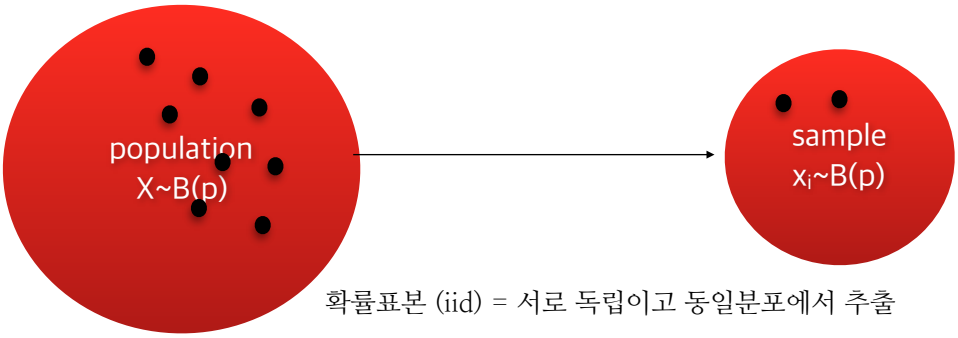
모비율 추론 개념



확률표본 (iid) = 서로 독립이고 동일분포에서 추출 $x_i \sim B(\theta = p)$

데이터 $x_1 = 0, x_1 = 1, x_3 = 1, ..., x_n = 0$

- 모집단 개체 확률변수의 분포는 모수가 p인 베르누이 분포를 갖는다.
- 표본 데이터의 관측치는 (성공, 실패) 두 값을 가진다.

MVUE for $\theta = p$

- 모수(모집단 비율 p)의 최소분산불편추정량은 표본비율 $\hat{\theta} = \hat{p} = \frac{\sum x_i}{n} = \frac{\#\ of\ 1's}{n}$
- 표본비율의 평균과 분산 : $E(\hat{p}) = p, V(\hat{p}) = \frac{pq}{n}$

MVUE 샘플링분포

- $\sum x_i \sim B(n,p)$ – 모집단 개체가 베르누이분포를 따르므로 베르누이 분포를 따르는 확률표본 합은 이항분포를 따른다.
- 대표본 이론 (Large Sample Theory) : 표본크기 n이 충분히 크면($\min(np, nq) \geq 5$) 표본 합(비율)의 분포는 정규분포에 근사한다.
- 그러므로 표본합~ $\sum x_i \sim N(np, npq)$, 표본비율~ $\hat{p} \sim N(p, \frac{pq}{n})$
- 피봇 통계량 $\frac{\hat{p} - p}{\sqrt{pq/n}} \sim N(0,1)$

통계적 추론 절차 (예제 데이터 이용)

CEREAL.csv : 아침으로 시리얼을 구입하는 고객 1,250명을 대상으로 시리얼 구입 할 때 고려하는 이유와 1) 건강음식을 위하여 2)다이어트 효과 3)식이요법 4)가격고려, 그리고 시리얼 구입 비용을 조사하였다.

(1) 현재 시리얼 구매고객이 125만명이라고 할 때 다이어트 효과를 위하여 구입한다고 응답한 고객 수의 95% 신뢰구간을 구하시오.

(2) 작년 조사에서 "다이어트 효과"를 고려하여 시리얼을 구매한다고 한 고객 비율 35%였다. 올해 비율이 증가하였다고 할 수 있나? 유의수준 5%

1. 연구문제 및 통계적 문제 정의

- 모집단 비율(다이어트 효과 고려하여 구입하는 고객 비율) 추정, 비율 추정 후 125만을 곱하면 고객 수가 됨 – 모비율 95% 신뢰구간 구하기
- 작년 비율 35% 비해 올해 "다이어트 효과" 고려한 구입 비율은 증가? – 모비율 가설검정 문제

2. 수집 데이터 정리 및 검증

```
> c=read.csv("cereal.csv")
> names(c)
[1] "Group" "Spend"
> length(c$Group)
[1] 1250
> table(c$Group)

  1   2   3   4
269 484 241 256
```

- 표본크기 n=1,250명, "다이어트 효과"(Group=2) 응답 고객 수 계산
- 표본비율 $\hat{p} = 484/1250 = 0.3872$
- 대표본 이론 검증 : $\min(1250*0.35, 1250*0.65) \geq 5$: 대표본 조건 만족

3. 통계적 가설 설정

- 귀무가설 : $p = 0.35$
- 대립가설 : $p > 0.35$ (단측가설, 고객비율이 증가)

```
> prop.test(484,1250,p=0.35,alternative=c("greater"),conf.level =0.95 )

        1-sample proportions test with continuity correction

data:  484 out of 1250, null probability 0.35
X-squared = 7.4409, df = 1, p-value = 0.003188
alternative hypothesis: true p is greater than 0.35
95 percent confidence interval:
 0.3644099 1.0000000
sample estimates:
     p
0.3872
```

4. 검정통계량, 유의확률 계산

- 검정 통계량 : $TS = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \dfrac{0.3872 - 0.35}{\sqrt{0.35*(1-0.35)/1250}} = 2.728 \sim N(0,1)$

- 유의확률 = 0.0031 – 유의수준 5%보다 작으므로 귀무가설 기각
- 95% 신뢰구간 : $-1.96 < \frac{0.3872 - p}{\sqrt{0.3872 * (1-0.3872)/1250}} < 1.96$ – 신뢰구간 구할 때는 분모 아래가 p를 사용해야 하나, 계산이 복잡해져 그냥 $\hat{p}$ 사용한다.

5. 결론 및 활용

표본비율	검정통계량	유의확률	95% 신뢰구간
0.3872	2.72	0.0031	(0.36, 0.42)

- 귀무가설 기각, 매우 유의 – 작년에 비해 "다이어트 효과"를 고려한 고객 비율은 증가하여 38.7%이다
- (0.36, 0.42), 각각에 125만명을 양측에 곱하면 고객 수의 95% 신뢰구간이다. 〉prop.test(484,1250,p=0.35)
- (활용) 기업의 시장 점유율을 고려하여 생산량을 결정할 수 있고 "다이어트 효과" 고려하는 고객들이 증가하고 있으므로 다이어트 효과 있는 시리얼을 개발해야 할 것이다.

소표본일 경우 추론

(문제)

소표본일 경우에는 대표본 이론(Large Sample theory)을 사용할 수 없으므로 MVUE 추정량 $\hat{p}$의 샘플링 분포를 알 수 없다.

(해결)

유의확률 계산 개념으로 접근 : $\sum x_i \sim B(n,p)$ 이용하자.

(예제)

H대 학생 흡연 비율이 20% 미만이라고 발표했다. 맞는지 알아보기 위하여 학생 20명을 확률 추출하여 흡연 여부를 알아본 결과 3명이 흡연하고 있다고 조사되었다. 발표가 맞는지 검정하시오.

- 귀무가설 : $H_0 : p = 0.2$ vs. 대립가설 : $H_0 : p < 0.2$
- 검정통계량 : $\sum x_i = 3 \sim B(n = 20, p = 0.2)$
- 유의확률 = $P(\sum x_i \le 3) = 0.41$ – 귀무가설 채택, H대 학생 흡연율은 20% 미만이라고 할 수 없다.

Wilson 추정량

표본크기 n에 비해 성공의 회수가 매우 작은 경우 : $\tilde{p} = \frac{\sum x + 2}{n + 4}$

문제 #1 (Keller 9판 12.5 예제문제) vote.csv

출구조사 : 민주당 후보 code=1, 공화당 후보 code=2 조사하였다. 공화당 후보가 선거에서 이길 지 검정하시오. 그리고 공화당 후보의 득표율의 95% 신뢰구간을 구하시오.

문제 #2 (Keller 9판 12.6 예제문제) cereal.csv

시리얼 고객 segmentation : 구입시 고려 사항, Group 1="건강식" 2="다이어트" 3="지병고려" 4="아무거나" 항목과 아침 식사 비용을 조사하였다. 시리얼 구입자 24,000,000명이라고 하자. 건강식을 고려하여 시리얼을 구입하는 사람 수에 대한 95% 신뢰구간을 구하시오.

작년 조사에서 "건강식"을 고려하여 시리얼을 구입한다는 비율이 20.7%였다. 올해 상승했다고 할 수 있나? (유의수준 5%)에서 검정하시오.