10. 두 모집단의 추론

담당교수 : 김 덕 기

toby123@cbnu.ac.kr

두 모집단의 평균비교 분석

Z, T ($H_0: \mu_A - \mu_B = 0$
$H_1:$ 〃 $\neq 0$

B, Z ($H_0: p_A - p_B = 0$
$H_1:$ 〃 $\neq 0$

X^2, F ($H_0: \sigma_A - \sigma_B = 0$
$H_1:$ 〃 $\neq 0$

$X_1, \cdots, X_m \sim m$
$Y_1 \cdots Y_n \sim n$
$\bar{X} \sim (\mu_X, \frac{\sigma^2}{m})$
$\bar{Y} \sim ($ 〃 $)$
$H_0\ \sigma_A^2 = \sigma_B^2$

$H_0: \mu_A \quad \mu_B$

두 표본집단
~~두 모집단~~의 평균 비교 분석

모집단 2개
독립표본인 경우

분산의 동일성 검정 수행
F-검정 : 분산에 대한 두 집단

등분산 $= \sigma^2 \neq$ 이분산

t-검정 : 등분산 가정 두 집단

t-검정 : 이분산 가정 두 집단

모집단 1개
대응
종속
쌍체표본인 경우

t-검정 : 쌍체비교

독립표본과 대응표본.

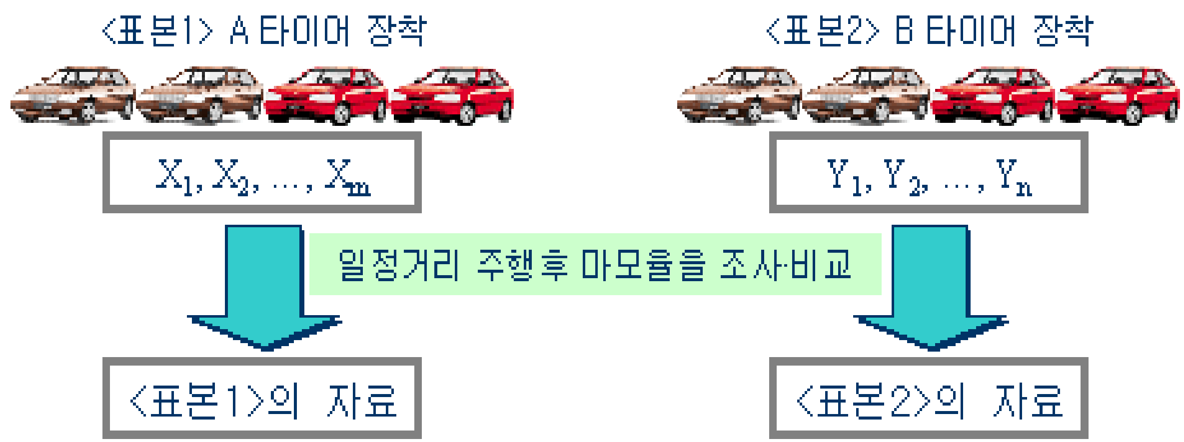
- 독립표본과 대응표본
  - 두 모집단의 비교를 위해서는 각 모집단에서 하나씩의 표본을 추출하게 되는데, 이 두 표본을 서로 독립적으로 추출할 것인지 아닌지에 따라 분석방법이 달라짐

  [예1] 두 회사 타이어의 마모율을 비교한다고 할 때, 두 가지의 실험방법이 가능

- **방법1** 두 표본자료의 차량이 서로 다르며 상관관계가 없다. 이러한 방법으로 만들어진 두 표본을 독립표본 이라 함.

$X_1 \cdots X_{40} \sim A$

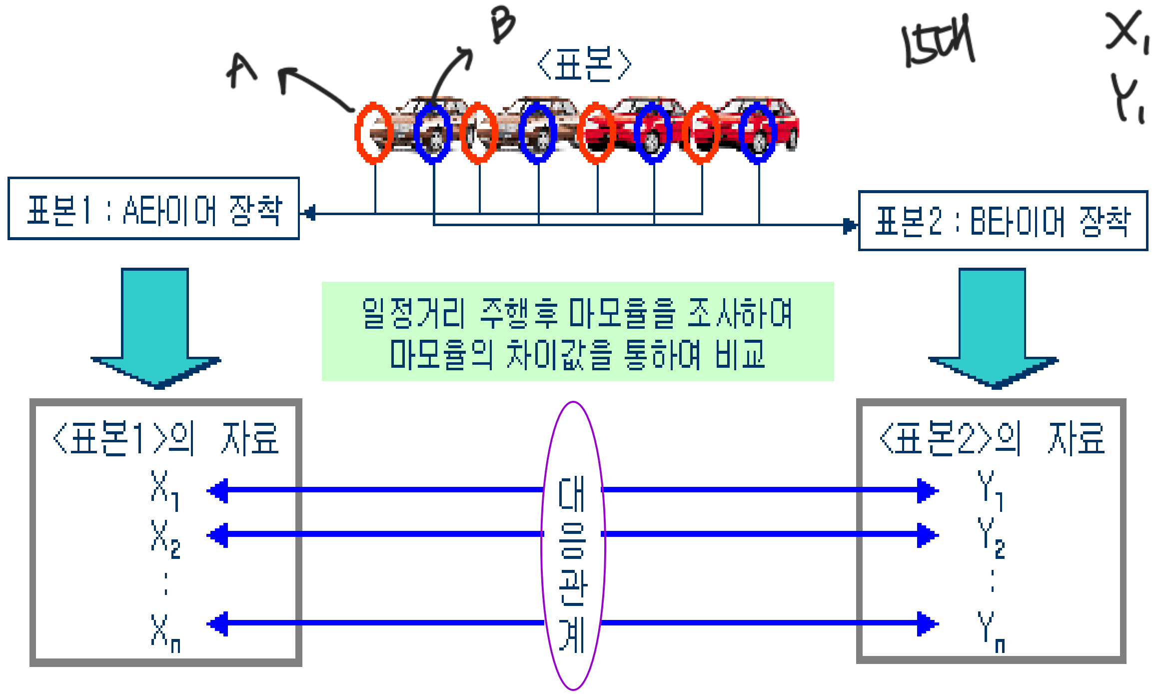
CLT.



타이어 품질에 따른 순수한 마모율 외에 운전자 체중, 운전습관, 차량무게, 차량품질 등 다른 요인의 영향때문에 타이어 품질의 차이에 의한 영향을 분석해 내기 어렵다.

(...계속)

- **방법2** | 두 표본자료의 차량이 같고 두 회사 타이어를 한쪽씩 각각 장착하는 방법으로 실험한 경우.



15대

$X_1 \cdots X_{30}$

$Y_1 \cdots Y_{30}$

CLT $\bar{X} - \bar{Y} \sim$ Normal

➡ 이러한 방법으로 만들어진 경우의 두 표본을 **쌍체 또는 대응표본(paired sample)**이라 함

독립표본(예)

(방법1) 16명을 임의로 두 그룹으로 나눈 후, 각각 학습방법 A와 학습방법 B에 의한 교육을 모두 받게 한 후, 각각의 이해도를 측정. 데이터의 형태는 다음과 같다.

사람	학습방법 A	학습방법 B
1	X_1	
2	X_2	
3		Y_1
4	X_3	
5		Y_2
6	X_4	
7		Y_3
8		Y_4
9		Y_5
10		6
11	X_5	
12	X_6	
13		Y_7
14	X_7	
15		Y_8
16	X_8	

➡ 이 경우 학습방법 A와 학습방법 B 교육을 받은 대상자가 다르다. 즉, 데이터가 서로 독립

➡ 이러한 방법으로 만들어진 두 표본을 <u>독립표본(independent sample)</u>이라 함

➡ 개인의 능력차로 인해 실제로 차이가 없음에도 불구하고 차이가 있다고 판단할 가능성이 있음

(독립 표본 시행하기 가능)

동질적 가능

사전 A=B ~ 사후 A≠B

70대 70대

독립 표본, 대응 표본으로 각각 분석한 결과에 대해 장단점 서술 문제 나옴.

대응표본(예)

(방법2) 8명을 임의로 선정해 학습방법 A와 학습방법 B에 의한 교육을 모두 받게 한 후, 각각의 이해도를 측정. 데이터의 형태는 다음과 같다.

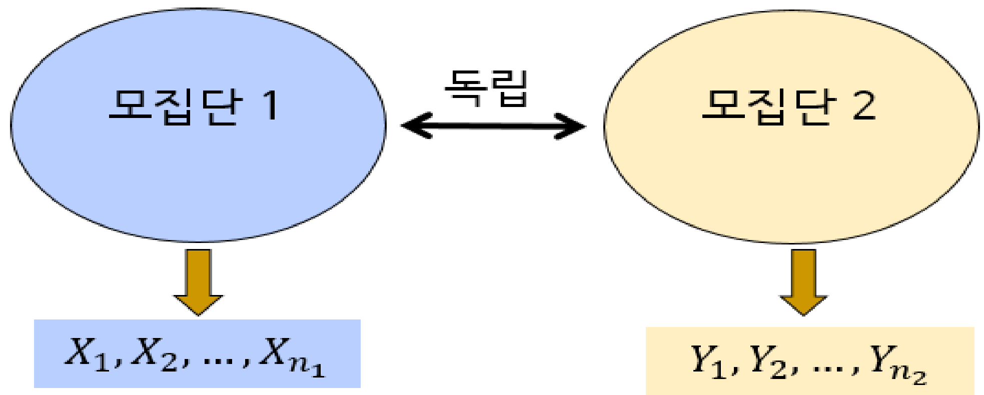
사람	학습방법 A	학습방법 B
1	X_1	Y_1
2	X_2	Y_2
3	X_3	Y_3
4	X_4	Y_4
5	X_5	Y_5
6	X_6	Y_6
7	X_7	Y_7
8	X_8	Y_8

> 한 사람이 학습방법 A 및 학습방법 B 교육을 모두 받았다. 데이터가 쌍의 형태를 띠고 있음

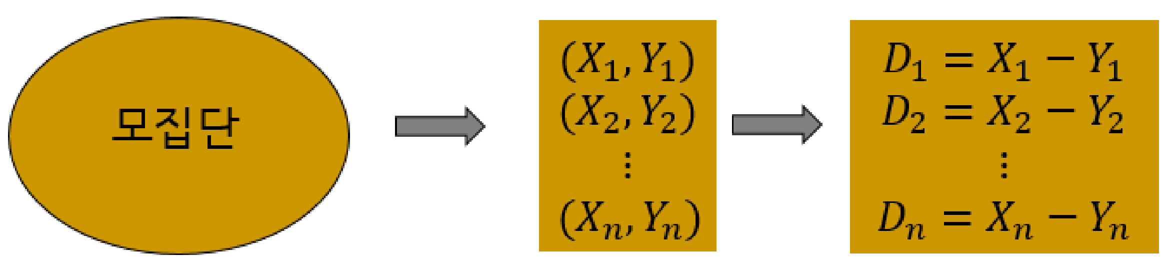
> 이러한 방법으로 만들어진 경우의 두 표본을 <u>쌍체 또는 대응표본(paired sample)</u>이라 함

독립 표본, 대응 표본

- 별개의 모집단에서 얻어진 두 개의 독립표본으로 간주한다.



- 각 쌍에 속한 두 실험 단위는 독립이 아니다.



두 개의 독립 표본의 점 추정

$$\begin{cases} X_1, X_2, \dots, X_{n_1}: \text{평균이 } \mu_1 \text{이고 분산이 } {\sigma_1}^2 \text{인 모집단에서 임의추출한 자료} \\ Y_1, Y_2, \dots, Y_{n_2}: \text{평균이 } \mu_2 \text{이고 분산이 } {\sigma_2}^2 \text{인 모집단에서 임의추출한 자료} \end{cases}$$

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \qquad {S_1}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1} \left(X_i - \overline{X}\right)^2$$

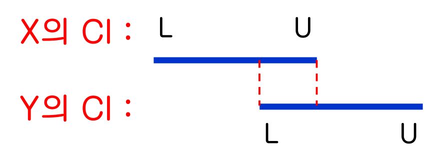
$$\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i, \qquad {S_2}^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2} \left(Y_i - \overline{Y}\right)^2$$

- 점추정
  - $\mu_1 - \mu_2$의 추정량: $\hat{\mu}_1 - \hat{\mu}_2 = \overline{X} - \overline{Y}$
  - 추정량의 표준오차: $S.E.(\hat{\mu}_1 - \hat{\mu}_2) = S.E.\left(\overline{X} - \overline{Y}\right) = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$
  - 추정된 표준오차: $S.E.(\widehat{\hat{\mu}_1 - \hat{\mu}_2}) = \sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}}$

표본의 크기가 충분히 클 때 $\boldsymbol{\mu_1 - \mu_2}$에 대한 추론

- 두 표본의 크기 n_1과 n_2가 충분히 크면 $(n_1 \geq 30, n_2 \geq 30)$

$$\bar{X} \sim N\left(\mu_1, \frac{{\sigma_1}^2}{n_1}\right), \bar{Y} \sim N\left(\mu_2, \frac{{\sigma_2}^2}{n_2}\right)$$



- 두 표본은 독립이므로

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}\right) \rightarrow \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}} \sim N(0, 1)$$

- 표본의 크기가 충분히 크면 ${S_1}^2 \approx {\sigma_1}^2, {S_2}^2 \approx {\sigma_2}^2$이므로

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}}} \sim N(0, 1)$$

$\mu_1 - \mu_2$에 대한 신뢰구간

- 표본의 크기가 충분히 클 때, $\mu_1 - \mu_2$에 대한 $100(1-\alpha)\%$ 신뢰구간:

$$(\overline{X} - \overline{Y}) \pm z_{\alpha/2}\sqrt{\frac{S_1{}^2}{n_1} + \frac{S_2{}^2}{n_2}}$$

두 평균 차이의 분포를 이용해 하나의 신뢰구간(CI)을 설정하여 판정

- 위의 신뢰구간은 표본의 크기가 클 때의 신뢰 구간의 형태 :

$$(\text{추정량}) \pm z_{\alpha/2} \times (\text{추정량의 표준오차})$$

예제 3) 과자를 담는 두 기계 A와 B에서 생산한 과자 한 봉지의 평균 무게를 각각 μ_A와 μ_B라고 할 때, $\mu_A - \mu_B$의 95% 신뢰구간은?

- 자료: $\begin{cases} \text{기계 } A: \ n_1 = 50,\ \overline{x} = 453,\ s_1 = 80 \\ \text{기계 } B: \ n_2 = 100,\ \overline{y} = 401,\ s_2 = 60 \end{cases}$

H_0 : 두 기계의 평균 차이가 없다.
H_1 : 두 기계의 평균 차이가 있다.

- $\mu_A - \mu_B$에 대한 95% 신뢰구간:

$$(\overline{x} - \overline{y}) \pm z_{0.025}\sqrt{\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}} = (453 - 401) \pm 1.96 \times \sqrt{\frac{80^2}{50} + \frac{60^2}{100}}$$

$$= 52 \pm 25.1 \quad \Rightarrow \ (\mathbf{25.9}, \mathbf{77.1})$$

$\mu_1 - \mu_2$에 대한 가설검정

- 가설:

$$H_0: \mu_1 - \mu_2 = \delta_0 \quad vs \quad \begin{cases} (i)\ H_1: \mu_1 - \mu_2 < \delta_0 \\ (ii)\ H_1: \mu_1 - \mu_2 > \delta_0 \\ (iii)\ H_1: \mu_1 - \mu_2 \neq \delta_0 \end{cases}$$

➔ 평균 차이가 작다.
➔ 평균 차이가 크다.
➔ 평균 차이가 있다.

➔ 보통 $\boldsymbol{\delta_0 = 0}$

- 검정통계량(표본의 크기가 클 때)

$$Z = \frac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}} \qquad Z = \frac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}}} \sim N(0,1)$$

위 식에서 $\mu_1 - \mu_2$ 대신 δ_0를 사용한 점에 유의할 것

기각력 판정 : example

- (i) $H_1: \mu_1 - \mu_2 < \delta_0$일 때, $R: Z = \dfrac{(\overline{X}-\overline{Y})-\delta_0}{\sqrt{\frac{s_1{}^2}{n_1}+\frac{s_2{}^2}{n_2}}} \le -z_\alpha$
- (ii) $H_1: \mu_1 - \mu_2 > \delta_0$일 때, $R: Z = \dfrac{(\overline{X}-\overline{Y})-\delta_0}{\sqrt{\frac{s_1{}^2}{n_1}+\frac{s_2{}^2}{n_2}}} \ge z_\alpha$
- (iii) $H_1: \mu_1 - \mu_2 \neq \delta_0$일 때, $R: |Z| = \dfrac{|(\overline{X}-\overline{Y})-\delta_0|}{\sqrt{\frac{s_1{}^2}{n_1}+\frac{s_2{}^2}{n_2}}} \ge z_{\alpha/2}$

➔ n이 충분히 큰 경우

➔ 분산 모름 : T 통계량 ~ T-표준화 ~Normal

예제 4) 예제 3에서 두 기계에서 생산한 과자의 무게에 차이 유의수준 1% 검정

자료: $\begin{cases} \text{기계 } A:\ n_1 = 50,\ \overline{x} = 453,\ s_1 = 80 \\ \text{기계 } B:\ n_2 = 100,\ \overline{y} = 401,\ s_2 = 60 \end{cases}$ $H_0: \mu_1 - \mu_2 = 0 \quad vs \quad H_1: \mu_1 - \mu_2 \neq 0$

- 검정 통계량:

$$Z = \frac{(\overline{X}-\overline{Y})-\delta_0}{\sqrt{\frac{s_1{}^2}{n_1}+\frac{s_2{}^2}{n_2}}} = \frac{453-401}{\sqrt{\frac{80^2}{50}+\frac{60^2}{100}}} = 4.06 > z_{0.005} = 2.58 \ \rightarrow \text{귀무가설 기각.}$$

➔ 예제3) CI 판정
➔ 예제4) 기각력 판정
➔ 추가로 P-value 판정

➔ **즉 두 기계에서 생산한 과자의 무게에 차이가 있다.**

두 집단의 공통분산 : σ^2의 합동추정량

- 가정
  - 두 모집단은 모두 정규분포를 따른다. (정규성 가정)
  - 두 모집단의 표준편차는 같다. ($\sigma_1 = \sigma_2 = \sigma$) (등분산성 가정)

표본표준편차를 이용하여 $\frac{1}{2} \le \frac{S_1}{S_2} \le 2$인지 확인 ➡ F-test 필요

$$\begin{cases} X_1, X_2, \dots, X_{n_1}: \text{정규모집단 } N(\mu_1, \sigma^2)\text{에서 임의추출한 자료} \\ Y_1, Y_2, \dots, Y_{n_2}: \text{정규모집단 } N(\mu_2, \sigma^2)\text{에서 임의추출한 자료} \end{cases} \Rightarrow \overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

σ^2의 합동 추정량(pooled estimator):

$$S_p{}^2 = \frac{1}{n_1+n_2-2}\left(\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{i=1}^{n_2}(Y_i - \overline{Y})^2\right)$$

$$= \frac{(n_1-1)S_1{}^2 + (n_2-1)S_2{}^2}{n_1+n_2-2}$$

등 분산의 경우 검정 통계량 & 신뢰구간

- [정리] 두 모집단이 정규분포를 따를 때

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

- 두 모집단이 정규분포를 따르고 분산이 같을 때, $\mu_1 - \mu_2$에 대한 $100(1-\alpha)\%$ 신뢰구간:

$$(\overline{X} - \overline{Y}) \pm t_{\alpha/2}(n_1 + n_2 - 2)S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

등 분산의 경우 신뢰구간 : example

예제 6) 목초의 종류에 따른 우유 생산량의 차에 대한 95% 신뢰구간은?

- 자료: $\begin{cases} \text{들판에서 말린 목초: } n_1 = 13,\ \bar{x} = 45.15,\ s_1 = 7.998 \\ \text{인공적으로 말린 목초: } n_2 = 12,\ \bar{y} = 42.25,\ s_2 = 8.740 \end{cases}$

- 표본의 크기가 작으므로 정규성 가정을 해야 한다.
- 두 표준편차의 비는 $\frac{s_1}{s_2} = \frac{7.998}{8.740} = 0.92$로 1에 가깝다.
- 합동추정량: $S_p = \sqrt{\frac{12\times 7.998^2 + 11\times 8.740^2}{13+12-2}} = 8.361$
- $\mu_1 - \mu_2$에 대한 95% 신뢰구간:

  $(\bar{x} - \bar{y}) \pm t_{0.025}(23) S_p \sqrt{\frac{1}{13} + \frac{1}{12}}$

H_0 : 두 사료에 평균 생산량에 차이가 없다.
H_1 : 두 사료에 평균 생산량에 차이가 있다.

$$= (45.15 - 42.25) \pm 2.069 \times 8.361 \times \sqrt{\frac{1}{13} + \frac{1}{12}} = 2.90 \pm 6.925$$

기각력 판정 : example

$$H_0: \mu_1 - \mu_2 = \delta_0 \quad vs \quad \begin{cases} (i)\ H_1: \mu_1 - \mu_2 < \delta_0 \\ (ii)\ H_1: \mu_1 - \mu_2 > \delta_0 \\ (iii)\ H_1: \mu_1 - \mu_2 \neq \delta_0 \end{cases}$$

$$t = \frac{(\overline{X} - \overline{Y}) - \delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

- (i) $H_1: \mu_1 - \mu_2 < \delta_0$일 때

  $$R: t = \frac{(\overline{X}-\overline{Y})-\delta_0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \le -t_\alpha(n_1 + n_2 - 2)$$

- (ii) $H_1: \mu_1 - \mu_2 > \delta_0$일 때

  $$R: t = \frac{(\overline{X}-\overline{Y})-\delta_0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \ge t_\alpha(n_1 + n_2 - 2)$$

- (iii) $H_1: \mu_1 - \mu_2 \neq \delta_0$일 때

  $$R: |t| = \frac{|(\overline{X}-\overline{Y})-\delta_0|}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \ge t_{\alpha/2}(n_1 + n_2 - 2)$$

예제 7) 들판에서 말린 목초를 먹은 젖소의 우유 생산량이 인공적으로 말린 목초를 먹은 젖소의 우유 생산량보다 많은지 유의수준 5%에서 검정

$$\begin{cases} \text{들판에서 말린 목초: } n_1 = 13, \overline{x} = 45.15, s_1 = 7.998 \\ \text{인공적으로 말린 목초: } n_2 = 12, \overline{y} = 42.25, s_2 = 8.740 \end{cases}$$

가설: $H_0: \mu_1 - \mu_2 = 0 \quad vs \quad H_1: \mu_1 - \mu_2 > 0$

$$t = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{45.15 - 42.15}{8.361 \times \sqrt{\frac{1}{13} + \frac{1}{12}}} = 0.866$$

$\boxed{t_{0.05}(23) = 1.714\text{ , }H_0\text{ 채택. 즉 우유 생산량에는 차이가 없다.}}$

두 정규모집단의 분산이 같지 않은 경우

- 두 정규모집단의 표준편차가 같지 않을 때 간단한 추론법 $(\sigma_1 \neq \sigma_2)$

$$t^* = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}}} \sim t(n^*)$$

단 $n^* = \min(n_1 - 1, n_2 - 1)$

- $\sigma_1 \neq \sigma_2$인 경우 분산의 합동추정량 ${S_p}^2$을 사용할 수 없다.
- 이 경우 사용하는 n^*는 보수적인 결과를 준다.
- 따라서 보다 사실에 근접한 결과를 얻으려면 다음의 자유도 근사법을 사용한다.
- 통계량 t^*의 자유도 ν^* 를 근사적으로 구하는 대표적 방법으로는 Satterthwaite 근사식이 있다. 보통 $n^* < \nu^* < n_1 + n_2 - 2$ 범위의 값을 갖는다.

기말 출제 X.

이 분산의 경우 가설검정 : example

예제 8) 두 지역의 집값에 차이가 있는지 유의수준 5%에서 검정

- 자료: $\begin{cases} \text{남쪽: } n_1 = 13,\ \bar{x} = 2.4,\ s_1 = 0.72 \\ \text{북쪽: } n_2 = 11,\ \bar{y} = 2.15,\ s_2 = 0.35 \end{cases}$

- 두 표준편차의 비는 $\frac{s_1}{s_2} = \frac{0.72}{0.35} = 2.06$으로 2배가 넘는다.
- 가설: $H_0: \mu_1 - \mu_2 = 0 \quad vs \quad H_1: \mu_1 - \mu_2 \neq 0$
- 검정통계량:

$$t^* = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1{}^2}{n_1} + \frac{S_2{}^2}{n_2}}} = \frac{2.4 - 2.15}{\sqrt{\frac{0.72^2}{13} + \frac{0.35^2}{11}}} = 1.107$$

- $n^* = \min(n_1 - 1, n_2 - 1) = \min(13 - 1, 11 - 1) = 10, t_{0.025}(10) =$ 2.228이므로 귀무가설을 기각하지 못함. 즉 두 지역의 집값에는 차이가 없다.

대응(쌍체)표본의 추론

- 자료:

쌍	처리 1	처리 2	차이
1	X_1	Y_1	$D_1 = X_1 - Y_1$
2	X_2	Y_2	$D_2 = X_2 - Y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	X_n	Y_n	$D_n = X_n - Y_n$

- $D_1, D_2, \dots, D_n$은 두 처리 효과의 차이를 나타낸다.
- $D_1, D_2, \dots, D_n$은 평균이 δ(두 처리 효과의 차이)이고 분산이 ${\sigma_D}^2$인 모집단에 임의추출한 표본이라고 할 수 있다.

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i\,, \qquad {s_D}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \overline{D}\right)^2$$

$\mu_D = \delta$ 에 대한 추론(표본의 크기가 클 때)

- 표본의 크기 n이 충분히 클 때

➔n이 충분히 크면 ➔ Z-통계량
➔n이 작고, 분산 모름 ➔ T-통계량

$$\frac{\sqrt{n}(\overline{D}-\delta)}{s_D} \sim N(0,1)$$

- δ에 대한 $100(1-\alpha)\%$ 신뢰구간:

$$\left(\overline{D} - z_{\alpha/2}\frac{s_D}{\sqrt{n}},\ \overline{D} + z_{\alpha/2}\frac{s_D}{\sqrt{n}}\right) \text{ or } \overline{X} \pm z_{\alpha/2}\frac{s_D}{\sqrt{n}}$$

- 가설 $H_0: \delta = \delta_0$ 에 대한 검정통계량:

$$Z = \frac{\sqrt{n}(\overline{D}-\delta_0)}{s_D} \sim N(0,1) \qquad (under\ H_0)$$

유의수준 α에서 기각역: $Z \le -z_\alpha, Z \ge z_\alpha, |Z| \ge z_{\alpha/2}$

$\mu_D = \delta$ 에 대한 추론(표본의 크기가 작을 때)

- 표본의 크기가 작을 때는 모집단에 대한 추가적인 가정이 필요하다.

- 가정: $D_1, D_2, \ldots, D_n$이 정규모집단 $N(\delta, {\sigma_D}^2)$에서 임의추출한 자료

$$\frac{\sqrt{n}(\overline{D}-\delta)}{s_D} \sim t(n-1) \sim \begin{cases} R: t = \frac{\sqrt{n}(\overline{D}-\delta_0)}{s_D} \le -t_\alpha(n-1) \\ R: t = \frac{\sqrt{n}(\overline{D}-\delta_0)}{s_D} \ge t_\alpha(n-1) \\ R: |t| = \frac{\sqrt{n}\,|\overline{D}-\delta_0|}{s_D} \ge t_{\alpha/2}(n-1) \end{cases}$$

- δ에 대한 $100(1-\alpha)\%$ 신뢰구간:

$$\left(\overline{D} - t_{\alpha/2}(n-1)\frac{s_D}{\sqrt{n}},\ \overline{D} + t_{\alpha/2}(n-1)\frac{s_D}{\sqrt{n}}\right) \text{ or } \overline{D} \pm t_{\alpha/2}(n-1)\frac{s_D}{\sqrt{n}}$$

$\mu_D = \delta$ 에 대한 신뢰구간 : example

예제 9) 15명의 환자를 대상으로 약의 복용 전과 후의 혈압 측정

환자	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
전(x)	70	80	72	76	76	76	72	78	82	64	74	92	74	68	84
후(y)	68	72	62	70	58	66	68	52	64	72	74	60	74	72	74
$d=(x-y)$	2	8	10	6	18	10	4	26	18	−8	0	32	0	−4	10

- D = 각 환자에 대하여 약의 복용 전과 복용 후의 혈압의 차
- $n = 15,\ \overline{d} = \frac{1}{15}\sum_{i=1}^{15} d_i = 8.80, s_D = \sqrt{\frac{1}{14}\sum_{i=1}^{15}\left(d_i - \overline{d}\right)^2} = 10.98$
- 혈압 강하량(δ)에 대한 95% 신뢰구간:

$$\overline{D} \pm t_{0.025}(14)\frac{s_D}{\sqrt{15}} = 8.80 \pm 2.145 \times \frac{10.98}{\sqrt{15}} = 8.80 \pm 6.08$$
$$= (2.72, 14.98)$$

$\mu_D = \delta$ 에 대한 검정 : example

예제 9의 연속)

- 약이 혈압을 떨어뜨리는지 유의수준 5%에서 검정

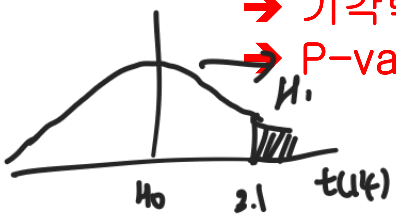
$\mu_D > 0$

- 가설: $H_0: \delta = 0 \quad vs \quad H_1: \delta > 0$
- 검정통계량:

우측검정

➔ CI 판정
➔ 기각력 판정
➔ P-value 판정

p-value



$$t = \frac{\sqrt{n}\,\overline{D}}{s_D} = \frac{\sqrt{15} \times 8.80}{10.98} = 3.10 > t_{0.05}(14) = 2.624$$

따라서 귀무가설을 기각. 즉 약이 혈압을 떨어뜨린다.

$p = P[t \geq 3.1 \mid H_0]$

이후 생략

두 모비율 차이에 대한 추론

두 모비율: $\begin{cases} p_1\text{: 모집단 1에서 특성 } A\text{를 가지는 비율} \\ p_2\text{: 모집단 2에서 특성 } A\text{를 가지는 비율} \end{cases}$ $\quad X \sim B(n_1, p_1)$, $Y \sim B(n_2, p_2)$

- 두 표본의 크기 n_1과 n_2가 충분히 크면

$$\hat{p}_1 \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \text{ and } \hat{p}_2 \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

- 두 표본은 독립이므로

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

- 표본의 크기가 충분히 클 때, $\hat{p}_1 \approx p_1, \hat{p}_2 \approx p_2$이므로

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1)$$

$p_1 - p_2$에 대한 신뢰구간

- 표본의 크기가 충분히 클 때, $p_1 - p_2$에 대한 $100(1-\alpha)\%$ 신뢰구간:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- 위의 신뢰구간은 표본의 크기가 클 때의 신뢰구간의 형태 :

$$(\text{추정량}) \pm z_{\alpha/2} \times (\text{추정량의 표준오차})$$

예제 10) 화학 처리된 씨의 발아 비율(p_1)과 화학 처리되지 않은 씨의 발아 비율(p_2)의 차에 대한 95% 신뢰구간

H_0 : 두 비율 차이가 없다.
H_1 : 두 비율 차이가 있다.

- 자료: $\begin{cases} \text{화학 처리된 씨: } n_1 = 100, X = 88 \\ \text{화학 처리되지 않은 씨: } n_2 = 150,\ Y = 126 \end{cases}$

$$\hat{p}_1 = \frac{88}{100} = 0.88, \hat{p}_2 = \frac{126}{150} = 0.84$$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{0.025}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= 0.04 \pm 0.0866 = (-0.0466, 0.1266)$$

$H_0: p_1 = p_2$에 대한 가설검정

$$H_0: p_1 = p_2 \quad vs \quad \begin{cases} (i)\ H_1: p_1 < p_2 \\ (ii)\ H_1: p_1 > p_2 \\ (iii)\ H_1: p_1 \neq p_2 \end{cases}$$

$$(p_1 = p_2 = p) \rightarrow \hat{p} = \frac{X+Y}{n_1+n_2}$$

(i) $H_1: p_1 < p_2$일 때, $R: Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \le -z_\alpha$

(ii) $H_1: p_1 > p_2$일 때, $R: Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \ge z_\alpha$

(iii) $H_1: p_1 \neq p_2$일 때, $R: |Z| = \dfrac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \ge z_{\alpha/2}$

$\hat{p}_1 - \hat{p}_2 \sim N\left(0,\ p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right)$이므로

$$\rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N(0,1)$$

예제 11) 예제 10에서 화학 처리가 씨의 발아 비율을 높이는지 유의수준 5%에서 검정. p-값은?

자료: $\begin{cases} \text{화학 처리된 씨: } n_1 = 100, X = 88 \\ \text{화학 처리되지 않은 씨: } n_2 = 150,\ Y = 126 \end{cases}$

$$H_0: p_1 = p_2 \quad vs \quad H_1: p_1 > p_2$$

$$\hat{p}_1 = \frac{88}{100} = 0.88, \hat{p}_2 = \frac{126}{150} = 0.84$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.88 - 0.84}{\sqrt{0.856 \times 0.144 \times \left(\frac{1}{100}+\frac{1}{150}\right)}} = 0.883$$

$Z < z_{0.05} = 1.645 \qquad p - \text{값} = P(Z > 0.833) = 0.1894$