제 17 장 평균의 정확성

1. 표본분포와 표준오차
2. 표본평균으로 모평균 추정하기
3. 표준오차
4. 측정오차
5. 모집단과 표본의 관계

1. 표본분포와 표준오차

기대값과 표준오차

우연 → 추출할 때마다 표본 내 25개의 숫자가 달라짐 → 표본평균이 달라짐

→ 표본평균의 기대값과 표준오차를 구해 보면 표본평균이 어떻게 실현될지 어느 정도 "그림"이 그려짐

1. 표본분포와 표준오차

표본 별로 그때그때 다르게 실현되는 표본평균

상자로부터 무작위로 추출한 숫자들의 평균과 상자의 평균이 확률오차에 의해서 얼마나 차이가 날까?

궁금하면 1 부터 7 까지의 카드가 한 장씩 들어 있는 상자로부터 25회 무작위 복원추출을 일단 두 번만 반복 시행해 보자. (n=25, R=2)

(i) 4 5 3 2 5 7 5 6 2 4 1 5 2 4 7 7 6 4 2 4 7 2 4 4 3

표본합 = 105, 표본평균 =105/25=4.2

(ii) 5 2 4 3 4 5 2 3 7 7 1 2 3 3 4 7 2 6 5 3 6 6 1 5 4

표본합 = 100, 표본평균 =100/25=4.0

1. 표본분포와 표준오차

모표준편차, 표본표준편차, 표본평균의 표준오차

모표준편차=상자의 표준편차 (σ)

- 상자로부터 하나의 값을 추출할 때 이 값이 상자의 평균(모평균)으로부터 얼마나 떨어져 있는가를 나타내는 지표

표본표준편차 (SD)

- 표본 내 하나의 값이 표본평균과 얼마나 떨어져 있는가를 나타내는 지표. 이는 상자의 표준편차에 대한 추정량임

표본평균의 표준오차 (표본평균의 SE)

- 추출한 값들의 평균(표본평균)이 모평균과 얼마나 떨어져 있는가를 나타내는 지표

1. 표본분포와 표준오차

표본평균의 표준오차 구하기

$$\text{표본평균의 표준오차} = \text{모표준편차}(\sigma) \div \sqrt{\text{표본크기}}$$
$$\approx \text{표본표준편차}(SD) \div \sqrt{\text{표본크기}}$$

1. 표본분포와 표준오차

표본평균의 기대값과 표준오차

1부터 7까지의 카드가 한 장씩 들어 있는 상자로부터 25회 무작위 복원추출할 경우, 표본평균의 기대값과 표준오차를 구하라.

- 표본평균의 기대값=상자의 평균 = 4
- 상자의 표준편차 = 2 → 표본평균의 표준오차 = $2/\sqrt{25}=0.4$

→ 표본평균: 4 ± 0.4

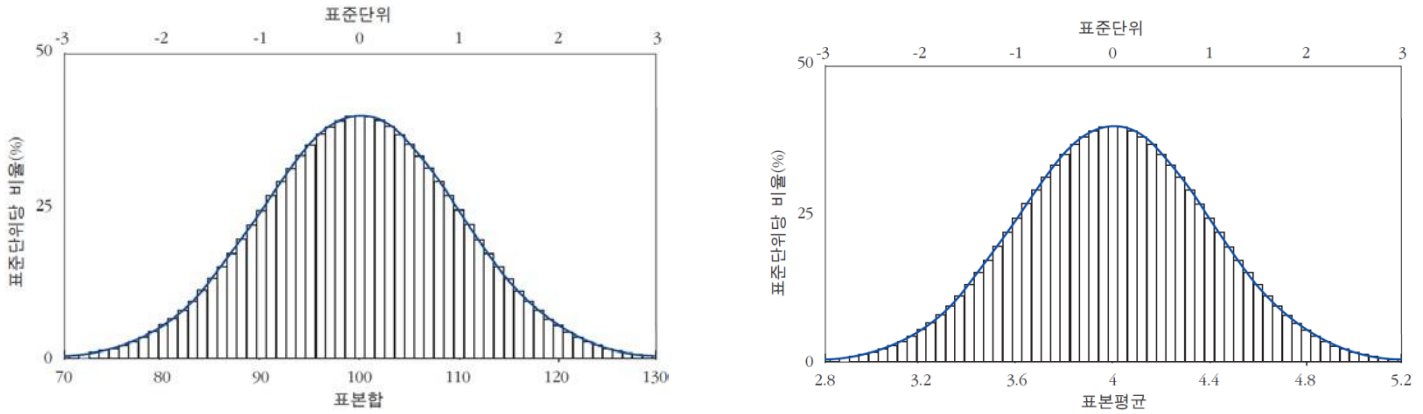
1. 표본분포와 표준오차

표본분포 (sampling distribution)

표본평균의 표본분포란 표본평균의 확률히스토그램을 말한다.

표본평균을 얻는 과정을 무한히 반복 ($R=\infty$) 시행하여 얻은 수없이 많은 표본평균들을 가지고 그린 표본평균의 히스토그램

그림 17-1 표본합과 표본평균의 확률히스토그램



- 표본합과 표본평균의 히스토그램은 스케일만 다를 뿐 전반적 모양은 똑 같다.
- 표본크기(n)가 크면 중심극한정리에 의해 각각의 히스토그램에 대한 정규 근사 가능함

1. 표본분포와 표준오차

표본평균의 기대값과 표준오차

1부터 7까지의 카드가 한 장씩 들어 있는 상자로부터 100회 무작위 복원추출할 경우, 표본평균의 기대값과 표준오차를 구하라.

- 표본평균의 기대값 = 상자의 평균 = 4
- 상자의 표준편차 = 2 → 표본평균의 표준오차=$2/\sqrt{100}=0.2$

→ 표본평균 : 4 ± 0.2

앞서 1부터 7까지의 카드가 한 장씩 들어 있는 상자로부터 25회 무작위 복원추출했던 경우와 비교해 보면

- 표본평균의 표준오차는 표본크기의 제곱근에 반비례하여 감소

2. 표본평균으로 모평균 추정하기

표본평균으로 모평균 추정하기

상자 안 카드에 대한 정보가 없을 때, 표본평균으로부터 상자의 평균을 추정하는 과정

예시) 2만5천 가구가 사는 도시에서 1,000 가구를 무작위로 추출한 결과 그 평균소득이 3,240만원 이었을 때, 도시 전체가구의 평균소득을 추정하라.

- 추출한 천 가구의 표준편차=1,900만원
- 표본평균의 표준오차
  =1,900만/$\sqrt{1,000}$=60만원
- 전체 2만5천 가구의 평균소득은 3,240±60(만원)

2. 표본평균으로 모평균 추정하기

모평균에 대한 신뢰구간 구하기

전체 2만5천 가구의 평균소득에 대한 95%신뢰구간은 표본평균으로부터 그 표준오차의 2배를 가감함으로써 얻는다.

3,240±120만원 ➔ (3,120만원, 3,360만원)

왜 표준편차대신 표준오차를 쓰는가?

- 표준편차 : 추출한 한 장의 카드가 모평균으로부터 떨어져 있는 정도
- 표본평균의 표준오차 : 추출한 카드들의 평균이 모평균으로부터 떨어져 있는 정도
- 모평균에 대해 추론할 때 카드 한 장에 든 정보를 이용하지 않고 카드들의 평균에 든 정보를 이용하므로 표준편차가 아닌 표본평균의 표준오차를 사용해야 한다.

2. 표본평균으로 모평균 추정하기

모평균에 대한 신뢰구간 구하기

표준정규분포곡선에서 95%신뢰구간은 - 2~+2에 해당하는 구간

신뢰구간을 구할 때 표준정규분포곡선을 쓰는 근거는?

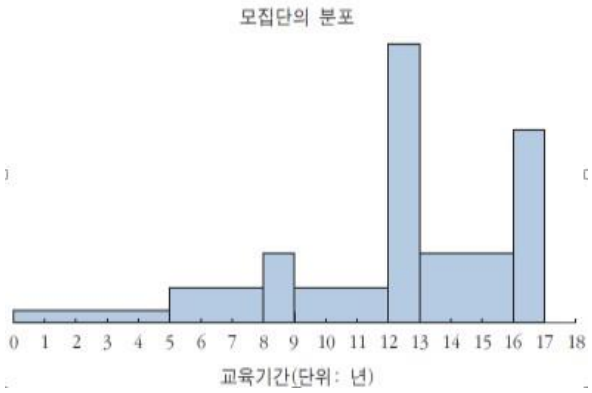
- 중심극한정리 : 개별관측치의 히스토그램이 정규분포곡선과 다르더라도 표본평균의 확률히스토그램은 표본크기가 커지면 그 모양이 정규분포곡선과 유사해진다. 모평균에 대한 추론에 있어서 표본평균을 사용함에 주목하라.

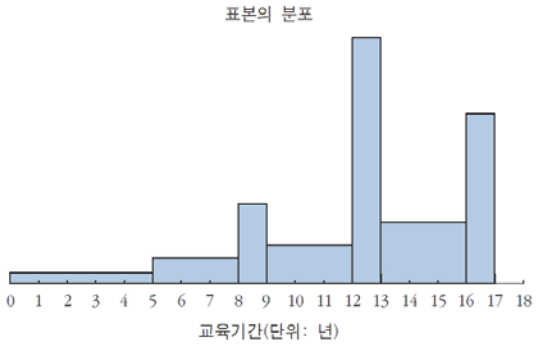
2. 표본평균으로 모평균 추정하기

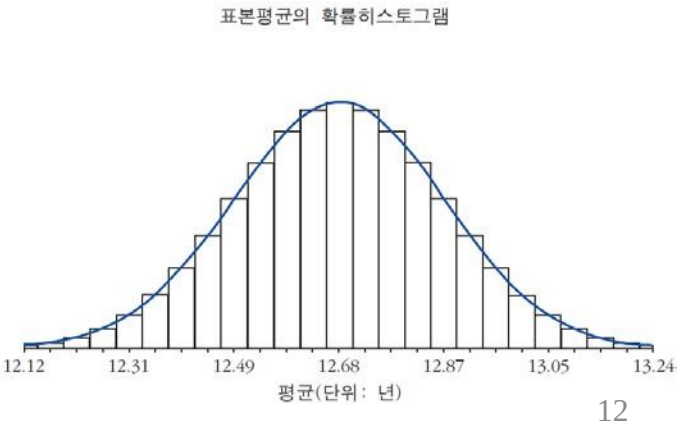
모집단의 분포, 표본의 분포, 표본평균의 확률히스토그램

어느 도시에 사는 25세 이상 인구 전체의 교육기간 분포

- 400명을 무작위로 추출하여 구한 표본의 분포가 모집단의 분포와 비슷함
- 두 분포가 닮았다는 사실은 표본이 모집단을 대표한다는 사실을 알려준다.
- 모집단의 분포나 표본의 분포는 개개인의 교육수준의 분포로서 정규분포와 아주 다르나 표본평균의 확률히스토그램은 정규분포곡선과 비슷하다.







3. 표준오차

어느 표준오차를 사용하나?

합에 대한 추론 → 표본합의 표준오차 평

균에 대한 추론 → 표본평균의 표준오차

개수에 대한 추론 → 표본개수의 표준오차

비율에 대한 추론 → 표본비율의 표준오차

표본합의 표준오차 = $\sqrt{\text{표본크기}}$ × (상자의 표준편차)

표본평균의 표준오차 = (상자의 표준편차) / $\sqrt{\text{표본크기}}$

표본개수의 표준오차 = 0-1 상자로부터의 표본합의 표준오차

표본비율의 표준오차= (0-1 상자의 표준편차)/$\sqrt{\text{표본크기}}$

일반적으로 각종 표준오차 공식에 등장하는 상자의 표준편차는 모르므로 이를 계산해서 아는 표본표준편차로 대체해서 공식을 사용한다.

3. 표준오차

표준오차

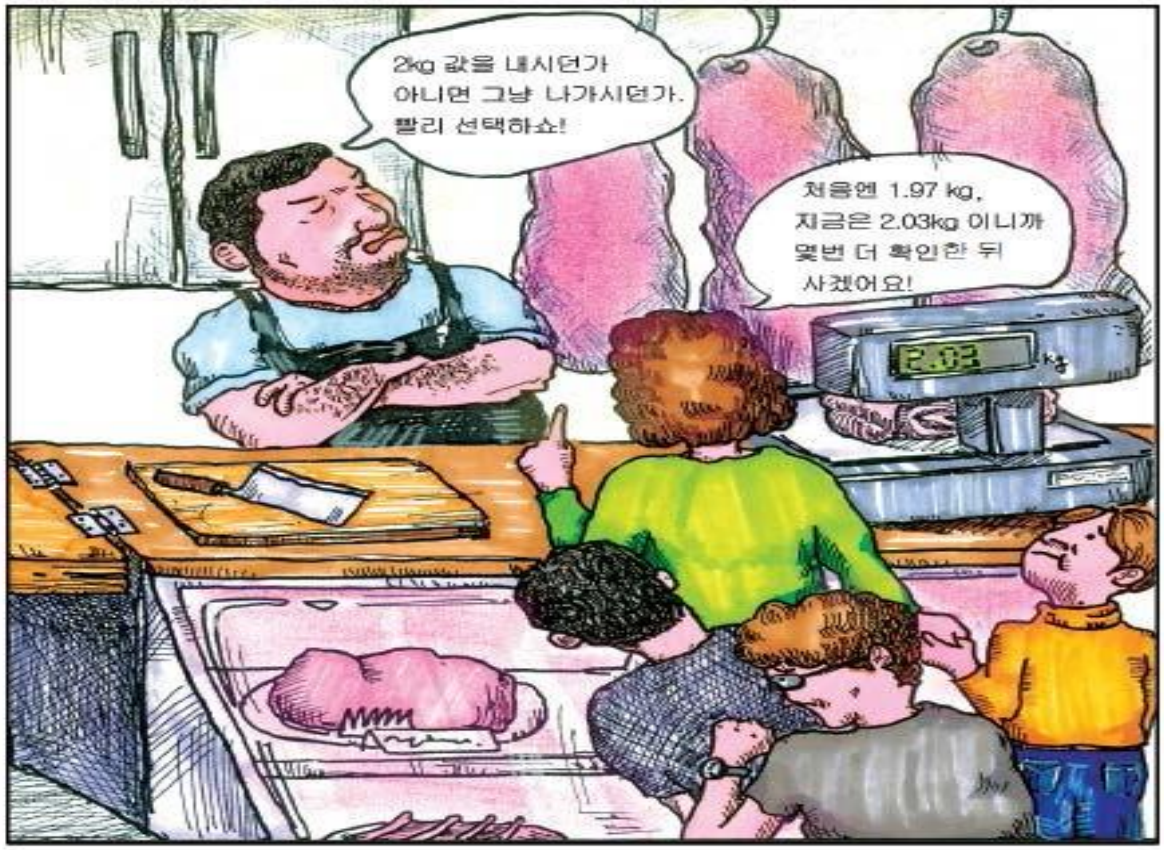
표본평균과 모평균 사이에 확률오차만큼 차이 존재
확률오차가 작으면 신뢰도가 높고 확률오차가 크면 신뢰도가 낮다.

확률오차의 전반적 크기는 표준오차로부터 알 수 있다.

표본평균의 표준오차는 표본표준편차를 표본크기의 제곱근으로 나누어 추정한다.

4. 측정오차

측정오차(measurement error)



아주머니 측정오차라는 게 있습니다.

4. 측정오차

측정오차(measurement error)

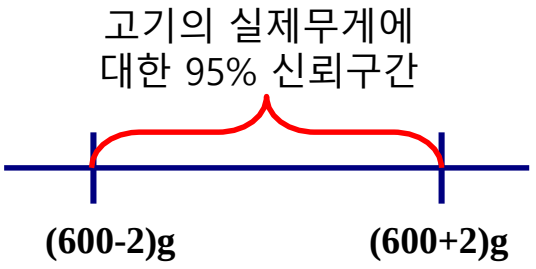
측정과정에 확률오차가 개입되기 때문에, 측정치는 참값과 확률오차(측정오차)만큼의 차이를 보인다.

→ 반복측정을 하여 측정치들을 평균하면 측정오차가 줄어 측정의 신뢰성이 증가한다.

예) 고기 한 덩어리의 무게를 100번 반복 측정한 결과

평균= 600g, 표준편차=10g

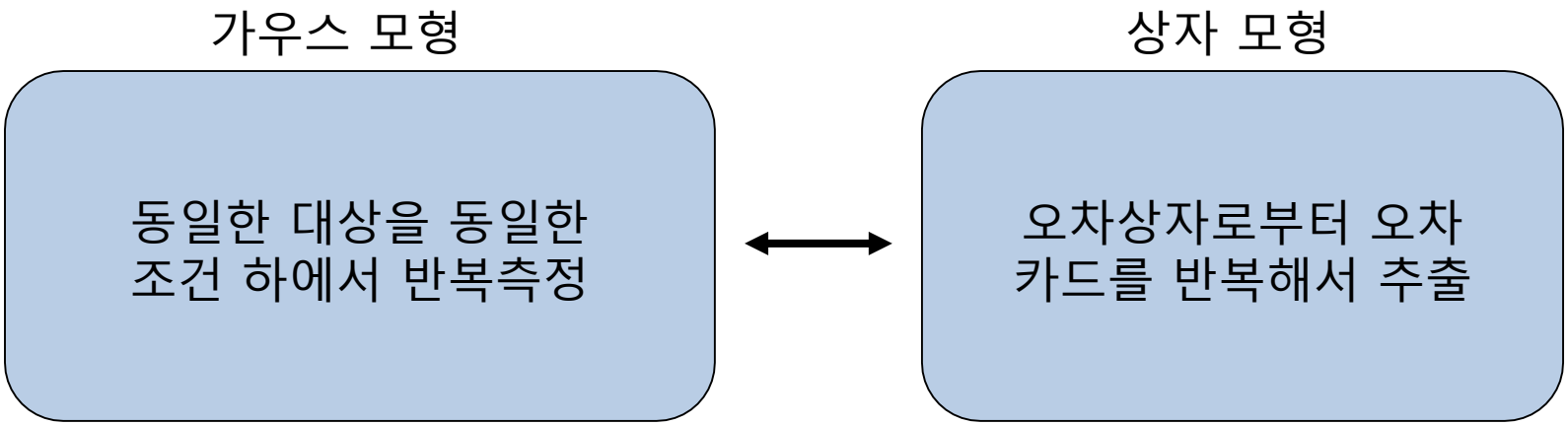
- 단일 측정치에 내재하는 측정오차의 크기: 표본표준편차인 10g 정도
- 측정치의 평균에 내재하는 측정오차의 크기: 표본평균의 표준오차인 1g(=10g / $\sqrt{100}$) 정도



4. 측정오차

가우스 모형

측정오차와 관련하여 가우스 모형은 하나의 측정치를 오차상자에서 하나의 오차카드를 꺼내 이를 참값에 더한 것으로 모형화한다.



4. 측정오차

가우스 모형

첫 번째 측정치
= 무게의 참값+오차상자에서 첫 번째로 추출한 값

두 번째 측정치
= 무게의 참값+오차상자에서 두 번째로 추출한 값

:

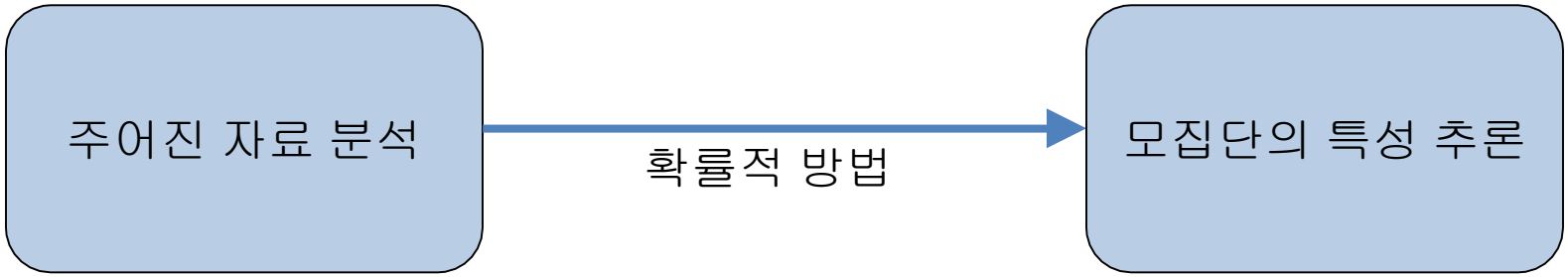
백 번째 측정치
= 무게의 참값+오차상자에서 백 번째로 추출한 값

반복측정치의 표준편차 = 확률오차의 표준편차

가우스 모형에서 반복측정치들의 표준편차는 오차상자의 표준편차에 대한 추정량이 된다. 표본의 크기가 커질수록, 즉 반복 측정의 횟수가 증가할수록 그 추정량의 신뢰도는 증가한다.

5. 모집단과 표본의 관계

상자모형과 통계적 추론



지금까지 배운 공식은 상자모형으로 환원되지 않는 모형에서는 무의미하다. 통계적 추론을 위해서는 상자모형과 같은 확률모형이 필요하다.

자료에 추세나 경향성이 있으면 적절한 변환을 통하여 이를 상자모형으로 환원시켜야 한다. 그렇지 않으면 의미 있는 통계적 추론을 하기가 어려워진다.

5. 모집단과 표본의 관계

모집단과 표본

모집단

- 관심의 대상이 되는 개념상의 집단

표본

- 실제로 분석하게 되는 자료

모평균

- 모집단으로부터 자료 하나를 추출할 때의 기대값

모표준편차

- 모평균으로부터의 편차를 제곱하여 그 기대값을 구한 뒤 제곱근을 취한 것

$$\sigma = \sqrt{E(Y_i - \mu)^2}$$

5. 모집단과 표본의 관계

표본평균

표본평균은 개별 관측치가 표본에서 어떤 값을 취할지 그 기대값을 알려준다. 이때의 기대값은 표본 내 자료 분포에 따라 구한 기대값이다.

- 모집단의 분포에 따라 기대값을 구하라는 연산자를 통상 E 로 표기하는 것처럼 표본 내 자료의 분포에 따라 기대값을 구하라는 연산자를 $\hat{E}$ 으로 표기하자.
- 즉, $\hat{E}$ 은 표본 내 관측치 각각에 1/n의 동일한 가중치를 주어 평균을 구하라는 연산자(operator)이다.
- 그러면 $EY_i = \mu$ 인 것처럼 $\hat{E}Y_i = Y_1 \times (1/n) + \cdots + Y_n \times (1/n) = \bar{Y}$ 가 된다.

5. 모집단과 표본의 관계

표본분산

개별 관측치를 표본평균으로부터의 편차로 표현해 보자. 표본분산은 편차 제곱의 표본 내 기대값에 해당된다. 이때의 기대값은 표본 내 자료 분포에 따라 구한 기대값이다.

- 모집단의 분포에 따라 기대값을 구하라는 연산자를 통상 E 로 표기하는 것처럼 표본 내 자료의 분포에 따라 기대값을 구하라는 연산자를 $\hat{E}$ 으로 표기하자.
- 즉, $\hat{E}$ 은 표본 내 관측치 각각에 1/n의 동일한 가중치를 주어 평균을 구하라는 연산자(operator)이다.
- 그러면 $E(Y_i - \mu)^2 = \sigma^2$ 인 것처럼
  $\hat{E}(Y_i - \overline{Y})^2 = (Y_1 - \overline{Y})^2 \times (1/n) + \cdots + (Y_n - \overline{Y})^2 \times (1/n) = \hat{\sigma}^2$ 가 성립한다.
- 잘 알려진 것처럼 $\hat{\sigma}^2$ 은 σ^2 을 체계적으로 과소평가하는 경향이 있다.
- σ^2 에 대한 불편추정량을 얻기 위하여 자유도 조정을 해준 것이 표본분산이다.

5. 모집단과 표본의 관계

표본분산과 표본표준편차

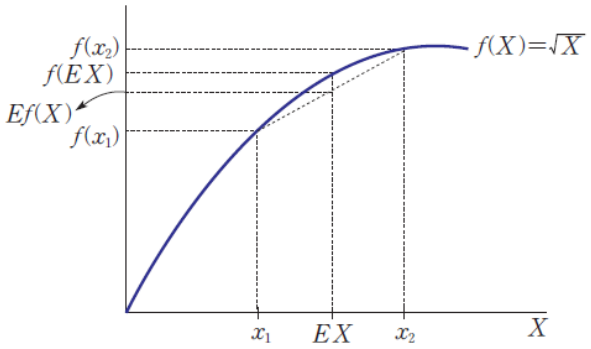
자유도로 나누어 구한 표본분산 S^2 은 모분산에 대한 불편추정량이 된다.

$S^2 = (Y_1 - \bar{Y})^2/(n-1) + \cdots + (Y_n - \bar{Y})^2/(n-1)$. 그러면 $ES^2 = \sigma^2$.

표본분산이 모분산의 불편추정량이라고 해서 표본표준편차마저 모표준편차의 불편추정량이 되는 것은 아니다.

$$ES = E\sqrt{S^2} < \sqrt{ES^2} = \sqrt{\sigma^2} = \sigma$$

젠슨의 부등식 예시



5. 모집단과 표본의 관계

표본분산과 표본표준편차

즉, 자유도 조정을 하더라도 표본표준편차는 여전히 모표준편차를 체계적으로 과소평가하는 경향이 있다.

그럼에도 불구하고 자유도 조정을 가함으로써 모분산에 대한 불편추정량을 얻었고 모표준편차 추정 시의 편의도 자유도 조정 이전보다는 줄이게 되었다.

표 17-1 모집단과 표본의 관계

	모집단	표본
자료	상자에 든 카드 일체	상자로부터 추출한 카드
자료 표기	$Y_1, \cdots, Y_N$ (N=모집단 크기)[1]	$Y_1, \cdots, Y_n$(n=표본크기)
평균	모평균 $\mu=EY_i$	표본평균 $\overline{Y}=\hat{E}Y_i$
분산	모분산 $\sigma^2=E(Y_i-\mu)^2$	가상적 표본분산 $\hat{\sigma}_n^2=\hat{E}(Y_i-\mu)^2$ [2]
		실제의 표본분산 $s^2=\hat{E}_{df}(Y_i-\overline{Y})^2$ [3]

1) 여기서는 모집단의 크기가 N으로 한정되어 있다고 가정한다.
2) $\hat{\sigma}_n^2=\hat{E}(Y_i-\mu)^2=(Y_1-\mu)^2\times(1/n)+\cdots+(Y_n-\mu)^2\times(1/n)$
3) $s^2=\hat{E}_{df}(Y_i-\overline{Y})^2=(Y_1-\overline{Y})^2/(n-1)+\cdots+(Y_n-\overline{Y})/(n-1)$
여기서 $\hat{E}_{df}$은 표본 내 자료의 분포에 따른 기대값을 구하되 자유도를 감안하여 구하는 연산자이다. 즉 표본 내 자료에 걸쳐 합친 다음에 n으로 나누지 않고 자유도인 $n-1$로 나누는 연산자이다.