제 16 장 표본추출과 확률오차

1. 표본추출과 확률오차
2. 기대값과 표준오차
3. 정규분포곡선의 활용
4. 모비율 추정의 정확도
5. 보정계수
6. 표본비율의 표준오차
7. 신뢰구간
8. 신뢰구간의 해석

1. 표본추출과 확률오차

단순 무작위 추출 (무작위 비복원 추출)

• 어느 대학 신입생 4,738명의 남녀 구성비를 알아보기 위해 그 대학 신입생 100명을 무작위로 추출

• 신입생 개개인에게 각각 1부터 4,738까지의 숫자를 부여한 뒤 컴퓨터의 난수발생기(random number generator)를 이용하여 그 중 100개의 숫자를 무작위로 추출

• 무작위 비복원 추출: 편의(bias) 발생하지 않음

1. 표본추출과 확률오차

표본추출과 확률오차

모집단의 남학생 구성비: 64%

100명(n=100) 씩으로 이루어진 표본을 250번(R=250) 반복 추출한 결과, 남학생을 정확히 64명 뽑은 경우는 단 18번에 불과 (표 16-1)

표 16-1 100명씩의 표본을 250번 반복 추출한 결과

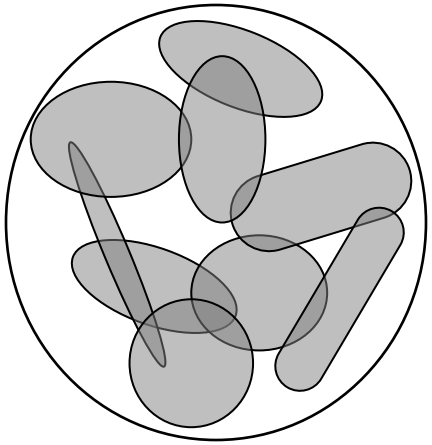
63 59 64 70 71	66 71 61 65 65	74 65 62 68 65	69 74 65 62 62	59 59 64 68 57
67 60 60 60 65	65 58 72 66 65	66 65 68 68 63	57 59 61 63 67	63 55 61 69 69
59 60 64 66 59	65 67 63 56 58	61 67 63 61 65	61 62 65 64 69	67 65 65 68 66
59 65 62 63 67	68 51 63 62 68	65 69 63 59 61	68 69 75 67 65	62 63 60 65 67
64 71 66 65 60	65 64 65 68 59	66 71 64 66 57	59 60 61 61 61	63 60 70 73 66
66 70 66 67 63	63 60 60 56 62	68 66 61 64 47	58 66 66 62 73	70 62 60 64 72
66 65 72 63 62	66 61 64 59 58	58 68 67 59 60	66 60 66 62 67	65 64 58 66 59
68 64 65 66 64	63 70 58 55 54	70 58 60 74 63	63 73 65 61 66	60 62 70 66 66
63 58 60 64 58	61 59 54 61 55	64 65 64 66 58	68 66 62 69 65	67 69 62 61 65
68 73 68 58 58	62 72 65 69 61	65 72 64 70 67	63 67 62 66 63	58 64 67 69 72

무작위로 표본을 추출하였으므로 표본의 남학생 구성비율은 그 기대값이 모비율과 같은 64%이지만 매번 실현되는 표본의 남학생 구성비는 확률오차로 인해 그 기댓값 64%와 다르게 된다.

1. 표본추출과 확률오차

표본추출

- 어떻게 4,738명에서 크기 100인 표본을 250개나 뽑을 수 있는가?



큰 원은 4,738명을 나타내고 각각의 빗금친 도형은 100명으로 이루어진 하나하나의 표본을 나타낸다. 도형간 서로 겹치는 부분은 있지만 어느 두 도형도 전체가 같지는 않다.

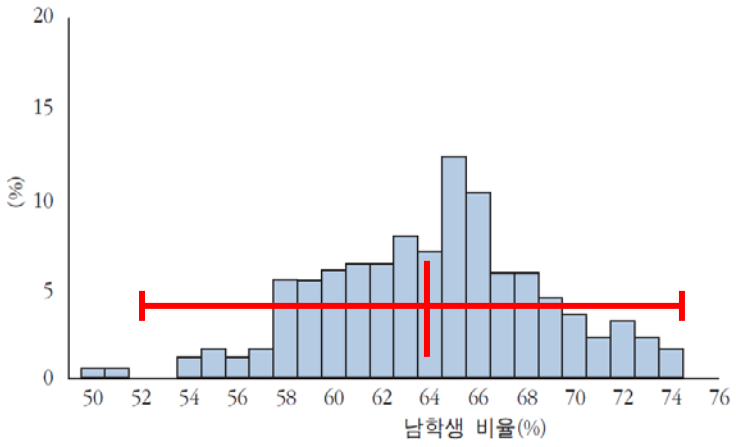
- 4,738명으로부터 100명을 뽑는 경우의 수
- ${}_{4738}C_{100}$가지의 서로 다른 표본이 가능함

1. 표본추출과 확률오차

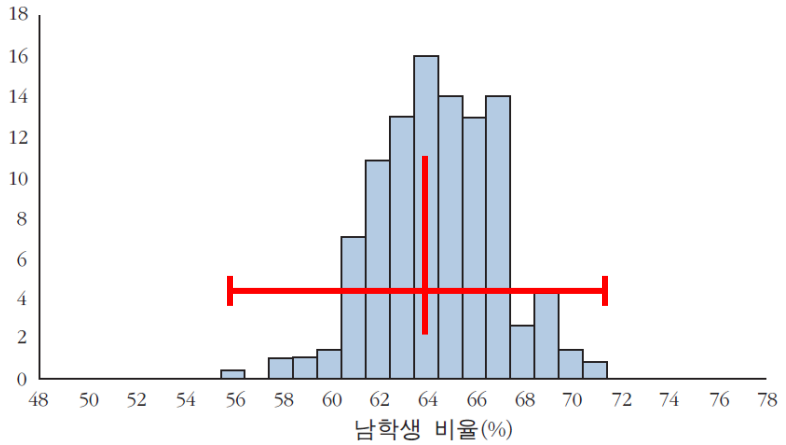
표본크기와 확률오차

- 표본크기를 늘리면 모집단과 비슷한 결과가 나오는가

표본크기가 100일 때 남학생 비율의 히스토그램(250개의 표본을 이용하여 그림)



표본크기가 400일 때 남학생 비율의 히스토그램(250개의 표본을 이용하여 그림)



- 평균의 법칙에 의해, 표본크기가 커지면 표본비율의 확률오차는 감소
- (표본 비율) = (모비율) + (확률오차)
- 표본크기가 커지면 표본비율은 점차 모비율로 수렴하게 된다.

2. 기대값과 표준오차

표본크기와 표준오차

- 단순무작위표본에서 표본의 구성비는 그 기대값이 모집단의 구성비와 같다.
- 하지만 실제 표본의 구성비는 모집단의 구성비와 확률오차만큼 차이가 난다.
- 이때 확률오차의 크기는?
  - 확률적 표본에서 표준오차는 확률오차의 표준적인 크기를 나타낸다.
- 표본크기가 커지면 표준오차는 어떻게 되는가?
  - 표본합의 표준오차는 표본크기의 제곱근으로 곱해져 증가
  - 표본비율의 표준오차는 표본크기의 제곱근으로 나누어져 감소
  - 표본크기의 제곱근이 중요: 제곱근 법칙

2. 기대값과 표준오차

개수의 표준오차

(i) 표본에서의 남학생 수에 대한 표준오차

상자모형 설정 : 상자 안에 0(여학생)과 1(남학생)만 넣음

표본에서의 남학생 수는 상자에서 100장의 카드를 뽑아 카드에 적힌 숫자들을 더한 것과 같다.

상자의 표준편차= $\sqrt{0.64 \times 0.36} = 0.48$

100번 추출한 합에 대한 표준오차= $\sqrt{100} \times 0.48 = 4.8$

2. 기대값과 표준오차

비율의 표준오차

(ii) 표본에서의 남학생 비율에 대한 표준오차

표본구성비의 표준오차(%)

= (상자의 표준편차)/$\sqrt{\text{표본크기}}$ × 100(%)

= $0.48/\sqrt{100}\times 100\% = 4.8\%$

표본에서의 남학생 비율에 대한 표준오차는 4.8%

3. 정규분포곡선의 활용

정규분포곡선의 활용

고객 100만 명 중 20%가 연소득 5천만 원 이상이라고 하자.

이 가운데 400명을 무작위로 추출한다.

(i) 상자모형, 표본비율의 기대값

연소득 5천만원 이상 고객=1, 나머지=0

상자의 기대값= 0.2 , 표본합의 기대값=400×0.2=80

표본비율의 기대값=모비율=20%

(ii) 표본비율의 표준오차

상자의 표준편차 $=\sqrt{0.2 \times 0.8} = 0.4$

표본비율의 표준오차= $0.4/\sqrt{400} = 0.02 = 2\%$

3. 정규분포곡선의 활용

정규분포곡선의 활용

- 표본에서 고소득자의 비율은 2% 정도의 오차를 두고 20% 근처에 있을 것이다.

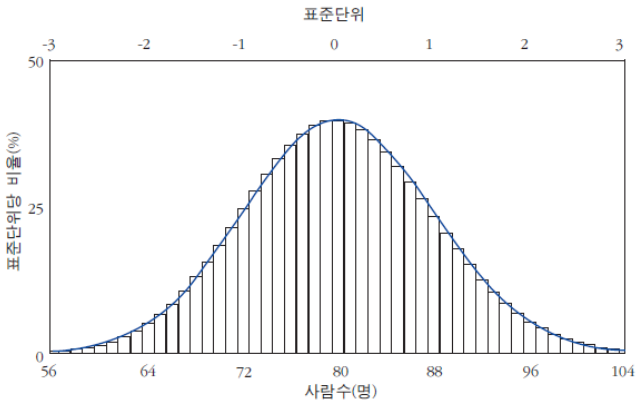
- 일반적으로 0과 1이 든 상자에서 무작위로 표본을 추출할 때, 표본비율은 그 표준오차 정도의 오차를 두고 모비율 근처에 있게 된다.

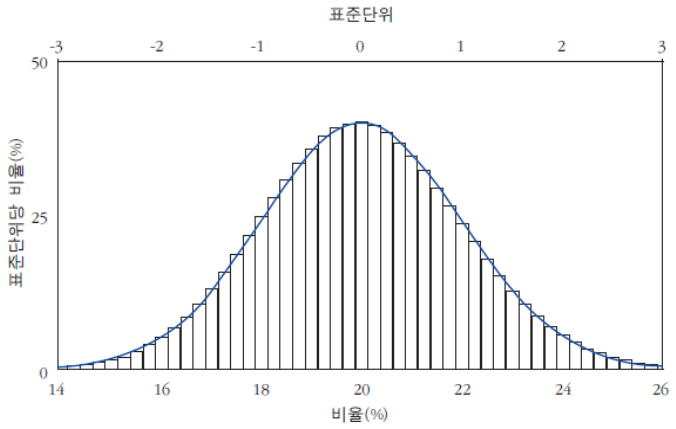
3. 정규분포곡선의 활용

정규분포곡선의 활용

표본에서 연소득 5천만원 이상인 고객의 비율이 18%에서 22% 사이에 있을 확률은?

표본비율은 기대값이 20%이고, 표준오차가 2%이다. 이를 이용하여 [18%, 22%]의 구간을 표준단위로 환산하면 [-1, 1]의 구간이 된다. 정규분포곡선을 활용하면 구하는 확률은 68%로 얻어진다.





4. 모비율 추정의 정확도

모비율 추정의 정확도

민주당 Obama 후보의 지지율 추정

- 유권자수 : 뉴멕시코 120만 명, 텍사스 1,250만 명
- 두 지역에서 각각 2,500명씩 무작위 비복원추출

어느 지역에서 Obama 지지율 추정치의 확률오차가 더 작을까?

- 비록 텍사스의 인구가 뉴멕시코보다 10배 이상 많지만, 크기가 2,500명인 표본은 텍사스에서나 뉴멕시코에서나 비슷한 양의 정보를 제공한다.
- 모집단이 표본에 비해 충분히 클 경우, 모비율 추정의 정확도를 결정하는 것은 표본의 절대적 크기이지 그 상대적 크기가 아니다.

4. 모비율 추정의 정확도

모집단의 크기와 모비율 추정의 정확도

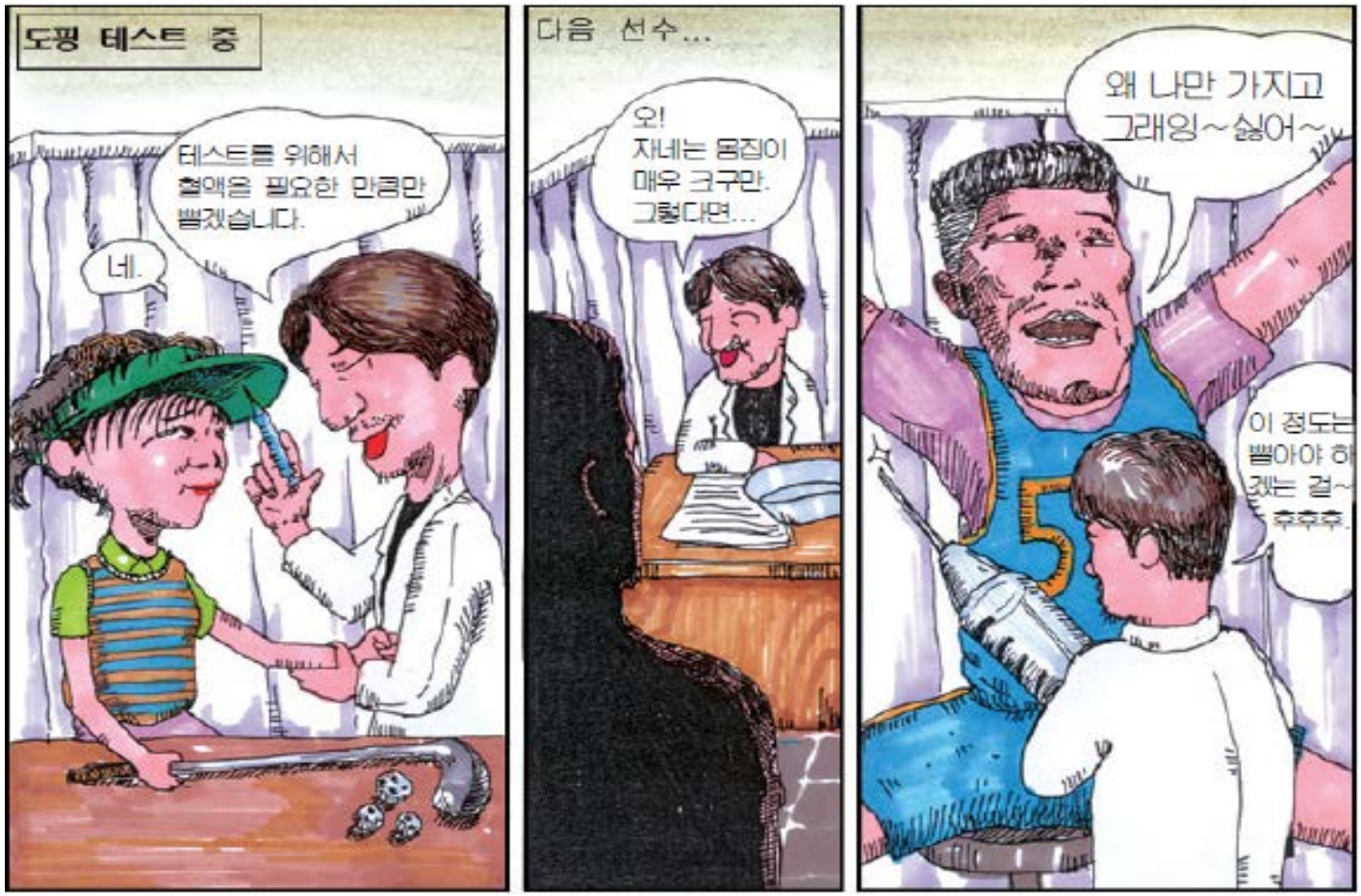
- 상자모형: 뉴멕시코와 텍사스 각주에 대해 하나씩 총 두 개의 상자모형 설정 (각 카드에는 Obama 지지자면 1, 아니면 0 기입)
- 모비율은 두 지역에서 모두 50%라고 가정 (하나의 보수적인 가정)
- 두 상자에서 각각 2,500명씩 무작위로 복원추출하면 표본 내 Obama 지지자 비율의 표준오차는 사실상 두 지역에서 같게 된다.

복원추출	비복원추출
어느 상자로부터 추출하든 차이가 없다. 매번 0 이나 1 을 뽑을확률은 50 대 50 이고 상자의 절대적 크기는 전혀 전혀 중요하지 않다.	추출되는 카드의 수는 상자에 든 카드의 수에 비해 훨씬 적다. 뽑힌 카드를 다시 넣든 말든 큰 차이가 없다. 비복원추출이 복원추출과 사실상 별로 다를 게 없다.

- 모집단이 표본에 비해 충분히 큰 경우, 모집단의 크기는 모비율 추정치의 정확도와 무관하다.

4. 모비율 추정의 정확도

덩치가 크다고 혈액 샘플까지 더 많은 양을 추출할 필요는 없습니다.



5. 보정계수

비복원추출과 보정계수

비복원추출의 경우, 추출할 때마다 상자에 든 카드의 수가 하나씩 줄어들게 되고 그 결과 상자 안 불확실성도 조금씩 줄어든다. 표준오차도 작아진다.

(비복원추출의 표준오차)=(복원추출의 표준오차)×(보정계수)

여기서 (보정계수) = $\sqrt{\frac{N-n}{N-1}}$

상자 안의 카드 수	보정계수
5,000	0.7072
10,000	0.8661
100,000	0.9874
500,000	0.9975
1,000,000	0.9988
12,500,000	0.9999

*n = 뽑은 카드의 수 = 2,500 으로 고정되어 있음

5. 보정계수

보정계수

- 보정계수는 표본에 비해 모집단이 충분히 큰 경우 1에 가깝다. 사실상 이때는 보정하지 않아도 된다.
- 그 경우 표본비율의 표준오차는 표본의 절대적 크기에 의존하며, 모집단 크기에 대비한 표본의 상대적 크기에 의존하지 않는다.
- 콜레스테롤 검사를 할 때 몸무게 200kg의 씨름 선수에게서 채취하는 혈액의 양과 몸무게 20kg의 유치원생에게서 채취하는 혈액의 양이 다르지 않다.

6. 표본비율의 표준오차

0-1 상자의 표준편차 구하기

(0-1 상자의 표준편차) = $\sqrt{(\text{상자 안 1의 비율})\times(\text{상자 안 0의 비율})}$

현실적으로 상자 내 1의 비율은 알려져 있지 않음. 두 가지 선택이 가능함

- (i) 상자 안 1의 비율(모비율)을 표본비율로 대체한 공식 사용
- (ii) 좀 더 보수적인 방법으로 0-1 상자의 표준편차를 그 최대값인 ½로 대체한 공식 사용. 이 공식의 사용이 보수적인 이유는 상자로부터 한 장의 카드를 뽑을 때 그 결과의 불확실성은 1이 나올지 안 나올지 반신반의할 때 가장 커지기 때문이다. 표준편차의 최대값=$\sqrt{(1/2)\times(1/2)}=1/2$.

6. 표본비율의 표준오차

표본비율의 표준오차

한 후보의 지지율 조사

- 모집단 : 유권자 10만 명의 선거구
- 표본 : 유권자 2,500명 무작위 추출
- 표본 내 지지율 : 53% (2,500명 중 1,328명이 지지)
- 추정된 표본 지지율은 모집단에서의 지지율과 얼마나 다를 것인가?
- 표본비율은 확률오차 때문에 모비율과 다름. 이 둘간의 차이를 대략적으로 알려주는 것이 바로 "표본비율의 표준오차"임

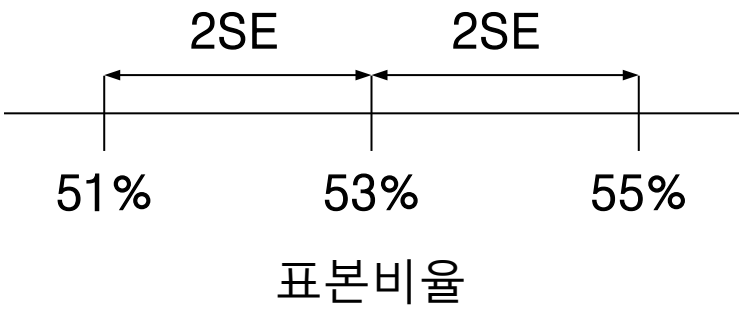
6. 표본비율의 표준오차

표본비율의 표준오차 계산

- 상자모형: 유권자마다 한 장씩 10만장의 카드가 든 상자
  (특정 후보를 지지하면=1, 지지하지 않으면=0)
- 표본 : 무작위로 추출된 2,500장의 카드
- 상자의 표준편차=$\sqrt{0.53\times 0.47}\approx 0.50$
- 표본지지율의 표준오차= $0.50/\sqrt{2500}=0.01=1\%$
- 표본지지율 53%는 당선에 충분한 50%보다 표본지지율의 표준오차 단위로 측정해서 3단위나 되는 큰 차이임 → 이 후보의 당선은 거의 확실하다.

7. 신뢰구간

신뢰도(confidence level)와 신뢰구간(confidence interval)



표본비율 = 모비율 + 확률오차

→ 모비율에 대한 95% 신뢰도의

신뢰구간 = 53%±2%

→ (51%, 55%): 95%신뢰도의 신뢰구간

표본비율 ± 1 SE의 구간은 68%의 신뢰도를 가진 신뢰구간

표본비율 ± 2 SE의 구간은 95%의 신뢰도를 가진 신뢰구간

표본비율 ± 3 SE의 구간은 99.7%의 신뢰도를 가진 신뢰구간

7. 신뢰구간

신뢰도의 해석, 정규근사 관련

신뢰도는 '정확히~' 가 아니라 '대략~' 이라고 말해야 한다.

- 표본비율의 표준오차 공식에 모비율 대신 표본비율 내지 1/2을 대입했다.
- 표본비율의 확률히스토그램을 정규분포로 근사시켜 사용했다.

표본비율이 0% 또는 100%에 가까울 경우 표본비율의 분포를 정규분포로 근사시키려면 표본크기가 충분히 커야 한다.

표본비율이 50%에 가까운 경우는 표본크기가 아주 크지 않아도 그 분포를 정규분포로 잘 근사시킬 수 있다.

8. 신뢰구간의 해석

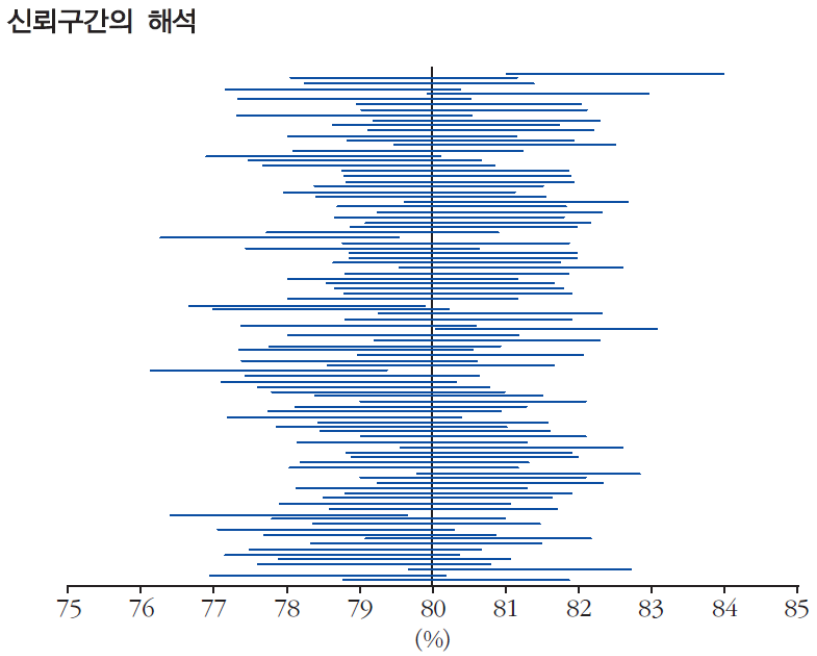
신뢰구간의 해석

신뢰구간은 표본에 의해 결정되며, 표본이 바뀌면 신뢰구간도 바뀌게 된다. 표본이 바뀌면 신뢰구간의 중심도 달라지고 신뢰구간의 길이도 달라진다.

표본을 반복적으로 추출하여 신뢰구간을 반복적으로 구해 보는 경우, 구한 전체 신뢰구간의 95% 정도가 '표본비율±2SE'라는 신뢰구간 내에 모비율을 포함하고 나머지 5%는 포함하지 않는다.

8. 신뢰구간의 해석

신뢰구간의 해석



100개의 서로 다른 표본에서 구한 100개의 95% 신뢰구간이 그려져 있다. 신뢰구간은 표본마다 다르게 나타난다. 100개의 신뢰구간 중 94개의 신뢰구간이 모비율(수직선)을 포함하고 있다.