제9장 두 집단의 비교

9.1 서론

(1) 두 모집단 간의 차이에 관심이 있다면 → 두 모집단의 비교를 위한 추론과정은 자료를 어떻게 수집하느냐에 따라 추론방법이 달라진다.

(2) 대표적인 두 종류의 자료수집과정

① 독립표본 (independent sample)

어떤 질병을 치료하기 위해 새롭게 개발된 약은 기존의 약보다 이론적으로 더 효과적이라고 생각된다고 한다. 그러나 이러한 가설을 증명하기 위해서는 동물이나 인체를 대상으로 한 실험이 필수적이다. 따라서 두 약의 효과를 비교하기 위해 건강상태가 비슷한 19마리의 쥐를 대상으로 병균을 투입한 후, 그 중에서 임의로 10마리의 쥐를 추출하여 그들에게는 기존의 약을 투입하고 나머지 쥐들에게는 새로운 약을 투약하였다. 그 후에 쥐들이 완치될 때까지 걸린 시간을 기록하였다.
▫ 완전확률화(completely randomized design) ▫ 서로 관련이 없는 두 집단에 각각 처리를 하고 반응값을 측정하므로 두 처리의 반응값들은 서로 관련이 없다.

② 대응표본 (paired sample)

어떤 질병에 걸린 환자들을 대상으로 기존의 약과 새로운 약의 효과를 비교하고자 한다. 환자들의 외적인 조건, 즉 나이, 성별, 건강상태 등이 아주 다양하다고 하자. 이럴 때는 우선 외적인 조건과 병의 경중 등을 고려하여 비교적 비슷한 조건을 갖는 환자끼리 짝을 짓는다. 즉 각 쌍 내의 환자들은 서로 비슷한 조건을 갖는 반면, 각 쌍들 간에는 서로 다른 조건들을 갖도록 한다. 이렇게 짝이 지어지면 각 쌍의 환자 중에서 임의로 한 환자를 추출하여 그 환자에게는 기존의 약을, 다른 한 명의 환자에게는 새로운 약을 투약한 후, 완치 때까지 걸린 시간을 기록한다.
▫ 쌍내에서의 확률화 $[(x_i, y_i)$ where $i = 1, \cdots, n]$ ▫ 실험대상은 대응쌍으로 선택되므로 대응쌍에서의 원소는 비슷하지만 다른 쌍에서의 원소를 실질적으로 다르다.

(3) 모집단 간 차이 비교연구에 사용되는 용어

① 처리(treatment): 비교하고자 하는 특성

② 실험단위[대상](experimental unit): 실험의 대상

③ 반응치(response value): 실험 후에 얻어지는 수치

9.2 독립확률표본

모집단 1: 평균 μ_1, 표준편차 σ_1	→	n_1개의 표본을 추출
모집단 2: 평균 μ_2, 표준편차 σ_2	→	n_2개의 표본을 추출

표본	통계량
모집단 1에서의 표본 $X_1, X_2, \cdots, X_{n_1}$	$\overline{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i$, $s_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(X_i-\overline{X}\right)^2$
모집단 2에서의 표본 $Y_1, Y_2, \cdots, Y_{n_2}$	$\overline{Y}=\frac{1}{n_2}\sum_{i=1}^{n_2}Y_i$, $s_2^2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}\left(Y_i-\overline{Y}\right)^2$

(1) 모평균의 차($\mu_1-\mu_2$)에 대한 추론 (대표본)

① $\overline{X}-\overline{Y}$의 분포

□ 대표본에서 $\mu_1-\mu_2$에 관한 추론은 $\overline{X}-\overline{Y}$에 근거를 둔다.

□ $\overline{X}$와 $\overline{Y}$는 점근적으로 정규분포를 따른다.

□ $\overline{X}-\overline{Y}$도 점근적으로 정규분포를 따른다.

② $\overline{X}-\overline{Y}$의 통계량

□ 평균: $E(\overline{X}-\overline{Y})=\mu_1-\mu_2$ □ 분산: $Var(\overline{X}-\overline{Y})=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$ 1)

□ 표준오차: $S.E.(\overline{X}-\overline{Y})=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$

1) $Var(\overline{X}-\overline{Y})=Var(\overline{X})+Var(\overline{Y})-2Cov(\overline{X},\overline{Y})$
독립표본에서 위 식은 $Var(\overline{X}-\overline{Y})=Var(\overline{X})+Var(\overline{Y})$이 되다. 왜냐하면 $Cov(\overline{X},\overline{Y})=0$.

③ 대표본에서 $H_0 : \mu_1 - \mu_2 = \delta_0$의 검정법

□ 검정통계량

$$Z = \frac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

□ H_1의 형태에 따른 기각역 설정

H_1	기각역
$H_1 : \mu_1 - \mu_2 > \delta_0$	$R : z \geq z_\alpha$
$H_1 : \mu_1 - \mu_2 < \delta_0$	$R : z \leq -z_\alpha$
$H_1 : \mu_1 - \mu_2 \neq \delta_0$	$R : \lvert z \rvert \geq z_{a/2}$

[예제 9.4] 두 종족 A와 B에서 여성의 초혼연령을 비교하기 위해 각 종족에서 100명의 기혼여성을 확률표본으로 택하여 초혼연령의 평균과 표준편차를 구하였다.

	A	B
평균	20.7	18.5
표준편차	6.3	5.8

(1) $\mu_A - \mu_B$의 95% 신뢰구간을 구하라.

【풀이】

주어진 정보: $n_A = n_B = 100$, $\overline{x_A} = 20.7$, $\overline{x_B} = 18.5$, $s_A = 6.3$, $s_B = 5.8$

$\alpha = 0.05$, $z_{0.05/2} = z_{0.025} = 1.96$

$$Var(\overline{X} - \overline{Y}) = Var(\overline{X}) + Var(\overline{Y}) = Var\left(\frac{1}{n_1}\sum_{i=1}^{n_1} X_i\right) + Var\left(\frac{1}{n_2}\sum_{i=1}^{n_2} Y_i\right)$$
$$= \frac{1}{n_1^2}\left[Var(X_1) + \cdots + Var(X_{n_1})\right] + \frac{1}{n_2^2}\left[Var(Y_1) + \cdots + Var(Y_{n_2})\right]$$
$$= \frac{n_1\sigma_1^2}{n_1^2} + \frac{n_2\sigma_2^2}{n_2^2}$$
$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

< $\mu_A-\mu_B$에 대한 $95\%[=100(1-0.05)\%]$ 신뢰구간 >

$$\left(\bar{x}_A-\bar{x}_B-z_{\alpha/2}\sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}},\ \bar{x}_A-\bar{x}_B+z_{\alpha/2}\sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}}\right)$$
$$=\left(20.7-18.5-1.96\sqrt{\frac{6.3^2}{100}+\frac{5.8^2}{100}},\ 20.7-18.5+1.96\sqrt{\frac{6.3^2}{100}+\frac{5.8^2}{100}}\right)$$
$$=(0.52,\ 3.88)$$

$\Rightarrow$ 종족 B의 여성들의 초혼연령이 종족 A보다 평균적으로 0.52년에서 3.88년 앞선다.

(2) 두 종족 간의 평균 초혼 연령이 다르다는 확신을 가질 수 있는가? $\alpha=0.02$에서 검정하라.

【풀이】

□ $H_0: \mu_A=\mu_B$, $H_1: \mu_A\neq\mu_B$ (양측검정)

□ 검정통계량

$$Z=\frac{(\bar{x}_A-\bar{x}_B)-(\mu_A-\mu_B)}{\sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}}}=\frac{(\bar{x}_A-\bar{x}_B)}{\sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}}}=\frac{(20.7-18.5)}{\sqrt{\frac{6.3^2}{100}+\frac{5.8^2}{100}}}=2.569$$

□ 기각역 설정: $\alpha=0.02 \rightarrow$ 기각치: $z_{\alpha/2}=z_{0.02/2}=z_{0.01}=2.33$

$R: |z|\geq 2.33$

□ 검정통계치가 기각역에 포함되므로 귀무가설을 기각하고 대립가설이 맞다고 말할 수 있다. 즉, 두 종족 간의 평균 초혼연령은 다르다고 말할 수 있다.

(2) 모평균의 차$(\mu_1-\mu_2)$에 대한 추론 (소표본)

① 소표본일 때 필요한 가정

□ 두 모집단이 모두 정규분포를 따른다.

□ 두 모집단의 표준편차가 일치한다. $(\sigma_1=\sigma_2=\sigma)$

② 공통표준편차($\sigma_1 = \sigma_2 = \sigma$) 가정의 적용

두 표본표준편차 s_1과 s_2의 상대적 크기가 매우 중요하다. $\sigma_1 = \sigma_2$ 가정은 $\frac{s_1}{s_2}$가 1과 크게 다르지 않을 때 받아들여지게 된다. 실제로 사용할 때는

□ $\frac{1}{2} \le \frac{s_1}{s_2} \le 2 \Rightarrow \sigma_1 = \sigma_2$ (합리적) → 합동추정치를 구함

□ $\frac{s_1}{s_2} < \frac{1}{2}$ or $\frac{s_1}{s_2} > 2 \Rightarrow \sigma_1 = \sigma_2$(?) → $\mu_1 - \mu_2$에 대한 근사적 추론방법을 쓴다.

③ $\overline{X} - \overline{Y}$의 통계량

□ 평균: $E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2$

□ 분산: $Var(\overline{X} - \overline{Y}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

□ 표준오차: $S.E.(\overline{X} - \overline{Y}) = \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

④ 두 모집단은 정규분포를 따르므로 두 표본평균은 각각 정규분포를 따른다. 따라서 두 표본평균의 차($\overline{X} - \overline{Y}$)도 정규분포를 따른다.

$$(\overline{X} - \overline{Y}) \sim N\left[\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]$$

※ 공통분산의 합동추정치(pooled estimator)
(두 집단의 전체 자료를 이용하여 구함)

$$s_p^2 = \frac{\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{i=1}^{n_2}(Y_i - \overline{Y})^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

여기서 $s_1^2 = \frac{\sum_{i=1}^{n_1}(X_i - \overline{X})^2}{n_1 - 1}$, $s_2^2 = \frac{\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2}{n_2 - 1}$

[예제 9.5] 다음 두 표본에서 s_p^2의 값을 구하라.

모집단 1에서의 표본: 8, 5, 7, 6, 9, 7
모집단 2에서의 표본: 2, 6, 4, 7, 6

【풀이】

$$\bar{x}=\frac{1}{n_1}\sum_{i=1}^{n_1}x_i=\frac{1}{6}\sum_{i=1}^{6}x_i=\frac{(8+5+7+6+9+7)}{6}=\frac{42}{6}=7 \text{ 여기서 } n_1=6$$

$$\bar{y}=\frac{1}{n_2}\sum_{i=1}^{n_2}y_i=\frac{1}{5}\sum_{i=1}^{5}y_i=\frac{(2+6+4+7+6)}{5}=\frac{25}{5}=5 \qquad \text{여기서 } n_2=5$$

$$s_1^2=\frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2}{n_1-1}=\frac{10}{5}=2, \qquad s_2^2=\frac{\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2}{n_2-1}=\frac{16}{4}=4$$

$$\rightarrow \frac{s_1}{s_2}=\frac{\sqrt{2}}{\sqrt{4}}=\frac{1}{\sqrt{2}}=0.7071$$

$$\Rightarrow \frac{1}{2}\le\frac{s_1}{s_2}\le 2 \Rightarrow \sigma_1=\sigma_2 \text{ (합리적) } \rightarrow \text{ 합동추정치를 구함}$$

$$s_p^2=\frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2+\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2}{n_1+n_2-2}=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}=\frac{10+16}{6+5-2}=\frac{26}{9}=2.89$$

※ s_p^2은 s_2^2보다 s_1^2에 더 가깝다. (이유: n_1이 n_2보다 더 큼)

⑤ 두 정규모집단에서 독립적으로 추출된 두 표본으로부터 얻게 되는 표준화된 확률변수는 자유도가 (n_1+n_2-2)인 $t-$분포를 따른다.

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

⑥ 모평균의 차$(\mu_1-\mu_2)$에 대한 신뢰구간 (소표본, 공통표준편차)

$$\left(\overline{X}-\overline{Y}-t_{\alpha/2}(n_1+n_2-2)\times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\ \overline{X}-\overline{Y}+t_{\alpha/2}(n_1+n_2-2)\times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

[예제 9.6] 사료 1과 사료 2의 우유 생산량 증가효과를 비교하기 위해 실험용 젖소 25마리를 택하여 그 중 13마리에는 사료 1을 공급하고 나머지 12마리에는 사료 2를 공급하였다. 3주일 후 젖소의 우유 생산량 검사를 해본 결과 다음의 결과를 얻었다.

사료 1: 44, 44, 56, 46, 47, 38, 58, 53, 49, 35, 46, 30, 41
사료 2: 35, 47, 55, 29, 40, 39, 32, 41, 42, 57, 51, 39

(1) 사료 1과 사료 2의 평균 우유 생산량의 차이($\mu_1-\mu_2$)에 대한 95% 신뢰구간을 구하라.

【풀이】

□ 사료 1: $n_1=13,\ \bar{x}=\frac{1}{n_1}\sum_{i=1}^{n_1}x_i=\frac{1}{13}\sum_{i=1}^{13}x_i=45.15$

□ 사료 2: $n_2=12,\ \bar{y}=\frac{1}{n_2}\sum_{i=1}^{n_2}y_i=\frac{1}{12}\sum_{i=1}^{12}y_i=42.25$

□ $s_1^2=\frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2}{n_1-1}=\frac{767.69}{12}=63.9742,$

$s_2^2=\frac{\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2}{n_2-1}=\frac{840.25}{11}=76.3864$

$\rightarrow\ \frac{s_1^2}{s_2^2}=\frac{63.9742}{76.3864}=0.8375,\ \frac{s_1}{s_2}=0.9152$

$\Rightarrow\ \frac{1}{2}\le\frac{s_1}{s_2}\le 2\ \Rightarrow\ \sigma_1=\sigma_2$ (합리적) $\rightarrow$ 합동추정치를 구함

□ pooled estimator s_p^2

$$s_p^2=\frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2+\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2}{n_1+n_2-2}=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}=\frac{767.69+840.25}{13+12-2}=\frac{1607.94}{23}$$

$$=69.91$$

$$s_p=\sqrt{69.91}=8.3612$$

□ $t_{\alpha/2}(n_1+n_2-2)=t_{0.05/2}(13+12-2)=t_{0.025}(23)=2.069$

□ $\mu_1-\mu_2$에 대한 95% 신뢰구간

$$\left(\overline{X}-\overline{Y}-t_{\alpha/2}(n_1+n_2-2)\times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\ \overline{X}-\overline{Y}+t_{\alpha/2}(n_1+n_2-2)\times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

$$=\left(45.45-42.25-2.069\times 8.3612\sqrt{\frac{1}{13}+\frac{1}{12}},\ 45.15-42.25+2.069\times 8.3612\sqrt{\frac{1}{13}+\frac{1}{12}}\right)$$

$$=(-4.025,\ 9.825)$$

⑦ $H_0:\mu_1-\mu_2=\delta_0$에 대한 가설검정 (소표본, 공통표준편차)

□ 검정통계량: $t=\dfrac{(\overline{X}-\overline{Y})-\delta_0}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2)$

□ H_1의 형태에 따른 기각역 설정

H_1	기각역
$H_1:\mu_1-\mu_2>\delta_0$	$R: t\geq t_\alpha$
$H_1:\mu_1-\mu_2<\delta_0$	$R: t\leq -t_\alpha$
$H_1:\mu_1-\mu_2\neq\delta_0$	$R: \lvert t\rvert\geq t_{a/2}$

[예제 9.6 계속]

(2) 사료 1에서의 우유 생산량이 사료 2에서의 생산량보다 더 많다고 결론을 내릴 수 있는가? 유의수준 $\alpha=0.05$에서 검정하라.

【풀이】

□ $H_0:\mu_1-\mu_2=0$, $H_1:\mu_1-\mu_2>0$ (단측검정)

□ 검정통계량

$$t=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}=\frac{45.15-42.25}{8.3612\sqrt{\dfrac{1}{13}+\dfrac{1}{12}}}=0.866$$

□ 기각역 설정

$\alpha=0.05\ \rightarrow$ 기각치: $t_\alpha(n_1+n_2-2)=t_{0.05}(23)=1.714$

$R: t\geq 1.714$

□ 검정통계치가 기각역에 포함되지 않으므로 귀무가설을 기각할 수 없다. 따라서 대립가설은 입증되지 않는다. 즉 사료 1에서의 우유 생산량이 사료 2에서의 생산량보다 더 많다고 말할 수 없다.

⑧ 모평균의 차($\mu_1-\mu_2$)에 대한 추론 (소표본, 두 모표준편차가 다른 경우)

□ 가정: A1. 모집단은 정규분포를 따른다. A2. $\sigma_1 \neq \sigma_2$

□ $\mu_1-\mu_2$에 대한 $100(1-\alpha)\%$ 신뢰구간

$$\left(\bar{x}-\bar{y}-t^*_{\alpha/2}\sqrt{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)},\ \bar{x}-\bar{y}+t^*_{\alpha/2}\sqrt{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)} \right)$$

□ $H_0 : \mu_1-\mu_2=\delta_0$

□ 검정통계량

$$t^* = \frac{(\overline{X}-\overline{Y})-\delta_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t$$

자유도: (n_1-1)과 (n_2-1) 중 작은 값

[예제] 지방의 한 도시에서 작은 개천을 사이에 두고 남북으로 나뉘어 있는 두 지역의 집값을 비교하고자 최근에 매매가 이루어진 집을 대상으로 남쪽으로 13가구, 북쪽으로 11가수의 집값을 조사하였더니 다음과 같았다.

남쪽: $n_1=13$, $\bar{x}=2.4$억 원, $s_1=0.72$억원
북쪽: $n_2=11$, $\bar{y}=2.15$억 원, $s_2=0.35$억원

유의수준 5%로 두 지역의 집값에 차가 있다고 할 수 있는지 검정하라.

【풀이】

□ $H_0 : \mu_1-\mu_2=0$, $H_1 : \mu_1-\mu_2 \neq 0$ (양측검정)

□ $\frac{s_1}{s_2}=\frac{0.72}{0.35}=2.0571$

$\rightarrow \frac{s_1}{s_2}>2 \Rightarrow \sigma_1=\sigma_2(?) \rightarrow \mu_1-\mu_2$에 대한 근사적 추론방법을 쓴다.

□ 검정통계량

$$t^* = \frac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{2.4 - 2.15}{\sqrt{\frac{0.5184}{13} + \frac{0.1225}{11}}} = 1.107$$

□ 기각역 설정

$\alpha = 0.05$

→ 기각치: $t_{\alpha/2}(n_2 - 1) = t_{0.025}(10) = 2.228$

[자유도: $(n_1 - 1)$과 $(n_2 - 1)$ 중 작은 값]

⇒ 기각역 $R : |t| \geq 2.228$

□ 결론: 검정통계치가 기각역에 포함되지 않으므로 귀무가설을 기각할 수 없다. 따라서 대립가설은 입증되지 않는다. 즉 두 지역 간에 집값 차이가 있다고 말할 수 없다.

11.3 대응[짝]비교

(1) 대응[짝]비교에서의 자료구조

① 자료구조

대응쌍	처리 1	처리 2	차
1	X_1	Y_1	$D_1 = X_1 - Y_1$
2	X_2	Y_2	$D_2 = X_2 - Y_2$
⋮	⋮	⋮	⋮
n	X_n	Y_n	$D_n = X_n - Y_n$

* 차 $D_1, D_2, \cdots, D_n$은 확률표본이다.

② 통계량

□ $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$

□ $S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$

(2) 차의 평균 δ에 대한 소표본 추론

차 $D_i = X_i - Y_i$가 $N(\delta, \sigma_D^2)$에서의 확률표본이라고 가정한다.

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \qquad S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2}$$

으로 두고 관측값을 $\bar{d}$, s_D라 한다.

(a) δ에 대한 $100(1-\alpha)\%$ 신뢰구간: $\left(\bar{d} - t_{\alpha/2}\frac{s_D}{\sqrt{n}},\ \bar{d} + t_{\alpha/2}\frac{s_D}{\sqrt{n}} \right)$

(b) $H_0 : \delta = \delta_0$에 대한 검정은 다음의 검정통계량을 이용한다.

$$T = \frac{\overline{D} - \delta_0}{S_D/\sqrt{n}} \qquad \text{자유도} = (n-1)$$

[예제 9] 어느 제약회사에서는 피임약이 사용자의 혈압을 떨어뜨리는 부작용이 있는지 알아보고자 한다. 15명의 주부를 택하여 혈압을 측정하고, 피임약을 6개월 동안 복용하게 한 후 혈압을 측정한 결과가 다음의 표에 기록되어있다. 이때 주부들의 혈압은 정규분포를 따른다고 가정한다.

	주부														
	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

(1) 혈압의 평균 감소량에 대한 95% 신뢰구간을 구하라.

【풀이】

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{15}\sum_{i=1}^{15} d_i = \frac{132}{15} = 8.80 \quad \text{여기서 } n = 15$$

$$s_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2} = \sqrt{\frac{1}{14}\sum_{i=1}^{15}(d_i - 8.80)^2} = 10.98$$

$$t_{\alpha/2}(n-1) = t_{0.05/2}(15-1) = t_{0.025}(14) = 2.145$$

< δ에 대한 95% 신뢰구간 >

$$\left(\bar{d} - t_{\alpha/2}\frac{s_D}{\sqrt{n}},\ \bar{d} + t_{\alpha/2}\frac{s_D}{\sqrt{n}} \right) = \left(8.80 - 2.145\frac{10.98}{\sqrt{15}}, 8.80 + 2.145\frac{10.98}{\sqrt{15}} \right) = (2.72, 14.88)$$

(2) 피임약이 혈압을 감소시킨다고 주장할 수 있는가? $\alpha = 0.01$에서 검정하라.

【풀이】

□ $H_0 : \mu_A - \mu_B = 0 \Rightarrow \delta = 0$ 여기서, $\delta = \delta_0,\ \delta_0 = 0$

$H_1 : \mu_A - \mu_B > 0 \Rightarrow \delta > 0$ (단측검정)

□ 검정통계량: $t = \dfrac{\bar{d} - \delta_0}{s_D/\sqrt{n}} = \dfrac{8.80}{10.98/\sqrt{15}} = 3.10$

□ 기각역 설정

$\alpha = 0.01 \rightarrow$ 기각치: $t_\alpha(n-1) = t_{0.01}(14) = 2.624$

기각역 $R : t \geq 2.624$

□ 결론

귀무가설 기각 → 대립가설이 맞다고 말할 수 있다.

⇒ 약이 혈압을 내린다는 주장을 뒷받침한다고 할 수 있다.

(3) 랜덤화 (randomization)

환자에게 있는 여러 조건들이 어느 한 쪽의 처리에만 영향을 주지 않고 확률적으로 같은 정도로 영향을 미치도록 해야 한다. 예를 들면 각 쌍에서 한 환자에게 동전 던지기로 A, B 중 하나의 약을 처방하고 남은 환자에게는 다른 약을 처방함으로써 환자 간에 어떤 차이가 있다고 하더라도 그 차이가 한 종류의 약에만 영향을 주지는 않도록 하는 것이다. 이와 같이 무작위로 배정하는 것은 랜덤화 (randomization)라고 표현한다.

11.4 두 모비율의 차에 대한 추론

(1) 모비율에 대한 점추정

① 설정

□ 두 모비율의 비교

p_1: 모집단 1에서의 특성치에 대한 비율

p_2: 모집단 2에서의 특성치에 대한 비율

□ 해야 할 일

$p_1 - p_2$에 대한 신뢰구간 구축

$H_0 : p_1 = p_2$에 대한 가설검정

② 예: 두 모집단으로부터 추출된 독립된 두 표본

	특성 A인 것(성공)	특성 A가 아닌 것(실패)	표본의 크기
모집단 1	X	$n_1 - X$	n_1
모집단 2	Y	$n_2 - Y$	n_2

□ 성공에 대한 모비율 p_1과 p_2는 다음의 표본비율로 추정된다.

$\hat{p_1} = \frac{X}{n_1}$, $\hat{p_2} = \frac{Y}{n_2}$

□ $\hat{p_1} - \hat{p_2}$는 $p_1 - p_2$의 추정치

□ $S.E.(\hat{p_1} - \hat{p_2}) = \sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}}$ 여기서 $\hat{q_1} = 1 - \hat{p_1}$, $\hat{q_2} = 1 - \hat{p_2}$

□ 만약 n_1과 n_2가 충분히 크면, $\hat{p_1} - \hat{p_2}$은 근사적으로 정규분포를 따른다.

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{S.E.(\hat{p_1} - \hat{p_2})} \sim N(0, 1^2)$$

여기서, $S.E.(\hat{p_1} - \hat{p_2}) = \sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}}$ [$\hat{q_1} = 1 - \hat{p_1}$, $\hat{q_2} = 1 - \hat{p_2}$]

(2) 모비율의 차 $(p_1 - p_2)$에 대한 신뢰구간 (표본의 크기 n_1과 n_2가 클 때)

$$(\hat{p_1} - \hat{p_2}) \pm z_{a/2} \sqrt{\frac{\hat{p_1}(1 - \hat{p_1})}{n_1} + \frac{\hat{p_2}(1 - \hat{p_2})}{n_2}}$$

[예제 9.9] 종자의 발아율에 대한 화학적 처리의 효과가 있는지를 알아보기 위하여 100개의 종자에는 화학처리를 하고 150개의 종자에는 화학처리를 하지 않았다. 화학처리를 한 종자 중 88개가 발아하였고 화학처리를 하지 않은 종자 중 126개가 발아하였다. 화학처리를 한 종자와 하지 않은 종자의 발아율에 대한 차의 95% 신뢰구간을 구하라.

【풀이】

	발아된 종자의 수	발아 안된 종자의 수	합
화학처리 O	88	12	100
화학처리 X	126	24	150

$$\hat{p_1}=\frac{X}{n_1}=\frac{88}{100}=0.88,\ \hat{p_2}=\frac{Y}{n_2}=\frac{126}{150}=0.84$$

$$S.E.(\hat{p_1}-\hat{p_2})=\sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1}+\frac{\hat{p_2}\hat{q_2}}{n_2}}=\sqrt{\frac{0.88\times0.12}{100}+\frac{0.84\times0.16}{150}}=0.044$$

$$\alpha=0.05\ \rightarrow\ z_{\alpha/2}=z_{0.05/2}=z_{0.025}=1.96$$

< p_1-p_2에 대한 95% 신뢰구간 >

▷ $(\hat{p_1}-\hat{p_2})\pm z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}=(\hat{p_1}-\hat{p_2})\pm z_{\alpha/2}S.E.(\hat{p_1}-\hat{p_2})$

$=(0.88-0.84)\pm1.96(0.044)=0.04\pm1.96(0.044)=0.04\pm0.09=(-0.05,\ 0.13)$

(3) $H_0: p_1=p_2$에 대한 가설검정 (대표본)

① 검정통계량: $Z=\frac{\hat{p_1}-\hat{p_2}}{S.E.(\hat{p_1}-\hat{p_2})}=\frac{\hat{p_1}-\hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

여기서 합동추정량(pooled estimator) $\hat{p}=\frac{X+Y}{n_1+n_2}=\frac{n_1\hat{p_1}+n_2\hat{p_2}}{n_1+n_2}$

$$S.E.(\hat{p_1}-\hat{p_2})=\sqrt{\hat{p}\hat{q}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

※ $\hat{p}$: 귀무가설 하에서의 공통비율 p (모비율이 같은 두 모집단의 공통비율)의 추정량

② H_1의 형태와 상응하는 기각역

H_1	기각역
$H_1 : p_1 > p_2$	$R: Z \geq z_\alpha$
$H_1 : p_1 < p_2$	$R: Z \leq -z_\alpha$
$H_1 : p_1 \neq p_2$	$R: \lvert Z \rvert \geq z_{\alpha/2}$

[예제 9.9 계속] 앞의 예제 9.9에서 화학적인 처리가 씨의 발아율을 높인다고 할 수 있는지 유의수준 5%로 검정하라. 또 $P-$값도 구하라.

【풀이】

① 가설설정

$H_0 : p_1 - p_2 = 0,\ H_1 : p_1 - p_2 > 0$ (단측검정)

② $n_1 = 100,\ n_2 = 150$ → 대표본

③ $\hat{p_1} = \frac{X}{n_1} = \frac{88}{100} = 0.88,\ \hat{p_2} = \frac{Y}{n_2} = \frac{126}{150} = 0.84$

합동추정량 $\hat{p} = \frac{X+Y}{n_1+n_2} = \frac{88+126}{100+150} = 0.856$

검정통계량 $Z = \frac{\hat{p_1} - \hat{p_2}}{S.E.(\hat{p_1} - \hat{p_2})} = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

$= \frac{0.88 - 0.84}{\sqrt{0.856 \times 0.144}\sqrt{\frac{1}{100}+\frac{1}{150}}} = 0.883$

④ 기각역의 설정

- $\alpha = 0.05$ → 기각치: $z_\alpha = z_{0.05} = 1.645$
- 기각역 $R: Z \geq 1.645$

⑤ 결론: 검정통계치가 기각역에 포함되지 않으므로 귀무가설을 기각할 수 없다. → 대립가설이 맞다고 말할 수 없다.

⇒ 화학적 처리가 씨의 발아율을 높인다는 충분한 근거가 없다.

⑥ $p-value = P[z \geq 0.883] = 0.1894$

[예제 9.10] 홍콩 독감의 백신을 5세에서 9세까지의 어린이에게 접종하였다. 남자 어린이 113명 가운데 34명이 항체가 생겼고 여자 어린이 139명 가운데 54이 항체가 생겼다.

(1) 항체의 생성 비율이 남자 어린이보다 여자 어린이쪽이 더 높다고 할 수 있는가? 유의수준 $\alpha = 0.05$에서 검정하라.

【풀이】

□ 설정

p_1: 남자 어린이의 항체 생성 모비율, p_2: 여자 어린이의 항체 생성 모비율

□ 연구가설

$H_0 : p_1 = p_2 \ [p_1 - p_2 = 0],\ H_1 : p_1 < p_2 \ [p_1 - p_2 < 0]$ (단측검정)

□ 검정통계량

$\hat{p_1} = \frac{X}{n_1} = \frac{34}{113} = 0.301,\ \hat{p_2} = \frac{Y}{n_2} = \frac{54}{139} = 0.388$

합동추정량 $\hat{p} = \frac{X+Y}{n_1+n_2} = \frac{34+54}{113+139} = 0.349$

$$Z = \frac{\hat{p_1} - \hat{p_2}}{S.E.(\hat{p_1} - \hat{p_2})} = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{0.301 - 0.388}{\sqrt{0.349 \times 0.651}\sqrt{\frac{1}{113} + \frac{1}{139}}} = -1.44$$

□ 기각역의 설정

$\alpha = 0.05 \rightarrow$ 기각치: $z_\alpha = z_{0.05} = 1.645$

기각역 $R : Z \le -z_{0.05} \Rightarrow R : Z \le -1.645$

□ 귀무가설을 기각할 수 없다 → 대립가설이 맞다고 말할 수 없다.

⇒ 항체의 생성 비율이 남자 어린이보다 여자 어린이쪽이 더 높다고 볼 수 없다.

(2) p-값을 구하라.

【풀이】

$p - value = P[z \le -1.44] = 0.0749$