1-3. 두 개의 모평균 사이의 차이 추정

!질문! 다음을 어떻게 추정할까요?

i) 두 개의 서로 다른 영양제로 자란 식물 줄기의 평균 지름은?

ii) 수학이나 물리를 전공한 학생들의 의학전문대학원 입학시험의 평균성적은?

	수학전공 학생 (모집단1)	물리전공 학생 (모집단2)
평균	μ_1	μ_2
분산	${\sigma_1}^2$	${\sigma_2}^2$

	수학전공 50명 (표본1)	물리전공 50명 (표본2)
평균	$\overline{X_1}$	$\overline{X_2}$
분산	${s_1}^2$	${s_2}^2$

① 모평균의 차 $(\mu_1-\mu_2)$에 대한 최우점추정량은 $(\overline{X_1}-\overline{X_2})$

② $(\overline{X_1}-\overline{X_2})$의 표본분포

㉠ 평균 = $\mu_1-\mu_2$

㉡ 표준오차 $SE=\sqrt{\frac{{\sigma_1}^2}{n_1}+\frac{{\sigma_2}^2}{n_2}}$

또는, 표본의 크기가 충분히 클 때, $SE=\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}$

③ 모집단들이 정규분포일 때, 표본의 크기에 관계없이 정확히 정규분포

④ 모집단들이 정규분포를 따르지 않으면 n_1과 n_2가 둘 다 30 이상일 때, 중심극한정리에 의해 근사적으로 정규분포

〖예〗 두 종류의 자동차 타이어의 품질을 비교하기 위하여 각 종류마다 $n_1=n_2=100$개씩 선택하여 주행검사를 실시하고 타이어가 닳을 때까지 달린 거리를 타이어의 마모도로 정의하였다. 그 검사결과가 아래 표와 같을 때, 타이어가 닳을 때까지 걸린 평균거리의 차이인 $\mu_1-\mu_2$를 99% 신뢰구간으로 추정하시오.

종류 1	종류 2
$\overline{X_1}=26,400$	$\overline{X_2}=25,100$
${s_1}^2=1,440,000$	${s_2}^2=1,960,000$

[풀이]

㉠ $(\mu_1-\mu_2)$의 점추정값은

$$(\overline{X_1}-\overline{X_2})=26,400-25,100=1,300$$

㉡ $(\overline{X_1}-\overline{X_2})$의 표준오차는

$$SE=\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}=\sqrt{\frac{1,440,000}{100}+\frac{1,960,000}{100}}=184.4$$

㉢ 99% 신뢰구간은

$$1,300-2.58(184.4)<\mu_1-\mu_2<1,300+2.58(184.4)$$

㉣ 두 종류의 타이어가 닳을 때까지 걸린 평균 거리의 차이는 824.2와 1775.8 사이에 있는 것으로 추정된다. 그리고 두 모집단 평균 사이에 차이가 없다면 $\mu_1 - \mu_2 = 0$일 것이다. 그런데 우리가 구한 신뢰구간에 0이 포함되지 않으므로 평균이 같지 않을 것이며, 두 종류의 타이어의 평균거리에는 차이가 있다고 결론내릴 수 있다.

〖예〗 남자와 여자 사이에 유제품의 일일 평균섭취량에 차이가 있는지 알아보고자 성인남녀 각각 50명씩 표본을 추출하여 일일 유제품 섭취량을 측정한 결과 아래표와 같다.

	남자	여자
표본크기	50	50
표본평균	756	762
표본표준편차	35	30

남자와 여자 사이에 유제품의 일일 평균섭취량에 대한 95% 신뢰구간을 구하고 성별에 차이가 있는지 설명하시오.

[풀이]

㉠ 모표준편차를 근사시키기 위하여 표본표준편차를 사용하면, 95% 신뢰구간은

$$(\overline{X_1} - \overline{X_2}) \pm 1.96\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$(756 - 762) \pm 1.96\sqrt{\frac{35^2}{50} + \frac{30^2}{50}}$$

-6 ± 12.78 즉, $-18.78 < \mu_1 - \mu_2 < 6.78$

㉡ 이 구간에서는 $(\mu_1 - \mu_2)$이 취할 수 있는 값으로 양의 값, 음의 값 그리고 0이 모두 가능하므로 성별에 따른 유제품의 일일 평균섭취량에 차이가 있다고 결론을 내릴 수 없다.

⑤ 통계량 $\dfrac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{{\sigma_1}^2}{n_1} + \dfrac{{\sigma_2}^2}{n_2}}}$ 의 표본분포는, 두 모집단이 모두 정규분포를 따를 때에는 모든 표본의 크기에 대하여 표준정규분포를 따르고, 만약 모집단이 정규분포를 가지지 않지만 표본의 크기가 크면($n_1 \geq 30,\ n_2 \geq 30$) 근사적으로 정규분포를 가진다.

⑥ σ_1^2과 σ_2^2가 알려져 있지 않고 표본추정치 s_1^2와 s_2^2로 추정될 때에는 이 통계량은 표본의 크기가 클 때 근사적으로 정규분포를 따른다.

⑦ 모분산이 알려져 있지 않고 표본의 크기가 작을 때(30미만)에는 $t-$분포에 따른다.

1-4. 두 이항비율 사이의 차이 추정

① 이항비율 p에 대한 추정을 확장하면 두 이항비율 사이의 차이를 추정하는 것이 된다.

㉠ 두 생산 라인에서 만들어지는 불량품의 비율

㉡ 어떤 후보자에 대한 남자와 여자 투표자의 비율

남자 투표자 비율 (모집단1)	여자 투표자 비율 (모집단2)	남자 투표자 n_1명의 비율(표본1)	여자 투표자 n_2명의 비율(표본2)
p_1	p_2	$\hat{p_1}=\frac{X_1}{n_1}$	$\hat{p_2}=\frac{X_2}{n_2}$

② 각 모집단 1(n_1개)과 모집단 2(n_2개)에서 독립임의표본을 추출하고 표본추정값 $\hat{p_1}$과 $\hat{p_2}$를 구하면 p_1-p_2의 불편추정량은 표본차이 $(\hat{p_1}-\hat{p_2})$

③ 두 표본비율의 차이 $(\hat{p_1}-\hat{p_2})$의 표본분포는

㉠ 평균= p_1-p_2

㉡ $SE=\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}$ OR $SE=\sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1}+\frac{\hat{p_2}\hat{q_2}}{n_2}}$

④ n_1과 n_2가 충분히 클 때, 중심극한정리에 의해 정규분포로 근사

〖예〗 A와 B 두 학교의 급식에 대해 만족하는지 조사하기 위하여 각각 50명의 학생과 100명의 학생을 랜덤하게 선택하여 의견을 물어보았다.

	A학교	B학교
표본크기	50	100
급식을 만족하는 학생 수	38	65

급식을 만족하는 학생의 실제 비율의 차이를 99% 신뢰구간으로 추정하여라.

[풀이]

㉠ (p_1-p_2)에 대한 점추정값

$$\hat{p_1}-\hat{p_2} = 0.76-0.65 = 0.11$$

㉡ $(\hat{p_1}-\hat{p_2})$의 표준오차

$$SE=\sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1}+\frac{\hat{p_2}\hat{q_2}}{n_2}}=\sqrt{\frac{(0.76)(0.24)}{50}+\frac{(0.65)(0.35)}{100}}=0.0770$$

㉢ 99% 신뢰구간

$$0.11-(2.58)(0.0770)<p_1-p_2<0.11+(2.58)(0.0770)$$

또는 $(-0.089, 0.309)$

㉣ 구간$(-0.089, 0.309)$은 (p_1-p_2)=0을 포함하기 때문에, 두 학교의 학생들이 급식에 만족하는 비율에 차이가 없다고 봄

㉤ 두 비율에 차이가 없기 때문에 두 표본은 실제로 다르지 않으므로 하나로 합쳐서 급식을 지지하는 전체 비율을 추정하는 데 사용 가능.

두 표본을 합치면 $n=150$, 급식을 만족하는 학생이 103명이므로 $\hat{p}=\frac{103}{150}=0.69$

㉱ p의 점추정치 = 0.69, 95%의 오차의 한계

$$\pm 1.96\sqrt{\frac{(0.69)(0.31)}{150}}=\pm 1.96(0.0378)=\pm 0.074$$

㉲ 따라서 급식을 만족하는 학생의 비율은 0.62와 0.76 사이에 있으며 이 값들은 0.5보다 큰 비율만 포함하므로 대다수의 학생들이 급식을 만족하는 것으로 볼 수 있다.

【참고】 표본의 크기 결정

Q. 한 표본에 얼마나 많은 측정값이 포함되어야 할까?
연구자는 얼마나 많은 정보를 원할까?

Why? 연구자는 표본의 크기를 결정해야 실험을 계획하는 데 진전을 이룰 수 있다. 표본에 포함된 정보의 총 양은 추론이 믿을 만한지 또는 얼마나 좋은지에 영향을 줄 것이다.

A. 통계적 추론문제에서, 추정치의 정확성은 오차의 한계나 신뢰구간의 너비로 측정된다. 두 방법은 표본의 크기에 대한 함수이기 때문에 정확성은 표본의 크기를 결정한다.

〖예〗 한 화학공정에서 일일 평균 산출량을 유의수준 0.05에서 추정하고 오차의 한계가 4톤 이하이기를 바란다. 만약 표본표준편차가 21톤이라 하면 표본의 크기는?

[풀이]

㉠ 약 95%에서 표본평균 $\overline{X}$와 모평균 μ의 거리가 $1.96SE$가 될 것이다. 즉,

$$1.96SE < 4 \text{ 또는 } (1.96)\left(\frac{\sigma}{\sqrt{n}}\right) < 4$$

㉡ 이제 n에 대하여 풀면,

$$n > \left(\frac{1.96}{4}\right)^2 \sigma^2 \text{ 또는 } n > 0.24\sigma^2$$

㉢ 만약 모표준편차 σ를 알면 그 값을 대입하여 풀면 되지만 모른다면
- 이전에 있던 표본에서 얻은 추정값 s
- 가장 큰 측정값과 가장 작은 측정값에서 얻은 범위 추정값 $\sigma \approx$ 범위$/4$

를 사용할 수 있다. 위의 예제에서 화학공정의 과거연구에서 표본표준편차 $s = 21$톤을 얻었다고 하자. 그러면

$$n > 0.24\sigma^2 = 0.24(21)^2 = 105.8$$

이므로 크기 $n = 106$의 표본을 사용하면 (확률 약 0.95로) 평균 산출량의 추정치는 실제 평균 산출량의 ± 4톤 이내에 있을 것이라고 확신할 수 있다.

㉣ 이 때, 평균의 표준오차를 구하기 위하여 σ의 근삿값을 사용하였기 때문에 $n = 106$은 근삿값임을 알 수 있다.

〖예〗 플라스틱 파이프 제조자들은 시장의 수요를 충족시키는데 충분한 파이프의 양을 유지하기 원한다. 플라스틱 파이프를 구매하는 도매상들을 설문조사하여 내년에 구매량을 증가하려고 하는 비율을 추정하고자 한다. 추정치가 확률 0.90으로 실제 비율의 0.04 이내에 있으려면 표본의 크기는 얼마여야 하는가?

[풀이]

㉠ $1.645SE = 1.645\sqrt{\frac{pq}{n}} = 0.04$ 이고 이 방정식을 n에 대하여 풀려면 p의 값을 대입해야 한다. 표본이 충분히 크다고 확신하기를 원하면 $p = 0.5$를 사용한다.

㉡ $1.645\sqrt{\frac{(0.5)(0.5)}{n}} \le 0.04$ 또는 $\sqrt{n} \ge \frac{(1.645)(0.5)}{0.04} = 20.56$ 즉, $n \ge (20.56)^2 = 422.7$

㉢ 따라서 제조자들은 비율 p를 0.04이내로 추정하고자 한다면 적어도 423명의 도매상을 조사에 포함시켜야 한다.