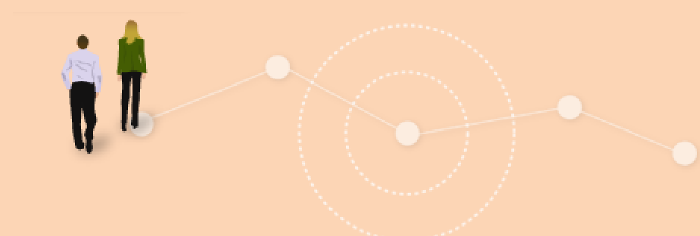
8 장

목 차

01 두 모집단의 평균 차이에 대한 가설검정(대응표본)

:: Keywords 대응표본 | 대응표본 t검정

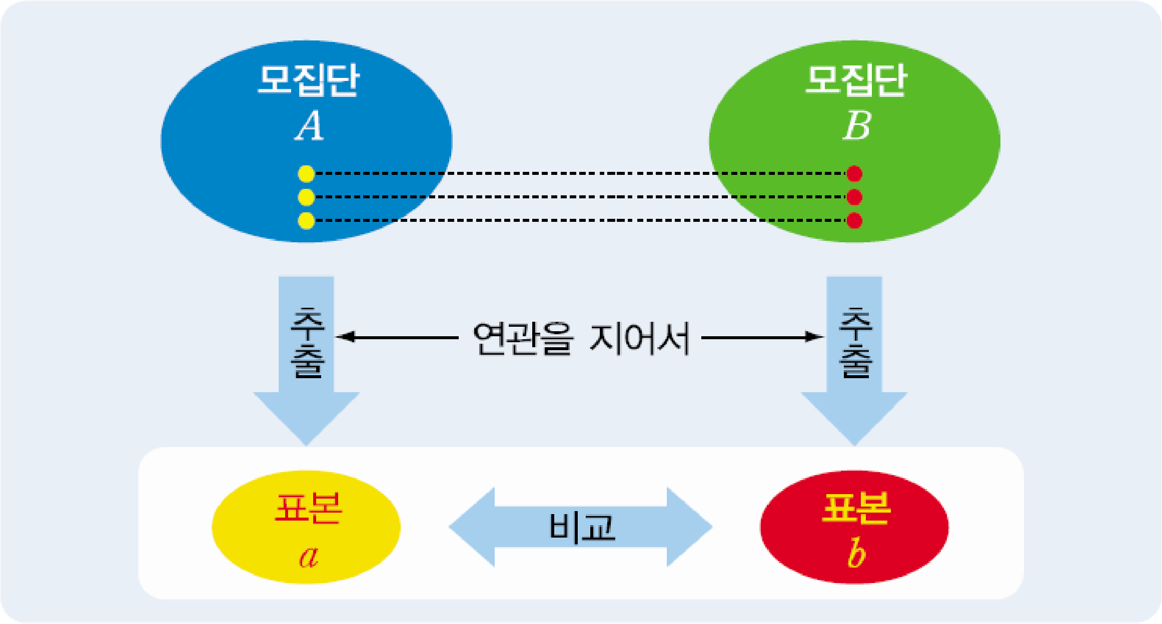


대응표본

■ 대응표본(paired sample)

두 모집단으로부터 표본을 각각 추출할 때,

표본을 구성하는 각각의 인자가 짝을 지어 서로 연관된 표본



→ 교육, 광고, 의료 등의 분야에서

사전(ex-ante)과 사후(ex-post)를 비교하여 효과를 평가할 때 주로 사용

대응표본 t검정

Ex. A제약회사

단위 : kg

번호	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
복용 전	75	74	75	75	83	77	82	62	77	82	72	75	78	71	68
복용 후	73	74	76	71	76	68	75	61	68	75	70	71	71	70	67
번호	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
복용 전	76	71	54	75	77	82	74	76	70	77	82	62	77	82	68
복용 후	73	74	50	76	68	75	74	73	69	68	75	61	68	75	67

→ 체중을 조절할 수 있는 건강식품을 개발했는데, 정말 체중을 조절하는 효과가 있는지 확인해보고자 한다.

대응표본 t검정

■ 가설 수립

조사의 목적이 체중 변화의 효과에 맞춰져 있으므로,

귀무가설 H_0는 체중 변화가 없다.

대립가설 H_1은 체중 변화가 없다.

$$H_0 \ : \ \overline{X_1} - \overline{X_2} = 0 \implies \text{체중 변화가 없다.}$$

$$H_1 \ : \ \overline{X_1} - \overline{X_2} \neq 0 \implies \text{체중 변화가 있다.}$$

대응표본 t검정

■ 표본통계량과 표준오차

복용 전의 표본을 X_1, 복용 후의 표본을 X_2, 표본의 크기를 n,

X_1 과 X_2 의 차이를 $X_1 - X_2 = d_x$ 라 하면,

복용 전/후 차이의 평균은 $\overline{d_X} = \frac{\sum(X_1 - X_2)}{n}$

복용 전/후 차이의 분산 $s^2_{d_X}$, 표준편차 s_{d_X}, 표준오차 $s_{\overline{d_X}}$ 는

$$s^2_{d_X} = \frac{\sum(d_X - \overline{d_X})^2}{n-1}, \quad s_{d_X} = \sqrt{\frac{\sum(d_X - \overline{d_X})^2}{n-1}}, \quad s_{\overline{d_X}} = \frac{s_{d_X}}{\sqrt{n}}$$

대응표본 t검정

■ 표본통계량과 표준오차 계산

복용 전/후 차이의 평균, 분산, 표준편차, 표준오차를 구하면

$$\overline{d_X} = \frac{(75-73)+(74-74)+\cdots+(68-67)}{30} = 3.900$$

$$s_{d_X}^2 = \frac{(2-3.9)^2+(0-3.9)^2+\cdots+(1-3.9)^2}{30-1} = 13.197$$

$$s_{d_X} = \sqrt{\frac{(2-3.9)^2+(0-3.9)^2+\cdots+(1-3.9)^2}{30-1}} = 3.633$$

$$s_{\overline{d_X}} = \frac{3.633}{\sqrt{30}} = 0.633$$

대응표본 t검정

■ 가설검정

수립된 가설에 대한 검정은 자유도를 (n-1)로 하는 t분포를 이용

신뢰구간 100(1-α)에서의 검정통계량은 X_1-X_2 에 대해

$$t = \frac{d_X - \overline{d_X}}{s_{d_X}/\sqrt{n}} = \frac{d_X}{s_{d_X}/\sqrt{n}}$$

대응표본 t검정

■ 가설검정

유의수준은 0.05이며 자유도가 29인 $t_{(29,0.025)}$ 는 2.045

그러므로 양측검정의 신뢰구간은

$$d_X - t_{\left(n-1,\frac{\alpha}{2}\right)} \cdot \frac{s_{d_X}}{\sqrt{n}} \le t \le d_X + t_{\left(n-1,\frac{\alpha}{2}\right)} \cdot \frac{s_{d_X}}{\sqrt{n}}$$

$$3.900 - 2.045 \cdot \frac{3.633}{\sqrt{30}} \le t \le 3.900 + 2.045 \cdot \frac{3.633}{\sqrt{30}}$$

$$3.900 - 2.045 \cdot 0.663 \le t \le 3.900 + 2.045 \cdot 0.663$$

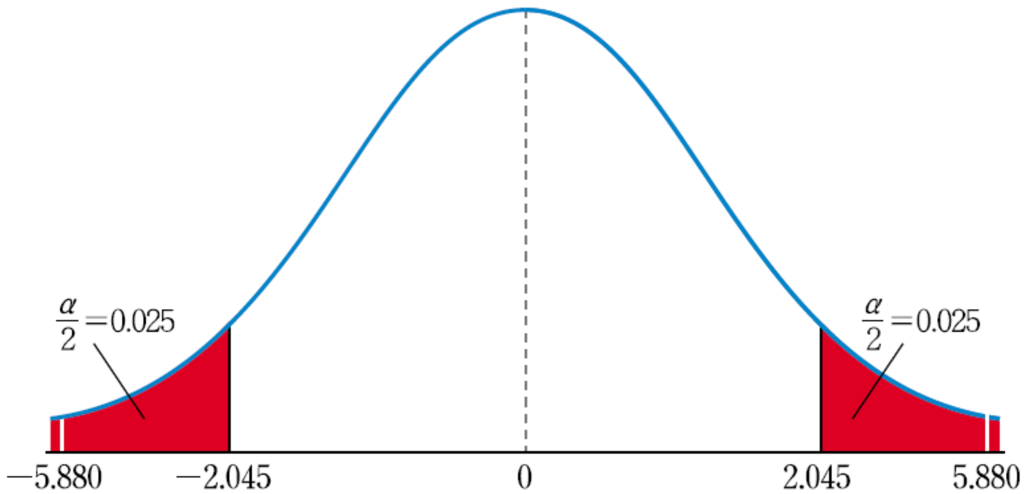
$$2.544 \le t \le 5.256$$

대응표본 t검정

■ 가설검정 Cont'd

95%의 신뢰구간에서 검정통계량을 구하면

$$t = \frac{d_X}{s_{d_X}/\sqrt{n}} = \frac{3.9}{0.663} = 5.880$$



5.880은 신뢰구간의 범위에 포함되지 않는다. 즉 귀무가설을 기각하고 대립가설을 채택

대응표본 t검정

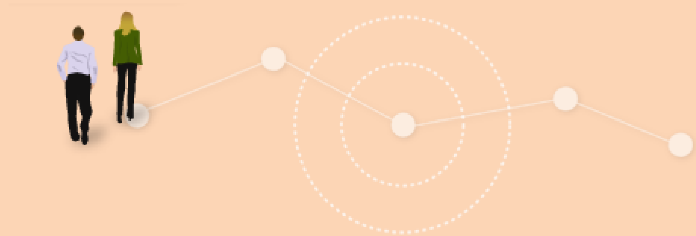
예제 8-1 대응표본의 가설검정

준비파일 | 8장_대응표본의 가설검정.xlsx

[표 8-1]을 바탕으로 A제약회사에서 개발한 건강식품이 체중조절에 효과가 있는지 Excel을 이용하여 검정하라. 단, 유의수준은 5%이다.

02 두 모집단의 평균 차이에 대한 가설검정(독립표본)

:: Keywords 독립표본 | 독립표본 t검정

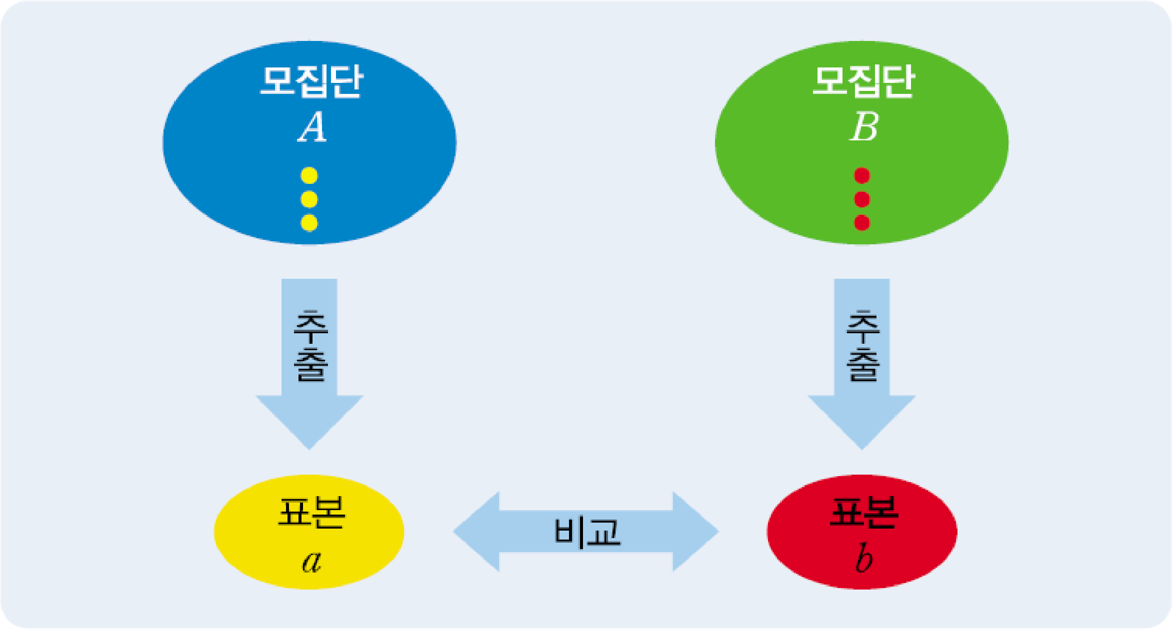


독립표본

■ 독립표본(independent sample)

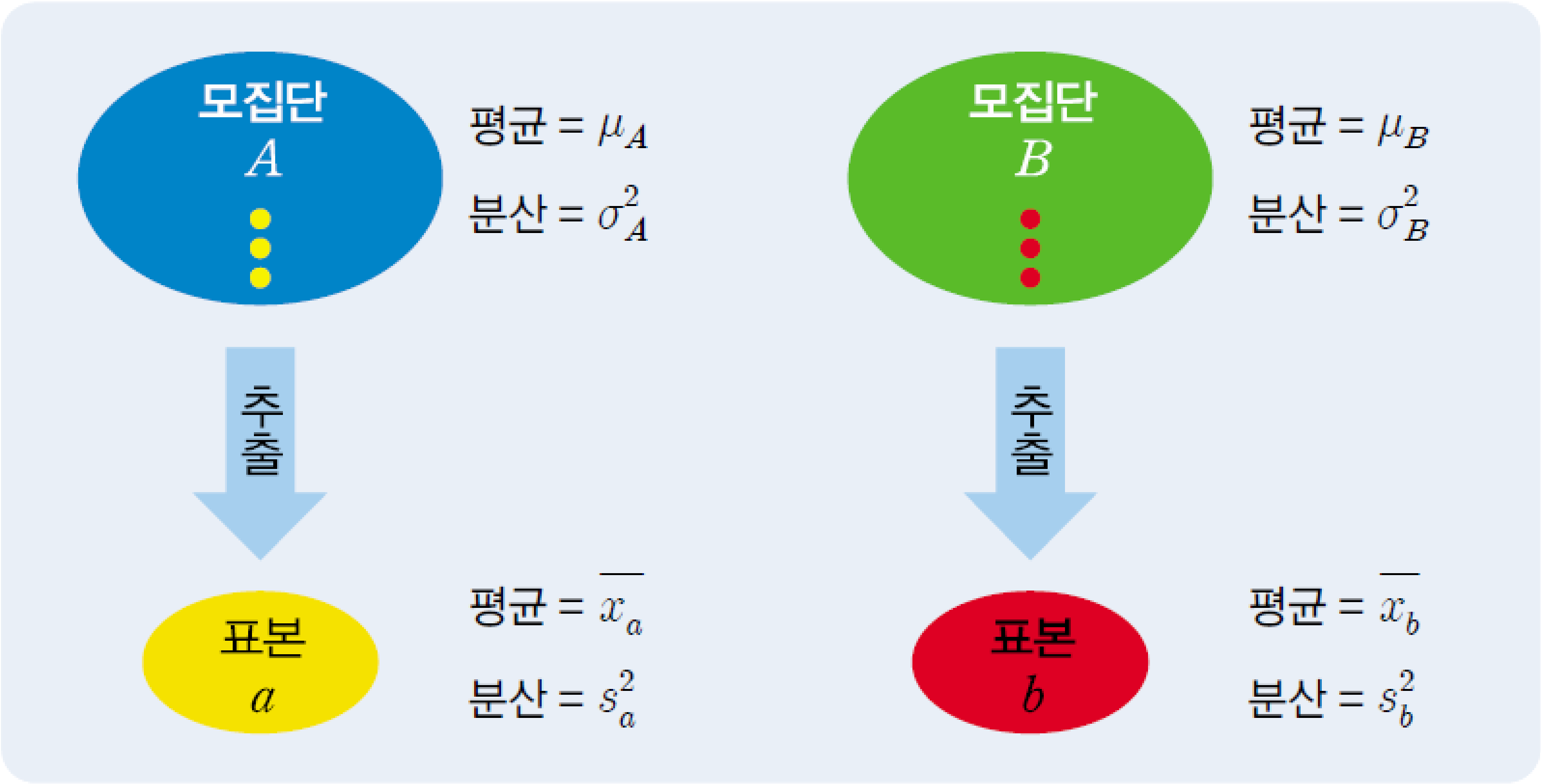
두 모집단으로부터 각각 추출할 때,

각 표본을 구성하는 인자가 서로 상관없는 표본



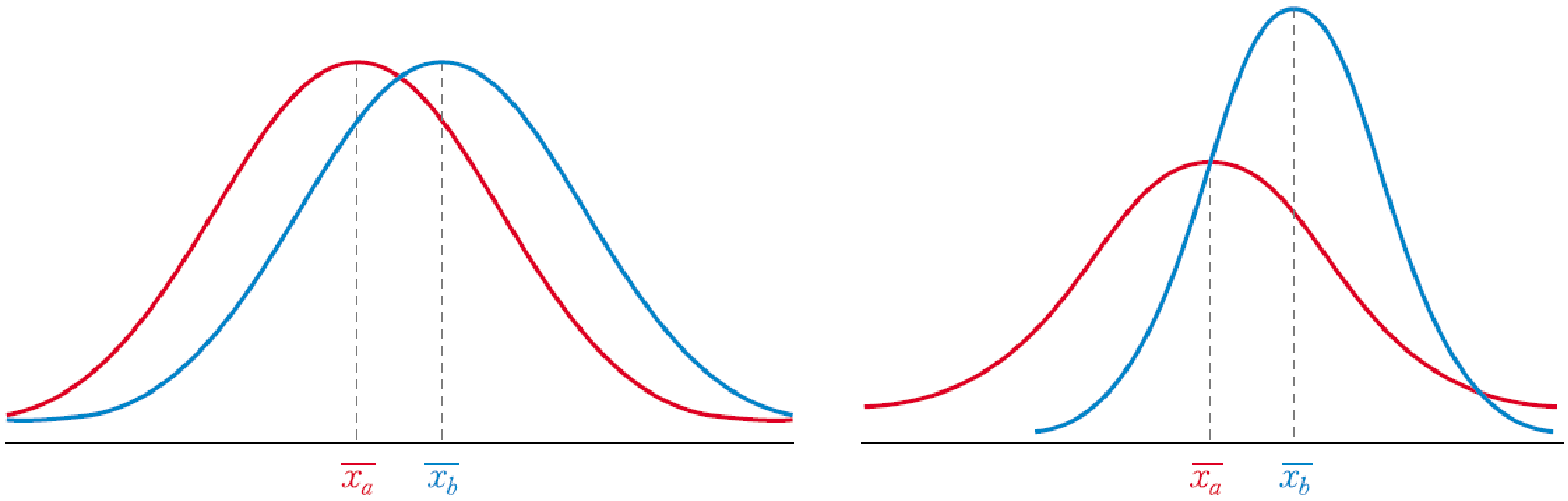
독립표본

■ 독립표본의 모수와 통계량



독립표본

■ 표본 a와 b의 분포는 분산이 같은 경우와 분산이 다른 경우로 나누어 생각



독립표본

독립표본은 표본의 개수와 분산의 동일성 여부에 따라 다음의 네 가지 경우로 구분

- 표본의 개수가 충분하고, 모분산이 동일한 경우
- 표본의 개수가 충분하고, 모분산 간의 동일성을 모르는 경우
- 표본의 개수가 충분하지 않고, 모분산이 동일한 경우
- 표본의 개수가 충분하지 않고, 모분산 간의 동일성을 모르는 경우

독립표본

■ 표본의 개수는 충분하고, 모분산이 동일한 경우

두 집단의 모평균 $\mu_{A,}$ $\mu_{B,}$ 를 비교할 때 σ_A^2, σ_B^2 을 알고 있다면,

$\mu_A - \mu_B = \delta$ 의 추정량은 $\overline{x_A} - \overline{x_B}$

표준화하면 $$z = \frac{(\overline{x_A} - \overline{x_B}) - (\mu_A - \mu_B)}{\sqrt{\dfrac{\sigma_A^2}{n_A} + \dfrac{\sigma_B^2}{n_B}}} ,\ N(0, 1)$$

독립표본

■ 표본의 개수는 충분하고, 모분산이 동일한 경우

$\mu_A - \mu_B$ 에 대한 100(1-α)%의 신뢰구간은

$$(\overline{x_A} - \overline{x_B}) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} \le \mu_A - \mu_B \le (\overline{x_A} - \overline{x_B}) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

독립표본

■ 표본의 개수는 충분하고, 모분산이 동일한 경우

가설을 수립하면

- $H_0: \mu_A - \mu_B = 0,\ \ H_1: \mu_A - \mu_B \neq 0 \ \Rightarrow |\mu_A - \mu_B| > z_{\frac{\alpha}{2}}$ 이면 H_0 기각, H_1 채택
- $H_0: \mu_A - \mu_B = 0,\ \ H_1: \mu_A - \mu_B > 0 \ \Rightarrow \mu_A - \mu_B > z_{\alpha}$ 이면 H_0 기각, H_1 채택
- $H_0: \mu_A - \mu_B = 0,\ \ H_1: \mu_A - \mu_B < 0 \ \Rightarrow \mu_A - \mu_B < -z_{\alpha}$ 이면 H_0 기각, H_1 채택

귀무가설인 $\mu_A - \mu_B$ =0의 가설검정을 위한 검정통계량은

$$z = \frac{(\overline{x}_A - \overline{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} = \frac{(\overline{x}_A - \overline{x}_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

독립표본

■ 표본의 개수는 충분하고, 모분산 간의 동일성을 모르는 경우

두 집단의 모평균 $\mu_{A,}\ \mu_B$ 를 비교할 때 각 분산 $\sigma_A^2,\ \sigma_B^2$을 모르는 경우

표본의 개수가 충분하다면 표본분산 $s_A^2,\ s_B^2$ 으로
분산 $\sigma_A^2,\ \sigma_B^2$ 을 대체할 수 있다.

표본의 개수가 충분히 크다는 가정하에 표본평균의 차 $\overline{x_A} - \overline{x_B}$ 에 대한
기대값과 분산은

$$E(\overline{x_A} - \overline{x_B}) = E(\overline{x_A}) - E(\overline{x_B}) = \mu_A - \mu_B$$

$$\begin{aligned} s^2(\overline{x_A} - \overline{x_B}) &= s^2(\overline{x_A}) + s^2(\overline{x_B}) - 2\,Cov(\overline{x_A},\ \overline{x_B}) \\ &= \frac{s_{x_A}^2}{n_A} + \frac{s_{x_B}^2}{n_B} \quad (n_A,\ n_B \geq 30) \end{aligned}$$

독립표본

■ 표본의 개수는 충분하고, 모분산 간의 동일성을 모르는 경우

$\mu_A - \mu_B = 0$의 추정량 $\overline{x}_A - \overline{x}_B$를 표준화하면

$$z = \frac{(\overline{x}_A - \overline{x}_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}, \; N(0, 1)$$

독립표본

■ 표본의 개수는 충분하고, 모분산 간의 동일성을 모르는 경우

$\mu_A - \mu_B$ 에 대한 100(1-α)% 의 신뢰구간은

$$(\overline{x_A} - \overline{x_B}) - z_{\frac{\alpha}{2}} \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} \leq \mu_A - \mu_B \leq (\overline{x_A} - \overline{x_B}) + z_{\frac{\alpha}{2}} \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

독립표본

■ 표본의 개수는 충분하고, 모분산 간의 동일성을 모르는 경우

가설을 수립하면

- $H_0: \mu_A - \mu_B = 0,\ \ H_1: \mu_A - \mu_B \neq 0 \ \Rightarrow |\mu_A - \mu_B| > z_{\frac{\alpha}{2}}$ 이면 H_0 기각, H_1 채택
- $H_0: \mu_A - \mu_B = 0,\ \ H_1: \mu_A - \mu_B > 0 \ \Rightarrow \mu_A - \mu_B > z_{\alpha}$ 이면 H_0 기각, H_1 채택
- $H_0: \mu_A - \mu_B = 0,\ \ H_1: \mu_A - \mu_B < 0 \ \Rightarrow \mu_A - \mu_B < -z_{\alpha}$ 이면 H_0 기각, H_1 채택

귀무가설인 $\mu_A - \mu_B = 0$의 가설검정을 위한 검정통계량은

$$z = \frac{(\overline{x_A} - \overline{x_B}) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{(\overline{x_A} - \overline{x_B})}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

독립표본

■ 표본의 개수는 충분하지 않고, 모분산이 동일한 경우

가정 : 두 모집단이 정규분포이거나 정규분포와 비슷하다.

모집단은 $N(\mu_A, \sigma_A^2)$과 $N(\mu_B, \sigma_B^2)$ 이며, $\sigma_A^2 = \sigma_B^2 = \sigma^2$ 이므로
두 모집단이 가지는 공통분산 σ^2 을 추정

모분산을 모르므로 표본분산으로 대체하여
$s_A^2 = s_B^2 = s^2$ 과 같이 분산이 서로 동일하다고 추정

독립표본

■ 표본의 개수는 충분하지 않고, 모분산이 동일한 경우

합동 분산 추정량 (pooled variance estimator)

서로 독립적인 모집단들의 공통 분산의 불편추정량(unbiased estimator)

s_p^2 으로 나타냄

$$s_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)\cdot s_i^2}{\sum_{i=1}^{k}(n_i - 1)} \quad (k: \text{모집단 개수},\ n: \text{표본 크기},\ s^2: \text{표본분산})$$

독립표본

■ 표본의 개수는 충분하지 않고, 모분산이 동일한 경우

모집단 A와 B를 대상으로 한 독립표본에 대한 검정을 다루고 있으므로,
표본 A와 표본 B의 합동 분산 추정량은

$$s_p^2 = \frac{(n_A - 1) \cdot s_A^2 + (n_B - 1) \cdot s_B^2}{(n_A - 1) + (n_B - 1)} = \frac{(n_A - 1) \cdot s_A^2 + (n_B - 1) \cdot s_B^2}{n_A + n_B - 2}$$

독립표본

■ 표본의 개수는 충분하지 않고, 모분산이 동일한 경우

$\mu_A - \mu_B$ 에 대한 $100(1-\alpha)\%$의 신뢰구간은

$$\begin{aligned}(\overline{x_A} - \overline{x_B}) - t_{\left(n_A + n_B - 2,\ \frac{\alpha}{2}\right)} s_p \sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} &\le \mu_A - \mu_B \\ &\le (\overline{x_A} - \overline{x_B}) + t_{\left(n_A + n_B - 2,\ \frac{\alpha}{2}\right)} s_p \sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\end{aligned}$$

독립표본

■ 표본의 개수는 충분하지 않고, 모분산이 동일한 경우

가설을 수립하면

- $H_0: \mu_A - \mu_B = 0,\ \ H_1: \mu_A - \mu_B \neq 0\ \Rightarrow |\mu_A - \mu_B| > z_{\frac{\alpha}{2}}$ 이면 H_0 기각, H_1 채택
- $H_0: \mu_A - \mu_B = 0,\ \ H_1: \mu_A - \mu_B > 0\ \Rightarrow \mu_A - \mu_B > z_{\alpha}$ 이면 H_0 기각, H_1 채택
- $H_0: \mu_A - \mu_B = 0,\ \ H_1: \mu_A - \mu_B < 0\ \Rightarrow \mu_A - \mu_B < -z_{\alpha}$ 이면 H_0 기각, H_1 채택

독립표본

■ 표본의 개수는 충분하지 않고, 모분산이 동일한 경우

귀무가설인 $\mu_A - \mu_B = 0$의 가설검정을 위한 검정통계량은

$$t_{\left(n_A + n_B - 2,\ \frac{\alpha}{2}\right)} = \frac{(\overline{x}_A - \overline{x}_B) - (\mu_A - \mu_B)}{s_p \sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{(\overline{x}_A - \overline{x}_B)}{s_p \sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

독립표본

■ 표본의 개수는 충분하지 않고, 모분산 간의 동일성을 모르는 경우

표본평균의 차 $\overline{x_A} - \overline{x_B}$ 에 대한 기대값과 분산은

$$E(\overline{x_A} - \overline{x_B}) = E(\overline{x_A}) - E(\overline{x_B}) = \mu_A - \mu_B$$

$$s^2(\overline{x_A} - \overline{x_B}) = s^2(\overline{x_A}) + s^2(\overline{x_B}) - 2Cov(\overline{x_A}, \overline{x_B}) = \frac{s_{x_A}^2}{n_A} + \frac{s_{x_B}^2}{n_B}$$

독립표본

■ 표본의 개수는 충분하지 않고, 모분산 간의 동일성을 모르는 경우

$\mu_A - \mu_B$ 에 대한 $100(1-\alpha)\%$의 신뢰구간은

$$(\overline{x}_A - \overline{x}_B) - t_{\left(n_A + n_B - 2,\ \frac{\alpha}{2}\right)} \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} \le \mu_A - \mu_B \le (\overline{x}_A - \overline{x}_B) + t_{\left(n_A + n_B - 2,\ \frac{\alpha}{2}\right)} \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

독립표본

■ 표본의 개수는 충분하지 않고, 모분산 간의 동일성을 모르는 경우

가설은

$$H_0: \mu_A - \mu_B = 0,\ \ H_1: \mu_A - \mu_B \neq 0$$

$$H_0: \mu_A - \mu_B = 0,\ \ H_1: \mu_A - \mu_B > 0$$

$$H_0: \mu_A - \mu_B = 0,\ \ H_1: \mu_A - \mu_B < 0$$

귀무가설인 $\mu_A - \mu_B = 0$의 가설검정을 위한 검정통계량은

$$t_{\left(n_A + n_B - 2,\ \frac{\alpha}{2}\right)} = \frac{(\overline{x_A} - \overline{x_B}) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{(\overline{x_A} - \overline{x_B})}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

'표본의 개수가 충분하지 않고, 모분산이 동일한 경우'와 같다.

독립표본 t검정

Ex. A사와 B사의 알카라인 건전지

단위 : 시간

번호	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A사	18	16	17	15	14	19	16	15	18	15	16	17	15	14	19	12	16
B사	16	15	16	14	15	18	13	15	12	17	15	16	18	12	14	15	16

→ A사와 B사의 알카라인 건전지의 작동 시간에 차이가 있는지 유의수준 5%에서 검정해보고자 한다.

독립표본 t검정

■ 가설의 수립

조사의 목적이 제조사별로 알카라인 건전지의 품질 차이에 맞춰져 있으므로,

귀무가설 H_0 는 제조사별 건전지의 사용 시간에 차이가 없다.
대립가설 H_1 은 제조사별로 건전지의 사용 시간에 차이가 있다.

A사의 건전지 표본의 평균 사용 시간을 $\overline{x_A}$
B사의 건전지 표본의 평균 사용 시간을 $\overline{x_B}$ 라 하면

H_0 : $\overline{x_A} = \overline{x_B}$ $\Longrightarrow$ A사와 B사의 건전지 사용 시간에 차이가 없다.
H_1 : $\overline{x_A} \neq \overline{x_B}$ $\Longrightarrow$ A사와 B사의 건전지 사용 시간에 차이가 있다.

독립표본 t검정

■ 표본통계량과 합동 분산 추정량

표본의 크기를 n이라 하고, $\overline{x}_A$과 $\overline{x}_B$ 의 차이를 $E(\overline{x}_A - \overline{x}_B) = \mu_A - \mu_B$ 라 하면,

σ^2에 대한 합동 분산 추정량 s_p^2 은

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A - 1) + (n_B - 1)}$$

독립표본 t검정

■ 표본통계량과 합동 분산 추정량

$\sigma^2(\overline{x_A}-\overline{x_B})$는 $s^2(\overline{x_A}-\overline{x_B})$ 이고, 이에 대한 불편 추정량은 $s^2(\overline{x_A}-\overline{x_B}) = s_p^2\left(\frac{1}{n_A}-\frac{1}{n_B}\right)$

$\overline{x_A}=16.1$, $\overline{x_B}=15.2$, $\overline{x_B}=15.2$, $s_B^2=2.800$ 이므로 합동 분산 추정량을 구하면

$$s_p^2 = \frac{(20-1)\cdot 3.253 + (20-1)\cdot 2.800}{40-2} = 3.026$$

이때 $s_p = \sqrt{3.026} = 1.74$는 s_A^2 과 s_B^2 을 각자의 자유도로 가중한 공통 분산의 불편추정량(unbiased estimator)

독립표본 t검정

■ 가설검정

수립된 가설에 대한 검정은 자유도를 $(n_A + n_B - 2)$ 로 하는 t분포를 이용 검정통계량은

$$t_{\left(n_A + n_B - 2,\, \frac{\alpha}{2}\right)} = \frac{(\overline{x_A} - \overline{x_B})}{s_p \sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

양측검정의 신뢰구간은

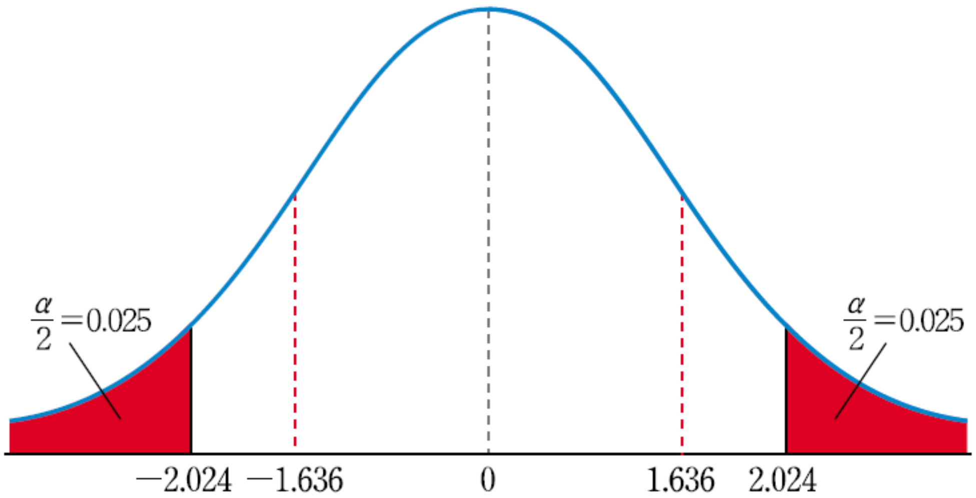
$$(\overline{x_A} - \overline{x_B}) - t_{\left(n_A + n_B - 2,\, \frac{\alpha}{2}\right)} s_p \sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} \le \mu_A - \mu_B$$
$$\le (\overline{x_A} - \overline{x_B}) + t_{\left(n_A + n_B - 2,\, \frac{\alpha}{2}\right)} s_p \sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

독립표본 t검정

■ 가설검정

신뢰구간 95%에서의 검정통계량을 구하면

$$t_{(38,\,0.025)} = \frac{16.1 - 15.2}{1.740 \cdot \sqrt{\frac{1}{20} + \frac{1}{20}}} = 1.636$$



독립표본 t검정

■ 가설검정

양측검정의 신뢰구간을 구하면

$$(16.1-15.2)-1.636\cdot 1.740\cdot\sqrt{\frac{1}{20}+\frac{1}{20}}\le \mu_A-\mu_B$$
$$\le (16.1-15.2)+1.636\cdot 1.740\cdot\sqrt{\frac{1}{20}+\frac{1}{20}}$$

$$0\le \mu_A-\mu_B\le 1.800$$

독립표본 t검정

예제 8-2 **독립표본의 가설검정** 준비파일 | 8장_독립표본의 가설검정.xlsx

[표 8-3]을 바탕으로 A사와 B사의 알카라인 건전지의 사용 시간에 정말 차이가 있는지 Excel을 이용하여 검정하라. 단, 유의수준은 5%이다.