의 기본 개념

정리
하기

추론통계학(Statistics Inference)

- ✓ **추론통계학(Statistical Inference)**
  - 표본을 추출하여 구한 표본 평균이나 표본 비율의 값을 근거로 모집단 평균이나 모집단 비율의 값이 얼마가 될 것이라고 추측하는 것
  - 모집단의 특성을 알고 표본의 특성에 대한 분석을 하는 표본 분포와는 그 과정이 반대임

정리
하기

추론통계학(Statistics Inference)

✓ **추정(Estimation)**

- **표본으로부터 통계량의 값**

예 표본 평균이나 표본 비율의 값을 구하여
그 값을 근거로 모수의 값

➡ **즉, 모집단 평균이나 모집단 비율의 값이
얼마나 될 것이라고 추정하는 방법**

정리
하기

추론통계학(Statistics Inference)

- ✓ 추정(Estimation)
  - 점 추정(Point Estimation)
    - : 표본의 특성을 나타내는 계산식(통계량) 중 모수를 유추하는 데 있어 최적의 계산식을 통해 구한 하나의 추정값을 구하는 방법
  - 구간 추정(Interval Estimation)
    - : 하나의 점(값)이 아닌 모수의 참값이 포함될 것으로 기대하는 구간을 추정하는 방법

정리
하기

추론통계학(Statistics Inference)

- ✓ 가설검정(Hypothesis Test)
  - 모수에 대해서 어떤 값을 가정하고 표본 정보를 이용하여 그 가정이 합당한가, 합당하지 않은가를 결정하는 방법


- 다음 시간에 살펴 볼 내용 -
10강 통계분석모형
수고하셨습니다.