

## 제 14 장

1. 비슷하지 않다. 질문이 동질적이지 못하다. ‘어떤 종류의 세제를 쓰는가?’에 대한 대답은 쉽게 나오기 어렵다. ‘특정세제를 쓰는가?’에 대한 답변은 상대적으로 쉽게 나올 수 있다. 나아가 이 경우 자신이 특정세제를 쓰지 않으면서도 질문에서 제시된 특정상표를 쓰고 있다고 대답할 가능성이 존재한다. 즉 응답편의의 문제가 발생한다.

2. “카지노에서는 돈을 딴 사람만 목소리가 크다”. 카지노 게임의 속성상 돈을 딴 경우보다 잃은 경우가 많음에도 불구하고 돈을 딴 사람만 자랑하기 때문에 일종의 표본추출편의(sample selection bias)가 발생한다. 주식시장의 경우는 카지노처럼 본질적으로 불공정한 게임은 아니지만 주식투자로 성공한 사례가 실패한 사례보다 더 많이 보도되고 부각되는 일은 흔히 볼 수 있다. 이런 경우 역시 표본추출편의의 한 예이다.

3. 매우 다를 것이다. 먼저 X-ray 검진에 응한 사람일수록 후에 X-ray 검진에 응한 사람보다 진폐증 증세를 더 심하게 느끼고 있어 자신의 진폐증 여부를 빨리 알고 싶어 하는 사람일 가능성이 크다.

4. 확률적 표본조사방법이라고 할 수 없다. 확률적 표본조사방법이라면 (i) 표본을 뽑는 데에 있어 조사원에게 주관적 판단의 여지가 없어야 하고, (ii) 모집단을 이루는 개개의 구성원이 표본으로 선택될 확률을 계산할 수 있어야 한다. 그런데 ‘특정 시간대에 그 대학의 학생회관 앞을 지나가는 학생들’을 표본으로 선택하는 것은 위 (i)과 (ii)의 조건 모두 충족하지 못한다. 지나가는 사람을 붙잡아 면담하는 조사 방식에는 조사원이 주관적 판단에 따라 표본을 선택할 여지가 얼마든지 있는데다가, 모집단을 이루는 개개의 구성원이 특정 시간대에 학생회관 앞을 지나가서 표본으로 선택될 확률 또한 계산할 수 없기 때문이다. 제멋대로 추출한 편의표본(convenience sample)의 전형적인 예이다. (편의표본에 대해서는 이 책의 제 22장을 참조하라.)

5. 그렇게 결론지을 수 없다. 초등학교 학생으로서 TEPS에 응시했다면 영어에 대한 관심이 높고 영어실력이 상당 수준에 이르는 학생일 가능성이 크다. 고등학교 학생의 경우 초등학교 학생들보다는 TEPS가 널리 잘 알려져 있고 남들이 TEPS를 치르니까 따라서 한 번 시험을 보거나 아니면 주위의 권유나 강요로 시험을 보거나 또는 대학을 진학하기 위한 목적 등으로 시험을 보게 되는 경우가 많을 것이다.

6. 집이 없어서 거리에서 생활하거나 수용시설에 들어간 적이 있을 정도의 빈곤층이라면 집에 전화가 없거나 아예 주거부정일 가능성이 높다. 이런 경우에는 표본으로 선택될 가능성이 희박하다. 이러한 빈곤층은 표본에 과소대표(過少代表, under-represented)되는 편의가 있을 것이다.

7. (1) 참

(2) 거짓: 다단계군집추출임

8.

(1) 참. 이 방법은 {1, 101, 201, …, 19901}, {2, 102, 202, …, 19902}, …, {100, 200, 300, …, 20000}이라는 100개의 부분집단을 만들어 놓고 이 100개의 부분집단 중 하나를 무작위로 선택하는 것과 같다. 주관이 개입할 여지가 없고, 각각의 학생들이 표본으로 추출될 확률이 동일하게 1/100임을 계산할 수 있다. 확률적 표본추출방법이다.

(2) 거짓. 단순무작위추출이라면 매번의 추출마다 모집단에 남아 있는 각 사람이 표본으로 뽑힐 확률이 같아야 하지만 이 실험에서는 그렇지 않다. 처음에 1번부터 100번까지 중 하나의 숫자(학생)를 선택하면, 예컨대 22번을 선택하면, 122번, 222번, …, 19,922번의 학생들이 뽑힐 확률은 100%가 되는 반면에, 나머지 번호의 학생들이 뽑힐 확률은 0%가 된다. 이러한 표본은 단순무작위표본이 아니다. 참고로 여기서의 표본추출법을 계통추출법(systematic sampling)이라 한다.

9. 12,000명 각각이 뽑힐 확률이 동일하지 않다.

10.

애초의 단순무작위추출을 유지해야 한다. 교장선생님께서 재분류하는 과정에 자의성이 개입될 수 있다. 결과적으로 편의를 초래할 수 있다. (교장선생님, 참으시지요!)

## 제 15 장

1. 경제활동인구조사의 표본추출방법은 다단계군집추출이다. 각 시도별 10%의 표본조사구로부터 각기 배정된 표본조사구를 추출할 때 가구수에 비례하는 확률로 추출을 하기 때문에 가구편의가 발생하지 않는다.

2. (1) 참

(2) 거짓, 표본의 대표성을 보장할 수 없음

(3) 거짓

3. (1) 추정치:  $(3030 + 2970) / 2 = 3000$ 만명

표준오차: 30만명

95% 신뢰구간: 3000만명  $\pm 2 * 30$ 만명

(2) 추정치:  $(3.8 + 4.2) / 2 = 4$

표준오차: 0.2

95% 신뢰구간: 4  $\pm 2 * 0.2$

(3) 표본조사는 센서스에 비해 표본추출편의의 문제가 발생할 수 있다. 편의가 있는 표본에서 자료를 살펴보는 것만으로는 편의의 존재를 알아내기 힘들다.

4. 평균 =  $(2.5+2.7+2.9+3.1) / 4 = 2.8$

분산 = 표준편차<sup>2</sup> =  $\{(2.5-2.8)^2+(2.7-2.8)^2+(2.9-2.8)^2+(3.1-2.8)^2\} / (4-1) = 0.067$

표준편차 = 0.26

전체 표본평균의 표준오차 = 표준편차 / 제곱근 4 = 0.13

95% 신뢰도의 신뢰구간 =  $2.8 \pm 2*0.13$

## 제 16 장

1. 상자의 표준편차  $\sqrt{0.5*0.5} = 0.5$

400회 추출의 표준편차  $\sqrt{400}*0.5 = 10$

맞게 된 것이다.

2.

특정 TV프로그램의 시청률 조사를 상자모형으로 표현하면 다음과 같다. 상자에는 해당 TV 프로그램을 시청하면 1, 그렇지 않으면 0이라고 적힌 카드가 시청자 수만큼 들어 있다. 이 상자로부터 표본크기만큼의 카드를 뽑아 1의 비율을 구한다. 상자 안 1의 비율인 모비율에 따라 상자의 표본편차가 달라진다. 다만 표준편차의 최대값은 모비율이 0.5인 경우에 0.5의 값을 갖는다. 그 결과 표본비율의 표준오차는  $0.5 * 1/\sqrt{\text{제곱근}n}$ 을 최대값으로 갖게 된다. 이 값이 0.01보다 작거나 같으려면 표본크기 n은  $n \geq 2,500$ 을 만족시켜야 한다. 즉 2,500명을 조사하면 시청률 추정치의 표준오차를 1% 이내로 통제하게 된다. 참고로 시청률 조사는 비복원추출에 해당하지만 통상 시청률 조사에서 모집단은 표본에 비해 충분히 크므로 보정계수는 무시해도 좋다.

3. ②

구성비에 대한 표준오차는 표본크기의 제곱근으로 나뉘어져 감소한다.

4.

상자 안에 250개의 구슬이 있는 경우에는 모집단이 표본에 비해 충분히 크다고 말하기 어렵다. 이때에는 구슬을 하나씩 꺼낼 때마다 상자의 크기가 점점 작아지면서 불확실성이 줄어들어 비복원추출이 우월하다. 그러나 상자 안에 10만개의 구슬이 있는 경우에는 모집단이 표본에 비해 충분히 크기 때문에 복원추출과 비복원추출의 차이는 무시해도 좋은 수준이다. 보정계수를 구해보면 거의 1에 가까워 비복원추출이라 해도 사실상 보정할 게 없다.

5.

(1) 10000개의 집단에서 500개를 추출하므로 추출 횟수에 비해 모집단이 훨씬 크다. 따라서 비복원 추출 초기하분포이지만 이항분포로 근사하여 생각할 수 있다.

$$n = 500 \quad p = \frac{4}{10} = \frac{2}{5}$$

또한  $n = 500 \gg 1$  이고  $p = \frac{2}{5}$  는  $\frac{1}{2}$  과 가까우므로 정규분포에 근사하여 생각할 수 있다.

$$N(500 \times \frac{2}{5}, 500 \times \frac{2}{5} \times \frac{3}{5})$$

$$N(200, 120)$$

따라서 표본에 있을 푸른 공의 개수에 대한 기대값 = 200

$$\text{관측값} = 218$$

$$\text{확률오차} = 18$$

$$\text{표준오차} = \frac{218 - 200}{\sqrt{120}}$$

$$(2) \text{ 기대값} = 200$$

$$\text{관측값} = 191$$

$$\text{확률오차} = 9$$

$$\text{표준오차} = \frac{9}{\sqrt{120}}$$

6.

95% 신뢰구간은 평균을 중심으로 2SE 차이에 해당하는 데 이 구간이  $49\% \pm 6\%$  로 나타났

으므로 3SE 구간에 해당하는 99.7% 신뢰구간은  $49\% \pm \frac{3}{2} \times 6\%$  이다.

답 ⑤

7.

기대값 : 950개,

표준오차 :  $\sqrt{1000 \times \sqrt{0.95 \times (1 - 0.95)}} \approx 6.89(\text{개}),$

$$z = \frac{960 - 950}{6.89} \approx 1.45,$$

$$0.5 - 0.4265 = 0.0735$$

960개 이상이 품질기준을 통과할 확률은 7.35%이다. 참고로 연속성 수정을 해주면 8.38%의 확률을 얻는데 이 값이 더 나은 답이다.

8.

$$2 \times 1.96 \times \frac{\sqrt{0.5 \times 0.5}}{n} \leq 0.05 \Rightarrow \frac{2 \times 1.96 \times 0.5}{0.05} \leq \sqrt{n} \Rightarrow n \geq \left(\frac{2 \times 1.96 \times 0.5}{0.05}\right)^2 = 1536.64$$

따라서 최소 1537명을 조사해야 한다.

9.

통계학자는 제곱근법칙에 의해 표준오차를 구했다. 이 계산이 올바른 계산이 되기 위해서는 매 거래일의 가격 등락이 독립(independent)이고 동일한 분포를 따른다(identical)는 가정이 필요하다.

10.

$$SE = \sqrt{\frac{0.5 \times 0.5}{1600}} = 0.0125$$

따라서 무작위 추출한 1600가구는  $N(0.5, (0.0125)^2)$  을 따르고  $\frac{0.52 - 0.5}{0.0125} = 1.6$

따라서  $Z = 1.6$

∴ 자동차가 없는 가구의 비율이 48 ~ 52% 가 될 확률은

$$0.4452 \times 2 = 0.8904$$

89.04% 이다.

11.

(1) 거짓. '판매한 TV의 불량품 비율'은 표본비율이 아니라 모비율이므로 주어진 진술은 옳지 않다.

(2) 참.  $\frac{\sqrt{0.0104 \times (1 - 0.0104)}}{\sqrt{193}} \times 100 \approx 0.73\%$

(3) 참.  $1.04\% \pm 2 \times 0.73\% = 1.04\% \pm 1.46\%$

12.

모집단의  $p = 0.02$  일 때 표본 만 가구의  $SE = \sqrt{\frac{0.02(0.98)}{10000}} = 0.0014$

따라서 6% 는 표본 평균 2@와  $\frac{0.04}{0.0014} \approx 28.57SE$

만큼 차이가 나므로 운에 의한 것이라 보기 어렵다. 이러한 현상이 나타나는 이유는 범외 피해를 입지 않은 가구의 무응답 편이 때문이라 유추해 볼 수 있다.

13.

(1) (표본 내 선거참여율의 표준오차) =  $\sqrt{0.7 \times 0.3 / \sqrt{10,000}} \approx 0.46\%$

(2) 확률오차로 설명할 수 없다. 설문조사를 통해 얻은 표본 내 선거참여율 70%와 실제 선거참여율 64%간의 차이 6%는 표본 내 선거참여율의 표준오차 0.46%의 무려 13배에 해당한다. 이 정도 또는 그 이상의 차이가 순전히 표본 추출상의 운에 의해 나타날 가능성은 거의 없다. 아마도 투표하는 것이 기권하는 것보다도 도덕적인 행위라는 인식 때문에 유권자들이 실제로는 투표하지 않고서도 서베이 응답 시에는 투표했다고 거짓으로 답하는 일종의 응답편의(response bias)가 존재했을 것으로 판단된다.

## 제 17 장

### 1. (1) 표본크기: 100

무작위 복원추출의 표본평균  $\overline{X}_{100}$ 의 기대치는 모집단의 평균과 동일하므로 50.

표본평균  $\overline{X}_{100}$ 의 표준오차는 "모집단의 표준편차/ $\sqrt{\text{표본크기}}$ "임. 따라서 표본평균의 표준오차는  $20/\sqrt{100} = 20/10 = 2$ 가 됨.

### (2)

비복원추출시 표본평균의 기대치는 복원추출의 경우와 동일하므로 50.

비복원추출시 표본평균의 표준오차는 보편추출의 경우와 다르다. 이 때 복원추출시의 표준오차에 곱해지는 보정계수는 다음과 같다. 그러나 그 보정계수는 거의 1에 가깝다. 아래의 보정계수를 2에 곱하면 비복원추출에서의 표본평균의 표준오차가 된다.

$$\frac{N-n}{N-1} = \sqrt{\frac{10000-100}{10000-1}} = \sqrt{\frac{9900}{9999}} \approx 1$$

### (3)

상자에 든 카드가 100장이라는 것은 모집단의 크기가 100이란 것이다. 여기서 표본의 크기도 100임에 유의한다.

복원추출시 표본평균의 기대치는 상자안의 평균 즉 모집단의 평균과 동일하다. 따라서 50.

복원추출시 표본평균의 표준편차는 "모집단의 표준편차/ $\sqrt{\text{표본크기}}$ "이다. 따라서 2이다.

비복원추출시 표본평균의 기대치는 복원추출의 경우와 동일하므로 50.

비복원추출시 표본평균의 표준오차는 보정계수를 곱해야 하는데,  $N=n=100$ 이므로 보정계수는 0이다. 따라서 표본평균의 표준오차는 0이다. 이는 정확히 모집단의 평균과 표본평균이 같다는 것을 의미한다. 왜 이럴까? 전수조사를 했기 때문이다.

### 2.

(1) 왼쪽 그림은 50회의 무작위 복원추출에 의해 얻은 표본의 분포를 나타내는 히스토그램이다. 따라서 음영으로 표시된 영역은 50회의 추출에서 숫자 4를 얻은 비율을 나타낸다.

(2) 오른쪽 그림은 표본평균의 확률히스토그램이다. 따라서 음영으로 표시된 영역은 표본평균이 (2.25, 2.75)의 구간에 존재할 확률을 나타낸다.

### 3.

(1) 1부터 7까지의 카드가 들어 있는 상자로부터 크기 25인 표본의 합이 50이상이

될 확률.

(2) 1부터 7까지의 카드가 들어 있는 상자로부터 크기 25인 표본의 평균이 3.6이상이 될 확률.

(3)  $90/25=3.6$

4.

(1) 의미 없다. 개념상 '상자의 표준편차'는 있어도 상자의 표준오차는 없다.

(2) 의미 있다. '상자의 표준편차'는 곧 모표준편차에 해당하며, 이는 상자 안에 들어있는 각각의 카드가 상자의 평균으로부터 전형적으로 떨어진 정도를 나타낸다.

(3) 의미 없다. '상자평균의 표준오차'란 곧 '모평균의 표준오차'를 말하는 셈인데, 모평균은 하나의 상수로서 그 확률오차도 표준오차도 없다.

(4) 의미 있다. '표본평균의 표준오차'란 추출한 카드들의 평균이 상자의 평균으로부터 전형적으로 떨어져 있는 정도를 나타낸다.

5.

-상자의 평균:5

-상자의 표준편차:  $\sqrt{\frac{20}{3}}$

(1) 표본크기:25

표본합의 기대값:  $5 \times 25 = 125$

표본합의 표준오차:  $5 \sqrt{\frac{20}{3}} = \frac{10 \sqrt{15}}{3} = 12.91$

표본평균의 기댓값:5

표본평균의 표준오차:  $\frac{\sqrt{\frac{20}{3}}}{5} = \frac{2 \sqrt{15}}{15} = 0.52$

(2) 표본크기:100

표본합의 기대값:  $5 \times 100 = 500$

표본합의 표준오차:  $20 \sqrt{\frac{5}{3}} = 20 \frac{\sqrt{15}}{3} = 25.82$

표본평균의 기댓값:5

표본평균의 표준오차:  $\frac{20 \frac{\sqrt{15}}{3}}{100} = \frac{\sqrt{15}}{15} = 0.26$

(3) 표본크기:400

표본합의 기대값:  $5 \times 400 = 2000$

표본합의 표준오차:  $20\sqrt{\frac{20}{3}} = 40\frac{\sqrt{15}}{3} = 51.64$

표본평균의 기댓값: 5

표본평균의 표준오차:  $\frac{40\frac{\sqrt{15}}{3}}{400} = \frac{\sqrt{15}}{30} = 0.13$

6.

모평균은  $1 \times \frac{1}{4} + 2 \times \frac{1}{2} + 3 \times \frac{1}{4} = 2$ 이고 표본평균의 기대값은 모평균과 같다.

- (1) 참
- (2) 거짓
- (3) 참

7.

- (1) 매 반복 시행마다 표본평균이 다르기 때문
- (2) 매 반복 시행마다 표본표준편차가 다르기 때문
- (3) 95개

8.

(1) 표본평균의 표준오차가  $4/\sqrt{100}=0.4$  이므로 평균교육기간에 대한 95% 신뢰구간은  $12 \pm 2 \times 0.4$ , 즉, (11.2, 12.8)이다.

(2)

(i) 보현과 혁린 두 명의 자료를 결합하면 표본크기는 200, 표본평균은 11.7. 그리고 표본평균의 표준오차는 대략  $4/\sqrt{200}=0.3$  이다. 따라서 95% 신뢰구간은  $11.7 \pm 2 \times 4/\sqrt{200}$ , 즉, (11.1, 12.3)이다.

(ii) 네 명 모두의 자료를 결합하면 표본크기는 400, 표본평균은 12, 그리고 표본평균의 표준오차는 대략  $4/\sqrt{400}=0.2$  이다. 따라서 95% 신뢰구간은  $12 \pm 2 \times 4/\sqrt{400}$ 이다. 즉, (11.6, 12.4)이다.

(iii) 우선 3개 표본평균의 평균과 표준편차를 구해보자. 평균은

$$\sqrt{\frac{(11.4-12)^2 + (12.2-12)^2 + (12.4-12)^2}{3-1}} = 0.53 \text{ 이다. 여기서 0.53은 표본}$$

평균 하나가 전반적인 평균으로부터 벗어난 정도를 측정해 준다. 따라서 표본평균 3개 평균의 표준오차는 0.53을  $\sqrt{3}$ 으로 나누어 구한다. 즉, 전반적인 평균의 표준오차는  $0.53/\sqrt{3}=0.3$ 이다. 결국 이들 3개 표본평균의 정보만을 이용하여 구한 교육기간 평균에 대한 95% 신뢰구간은  $12 \pm 2 \times 0.3$ , 즉 (11.4, 12.6)이다.

9.

(1) 표본평균의 표준오차 :  $\frac{\hat{s}}{\sqrt{n}} = \frac{1.75}{\sqrt{625}} = 0.07$  , 따라서 T

(2) 신뢰구간의 의미 상기 : F

(3) 95% 신뢰구간 :  $2.30 \pm 2 \times 0.07$ , 따라서 (2.16, 2.44), T

(4) 신뢰구간의 의미 상기 : F

10.

(1) 개별 숫자의 히스토그램은 상자의 히스토그램과 유사한 모양이 된다. 즉 표본의 분포는 모집단의 분포와 비슷해진다.

(2) 히스토그램의 모양은 중심극한정리에 의해 정규분포곡선과 비슷해진다.

(3) (1)에서의 모양은 상자의 히스토그램과 비슷한 반면 (2)에서의 모양은 정규분포곡선과 유사하다. (2)에서 구한 히스토그램은 표본평균의 확률히스토그램이다.

11. 정보량이 작아 추정의 표준오차가 크다.

12.

(i) 추출한 숫자들의 히스토그램

(ii) 표본평균의 히스토그램

(iii) 상자에 든 숫자들의 히스토그램

13.

(1) T

(2) F

(3) F

(4) F

(5) F

14.

(1) 참.

‘225명의 응답자의 평균’은 표본평균에 해당한다.

(이동거리에 대한 95% 신뢰구간) = (표본평균)  $\pm$  2SE = (488, 592)이므로,

$$(\text{표본평균}) = \frac{488 + 592}{2} = 540(\text{km})$$

(2) 참.

‘225명의 응답자의 표준편차’는 표본표준편차에 해당한다.

95% 신뢰구간의 길이는 표본평균으로부터 양쪽으로 2SE씩 총 4SE에 해당된다는 사실과 표본평균의 표준오차는 표본표준편차를 표본크기의 제곱근으로 나누어 구한다는 사실로부터

$$4 \times (\text{표본평균의 표준편차}) / \sqrt{225} = 592 - 488$$

의 식을 얻는다. 이를 풀면, 표본평균의 표준편차는 약 390km가 된다.

(3) 거짓. ‘225명의 응답을 히스토그램으로 그린 것’은 표본추출한 개별관측치의 히스토그램에 해당하고, 이는 모집단의 분포와 비슷할 것이다. 모집단의 분포가 정규분포와 거리가 멀면, 표본자료를 가지고 구한 히스토그램도 정규분포와 다를 것이다.

(4) 참. 개별관측치의 히스토그램이 정규분포곡선과 다르더라도 중심극한정리에 의해 표본평균의 확률분포를 히스토그램으로 그리면 그 모양은 정규분포곡선과

비슷해진다.

- (5) 거짓. 중심극한정리에 의해 정규분포로 수렴하는 것은 어디까지나 '표본평균의 확률분포'이지 '모평균의 확률분포'가 아니다. 모평균은 하나의 상수값으로 존재한다. '모평균의 확률분포'라는 개념은 존재하지 않는다.
- (6) 거짓. 동일한 신뢰도 하에서 모평균에 대한 신뢰구간의 길이는 표본평균의 표준오차에 정비례한다. 그리고 제곱근 법칙에 의해 표본크기( $n$ )가 증가하면 표본평균의 표준오차는  $1/n$ 배로 줄어드는 것이 아니라  $1/\sqrt{n}$ 배로 줄어든다. 450명을 이용하여 95% 신뢰구간을 구하면 225명을 이용하여 구했을 때보다 구간의 길이가  $1/\sqrt{2}$ 배로 감소한다.

15. 적절하지 않다. 날씨는 지난 과거의 날씨와 독립적으로 실현되지 않는다.

16.

모비율, 모평균 등의 신뢰구간을 구할 때는 표준정규분포곡선이 이용된다. 이때 표본비율이나 표본평균의 확률히스토그램 대신 표준정규분포곡선을 사용하는 것은 중심극한정리 때문이다. 표본이 충분히 크기만 하면 모집단의 분포가 정규분포를 따르지 않더라도 표본비율이나 표본평균의 확률히스토그램은 정규분포곡선과 비슷해진다.

17.

(1) 15

(2)  $\frac{15}{\sqrt{25}}$

18. 두 사람이 두 번씩 측정했으므로 총 4개의 측정치의 평균을 측정값으로 삼는 것이 좋다. 합의 여부는 자료를 왜곡시킬 뿐이다.

19. 생략