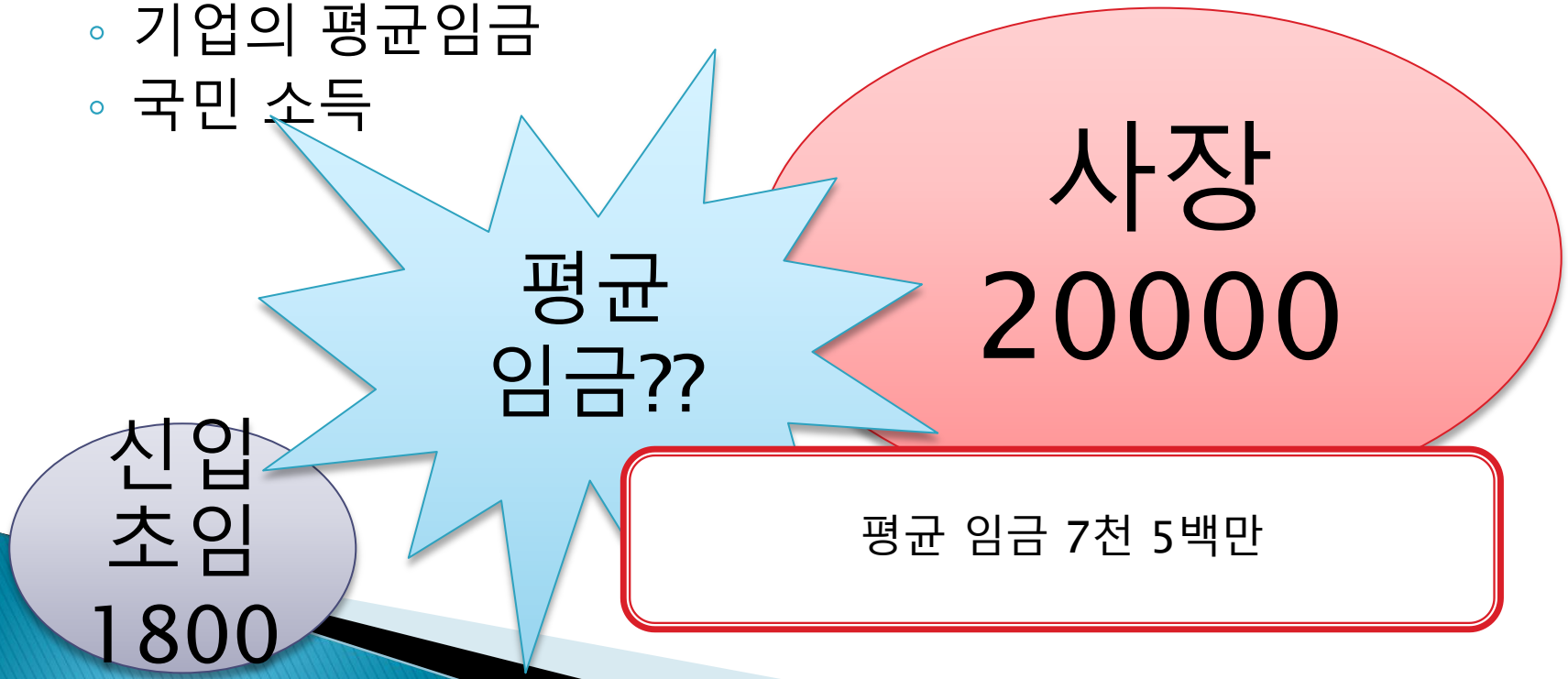
기초 통계 이론

평균

- 전체 데이터를 더한 후 개수로 나눈 값을 의미
- 평균의 맹점
  - 값은 모든 데이터를 다 더해서 나누는 것이라서 데이터가 왜곡될 수 있음.
  - 기업의 평균임금
  - 국민 소득

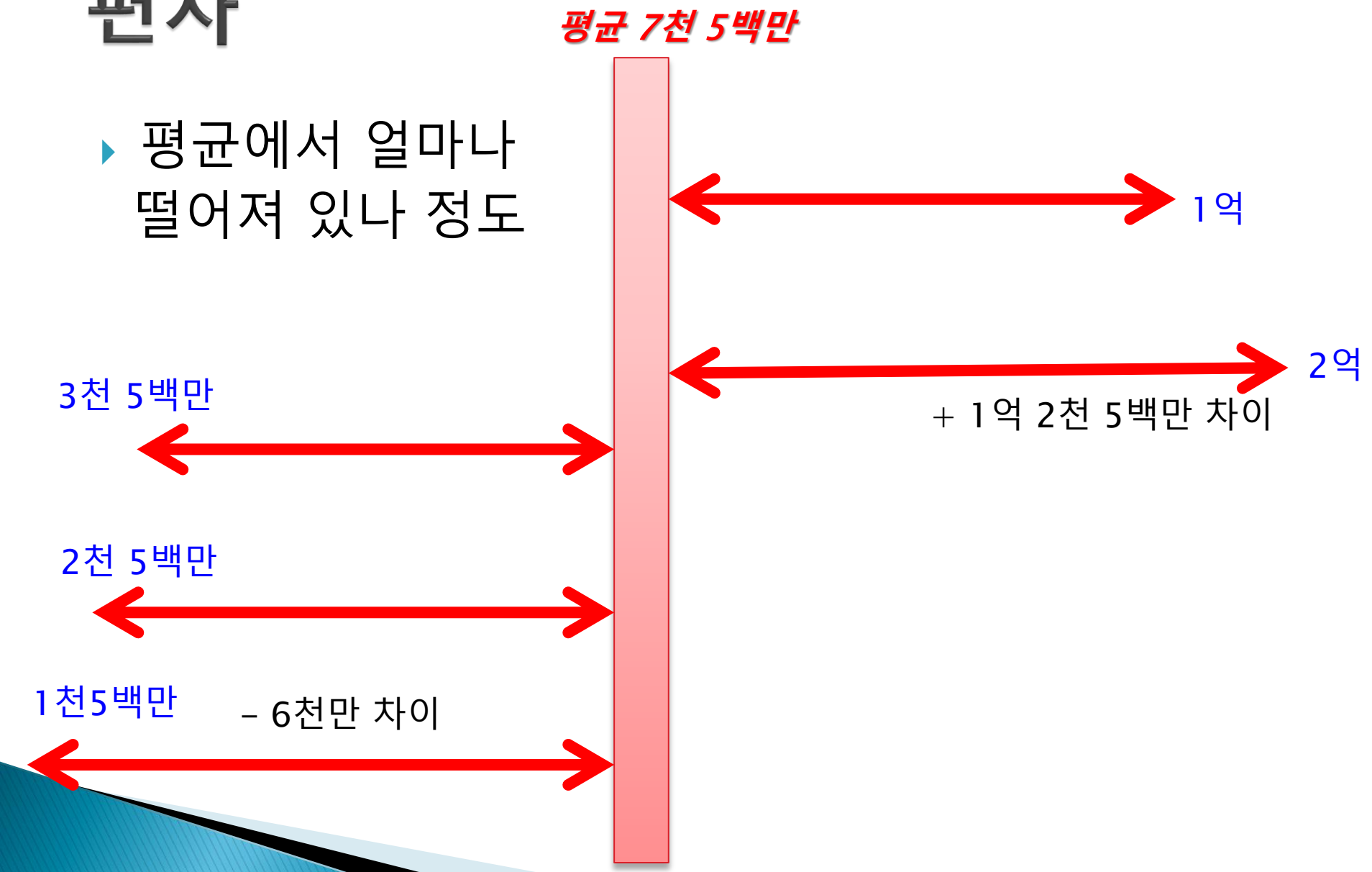


중앙값

- 최대값과 최소값의 차이가 많이 나는 경우
- 모든 데이터를 크기 순서대로 정렬시킨 후 가운데 있는 값을 의미

편차

- 평균에서 얼마나 떨어져 있나 정도



제곱평균(평균제곱)

- 편차간의 평균을 구하고자 하는 경우 편차가 +,- 가 있어 구할 수 없음.
- 각 편차에 제곱을 구해 평균을 구함.
- 제곱한 값의 평균
- 각 편차에 제곱을 구한 값=> **분산(Variance)**
- **표준 편차(Standard Deviation, SD)** => 분산은 제곱한 값이므로 분산에 루트를 씌워준 값(제곱을 제거한 값)

분산

- 각 편차에 제곱을 구한 평균값

표준 편차

- 분산은 제곱한 값이므로 분산에 루트를 씌워준 값 (제곱을 제거한 값)

월급의 표준 편차는? (단위: 천만원)

A 1500, 2500, 3500, 10000, 20000

7500

- 6^2 = 36
- 5^2 = 25
- 4^2 = 16
- 2.5^2 = 6.25
- 12.5^2 = 156.25
- ----------
- 합: 239.5
- 평균(분산): 47.9=>루트값 (표준편차) 6.9

B 1500, 3000, 5500, 6000, 10000

5200

- 3.7^2 = 13.69
- 1.2^2 = 1.44
- 3^2 = 9
- 8^2 = 64
- 4.8^2 = 23.4
- ----------
- 합: 111.53
- 평균(분산): 20 =>루트값 (표준편차) 4.5

A가 B보다 편차가 크다.!

모집단, 표본

- **모집단:** 계산에 사용한 원본 대상
- 모집단이 아주 클 경우: 국민, 세계 인구 대상
- 일부 대표성을 가진 데이터를 분석해서 전체도 그럴 것이다라고 추측을 하는 방법을 사용
- **표본:** 일부 대표성을 가진 데이터

자유도

- 통계학에서는 표본의 분산과 표준편차를 계산할 때 나누는 분모의 수를 모집단 -1개로 계산해서 사용 함.
- 모집단 -1 => 자유도!
- 주어진 데이터에서 표본을 자유롭게 뽑을 수 있는 경우의 수
- 1~100 중 1을 선택하면 나머지 99를 자유롭게 뽑을 수 있음.
- 모집단이 100개 일 때, 자유도 99
- 표본을 추출해서 표본의 분산과 표준 편차를 계산할 때는 항상 자유도를 분모로 사용하는 것이 통계학의 규칙

연습문제

학생 번호	국어	수학
100	88	66
200	77	100
300	44	99
400	99	77
500	100	88

1. 국어의 평균을 구하세요.
2. 수학의 평균을 구하세요.
3. 국어의 표준 편차를 구하세요.
4. 수학의 표준 편차를 구하세요.

평균의 종류

- **산술평균(평균)**
  - (1+2+3+4)/4
- **기하평균(상승평균)**
  - 2018년 10%증가, 2019년 2%감소 2년 평균 성장률
  - (10+2)/2 = 6%??
  - $\sqrt{1.1\ x\ 0.98} = \sqrt{1.078} = 1.04$ (4%성장)
- **조화평균**
  - 주로 평균 속도를 구할 때 많이 사용
  - 100 + 60 = 80??
  - 2xy/x+y = 2(100x80)/100+80 = 88.99
- **제곱평균**
  - 표준 편차를 구할 때 -값이 있으면 제곱하여 평균을 구하는 중간에 구할 때 사용.

평균의 종류

표준화, 표준값

이름	영어	수학
김아무개	55	77
송아무개	**88**	99
박아무개	77	**88**
정아무개	44	66
평균	66	82.5

영어 점수 88점과 수학 88점은 누가 더 잘한 것일까?
⇒ 평균에서 떨어진 정도로 비교해보자.!

영어 점수인 88점은 평균보다 많이 높은 점수이고,
수학 점수인 88점은 평균에서 가까운 점수이다.

표준화, 표준값

- **표준화:** 모든 값들의 표준값을 정해서 그 값을 기준으로 차이를 구해서 비교를 하는 방법
- 표준값 = (각데이터-평균)/표준편차
- 편차값 = 표준값 x10+50
- 만점이 기준이 달라지더라도 그 표준값의 평균은 반드시 50, 표준편차는 반드시 10이라고 가정.

표준화, 표준값

- c1 <- c(55,88,77,44)
- c2 <- c(77,99,88,66)
- sd(c1) #표준편차 20
- sd(c2) #표준편차 14

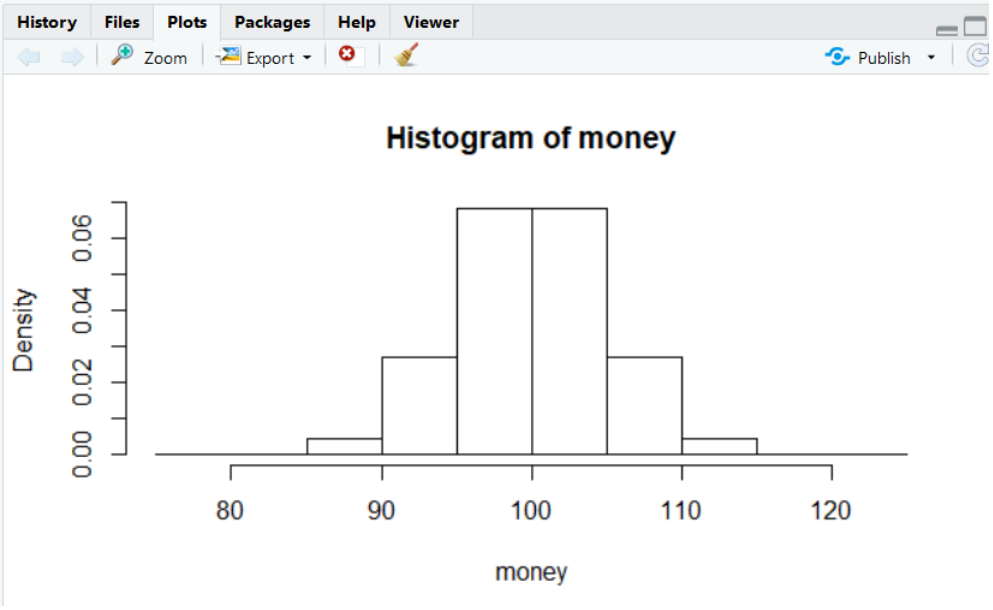
- 평균이 동일하다고 가정한 경우
- c1의 편차가 더 크다.
- 따라서, c1이 c2의 동일한 점수보다 더 잘했다고
- 할 수 있다.

- 따라서, 표준화란 비교해야 할 데이터의 기준이 서로 다를 경우에 같은 기준으로 만들어서 비교할 수 있게 해줌.

정규분포

- 다양한 데이터가 존재할 경우에 “데이터가 분포한다”라고 표현
- 데이터들이 평균값을 기준으로 좌우 대칭형으로 분포되어 있는 형태

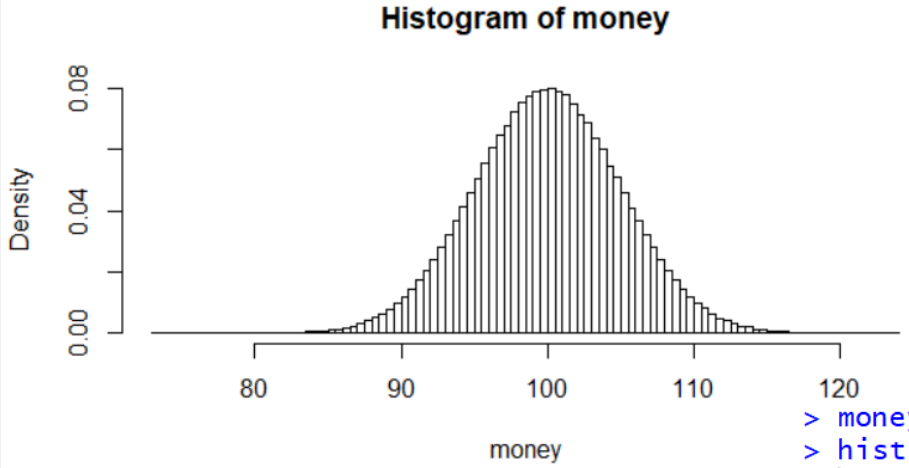
```
> money <- rnorm(n=1000000, mean=100, sd=5)
> hist(money, breaks = 10, probability = T)
>
```



```
> money <- rnorm(n=1000000, mean=100, sd=5)
> hist(money, breaks = 100, probability = T)
>
```

History | Files | Plots | Packages | Help | Viewer

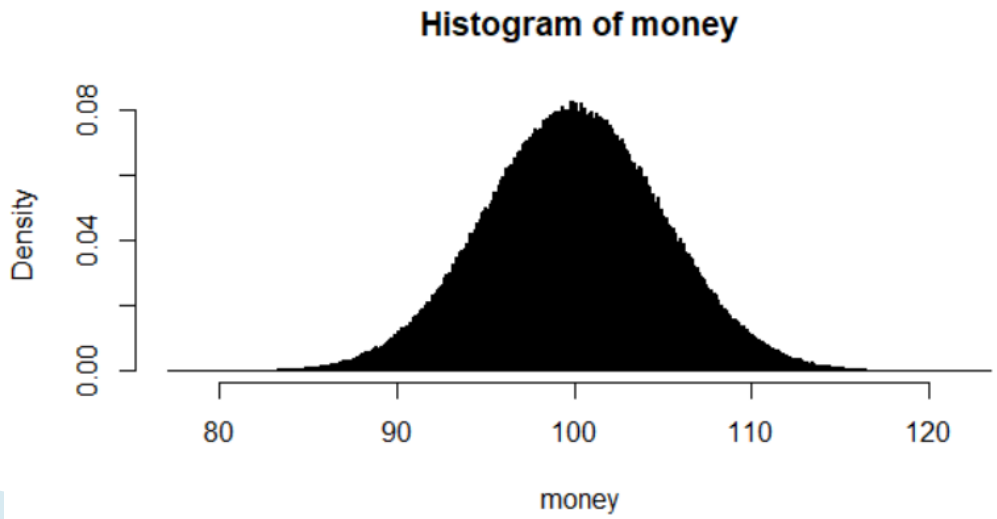
Zoom | Export | Publish




```
> money <- rnorm(n=1000000, mean=100, sd=5)
> hist(money, breaks = 1000, probability = T)
>
```

History | Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



T분포

- 정규분포와 유사
- 전체 데이터를 검증할 수 없어서 일부 데이터만 활용해서 사용하는 분포
- T분포에서 표본의 수량이 많아질수록 정규분포와 닮아감.

T검정

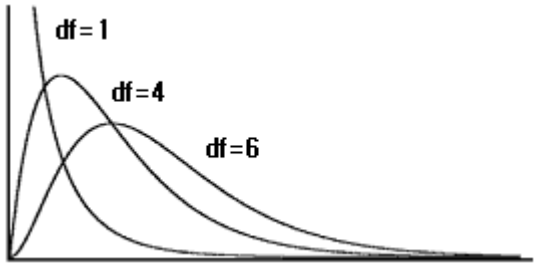
- **귀무가설**(0가설, 영가설): 두 가지 변수가 차이가 없다
- **대립가설**: 두 가지 변수가 차이가 있다.
- 전체 학생들의 평균을 추정
  - 100명 중 10명만 선택
  - 지난 학기의 평균 75점
  - 이번 학기와 지난 학기의 평균은 동일하다. 차이가 없다. (성적 변동 없다. 귀무가설)
  - 이번 학기와 지난 학기의 평균은 동일하지 않다. 차이가 있다. (성적 변동 있다. 대립가설)

T검정

- **유의확률(p-value):** 0.05%보다 크게 나오면 귀무가설 채택, 적게 나오면 대립 가설이 채택

카이 제곱 분포

- 분산값을 그래프로 만든 분포
- 1개의 그룹의 특정 데이터들의 분산값을 추측하거나 검증할 때 많이 사용
- 분산 자체가 편차의 제곱된 값을 다루기 때문에 카이 제곱 분포라고 불림.
- 음수가 없음.
- 자유도가 커질수록 오른쪽 종모양이 됨.(양수로만 이루어진 정규분포가 됨.)



카이 제곱 분포

- 평균값을 기준으로 양수값 중에서 왼쪽으로 올수록 값이 커지고, 오른쪽으로 갈 수록 값이 줄어듦.

모집단, 모평균, 표본평균

- 데이터 분석시 가장 정확한 것은 대상 데이터를 전부 조사해서 분석하는 것.
- 전국민을 대상으로 하는 선호도 조사 등.
- 비용과 시간이 많이 필요한 현실적인 문제.
- 일부 대상을 추출해서 표본으로 만들고 표본을 분석해서 그 결과로 전체 데이터를 추정하는 방법ㅇ을 사용함.
- **모집단**: 조사하고자 하는 데이터 전체
  - **무한 모집단**– 셀 수 없는 모집단
  - **유한 모집단**– 셀 수 있는 모집단
    - 국끓이기

무한 모집단의 모평균, 표본 평균

- 모평균은 구하기 어려워.
- 표본을 대상으로 평균을 구해서 추측을 하는 방법을 일반적으로 사용함.
- 표본 평균: 표본이 많을 수록 표본 평균은 모평균과 비슷 해짐.
- 표본 평균으로 모평균을 예측하는 일을 많음.