제 1 장 두 모집단의 처리에 대한 비교 분석

1 서 론

어떤 두 종류의 작물의 평균 수확량을 비교하던가, 새로 개발된 약품의 효능을 기존의 약품과 비교한다거나 또는 특정 지역의 토양 오염도를 비교하는 문제 등은 서로 다른 두 모집단의 특성을 비교하는 몇 가지 예이다. 두 모집단을 비교 분석하는 통계적 방법론은 비교대상이 되는 두 모집단에서 각각 하나씩 두 개의 독립인 표본을 추출하여 이로부터 두 모집단의 모평균, 모비율 차이와 그리고 모분산의 비를 추정하는 방법과 대응표본을 이용하여 두 모집단에 대한 평균 차이를 추정하는 방법으로 나누어 생각할 수 있다.

1. 실험 계획법(experimental design) 또는 표본설계법(sampling design)
; 실험 계획법이란 용어는 실험대상을 선택하여 각 처리에 할당하는 방법을 말한다. 두 처리를 비교하기 위한 계획의 두 가지 기본 형태는 독립표본(independent sample)(완전확률화) 과 대응표본(matched pairs sample)(각 대응 쌍내에서의 확률화)가 있다.

1) 실험단위(experimental unit) 또는 실험대상(experimental subject)
; 비교의 목적을 위하여 그 매개체로 사용되는 대상을 실험단위라고 한다.

2) 처리(treatment)
; 실험단위에 적용되어 특성치를 결정지어 주는 것 즉, 실험의 결과에 영향을 미치는 것을 처리라 한다.

3) 반응(response) 또는 처리효과(treatment effect)
; 실험 대상에 처리를 적용한 후에 나타나는 특성을 반응이라 한다.

2. 이표본 추론의 형태(두 모집단의 모평균 비교)

1) 독립표본(independent sample)-이표본(two sample)에 의한 평균비교
; 서로 다른 두 모집단이나 실험대상을 임의로 두 개의 집단으로 나누어 하나의 집단에 처리1을 적용시키고 나머지 집단에 처리2를 적용시켜 그 결과를 이용하여 두 처리를 비교하여 추론하는 방법 즉, 두 모집단에서 각각 독립적으로 표본을 추출하여 그들의 표본평균을 이용하여 추론하는 방법을 말한다.

2) 대응표본(correspondence sample)-쌍체비교(paired comparision)
; 실험대상을 동질적인 쌍(서로 비슷한 대상들이 각각 쌍을 이루고 쌍들 사이에는 실질적인 차이가 있도록 결합하는 것)으로 택하여 각 쌍에서 임의로 한 실험단위에는 처리1을 적용시키고 나머지 실험단위에는 처리2를 적용시켜 비교대상의 쌍들을 조사하여 각 쌍 내에서의 차를 이용하여 추론하는 방법을 즉, 두 모집단으로부터의 표본을 짝이 되도록 추출하고 짝이 되는 표본들끼리 서로 비교하여 추론하는 방법이다.

3. 확률화(randomization)의 원리
; 실험자가 마음대로 실험단위를 선택하여 처리를 적용할 수 있는 경우에는 이러한 점에 유의하여 두 처리의 공정한 비교를 할 수 있도록 해야 한다. 공정하고 객관적인 비교를 하기 위해서는 공정하게 실험단위를 택하는 것이 가장 중요하다. 실험자의 주관에 맡기면, 무의식적 일지라도 특정한 처리에 치우칠 수 있기 때문이다. 이와 같이 편견이 없도록 선택하

는 원리를 확률화의 원리라 한다. 확률화 원리로 실험단위를 선택하면 반응측정값에 나타나는 처리 효과 이외의 요인에 영향을 배제할 수 있어 효과적인 실험계획의 가장 기본적인 원리이다.

2 두 모집단에서의 독립확률표본

1. 통계적 모형

1) 통계적 모형

통계적 모형 : 독립인 확률표본

① $X_1,\ X_2,\ \ldots,\ X_{n_1}$은 평균이 μ_1이고 표준편차가 σ_1인 모집단 1에서의 크기 n_1인 확률표본이다.
② $Y_1,\ Y_2,\ \ldots,\ Y_{n_2}$는 평균이 μ_2이고 표준편차가 σ_2인 모집단 2에서의 크기 n_2인 확률표본이다.
③ 표본들은 독립이다. 다시 말해, 한 처리의 반응측정값들은 다른 처리에 의한 반응측정값들과 서로 무관하다.

2) 표본과 통계량

표본	통계량	
모집단1 : $X_1,\ X_2,\ \ldots,\ X_{n_1}$	$\overline{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i,$	$S_1^2=\frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2}{n_1-1}$
모집단2 : $Y_1,\ Y_2,\ \ldots,\ Y_{n_2}$	$\overline{Y}=\frac{1}{n_2}\sum_{i=1}^{n_2}Y_i,$	$S_2^2=\frac{\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2}{n_2-1}$

3 두 평균의 차이에 관한 대표본 추론

1. 두 모평균의 차 $\mu_1-\mu_2$에 대한 점추정

1) $\mu_1-\mu_2$의 점추정값 ; $\widehat{\mu_1-\mu_2}=\overline{X}-\overline{Y}$

2) ① $\overline{X}\approx N\left(\mu_1,\ \frac{\sigma_1^2}{n_1}\right),\ \overline{Y}\approx N\left(\mu_2,\ \frac{\sigma_2^2}{n_2}\right)$

② $E(\overline{X}-\overline{Y})=\mu_1-\mu_2,\ Var(\overline{X}-\overline{Y})=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$

③ $\overline{X}-\overline{Y} \approx N\left(\mu_1-\mu_2,\ \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$

3) $Z=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \approx N(0,\ 1),\ \hat{\sigma_1}=S_1,\ \hat{\sigma_2}=S_2$

2. 대표본에서 $\mu_1-\mu_2$의 신뢰구간

$\mu_1-\mu_2$의 근사적인 $100(1-\alpha)\%$ 신뢰구간

$$\left(\overline{X}-\overline{Y}-Z_{\alpha/2}\times\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}},\ \overline{X}-\overline{Y}+Z_{\alpha/2}\times\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right)$$

3. 대표본에서의 가설 검정 : Z-검정

대표본에서 H_0 : $\mu_1-\mu_2=\delta_0$의 가설검정

1) 검정 통계량

; $Z=\frac{\overline{X}-\overline{Y}-\delta_0}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$

2)

[대립가설]	[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 : $\lvert Z\rvert \geq Z_{\alpha/2}$

4 소표본에서의 추론 ; 분산이 같은 정규모집단들

1. 통계적 모형

; 일반적으로 소표본에서 적절한 추론을 하기 위해서는 모집단 분포에 대해 보다 많은 가정을 필요로 한다. 물론, 두 모집단이 정규모집단이 아니거나 또는 두 모집단의 분산이 공통이 아닌 경우에도 평균차에 대한 추론은 할 수 있다. 다만 아래에서와 같은 가정하에서는 정확한 추론이 이루어지지만 그렇지 않는 경우는 근사적인 방법으로 추론이 되기 때문에 여기에서는 모집단 분포가 다음 가정을 만족할 때 유효한 소표본에서의 추론 방법을 소개한다.

두 모집단의 평균 추론을 위한 독립표본

1) 가정

① 두 모집단이 정규분포를 따른다.

$X_1,\ X_2,\ \cdots,\ X_n$: $N(\mu_1,\ \sigma_1^2)$에서의 확률표본

$Y_1,\ Y_2,\ \ldots,\ Y_n$: $N(\mu_2,\ \sigma_2^2)$에서의 확률표본

$X_1,\ X_2,\ \ldots,\ X_n$ 와 $Y_1,\ Y_2,\ \ldots,\ Y_n$은 서로 독립이다.

② 두 모집단의 분산이 같다.

$\sigma_1^2 = \sigma_2^2$ (등분산성)

2) 통계량

① $\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$, $S_1{}^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\overline{X})^2$, $\overline{X} \sim N\left(\mu_1,\ \frac{\sigma_1^2}{n_1}\right)$

② $\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$, $S_2{}^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2$, $\overline{Y} \sim N\left(\mu_2,\ \frac{\sigma_2^2}{n_2}\right)$

2. 독립표본에서 모평균 차$(\mu_1-\mu_2)$의 점추정

1) $\mu_1-\mu_2$의 점추정값 ; $\widehat{\mu_1-\mu_2} = \overline{X}-\overline{Y}$

2) ① $E(\overline{X}-\overline{Y}) = \mu_1-\mu_2$

② $Var(\overline{X}-\overline{Y}) = \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2} = \sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$

③ $S.E(\overline{X}-\overline{Y}) = \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

④ 추정된 $S.E(\overline{X}-\overline{Y}) = S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

3) 공통분산 σ^2의 합동 추정량(pooled estimator)

; $S_p^2 = \frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2+\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2}{n_1+n_2-2} = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$

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

3. 소표본에서 $\mu_1-\mu_2$의 신뢰구간

$\mu_1-\mu_2$의 $100(1-\alpha)\%$신뢰구간

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

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

4. 소표본에서의 가설 검정 : 독립표본 T-검정(independent T-test)

소표본에서 $H_0\ :\ \mu_1-\mu_2=\delta_0$의 검정

1) 검정 통계량

; $T=\frac{\overline{X}-\overline{Y}-\delta_0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, $\quad df=\ n_1+n_2-2$

2)

[대립가설]	[기각역]
$H_1\ :\ \mu_1-\mu_2\ >\ \delta_0$	$R\ :\ T\geq\ t_\alpha(n_1+n_2-2)$
$H_1\ :\ \mu_1-\mu_2\ <\ \delta_0$	$R\ :\ T\leq\ -t_\alpha(n_1+n_2-2)$
$H_1\ :\ \mu_1-\mu_2\ \neq\ \delta_0$	$R\ :\ \lvert T\rvert\ \geq\ t_{\alpha/2}(n_1+n_2-2)$

5. 공통분산 가정의 결정

1) 정규모집단의 소표본들에 대해 통계량 $\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$이 $t-$분포를 따르지 않을 뿐더러(이론적으로 밝혀짐) 미지의 양 σ_1^2/σ_2^2에 의해 좌우되므로 소표본 추론에서는 등분산성$(\sigma_1^2=\sigma_2^2)$을 가정하면 $\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$이 $t-$분포를 따르게 되어 추론이 가능하다.

그러나 대표본의 경우는 정규근사에 의해 등분산성과 σ의 합동추정값을 계산할 필요가 없다.

2) 등분산성 가정이 힘든 경우의 추론에서는 대략적으로 두 표본분산에 대해 $1/4\leq S_1^2/S_2^2\leq 4$인 경우에는 합동 추정값을 이용하여 추정하는 것이 타당하다고 알려져 있다. 그렇지 않을 경우에는 $\sigma_1=\sigma_2$ 라는 가정을 의심해야 한다. 이 경우에는 $\mu_1-\mu_2$에 관한 추론 방법을 자유도를 적절히 조절하여 $t-$검정을 하면 된다.

예제 **1-1**. 사람은 일을 하는데 상당한 비율의 시간을 소모한다. 이때 직업에 대한 만족도와 삶에 대한 만족도는 서로 관련이 있는 경향이 있다. 직업 만족도는 일반적으로 4점 척도로 측정된다. 수치 척도로는 매우 불만족은 1점, 불만족은 2점, 만족은 3점, 매우 만족은 4점로 할당하였다. 226명의 소방관과 247명의 사무실 관리자들의 응답 결과를 요약하면 다음과 같다.

	소방관	관리자
평균	3.673	3.547
표준편차	0.7235	0.6089

a) 평균 직업 만족도 차에 대한 점추정값을 계산하여라.

b) 평균 직업 만족도 차에 대한 99% 신뢰구간을 구하여라.

c) 데이터는 소방관들의 평균 직업 만족도가 관리자들의 평균 직업 만족도와 다르다는 강력한 증거를 제공해주는지 유의수준 2%에서 검정하여라.

예제 **1-2.** 2020년 6월에 어느 도시에 있는 호수의 85개 지역에서 단위용적의 물을 표본으로 채취하여 염소 함유량을 측정하였다. 2년 후인 2022년 6월에 110개의 물을 표본으로 채취하여 분석을 통해 염소 함유량을 측정하여 기록하였다. 이들로부터 평균과 표준편차를 계산하여 다음을 얻었다.

	2020	2022
평균	18.3	17.8
표준편차	1.2	1.3

a) 평균 염소함유량의 차에 대한 점추정값을 계산하여라.

b) 평균 염소함유량의 차에 대한 90% 신뢰구간을 구하여라.

c) 위 데이터는 2010년 수준과 비교하여 2012년에 호수물의 평균염소 함유량이 감소되었다는 근거를 제시하는지 유의수준 5%에서 검정하여라.

예제 **1-3.** 성별에 따라 콜레스테롤 지수의 차이를 알기위해 플로리다주의 성인 107명을 대상으로 콜레스테롤 지수를 조사하였다. 조사자는 남성 58명과 여성 49명의 콜레스테롤 지수가 차이가 있다고 생각한다. 다음 물음에 답하여라.

	남	여
평균	216.07	219.82
표준편차	46.49	60.65

a) 성별에 따른 평균 콜레스테롤 지수의 차에 대한 95% 신뢰구간을 구하여라.

b) 성별에 따른 콜레스테롤 지수에 차이가 있다고 말할 수 있는지를 유의수준 5%에서 검정하여라.

c) 유의확률을 계산하여라.

예제 1-4. 회계학과 신입생 남학생과 여학생들을 대상으로 시험을 보아 그들의 답에 근거하여 컴퓨터 불안 평가점수(CARS)를 조사한 결과는 아래 표와 같다. 다음 물음에 답하여라.

성별	학생 수	평균	표준편차
남	15	2.514	0.773
여	20	2.963	0.525

a) 남학생과 여학생에 대한 불안 평가점수의 평균에 대한 차이의 95% 신뢰구간을 구하여라.

b) 여학생들의 평균 점수가 남학생들의 평균 점수보다 낮다고 주장할 수 있는지 유의수준 5%에서 검정하여라.

예제 **1-5.** 어느 대학교 휴게실 앞에 설치된 두 자판기에 대한 학생들의 만족도에 차이가 있는지 알고 싶어 학생 22명을 대상으로 만족도에 대한 설문 조사를 실시하였다. 기계 A와 기계 B에 대해 각각 10명과 12명의 학생이 만족도에 대해 답한 결과는 아래 표와 같다. 다음 물음에 답하여라.

기계	학생 수	평균	표준편차
A	10	5.38	1.59
B	12	5.92	0.83

a) 두 자판기에 대한 만족도 차이의 90% 신뢰구간을 구하여라.

b) 두 자판기에 대해 학생들의 만족도에 차이가 있다고 말할 수 있는지에 대하여 유의수준 10%에서 검정하여라.

예제 1-6. 트럭운송회사의 한 저장소에서는 다른 저장소로 상품을 운송하기 위해 두 도로 중 어느 하나를 선택하고자 한다. 10명중 5명을 선택하여 그들에게 도로 A를 적용하고 나머지 5명에게는 도로 B를 적용하여 운송시간을 측정하여, 도로 A와 도로 B에서의 평균운송시간에 차이가 있는지 알아보고자 한다.

트럭회사	운송시간(단위 : 시간)				
도로A	1.8	2.4	3.0	2.1	3.2
도로B	2.2	2.9	3.4	2.5	3.5

위 데이터에 대하여 분석한 결과는 다음과 같다.

집단통계량

	도로	N	평균	표준편차	평균의 표준오차
시간	도로A	5	2.500	.5916	.2646
	도로B	5	2.900	.5612	.2510

a) 도로에 따른 따른 평균 수송 시간의 차이에 대한 95% 신뢰구간을 구하여라.

b) 위의 데이터는 도로 A와 도로 B에서의 평균 운송시간에 차이가 있다고 할 수 있는지를 유의수준 5%에서 검정하여라.

예제 1-7. 다수의 사람들이 13일의 금요일을 불길한 날로 두려워한다. 영국의 한 연구소의 연구원들은 6일 금요일과 13일 금요일의 교통사고나 교통사고로 인한 병원 입원에 차이가 있는지 알아보기 위하여 조사한 결과 다음과 같았다. 교통사고나 교통사고로 인한 병원 입원에 차이가 있는지를 유의수준 1%에서 검정하여라.

년 월	6일 금요일	13일 금요일
1989년 10월	9	13
1990년 07월	6	12
1991년 09월	11	14
1991년 12월	11	10
1992년 03월	3	4
1992년 11월	5	12

예제 **1-8**. 다음은 인슐린에 의해 저혈당(hypoglycemia)상태 동안 발한 후 프로프라노롤(pro-pranolol)을 받은 환자와 대조군 환자에 대한 누적 체중감소량에 대한 자료이다. 두 집단의 평균 체중감소량에 차이가 있는지 유의수준 1%에서 검정하여라.

집단	시료 수	평균	표준편차
프로프라노롤	12	120	10
대조군	11	70	8

예제 1-9. 흡연은 혈중 COHb량을 증가시킨다고 알려져있다. 이를 검정하기 위하여 흡연자 12명과 비흡연자 12명의 혈중 COHb량을 비교한 결과는 아래와 같다. 흡연자들의 COHb량이 비흡연자에 비하여 높아고 주장할 수 있는가를 유의수준 1%에서 검정하여라.

흡연유무	N	평균	표준편차
흡연자	12	49.39	317.453
비흡연자	12	37.04	172.904

예제 **1-10**. 어느 두 지역의 호수 장흥지와 숭뢰지의 BOD 함유량을 비교 분석하기 위하여 두 호수 각각 24곳을 선택하여 BOD 함유량을 조사하여 SPSS로 분석한 결과는 아래 표와 같다. 다음 물음에 답하여라.

집단통계량

	분류	N	평균	표준편차	평균의 표준오차
BOD	장흥지	24	2.567	1.3027	.2659
	숭뢰지	24	1.767	1.0349	.2112

독립표본 검정

		Levene의 등분산 검정		평균의 동일성에 대한 t-검정						
									차이의 95% 신뢰구간	
		F	유의확률	t	자유도	유의확률 (양쪽)	평균차	차이의 표준오차	하한	상한
BOD	등분산이 가정됨	.023	.879	2.356	46	.023	.8000	.3396	.1164	1.4836
	등분산이 가정되지 않음			2.356	43.761	.023	.8000	.3396	.1154	1.4846

a) 평균 BOD 함유량 차이에 대한 95%신뢰구간을 구하여라.

b) 위의 결과로부터 장흥지의 BOD 함유량이 승뢰지의 BOD 함유량과 서로 다르다고 주장할 수 있는지를 유의수준 1%에서 검정하여라.

c) 위의 결과로부터 장흥지의 BOD 함유량이 숭뢰지에 비하여 높다고 주장 할 수 있는지를 유의수준 1%에서 검정하여라.

예제 **1-11**. 한 연구자는 혈액 응고 연구를 위해 60명을 대상으로 두 혈액 수집 방법을 비교하였다. 연구자는 방법1 30명과 방법2 30명의 APTT(Activated Partial Thromboplastin Time)값에 대한 결과는 아래와 같다. 방법1의 평균이 방법2에 비하여 높다고 주장할 수 있는 가를 유의수준 5%에서 검정하여라.

APTT	방법1	방법2
평균	33.369	32.000
표준편차	10.7299	10.4570

예제 **1-12.** 서로 다른 두 종류의 치료방법에 대한 환자의 회복시간의 차이에 대해 조사하고자 하여 82명을 대상으로 40명을 랜덤추출하여 치료방법 A를, 나머지는 치료방법 B를 적용한 결과 다음과 같은 결과를 얻었다.

치료방법	방법A	방법B
평균회복시간	7.5	8.8
표준편차	1.3	1.5

이 자료에 의해 두 치료방법에 의한 평균회복시간의 차이가 있는지 유의수준 1%에서 검정하여라.

5 두 개의 대응표본에서의 추론

1. 대응표본의 data 구조

1) 대응 쌍에 의한 표본 추출법(sampling by matched pairs)

; 지금까지는 독립인 두 확률표본에 근거를 둔 두 처리효과의 비교를 다루었다. 이와 같이 두 확률표본에 기초를 둔 추론에서 두 처리효과를 비교하기 위해서는 실험단위나 대상이 가능한 비슷해야만 두 집단의 반응 측정값 차이가 처리효과의 차이에서 기인된다 할 수 있다. 그러나 이러한 동질성 조건을 만족시킬 수 없는 경우가 많으며, 반응에 영향을 주는 여러 가지 요인을 실험자가 조절할 수 없게 되어 반응측정값의 변동이 심하게 되고, 따라서 처리효과의 차이를 모호하게 할 수 있다. 예를 들어, 두 종류의 진통제를 비교하기 위해 성별과 나이 그리고 종합적인 건강 조건이 같고 통증의 정도가 같은 환자를 많이 찾는다는 것은 현실적으로 어렵다. 이러한 환자들에게 처리를 적용할 때 이용할 수 있는 추론 방법이 대응 표본에서의 추론이다. 대응(matching) 또는 구획화(blocking)는 실험단위들이 비슷하면서 다른 종류들인 두 가지 양립되는 조건을 만족하기 위한 것이다. 구획화에서는 유사한 실험단위까지 여러 개의 구획 또는 쌍으로 나눈다. 물론 각 쌍에 있는 단위는 비슷하게 하고 다른 쌍에 있는 단위와는 차이가 있도록 구성을 한다. 이와 같이 함으로써, 각 쌍 내에서는 처리효과의 비교를 쉽게 하고, 서로 다른 쌍 간에는 이질성을 허용함으로써 추론의 범위를 확대할 수 있게 된다. 물론 실험단위를 선택할 때 편견을 피하기 위해 처리는 임의로 각 쌍에 적용시켜야 한다. 이 계획법을 대응 쌍에 의한 표본 추출법이라 한다.

2) 표본 대응비교에서의 데이터 구조

표본 대응비교에서의 데이터 구조

쌍	처리1	처리2	차이
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	$D_n = X_n - Y_n$

3) 표본 대응비교에서의 통계량

표본 대응비교에서의 통계량

$D_i = X_i - Y_i$ 들이 $N(\mu_D, \sigma^2)$ 분포에서의 확률표본이라 가정할 때 통계량은 다음과 같다.

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

2. 대응표본에서 모평균차(μ_D : 모집단에서의 처리효과의 평균차이)의 점추정

1) μ_D의 점추정값 ; $\widehat{\mu_D} = \overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$

2) $D_i = X_i - Y_i$: $N(\mu_D,\ \sigma_D^2)$ 분포에서의 확률표본

① $E(D_i) = \mu_D,\ Var(D_i) = \sigma_D^2$

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

3) ① $S.E.(\overline{D}) = \sigma_D/\sqrt{n}$

② 추정된 $S.E.(\overline{D}) = S_D/\sqrt{n}$

4) 대표본의 경우 ; $Z = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \approx N(0,\ 1)$ 소표본의 경우 ; $T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \sim t(n-1)$

3. 대응표본에서 μ_D의 구간추정

; 정규모집단으로부터 소표본($n < 30$)인 확률표본 $D_1,\ D_2,\ \cdots\ D_n$에 의해 정의되는 표본평균과 표본 분산은 각각 $\overline{D}$, S_D^2이라고 할 때, μ_D의 $100(1-\alpha)$ 신뢰구간은 다음과 같다.

μ_D의 근사적인 $100(1-\alpha)\%$신뢰구간
$(\overline{D} - t_{\alpha/2}(n-1) \times S_D/\sqrt{n},\ \overline{D} + t_{\alpha/2}(n-1) \times S_D/\sqrt{n})$

4. 소표본에서의 가설검정 - 대응표본 T-검정(paired T-test)

소표본에서 H_0 : $\mu_D = \mu_0$의 가설검정

1) 검정 통계량 : $T_D = \frac{\overline{D} - \mu_0}{S_D/\sqrt{n}}$, $d.f. = n-1$

2)

[대립가설]	[기각역]
H_1 : $\mu_D > \mu_0$	R : $T_D \geq t_\alpha(n-1)$
H_1 : $\mu_D < \mu_0$	R : $T_D \leq -t_\alpha(n-1)$
H_1 : $\mu_D \neq \mu_0$	R : $\lvert T_D \rvert \geq t_{\alpha/2}(n-1)$

$c.f.$ 위의 추론은 표본 $D_1,\ D_2,\ \cdots\ D_n$이 정규모집단으로부터의 랜덤표본이라는 전제하에 정확한 것이며, 표본의 크기 n이 충분히 큰 경우에는 모집단의 분포가 정규분포가 아니더라도 중심극한 정리에 의해 $Z = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \approx N(0,\ 1)$이므로 정규검정 $(Z-test)$이 가능하다.

예제 **1-13**. 산업안전교육은 사고에 기인한 노동시간의 손실을 감소하는데 효과가 있다고 주장한다. 다음의 데이터는 안전교육을 시작하기 전과 후에 6개 공장에서 발생한 사고로 인하여 1주일간에 손실된 작업시간이다.

	공 장					
	1	2	3	4	5	6
안전교육 전(x)	12	29	16	37	28	15
안전교육 후(y)	10	28	17	35	25	16
$d=(x-y)$						

안전교육을 실시하기 전과 실시 후의 손실된 작업시간의 차의 분포는 정규분포를 따른다고 가정할 때, 다음 물음에 답하여라.

a) μ_D에 대한 90%신뢰구간을 구하라.

b) 이 데이터에 의하면, 위의 주장이 입증되는지를 유의수준 1%에서 검정하여라.

예제 **1-14.** 의학연구원은 특정 암에 대한 진통제가 사용자의 혈압을 저하시키는지를 알고자 한다. 먼저 15명의 환자를 대상으로 진통제를 복용하기 전에 혈압을 측정하여 기록하였다. 그리고 그들에게 6개월 동안 정기적으로 진통제를 복용하게 한 후, 다시 혈압을 측정하였다. 표에 주어진 데이터에 의해 진통제의 영향에 관해 추론하고자 한다.

	실 험 대 상														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
복용 전(x)	70	80	72	76	76	76	72	78	82	64	74	92	74	68	84
복용 후(y)	68	72	62	70	58	66	68	52	64	72	74	60	74	72	74
$d=(x-y)$															

a) μ_D에 대한 95%신뢰구간을 구하라.

b) 위의 데이터로부터 진통제를 복용하면 혈압이 저하된다고 주장할 수 있는지를 $\alpha=0.05$에서 검정하라.

예제 **1-15.** 두 종류의 감기약이 수면시간의 증가에 미치는 상대적인 효과를 연구하기 위해 감기에 걸린 6명에게 첫째 날 저녁에는 약 A를 복용하고 둘째 날 저녁에는 약 B를 복용시켰다. 그리고 각 밤에 그들의 수면시간을 기록하여 아래의 데이터를 얻었다.

약 A	4.8	4.1	5.8	4.9	5.3	7.4
약 B	3.9	4.2	5.0	4.9	5.4	7.1

위 데이터에 대하여 SPSS로 분석한 결과는 다음과 같다.

대응표본 통계량

		평균	N	표준편차	평균의 표준오차
대응 1	약A	5.383	6	1.1374	.4643
	약B	5.083	6	1.1303	.4615

대응표본 검정

	대응차					t	자유도	유의확률 (양쪽)
	평균	표준편차	평균의 표준오차	차이의 95% 신뢰구간 하한	상한			
대응 1 약A - 약B	.3000	.4517	.1844	-.1740	.7740	1.627	5	.165

a) 이와 같은 경우에 적용되는 검정법을 무엇이라 부르는가?

b) 두 종류의 감기약에 대한 평균 수면시간의 차이의 95% 신뢰구간을 구하여라.

c) 평균 수면시간이 차이가 나는지 유의수준 5%로 검정하여라.

예제 **1-16.** 어느 식품학자는 PC라는 특정 형태의 박테리아의 예비 배양을 한 것과 하지 않은 탈지 우유로 만든 요구르트 사이에 품질의 차이가 있는지 연구하고자 한다. 7개의 낙농 농장에서 탈지 우유의 표본이 조달되었다. 각 농장에서 가져온 우유의 절반은 PC 처리를 하고 나머지 반은 PC 처리를 하지 않고 이 표본 우유로 요구르트를 만들어서 그 요구르트의 탄력을 측정한 결과 다음과 같다.

	1	2	3	4	5	6	7
PC처리함	68	75	62	86	52	46	72
PC처리 하지 않음	61	69	64	76	52	38	68

위 데이터에 대하여 SPSS로 분석한 결과는 다음과 같다.

대응표본 통계량

		평균	N	표준편차	평균의 표준오차
대응 1	PC처리함	65.86	7	13.741	5.194
	PC처리하지않음	61.14	7	12.628	4.773

대응표본 검정

	대응차					t	자유도	유의확률 (양측)
				차이의 95% 신뢰구간				
	평균	표준편차	평균의 표준오차	하한	상한			
대응 1 PC처리함 - PC처리하지않음	4.714	4.348	1.643	.693	8.735	2.869	6	.028

a) PC 처리로 인한 탄력의 평균증가에 대한 90% 신뢰구간을 구하여라.

b) 이 데이터는 PC 처리를 한 요구르트의 탄력이 더 높다는 추측을 입증하는가? $\alpha = 0.05$ 에서 검정하여라.

예제 **1-17.** 코로나19로 인한 2주간의 자가격리가 투수의 구속을 저하시키는 알아보기 위하여 투수 11명을 대상으로 자가격리 전과 자가격리 후의 구속을 조사하였다.

	1	2	3	4	5	6	7	8	9	10	11
자가격리 전	146.8	145.2	148.8	146.2	145.3	143.4	146.6	145.9	144.7	150.1	143.2
자가격리 후	146.9	146.7	144.1	145.9	142.3	145.2	144.5	145.2	143.0	147.6	141.9

위 데이터에 대한 SPSS 결과는 다음과 같다.

대응표본 통계량

		평균	N	표준편차	평균의 표준오차
대응 1	전	146.018	11	2.0764	.6261
	후	144.845	11	1.8933	.5709

대응표본 상관계수

	N	상관계수	유의확률
대응 1 전 & 후	11	.530	.093

대응표본 검정

	대응차					t	자유도	유의확률 (양쪽)
				차이의 95% 신뢰구간				
	평균	표준편차	평균의 표준오차	하한	상한			
대응 1 전 - 후	1.1727	1.9309	.5822	-.1244	2.4699	2.014	10	.072

a) 2주간의 자가격리가 투수의 구속을 저하 시키는 효과가 있는지 유의수준 5%에서 검정하여라.(단, 검정통계량과 기각역을 이용)

b) 2주간의 자가격리가 투수의 구속을 저하 시키는 효과가 있는지 유의수준 5%에서 검정하여라.(단, 유의확률과 유의수준을 이용)

예제 **1-18**. 두 종류의 사료가 젖소의 우유 생산량에 미치는 영향을 조사하기 위하여 25마리의 젖소를 대상으로 실험하였다. 25마리 중에서 무작위로 선택된 12마리에게는 인공건초를 사료로 사용하고 나머지 13마리에게는 자연건초를 사료로 사용하여 3주일간 우유 생산량을 조사하여 하루 우유의 평균 생산량을 아래 표에 기록하였다.

우유 생산량(단위:kg)												
자연건초	19.8	19.8	25.2	20.7	21.15	27.1	26.1	23.85	22.05	15.75	20.7	28.50	18.45
인공건초	15.75	21.15	24.75	13.05	18	17.55	14.4	18.45	18.9	25.65	22.95	17.55	

위 데이터에 대하여 SPSS로 분석한 결과는 다음과 같다.

집단통계량

	집단	N	평균	표준편차	평균의 표준오차
구분	자연건초	13	22.2423	3.69029	1.02350
	인공건초	12	19.0125	3.93297	1.13535

독립표본 검정

		Levene의 등분산 검정		평균의 동일성에 대한 t-검정						
		F	유의확률	t	자유도	유의확률 (양쪽)	평균차	차이의 표준오차	차이의 95% 신뢰구간 하한	상한
생산량	등분산이 가정됨	.007	.936	2.119	23	.045	3.22981	1.52453	.07607	6.38355
	등분산이 가정되지 않음			2.113	22.514	.046	3.22981	1.52859	.06390	6.39571

다음 물음에 답하여라.

a) 평균 우유 생산량의 차이에 대한 95%신뢰구간을 구하여라.

b) 위의 결과로부터 인공건초를 사료로 사용하면 자연건초를 사료로 사용한 경우보다 우유 생산량이 적다는 것을 주장할 수 있는가?($\alpha = 0.05$)

예제 **1-19.** 다음 자료는 어느 대학의 신입생 68명에 대상으로 학기초와 학기말의 몸무게의 자료를 조사한 것이다. 학기 말에 몸무게가 증가했다고 주장 할 수 있는가를 유의수준 1%에서 검정하여라. (단위 : kg)

ID	전	후	ID	전	후	ID	전	후	ID	전	후
1	77.6	76.2	18	57.6	57.6	35	67.1	68.0	52	60.8	60.8
2	49.9	50.3	19	46.3	47.6	36	74.4	74.8	53	68.5	68.5
3	60.8	61.7	20	56.7	56.7	37	62.1	62.6	54	57.6	59.0
4	52.2	54.0	21	71.2	71.7	38	89.8	91.2	55	48.1	49.0
5	68.0	70.3	22	54.0	57.2	39	55.3	56.2	56	83.9	85.3
6	47.2	48.1	23	51.3	51.7	40	66.2	66.2	57	56.7	58.1
7	64.4	67.1	24	54.4	58.1	41	68.0	68.5	58	56.7	57.2
8	54.4	56.2	25	61.2	63.1	42	84.8	87.1	59	70.3	71.7
9	65.3	67.1	26	67.1	68.0	43	42.6	43.5	60	53.5	54.4
10	70.8	69.9	27	49.9	50.8	44	47.6	47.6	61	67.6	68.0
11	51.7	51.7	28	72.6	73.9	45	57.6	59.0	62	67.6	67.6
12	54.9	55.8	29	99.8	101.6	46	64.4	65.3	63	55.3	54.9
13	55.3	57.2	30	59.9	60.3	47	63.5	64.9	64	70.3	71.7
14	54.4	52.2	31	65.8	66.7	48	48.5	48.5	65	72.6	73.0
15	52.2	53.5	32	64.0	64.0	49	47.2	47.6	66	52.2	54.0
16	49.9	51.3	33	71.7	72.6	50	50.3	50.8	67	75.8	77.1
17	64.4	66.2	34	61.2	60.8	51	72.6	73.5	68	59.4	59.4

예제 **1-20.** 야구부가 있는 B고등학교는 선수들의 타격 능력을 향상 시키시 위하여 트랙맨 데이터를 활용하였다. 10명을 대상으로 트랙맨 데이터를 활용하기 전과 후의 타율 데이터는 다음과 같다. 타율이 향상되었다고 주장할 수 있느가를 유의수준 5%에서 검정하여라.

구분	1	2	3	4	5	6	7	8	9	10
before	0.257	0.202	0.314	0.306	0.221	0.187	0.236	0.276	0.232	0.354
after	0.275	0.215	0.354	0.315	0.235	0.197	0.242	0.283	0.275	0.374

예제 **1-21.** 다음 자료는 면역 4주 전과 4주 후의 항체농도이다. 면역 전과 후에 대하여 평균 항체농도 간에 차이가 있는지를 유의수준 5%에서 검정하여라.

구분	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
before	0.4	0.4	0.4	0.4	0.5	0.5	0.5	0.5	0.5	0.6	0.7	0.7	0.8	0.9	0.9	1.0	1.0	2.0
after	0.4	0.5	0.5	0.9	0.5	0.5	0.5	0.5	0.5	0.6	1.1	1.2	0.8	1.2	1.9	0.9	2.0	3.7

예제 **1-22.** 왼손으로 글을 쓰는 사람 10명에 대하여 오른손과 왼손의 악력에 대한 측정을 한 결과가 다음과 같다. 왼손으로 글을 쓰는 사람은 왼손의 악력이 오른손에 보다 강하다고 주장할 수 있는지를 유의수준 5%에서 검정하여라.

구분	1	2	3	4	5	6	7	8	9	10
왼손	140	90	125	130	95	121	85	97	131	110
오른손	130	87	110	132	96	120	86	90	129	100

예제 **1-23.** 다음의 표는 15개 지경의 2001년과 2007년 소비자물가지수의 값이다. 2007년의 소비자물가지수의 2001년에 비하여 높다고 주장할 수 있는지를 유의수준 1%에서 검정하여라.

연도	앵커리지	애틀란타	보스턴	시카고	클리블랜드
2001년	155	176	191	178	173
2007년	181	198	227	198	186

연도	덴버	디트로이트	호놀룰루	휴스턴	캔자스시티
2001년	181	174	178	159	172
2007년	194	195	219	182	186

연도	로스앤젤러스	마이애미	미니아폴리스	뉴욕	필라델피아
2001년	177	173	177	187	181
2007년	210	210	195	221	216