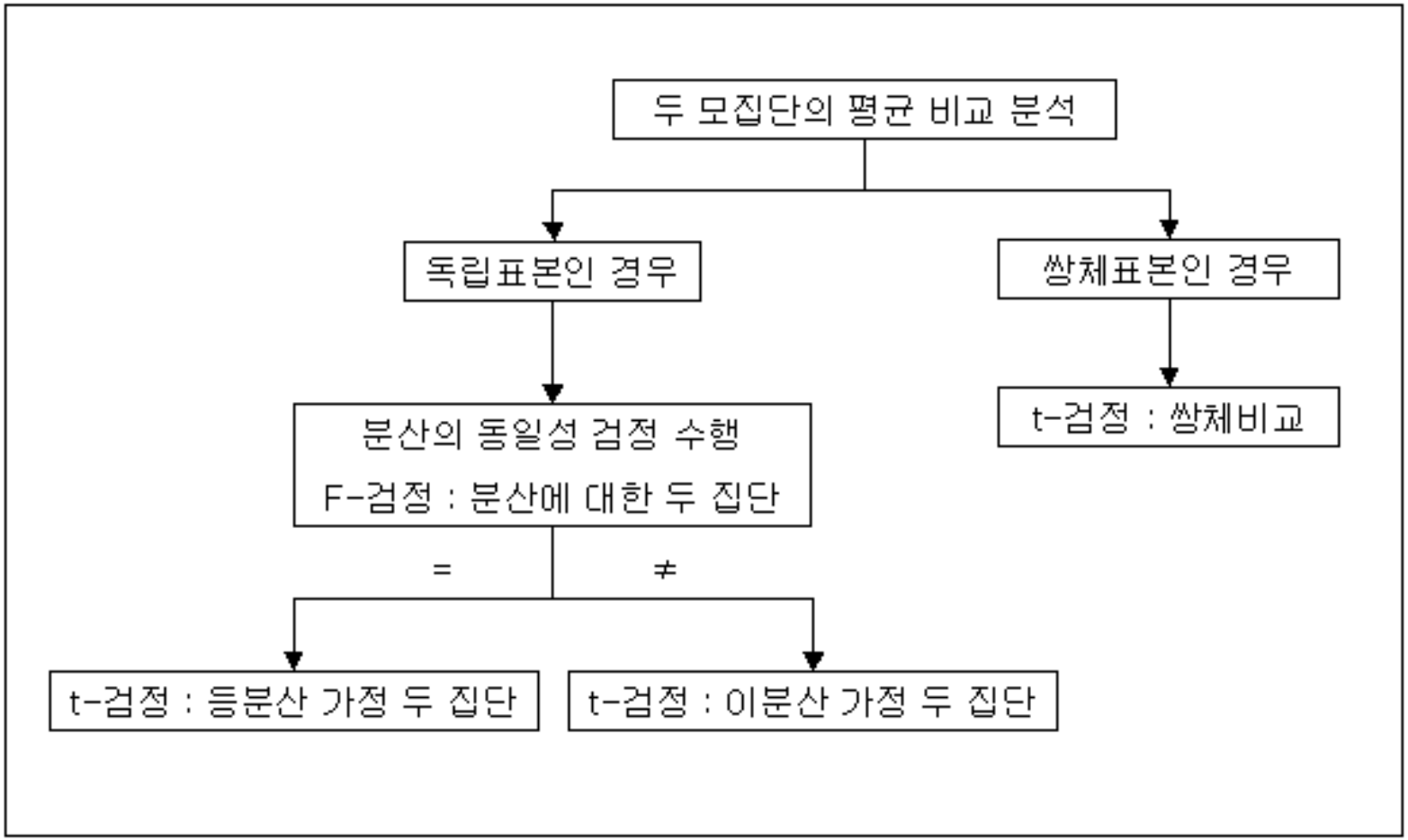
10. 두 모집단의 추론

정보통계학과 : 김 덕 기

http://cafe.daum.net/cb-stat

toby123@cbnu.ac.kr

두 모집단의 평균비교 분석



독립표본과 대응표본.

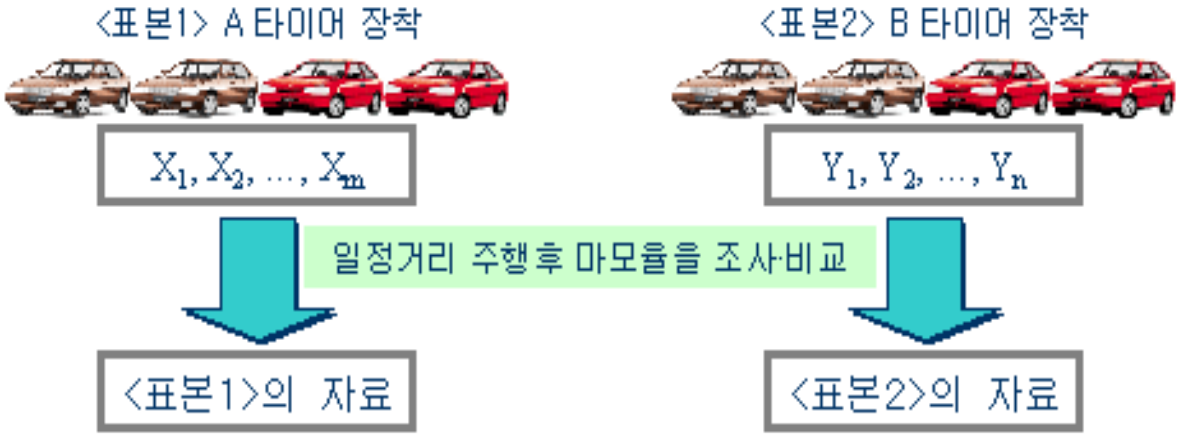
• **독립표본과 대응표본**

- 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**

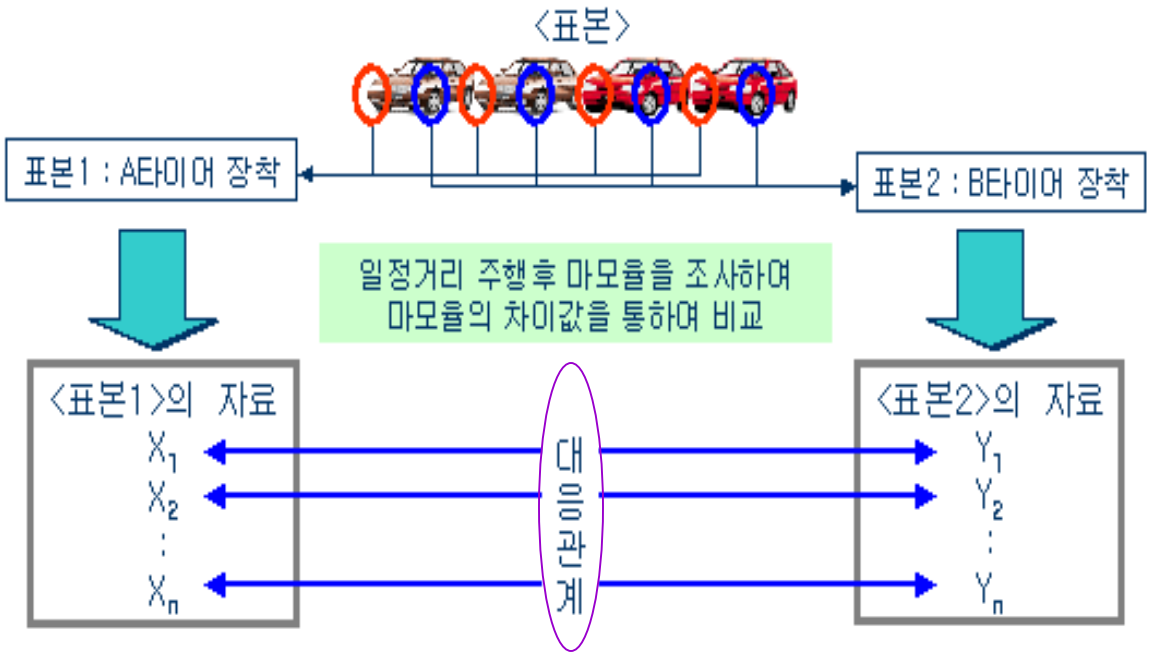
두 표본자료의 차량이 서로 다르며 상관관계가 없다. 이러한 방법으로 만들어진 두 표본을 독립표본 이라 함.



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.



➡ 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 독립표본(independent sample)이라 함

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

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 띄고 있음

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

두 모집단의 분산을 알고 있는 경우

(1) 두 모집단의 분산 σ_1^2, σ_2^2을 알고 있는 경우

$$X_1,\cdots,X_m \sim N(\mu_x, \sigma_x^2) \quad , \quad Y_1,\cdots,Y_n \sim N(\mu_y, \sigma_y^2), \quad (\overline{X}-\overline{Y}) \sim N(\mu_x-\mu_y, \frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n})$$

$$\text{검정통계량} : Z = \frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}}}$$

	귀무가설	대립가설	유의수준 α에 대한 기각역
(a)	$H_0 : \mu_x-\mu_y = 0$	$H_1 : \mu_x-\mu_y \neq 0$	$\lvert Z\rvert > z_{\alpha/2}$
(b)	$H_0 : \mu_x-\mu_y \geq 0$	$H_1 : \mu_x-\mu_y < 0$	$Z < -z_\alpha$
(c)	$H_0 : \mu_x-\mu_y \leq 0$	$H_1 : \mu_x-\mu_y > 0$	$Z > z_\alpha$

모평균차의 (1−alpha)*100 신뢰구간

$$Z=\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}}\sim N(0,1)$$

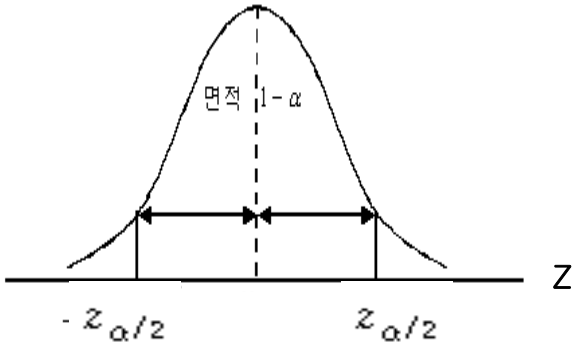
이 결과로부터 $\mu_X-\mu_Y$에 관한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같이 주어진다.

$$P\left(-z_{\alpha/2}\le Z\le z_{\alpha/2}\right)=1-\alpha$$

$$P\left(-z_{\alpha/2}<\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}}<z_{\alpha/2}\right)=1-\alpha$$

$$P\left((\bar{X}-\bar{Y})-z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}<\mu_X-\mu_Y<(\bar{X}-\bar{Y})+z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}\right)=1-\alpha$$

$$(\bar{X}-\bar{Y})\pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}$$



(1−alpha)*100신뢰구간

(...예제)

[예제 1] K대학 부속병원에서는 환자의 입원기간을 조사하여 남,여 환자간에 차이가 있는가를 알고자 한다. 남자환자 40명, 여자환자 35명을 표본으로 추출하여 평균입원기간을 조사한 결과 각각 9.08일, 7.11일이었다. (1) 모평균차에 대한 95% 신뢰구간을 구하여라. (2)남자환자의 평균입원기간이 여자환자의 평균입원기간보다 길다고 할 수 있는가를 유의수준 10%로 검정하라. (단, $\sigma_1 = 7.5$일 , $\sigma_2 = 6.8$일)

$$(\overline{X} - \overline{Y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}} = 1.97 \pm 1.96 \times (1.651) = 1.97 \pm 3.237$$

① $H_0 : \mu_1 - \mu_2 = 0$, $H_1 : \mu_1 - \mu_2 > 0$

② $\alpha = 0.1$

③ $\alpha = 0.1$의 우측검정이므로 $z_\alpha = z_{0.1} = 1.28$, 기각영역 : $Z > z_\alpha = 1.28$

④ $\overline{X}_1 - \overline{X}_2 = 9.08 - 7.11 = 1.97$,

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} = \frac{(1.97 - 0)}{\sqrt{7.5^2/40 + 6.8^2/35}} = 1.19 < z_\alpha = 1.28$$, H_0 기각하지 못함.

즉, 여자환자에 비해 남자환자의 평균입원기간이 길다는 근거가 없다.

두 모집단의 분산을 모르는 경우.

(2) 두 모집단의 분산 σ_1^2, σ_2^2을 모르는 경우 $\sigma_1^2 \neq \sigma_2^2$

(대표본) 검정통계량 : $Z = \dfrac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\dfrac{S_x^2}{m} + \dfrac{S_y^2}{n}}}$, (소표본) 검정통계량 : $t = \dfrac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\dfrac{S_x^2}{m} + \dfrac{S_y^2}{n}}}$

(3) 두 모집단의 분산 σ_1^2, σ_2^2을 모르는 경우 $\sigma_1^2 = \sigma_2^2$인 경우 (소표본)

검정통계량 : $t = \dfrac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{S_p \sqrt{\dfrac{1}{m} + \dfrac{1}{n}}}$, $S_p^2 = \dfrac{(m-1)S_x^2 + (n-1)S_y^2}{m+n-2}$

2집단의 표본분산 : $S_x^2 = \dfrac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X})^2$, $S_y^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$

	귀무가설	대립가설	유의수준 α에 대한 기각역
(a)	$H_0 : \mu_x - \mu_y = 0$	$H_1 : \mu_x - \mu_y \neq 0$	$\lvert t \rvert > t_{\alpha/2}(m+n-2)$
(b)	$H_0 : \mu_x - \mu_y \geq 0$	$H_1 : \mu_x - \mu_y < 0$	$t < -t_{\alpha}(m+n-2)$
(c)	$H_0 : \mu_x - \mu_y \leq 0$	$H_1 : \mu_x - \mu_y > 0$	$t > t_{\alpha}(m+n-2)$

두 모집단의 분산을 모르는 경우.

(2), (3)의 $\mu_X - \mu_Y$에 관한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같이 주어진다.

(2) 분산이 다를 때:

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}} \sim N(0,1)$$

$$(\overline{X} - \overline{Y}) \pm z_{\alpha/2} \sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}$$

(3) 분산이 같을 때:

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2)$$

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

(...예제)

[예제 2] 동일한 제품의 포장형태에 따라 판매량에 차이가 발생하는가를 조사하려고 한다. 150개의 지역에서 A, B 두 가지 형태의 포장을 한 제품을 시판한 결과 평균판매량이 각각 27.45개, 37.14개, 표준편차는 8.30개, 9.44개로 나타났다. 포장형태에 따라 평균판매량에 차이가 있다고 할 수 있는가를 1% 유의수준으로 검정하라.

CI 판정 : $CI \notin H_0 : \mu_1 - \mu_2 = 0 \Rightarrow H_0$ *accept*

$p-value$ 판정 : $P = p(z \geq z_0) \times 2 < \alpha \Rightarrow H_0$ *reject*

① $H_0 : \mu_1 = \mu_2 (\mu_1 - \mu_2 = 0)$, $H_1 : \mu_1 \neq \mu_2 (\mu_1 - \mu_2 \neq 0)$

② $\alpha = 0.01$

③ $\alpha = 0.01$의 양측검정이므로 $z_{\alpha/2} = z_{0.005} = 2.57$, 기각영역 : $|Z| > z_{\alpha/2} = 2.57$

④ $\overline{X_1} - \overline{X_2} = 27.45 - 37.14 = -9.69$,

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} = \frac{(-9.69 - 0)}{\sqrt{8.3^2/150 + 9.44^2/150}} = -9.44 < z_{\alpha/2} = -2.57$$, H_0 기각함.

즉, 포장형태별 평균판매량에 차이가 있다고 할 수 있다.

(...예제)

[예제 3] C대학 심리학과 교수는 자신의 과목을 수강하는 남,여학생들간에 IQ의 차이가 있는가를 조사하려고 한다. 이를 위해 남,여학생 각각 15명을 표본으로 추출하여 IQ테스트를 실시한 결과가 다음과 같다. 남,여학생들의 IQ가 정규분포를 이루고 동일한 분산을 갖는다고 할 때 남,여학생의 평균 IQ가 다르다고 할 수 있는가를 유의수준 5%로 검정하라.

남학생	여학생
130 109 117 127 120 125 100 118 124 130 126 118 131 107 130	106 107 122 144 120 104 131 101 122 119 134 133 109 124 114

① $H_0 : \mu_1 = \mu_2 (\mu_1 - \mu_2 = 0)$, $H_1 : \mu_1 \neq \mu_2 (\mu_1 - \mu_2 \neq 0)$

② $\alpha = 0.05$

③ $\alpha = 0.05$의 양측검정이므로 $t_{\alpha/2}(m+n-2) = t_{0.025}(28) = 2.048$, 기각영역 : $|t| > 2.048$

④ $\overline{X_1} = 120.8$, $\overline{X_2} = 119.3$, $s_1 = 9.767$, $s_2 = 12.591$, $S_p = 11.268$

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{120.8 - 119.3}{11.268 \sqrt{\frac{1}{15} + \frac{1}{15}}} = 0.36 < 2.048$$, H_0을 기각하지 못함.

즉, 남, 여학생들의 평균(IQ)가 다르다고 할 수 없다.

CI 판정 : CI ∉ $H_0 : \mu_1 - \mu_2 = 0$ ⇒ H_0 accept

p−value 판정 : $P = p(t \geq t_0) \times 2 < \alpha$ ⇒ H_0 reject

두 모집단 평균차의 검정절차.

- 검정 절차

[단계 1] 두 집단 간의 등분산성에 대한 검정 수행

[단계 2] [단계1]의 결과에 따라 두 가지의 검정 방법이 있음

❶ **두 집단 간의 분산이 동일한 경우** ➡ 두 집단의 공통분산에 대한 합동추정량(pooled estimator) S_p^2을 이용한 t-검정 수행

❷ **두 집단 간의 분산이 동일하지 않은 경우** ➡ 각 집단의 분산에 대한 개별 추정량 S_1^2과 S_2^2을 이용한 t-검정 수행

두 집단의 등분산 검정.

$H_0: \sigma_1^2 = \sigma_2^2$ vs $H_1: \sigma_1^2 \neq \sigma_2^2$ 또는 $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ vs $H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

$$F = \frac{\frac{(m-1)S_1^2}{\sigma_1^2} \Big/ (m-1)}{\frac{(n-1)S_2^2}{\sigma_2^2} \Big/ (n-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{(m-1,\,n-1)}$$

IF $S_1^2 > S_2^2$ THEN $F = \frac{S_1^2}{S_2^2}$

이를 바탕으로 가설 $H_0: \sigma_1^2 = \sigma_2^2$을 검정하여 보자. 귀무가설 H_0하에서

$$F_0 = \frac{S_1^2}{S_2^2} \sim F(m-1, n-1)$$

유의수준에 따른 기각역 : (양측검정)

$F_0 < F_{1-\alpha/2}(m-1, n-1)$ 또는 $F_0 > F_{\alpha/2}(m-1, n-1)$

(우측 검정) : $H_0: \sigma_1^2 = \sigma_2^2$ vs $H_1: \sigma_1^2 > \sigma_2^2$ ⇒ $F_0 > F_\alpha(m-1, n-1)$

(좌측 검정) : $H_0: \sigma_1^2 = \sigma_2^2$ vs $H_1: \sigma_1^2 < \sigma_2^2$ ⇒ $F_0 < F_{1-\alpha/2}(m-1, n-1)$

등 분산검정 (...예제)

[예제 4] 출근시간에 반포에서 시청으로 가는 방법은 A, B 두 가지의 방법이 있다. 각각 5일 동안 A, B 두 가지 방법으로 출근한 결과 소요시간의 평균과 표준편차가 다음과 같다.

$$\bar{x}_A = 38\text{분} \quad \bar{x}_B = 34.2\text{분} \quad S_A = 16.4\text{분} \quad S_B = 6.1\text{분}$$

A, B 두 가지 방법의 평균 소요시간의 차이에 대한 유의수준 10%로 가설검정을 하는 경우, 표본이 5로 소 표본이므로 두 집단간의 분산이 동일한지 여부를 먼저 검정해야 가설검정이 가능하다.

① $H_0 : \sigma_1^2 = \sigma_2^2 \quad vs \quad H_1 : \sigma_1^2 \neq \sigma_2^2$

② $\alpha = 0.1$

③ $\alpha = 0.05$의 양측검정이므로, 기각영역 : $F > 6.3883$ or $F < 0.1565$

우측 : $F_{\alpha/2}(m-1, n-1) = F_{0.05}(4,4) = 6.3883$, 좌측 : $F_{1-\alpha/2}(m-1, n-1) = F_{0.95}(4,4) = 0.1565$

④ $F = \dfrac{S_A^2}{S_B^2} = \dfrac{16.4^2}{6.1^2} = 7.23 > 6.3883 = F_{0.05}(4,4)$

$IF \ S_1^2 > S_2^2 \ THEN \ F = \dfrac{S_1^2}{S_2^2}$

H_0을 기각함. 즉, 두 가지 방법의 소요시간에 대한 분산이 동일하다고 할 수 없다.

두 집단의 평균차이 검정절차1

[예제5] 청주시청에서는 그 도시 안에 있는 A 지역과 B 지역의 가구 당 평균 월 수입 간에 차이가 있는지를 알아보기 위하여, A 지역과 B 지역에서 각각 20가구와 25가구를 랜덤 하게 추출하여 월 수입을 조사하여 다음을 얻었다. 이 자료를 바탕으로 두 지역의 가구 당 평균 월 수입이 다르다고 할 수 있는지를 유의수준 5%에서 검정하시오.

A-지역	B-지역
152 128 146 142 113 140 161 185 142 164 132 160 140 154 126 109 138 182 144 172	144 179 160 150 174 161 157 170 162 127 187 163 171 168 191 157 172 135 164 130 143 108 132 111 158

(1) 가설 설정

A 지역과 B 지역의 가구당 평균 월수입을 각각 μ_1, μ_2라고 할 때, 문제의 뜻에 적절한 가설을 설정하면 다음과 같다.

귀무가설 $H_0: \mu_1 = \mu_2$(두 지역의 가구당 평균 월수입은 같다.)

대립가설 $H_1: \mu_1 \neq \mu_2$(두 지역의 가구당 평균 월수입은 같지 않다.)

두 집단의 등 분산검정(검정절차2)

A, B 두 지역의 가구 당 월 수입의 등 분산 검정을 먼저 수행. (유의수준 5%)

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad vs \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

① $H_0 : \sigma_1^2 = \sigma_2^2 \quad v.s. \quad H_1 : \sigma_1^2 \neq \sigma_2^2$

② $\alpha = 0.05$

③ $F = \dfrac{S_B^2}{S_A^2} = \dfrac{21.642^2}{20.369^2} = 1.129 \quad (\because \ S_A^2 < S_B^2)$

(A자유도 : m-1=19, B자유도 : n-1=24)

④ $\alpha = 0.05$의 양측검정이므로, 기각영역 : $F > 2.45$ or $F < 0.426$

우측 : $F_{\alpha/2}(n-1, m-1) = F_{0.025}(24,19) = 2.45$

좌측 : $F_{1-\alpha/2}(n-1, m-1) = \dfrac{1}{F_{\alpha}(m-1, n-1)} = \dfrac{1}{2.33} = 0.426 = F_{0.975}(24,19)$

⑤ $F = 1.129 < 2.45 = F_{0.025}(24,19)$이므로 H_0 기각, 즉, 두 지역의 분산이 다르지 않다고 할 수 있다.

두 집단의 평균차이 검정절차3

A, B 두 지역의 가구 당 월 수입의 평균차이 검정을 수행 : 독립 t-검정. (유의수준 5%)

귀무가설 H_0: $\mu_1 = \mu_2$(두 지역의 가구당 평균 월수입은 같다.)

대립가설 H_1: $\mu_1 \neq \mu_2$(두 지역의 가구당 평균 월수입은 같지 않다.)

① $H_0 : \mu_1 = \mu_2 \quad v.s. \quad H_1 : \mu_1 \neq \mu_2$

② $\alpha = 0.05$

③ $\alpha = 0.05$의 양측검정이므로, 기각영역 : $|t| > t_{\alpha/2}(m+n-2)$

$|t| > t_{0.025}(20+25-2) = t_{0.025}(43) = 2.017 \fallingdotseq z_{0.025}$

④ 두 모집단의 분산을 모르는 경우($\sigma_1^2 = \sigma_2^2$) :

검정통계량 $T = \dfrac{(\overline{X} - \overline{Y})}{S_p\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}} = \dfrac{(146.5-155.0)}{21.1\sqrt{1/20+1/25}} = -1.34$

⑤ $|T| = 1.34 < 2.017$0이므로 H_0을 기각하지 못함. 즉, 두 지역의 가구당 평균 월수입이 다르다고 할 만한 확실한 근거가 없다는 결론을 내린다.

CI 판정 : $CI \ni H_0 : \mu_1 - \mu_2 = 0 \Rightarrow H_0$ accept

p-value 판정 : $P = p(t \geq t_0) \times 2 < \alpha \Rightarrow H_0$ reject

대응(쌍체)표본의 가설검정.

쌍	1	2	…	n
처리1 (X)	X_1	X_2	…	X_n
처리2 (Y)	Y_1	Y_2	…	Y_n
차이 $(D=X-Y)$	D_1	D_2	…	D_n

가설 설정 : $H_0 : \mu_D = \mu_1 - \mu_2 = 0\ (\mu_1 = \mu_2)$ $\quad vs \quad$ $H_1 : \mu_D = \mu_1 - \mu_2 \neq 0 (\mu_1 \neq \mu_2)$

$D_i = X_i - Y_i (i = 1, \cdots, n)$, D_i의 평균과 분산 : $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$, $S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$

검정통계량 : $t_0 = \frac{\overline{D}}{S_D/\sqrt{n}} \sim t(n-1)$

대립가설에 따른 판정 :

대립가설	기각역
좌측검정 : $H_1 : \mu_1 - \mu_2 < 0$	$t_0 < -t(n-1, \alpha)$
우측검정 : $H_1 : \mu_1 - \mu_2 > 0$	$t_0 > t(n-1, \alpha)$
양측검정 : $H_1 : \mu_1 - \mu_2 \neq 0$	$t_0 < -t(n-1, \alpha/2)$ or $t_0 > t(n-1, \alpha/2)$

대응(쌍체)표본의 가설검정절차 1

[예제] 어떤 과학자는 휘발유에 자신이 개발한 새로운 첨가제를 넣을 경우에 주행거리를 증가시킨다고 주장하였다. 이를 확인하기 위하여 10종류의 자동차에 대하여 실험을 실시한 결과 1리터당 주행거리가 다음과 같았다.

차번호	1	2	3	4	5	6	7	8	9	10
사용전	11.5	15.8	20.3	18.2	25.3	9.8	14.8	16.0	16.4	13.5
사용후	12.3	15.5	23.2	19.5	28.1	10.2	16.2	15.8	16.5	14.4

위의 자료를 바탕으로, 이 과학자의 주장을 받아들일만한 뚜렷한 증거가 있는지를 유의수준 5%에서 검정하시오.

귀무가설 $H_0: \mu_D = 0$ (첨가제의 사용전과 후의 1리터당 주행거리는 차이가 없다.)

대립가설 $H_1: \mu_D < 0$ (첨가제의 사용후의 1리터당 주행거리는 사용전 보다 증가시킨다.)

$$D_i = (X_i - Y_i), \quad \overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}, \quad S_D^2 = \frac{\sum_{i=1}^{n} (D_i - \overline{D})}{n-1} \quad \Rightarrow \quad T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \sim t(n-1)$$

$$CI: \ \overline{D} \pm t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}$$

대응(쌍체)표본의 가설검정절차 2

① $H_0 : \mu_1 = \mu_2 (\mu_D = 0)$ $\quad v.s. \quad$ $H_1 : \mu_1 < \mu_2 (\mu_D < 0)$

② $\alpha = 0.05$

③ $\alpha = 0.05$의 좌측검정이므로, 기각영역 : $t < -t_\alpha(n-1)$

$t < -t_{0.05}(9) = -1.833$

④ 검정통계량 $T = \dfrac{\bar{D} - \mu_D}{S_D \sqrt{\dfrac{1}{n}}} = \dfrac{-1.01 - 0}{1.128 \times \sqrt{1/10}} = -2.83$

⑤ $T = -2.83 < -1.833$이므로 H_0을 기각함. 즉, 첨가제를 사용한 후에 1리터 당 주행거리가 첨가제를 넣기 전보다 증가한다고 할 수 있다.

$P = p(t \leq t_0) = p(t \leq -2.83) = 0.01$

두 모비율의 가설검정.

모집단 I : 확률표본을 $X_1, X_2, \cdots, X_m$

모집단 II : 확률표본을 $Y_1, Y_2, \cdots, Y_n$

X : 모집단 I 로부터 크기 m인 확률표본을 택한 경우의 성공 횟수 $X \sim B(m, p_1)$

Y : 모집단 II 로부터 크기 n인 확률표본을 택한 경우의 성공 횟수 $Y \sim B(n, p_2)$

모 비율의 추정량 표본비율 : $\hat{p}_1 = \frac{X}{m}, \quad \hat{p}_2 = \frac{Y}{n}$

$$E(\hat{p_1}) = \frac{1}{m}E(X) = p_1, \quad E(\hat{p_2}) = \frac{1}{n}E(Y) = p_2$$

$$Var(\hat{p_1}) = \frac{p_1(1-p_1)}{m}, \quad Var(\hat{p_2}) = \frac{p_2(1-p_2)}{n}$$

표본비율의 차 $\hat{p}_1 - \hat{p}_2$의 평균과 분산은 각각

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}$$

표본의 크기 m과 n이 큰 경우에는 $\hat{p}_1 - \hat{p}_2$는 근사적으로 정규분포를 따르게 된다.

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}} \sim N(0,1)$$

(...계속)

$p_1 - p_2$에 대한 $100(1-\alpha)\%$ 신뢰구간은 정규분포에 의한 근사로부터 다음과 같이 구할 수 있다.

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

귀무가설 H_0 하에서 검정통계량 : $Z_0 = \dfrac{(\hat{p}_1 - \hat{p}_2) - p_0}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{m} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n}}} \sim N(0,1)$

	귀무가설	대립가설	근사적으로 유의수준 α인 기각역
(a)	$H_0 : p_1 - p_2 = p_0$	$H_1 : p_1 - p_2 \neq p_0$	$\lvert Z_0 \rvert \geq z_{\alpha/2}$
(b)	$H_0 : p_1 - p_2 \leq p_0$	$H_1 : p_1 - p_2 > p_0$	$Z_0 \geq z_{\alpha}$
(c)	$H_0 : p_1 - p_2 \geq p_0$	$H_1 : p_1 - p_2 < p_0$	$Z_0 \leq -z_{\alpha}$

(...계속)

특히 $p_0 = 0$인 경우 :

$$H_0 : p_1 = p_2 \quad \text{vs} \quad H_1 : p_1 \neq p_2$$

귀무가설 H_0하에서 $p_1 = p_2$이므로 $p_1 = p_2 = p$라 하면, 충분히 큰 m과 n에 대하여, $\hat{p}_1 - \hat{p}_2$는 근사적으로 정규분포를 따르며, 귀무가설 H_0하에서 검정통계량 :

$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{m}+\frac{1}{n}\right)}} \sim N(0,1)$$

$\hat{p} = \frac{X+Y}{m+n}$로 계산한다. 여기서 $\hat{p}$를 모비율 p의 합동추정량이라 한다.

	귀무가설	대립가설	근사적으로 유의수준 α인 기각역
(a)	$H_0 : p_1 = p_2$	$H_1 : p_1 \neq p_2$	$\lvert Z_0 \rvert \geq z_{\alpha/2}$
(b)	$H_0 : p_1 \leq p_2$	$H_1 : p_1 > p_2$	$Z_0 \geq z_{\alpha}$
(c)	$H_0 : p_1 \geq p_2$	$H_1 : p_1 < p_2$	$Z_0 \leq -z_{\alpha}$

두 모비율의 가설검정(예제1)

[예제10.6] 신장이 정상적인 사람 100명과, 신장에 이상이 있는 사람 100명을 대상으로 폐렴에 대한 항생제를 주사하였다. 알레르기성 반응이 정상적인 사람 중에서는 21명, 신장에 이상이 있는 사람들 중에서는 38명이 나타났다.

(1) 신장에 이상이 있는 사람의 경우에 알레르기성 반응이 일어날 비율이 더 높다고 할 수 있는가? (유의수준 $\alpha = 0.01$)

(2) 모비율의 차에 대한 99% 신뢰구간을 구하여라.

(1) 정상적인 사람의 경우 알레르기성 반응이 일어날 모비율을 p_1,비정상적인 사람의 경우 알레르기성 반응이 일어날 모비율을 p_2라 하자.

① $H_0 : p_1 = p_2$ vs $H_1 : p_1 < p_2$

② 귀무가설 하에서 모 비율 p의 합동 추정 값 : $\hat{p} = \frac{X+Y}{m+n} = \frac{21+38}{100+100} = 0.295$

(...계속)

③ 검정통계량 :

$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{m}+\frac{1}{n}\right)}} = \frac{-0.17}{\sqrt{0.295\times 0.705\times\left(\frac{1}{100}+\frac{1}{100}\right)}} = -2.6359$$

④ 유의수준 1%에서의 기각역 : $Z_0 \le -z_{0.01} = -2.326$

⑤ 이므로 귀무가설 H_0를 기각할 수 있다. 즉, 신장에 이상이 있는 사람의 경우에 알레르기성 반응이 일어날 비율이 더 높다고 할 수 있다.

(2) 모비율의 차에 대한 99% 신뢰구간은

$$(\hat{p}_1 - \hat{p}_2) \pm z_{0.005}\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{100}+\frac{1}{100}\right)}$$

$$= (0.21 - 0.38) \pm 2.575\sqrt{0.295\times 0.705\left(\frac{1}{100}+\frac{1}{100}\right)}$$

$$= -0.17 \pm 0.166$$

$$= (-0.336, -0.004)$$