5장 모집단 평균벡터에 관한 추론

덕성여자대학교 정보통계학과 김 재희

5.1 서론

- 다변량 정규 모집단으로부터 얻은 확률표본에 대해 모집단의 평균벡터에 대한 추론
- 다변량 정규분포 모집단 분포의 주요 관심 모수인 모평균벡터와 그의 성분에 대한 추론
- 일변량 t 통계량 사용한 경우의 모평균 문제를 다변량으로 확장
- Hotelling의 T^2 통계량
- 모집단이 다변량 정규분포를 따르는 경우 우도비를 이용한 검정법
- 모평균벡터에 관한 동시신뢰영역문제

5.2 Student t-검정과 다변량인 경우로의 확장

5.2.1 Student t-통계량과 Hotelling T^2 통계량

일변량 확률표본 $X_1, X_2, \cdots, X_n$이 서로 독립이며 $N(\mu_0, \sigma^2)$를 따를 때

- 모분산에 대한 불편추정량 $\hat{\sigma^2} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
- t-통계량은

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

- $$t^2 = \frac{(\overline{X} - \mu_0)^2}{(s/\sqrt{n})^2} = n(\overline{X} - \mu_0)(s^2)^{-1}(\overline{X} - \mu_0)$$

■ 다변량 확률표본 $X_1, X_2, \cdots, X_n$: 서로 독립이며 p-변량 정규분포 $N_p(\mu_0, \Sigma)$

여기서 모공분산행렬 Σ는 미지(unknown)일 때, $\hat{\Sigma} = S$를 얻는다.

모평균벡터에 대한 추론을 위한 통계량으로

$$T^2 = (\overline{X} - \mu_0)' \left(\frac{S}{n}\right)^{-1} (\overline{X} - \mu_0) = n(\overline{X} - \mu_0)' S^{-1} (\overline{X} - \mu_0)$$

: Hotelling의 T^2 통계량

여기서

$\mu_{p\times 1}{}' = (\mu_{10}, \mu_{20}, \ldots, \mu_{p0})$: 모평균벡터

j 번째 개체에 대한 확률벡터 $X_j' = (X_{1j}, X_{2j}, \ldots, X_{pj})$

$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$: $p\times 1$ 표본평균벡터

$S_{p\times p} = \frac{1}{n-1}\sum_{j=1}^{n} (X_j - \overline{X})(X_j - \overline{X})'$: $p\times p$ 표본공분산행렬이다.

5.2.2 Hotelling T^2 통계량의 분포

- Hotelling의 T^2 통계량을 살펴보면

$$\frac{(n-p)}{(n-1)p}T^2 = \frac{n(\overline{X}-\mu_0)'(\overline{X}-\mu_0)/p}{(n-1)S/(n-p)} \quad \begin{matrix} \sim \chi_p^2 \\ \sim \chi_{n-p}^2 \end{matrix}$$

분자와 분모가 서로 독립이므로 위 통계량은 자유도가 $p,\ n-p$인 F-분포, $F_{p,n-p}$따른다.

- Hotelling의 T^2 통계량의 분포

$$T^2 \sim \frac{(n-1)p}{(n-p)}F_{p,n-p}$$

- $$P\left[T^2 \geqq \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right] = \alpha$$

여기서 $F_{p,n-p}(\alpha)$는 $F_{p,n-p}$를 따르는 확률변수 F에 대해

$P(F \geqq F_{p,n-p}(\alpha)) = \alpha$를 만족하는 값.

5.3 Σ를 모를 때 $H_0 : \mu = \mu_0$에 대한 Hotelling T^2 검정

5.3.1 소표본 $(n \leq 30)$이며 모분산 σ^2을 모르는 경우

한 개의 정규분포 모집단 확률표본 $X_1, \cdots, X_n$

모평균에 대한 가설검정을 하기 위한 검정통계량은 H_0 하에서 t_{n-1}분포를 따른다.

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

■ 모평균에 관한 가설들에 대해 검정법

(i) $H_0 : \mu = \mu_0$ 에 대해 $H_1 : \mu \neq \mu_0$ ⇒ 유의수준 α에서 $|t| \geq t_{n-1}(\alpha/2)$이면 H_0 기각

(ii) $H_0 : \mu = \mu_0$에 대해 $H_1 : \mu > \mu_0$ ⇒ 유의수준 α에서 $t \geq t_{n-1}(\alpha)$이면 H_0 기각

(iii) $H_0 : \mu = \mu_0$에 대해 $H_1 : \mu < \mu_0$ ⇒ 유의수준 α 에서 $t \leq -t_{n-1}(\alpha)$이면 H_0 기각.

여기서 $t_{n-1}(\alpha)$는 $P(T \geq t_{n-1}(\alpha)) = \alpha$를 만족하는 값

5.3.2 다변량 표본의 경우

한 개 모집단, 서로 독립, 모평균 벡터 μ, 공분산행렬 Σ인 p-변량 정규분포를 따르는 n개 $p\times 1$ 확률벡터 $X_1,\cdots,X_n$을 확률표본으로 얻은 경우,

■ 모평균벡터가 어떤 특정한 모평균벡터 μ_0를 갖는지 검정법

(1) 통계적 가설

$H_0 : \mu = \mu_0$에 대해 $H_1 : \mu \neq \mu_0$

(2) 검정법 : 유의수준 α에서

$$T^2 = n(\overline{X}-\mu_0)'S^{-1}(\overline{X}-\mu_0) \geq T^2_{\alpha,p,n-1} = \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$$

이면 H_0를 기각한다.

[참고] ■ 일변량 t^2 통계량은 다음과 같이 분해하여

$$t^2 = \frac{(\bar{x}-\mu_0)^2}{(s/\sqrt{n})^2} = \sqrt{n}(\bar{x}-\mu_0)(s^2)^{-1}\sqrt{n}(\bar{x}-\mu_0)$$

$$= \begin{pmatrix} \text{일변량 정규} \\ \text{확률변수} \end{pmatrix}' \begin{pmatrix} \text{카이제곱 확률변수} \\ \text{자유도} \end{pmatrix}^{-1} \begin{pmatrix} \text{일변량 정규} \\ \text{확률변수} \end{pmatrix}$$

■ 다변량의 경우 T^2통계량

$$T^2 = \sqrt{n}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)' \left[\frac{\sum_{j=1}^{n}(\boldsymbol{X}_j-\bar{\boldsymbol{X}})(\boldsymbol{X}_j-\bar{\boldsymbol{X}})'}{n-1} \right]^{-1} \sqrt{n}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)$$

$$= \begin{pmatrix} \text{다변량 정규} \\ \text{확률벡터} \end{pmatrix}' \begin{pmatrix} \text{Wishart 확률행렬} \\ \text{자유도} \end{pmatrix}^{-1} \begin{pmatrix} \text{다변량 정규} \\ \text{확률벡터} \end{pmatrix}.$$

≪예제 5.1≫ 20명의 정상인 여성들의 땀에 관해 측정하여 [표 5.1]의 자료를 얻었다.

$$X_1 = \text{땀의 비율(sweat rate)}$$

$$X_2 = \text{sodium 양}$$

$$X_3 = \text{potassium 양}$$

$H_0 : \mu' = (4, 50, 10)$에 대해 $H_1 : \mu' \neq (4, 50, 10)$

가설 검정을 유의수준 5%와 10%에서 시행하고자 한다.

■ 표본평균벡터와 표본공분산행렬

$$\overline{X} = \begin{pmatrix} 4.640 \\ 45.400 \\ 9.965 \end{pmatrix}, \quad S = \begin{pmatrix} 2.879 & 10.010 & -1.810 \\ 10.010 & 199.788 & -5.640 \\ -1.810 & -5.640 & 3.628 \end{pmatrix}$$

■ 역행렬

$$S^{-1} = \begin{pmatrix} 0.586 & -0.022 & 0.258 \\ -0.022 & 0.006 & -0.002 \\ 0.258 & -0.002 & 0.402 \end{pmatrix}$$

■ T^2통계량

$$T^2 = n\,(\overline{X}-\mu_0)'S^{-1}(\overline{X}-\mu_0)$$

$$= 20(0.640,\ -4.6,\ -0.035)\begin{pmatrix} 0.586 & -0.022 & 0.258 \\ -0.022 & 0.006 & -0.002 \\ 0.258 & -0.002 & 0.402 \end{pmatrix}\begin{pmatrix} 0.640 \\ -4.6 \\ -0.035 \end{pmatrix} = 9.74$$

(1) 유의수준 5%에서의 기각값은

$$T^2_{\alpha,p,n-1} = \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha) = \frac{19\cdot 3}{20-3}F_{3,17}(0.05) = 3.353\cdot 3.20 = 10.73$$

이고 $T^2 = 9.74 \not\geq 10.73$이므로 H_0를 기각하지 못한다.

즉 기존 결과와의 차이가 유의하지 않다.음을 알 수 있다.

(2) 유의수준 10%에서의 기각값은

$$T^2_{\alpha,p,n-1} = \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha) = \frac{19\cdot 3}{20-3}F_{3,17}(0.10) = 3.353\cdot 2.44 = 8.18$$

이고 $T^2 = 9.74 \geq 8.18$이므로 H_0를 기각. 기존 결과와 차이가 있음을 알게 된다.

[표 5.1] 땅 자료

개인 번호	X_1	X_2	X_3
1	3.7	48.5	9.3
2	5.7	65.1	8.0
3	3.8	47.2	10.9
4	3.2	53.2	12.0
5	3.1	55.5	9.7
6	4.6	36.1	7.9
7	2.4	24.8	14.0
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3
11	3.9	36.9	12.7
12	4.5	58.8	12.3
13	3.5	27.8	9.8
14	4.5	40.2	8.4
15	1.5	13.5	10.1
16	8.5	56.4	7.1
17	4.5	71.6	8.2
18	6.5	52.8	10.9
19	4.1	44.1	11.2
20	5.5	40.9	9.4

5.3.3 Hotelling T^2 통계량의 불변성

T^2 통계량은 위치와 척도변환에 대해 불변(invariant): 통계량으로서 바람직한 성질.

- $X \sim N_p(\mu_0, \Sigma)$일 때, 정칙행렬 C 와 상수벡터 d 에 대해

$$Y = CX + d$$

- 변환된 변수 Y 의 기대값과 공분산행렬을 구하면

$$EY = \mu_Y = C\mu_0 + d$$

$$Var(Y) = CVar(X)C' = C\Sigma C'$$

이므로 $\widehat{Var}(Y) = CSC'$.

- $\overline{Y} - \mu_Y = C(\overline{X} - \mu_0)$

- Hotelling의 T^2통계량을 구하면

$$
\begin{aligned}
T_Y^2 &= n(\overline{Y}-\mu_Y)' S_Y^{-1}(\overline{Y}-\mu_Y) \\
&= n\{C(\overline{X}-\mu_0)\}'\{CSC'\}^{-1}\{C(\overline{X}-\mu_0)\} \\
&= n(\overline{X}-\mu_0)' C'(C')^{-1} S^{-1} C^{-1} C(\overline{X}-\mu_0) \\
&= n(\overline{X}-\mu_0)' S^{-1}(\overline{X}-\mu_0) \\
&= T_X^2
\end{aligned}
$$

5.4 두 모집단에 대한 Hotelling의 T^2 검정

5.4.1 일변량 소표본($n \leqq 30$)에서 모분산 $\sigma_1^2 = \sigma_2^2 = \sigma^2$이며 σ^2을 모르는 경우

[표 5.2] 두 개의 집단으로부터 얻어진 일변량의 표본 정보

	모집단 1	모집단 2
모평균	μ_1	μ_2
모분산	σ_1^2	σ_2^2
확률표본	$X_{11}, X_{12}, ..., X_{1n_1}$	$X_{21}, X_{22}, ..., X_{2n_2}$
표본평균	$\overline{X_1} = \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1}$	$\overline{X_2} = \frac{\sum_{i=1}^{n_2} X_{2i}}{n_2}$
표본분산	$s_1^2 = \frac{\sum_{i=1}^{n_1} (X_{1i} - \overline{X_1})^2}{n_1 - 1}$	$s_2^2 = \frac{\sum_{i=1}^{n_2} (X_{2i} - \overline{X_2})^2}{n_2 - 1}$

■ 합동분산추정량(pooled variance estimator): $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

■ 검정통계량

$$t = \frac{(\overline{X_1} - \overline{X_2}) - \delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

은 H_0 하에서 $t_{n_1+n_2-2}$분포를 따르며

■ 두 집단 모평균에 관한 다음의 가설들에 대해 검정법

(i) $H_0 : \mu_1 - \mu_2 = \delta_0$ 에 대해 $H_1 : \mu_1 - \mu_2 \neq \delta_0$

유의수준 α에서 검정법은 $| t | \geq t_{n_1+n_2-2}(\alpha/2)$이면 H_0를 기각한다.

(ii) $H_0 : \mu_1 - \mu_2 \leqq \delta_0$ 에 대해 $H_1 : \mu_1 - \mu_2 > \delta_0$

유의수준 α에서 검정법은 $t \geqq t_{n_1+n_2-2}(\alpha)$이면 H_0를 기각한다.

(iii) $H_0 : \mu_1 - \mu_2 \geq \delta_0$ 에 대해 $H_1 : \mu_1 - \mu_2 < \delta_0$

유의수준 α에서 검정법은 $t \leqq -t_{n_1+n_2-2}(\alpha)$이면 H_0를 기각한다.

5.4.2 다변량 표본이며 $\Sigma_1 = \Sigma_2 = \Sigma$ 인 경우

두 개 모집단으로부터 서로 독립이고 p변량 정규분포를 따르며 n_1개의 $p \times 1$ 확률벡터로 구성된 $X_{11}, \cdots, X_{1n_1}$ 확률표본과 n_2개의 $p \times 1$ 확률벡터로 구성된 $X_{21}, \cdots, X_{2n_2}$ 확률표본을 얻었다.

$X_{1j} \sim iid\ N_p(\mu_1, \Sigma_1), X_{2j} \sim iid\ N_p(\mu_2, \Sigma_2)$ 일 때

두 집단의 모평균벡터가 같은지 검정하고자한다.

공분산행렬에 대한 가정: $\Sigma_1 = \Sigma_2 = \Sigma$,

Σ는 알려져 있지 않다고 가정한다.

(1) 통계적 가설

$$H_0 : \mu_1 = \mu_2 \text{에 대해 } H_1 : \mu_1 \neq \mu_2$$

(2) 검정법 : 유의수준 α에서

$$T^2 = (\overline{X}_1 - \overline{X}_2)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{pl} \right]^{-1} (\overline{X}_1 - \overline{X}_2)$$

$$= \frac{n_1 n_2}{n_1 + n_2} (\overline{X}_1 - \overline{X}_2)' S_{pl}^{-1} (\overline{X}_1 - \overline{X}_2)$$

$$\geqq T^2_{\alpha, p, n_1 + n_2 - 2} = \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1}(\alpha)$$

이면 H_0를 기각한다. 여기서 $S_1 = \hat{\Sigma}_1, S_2 = \hat{\Sigma}_2$

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

은 $p \times p$ 합동공분산행렬(pooled covariance matrix)이다.

- H_0 하에서 $\overline{X}_1 - \overline{X}_2$의 기대값은

$$E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2 = 0$$

$$Cov(\overline{X}_1, \overline{X}_2) = 0$$

- 두 평균벡터 차의 공분산행렬을 구하면

$$\begin{aligned} Cov(\overline{X}_1 - \overline{X}_2) &= Cov(\overline{X}_1) + Cov(\overline{X}_2) \\ &= \frac{1}{n_1}\Sigma + \frac{1}{n_2}\Sigma \\ &= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma \end{aligned}$$

■ 공분산행렬 추정량: $\widehat{Cov}(\overline{X}_1 - \overline{X}_2) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{pl}$

≪예제 5.2≫ 남자(1) 32명과 여자(2) 32명에 대한 심리 시험 결과가 [표 5.3]과 같이 주어졌다. 성별에 따라 평균벡터간에 차이가 있는지 검정하고자 한다. 여기서 각 변수는 X_1 =그림 인식, X_2 =문제 인식, X_3 =도구 인식, X_4 =언어력으로 정의된다. 단, 각 집단은 공분산행렬이 같은 다변량 정규분포를 따른다고 가정한다.

■ $H_0 : \mu_1 = \mu_2$에 대해 $H_1 : \mu_1 \neq \mu_2$ 가설 검정을 유의수준 5%에서 하고자한다.

■ 두 집단의 표본평균벡터와 표본공분산행렬을 각각 구하면

$$\overline{X_1} = \begin{pmatrix} 15.97 \\ 15.91 \\ 27.19 \\ 22.75 \end{pmatrix}, \quad S_1 = \begin{pmatrix} 5.192 & 4.545 & 6.522 & 5.250 \\ 4.545 & 13.185 & 6.760 & 6.266 \\ 6.522 & 6.760 & 28.673 & 14.468 \\ 5.250 & 6.266 & 14.468 & 16.656 \end{pmatrix}$$

와

$$\overline{X_2}=\begin{pmatrix}12.34\\13.91\\16.66\\21.94\end{pmatrix},\quad S_2=\begin{pmatrix}9.136 & 7.549 & 4.864 & 4.151\\7.549 & 18.604 & 10.225 & 5.446\\4.864 & 10.225 & 30.039 & 13.494\\4.151 & 5.446 & 13.494 & 28.00\end{pmatrix}$$

■ 합동공분산행렬: $$S_{pl}=\frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}=\frac{(32-1)S_1+(32-1)S_2}{32+32-2}$$

$$=\begin{pmatrix}7.164 & 6.047 & 5.693 & 4.701\\6.047 & 15.89 & 8.492 & 5.856\\5.693 & 8.492 & 29.36 & 13.98\\4.701 & 5.856 & 13.98 & 22.32\end{pmatrix}$$

■ 검정통계량: $$T^2=(\overline{X_1}-\overline{X_2})'\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_{pl}\right]^{-1}(\overline{X_1}-\overline{X_2})=97.6015$$

■ 유의수준 5%에서의 기각치:

$$T^2_{n_1+n_2-2}(\alpha) = \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1}(\alpha)$$
$$= \frac{62 \cdot 4}{59} F_{4,59}(0.05) \approx \frac{248}{59} F_{4,60}(0.05) = 10.635$$

$T^2 = 97.6015 > 10.635$이므로 H_0를 기각. 즉 남녀간의 평균벡터 차이가 통계적으로 유의하다.

[표 5.3] 남자 32명, 여자 32명에 대한 4가지 심리 검사 자료

번호(남자)	X_1	X_2	X_3	X_4	번호(여자)	X_1	X_2	X_3	X_4
1	15	17	24	14	1	13	14	12	21
2	17	15	32	26	2	14	12	14	26
3	15	14	29	23	3	12	19	21	21
4	13	12	10	16	4	12	13	10	16
5	20	17	26	28	5	11	20	16	16
6	15	21	26	21	6	12	9	14	18
7	15	13	26	22	7	10	13	18	24
8	13	5	22	22	8	10	8	13	23
9	14	7	30	17	9	12	20	19	23
10	17	15	30	27	10	11	10	11	27
11	17	17	26	20	11	12	18	25	25
12	17	20	28	24	12	14	18	13	26
13	15	15	29	24	13	14	10	25	28
14	18	19	32	28	14	13	16	8	14
15	18	18	31	27	15	14	8	13	25
16	15	14	26	21	16	13	16	23	28
17	18	17	33	26	17	16	21	26	26
18	10	14	19	17	18	14	17	14	14
19	18	21	30	29	19	16	16	15	23
20	18	21	34	26	20	13	16	23	24
21	13	17	30	24	21	2	6	16	21
22	16	16	16	16	22	14	16	22	26
23	11	15	25	23	23	17	17	22	28
24	16	13	26	16	24	16	13	16	14
25	16	13	23	21	25	15	14	20	26
26	18	18	34	24	26	12	10	12	9
27	16	15	28	27	27	14	17	24	23
28	15	16	29	24	28	13	15	18	20
29	18	19	32	23	29	11	16	18	28
30	18	16	33	23	30	7	7	19	18
31	17	20	21	21	31	12	15	7	28
32	19	19	30	28	32	6	5	6	13

5.4.3 다변량 표본이며 $\Sigma_1 \neq \Sigma_2$인 경우

두 개 모집단으로부터 서로 독립이고 p-변량 정규분포를 따르는 n_1개의 $p \times 1$ 확률벡터 $X_{11}, \cdots, X_{1n_1}$와 n_2개의 $p \times 1$ 확률벡터 $X_{21}, \cdots, X_{2n_2}$를 얻은 경우 즉,

- $X_{1j} \sim N_p(\mu_1, \Sigma_1), X_{2j} \sim N_p(\mu_2, \Sigma_2)$를 따르는 경우,

두 집단의 모평균벡터가 같은지 검정하고자 한다. 여기서 $\Sigma_1 \neq \Sigma_2$ 이며 Σ_1과 Σ_2는 알려져 있지 않다고 가정한다.

- H_0 하에서 $\overline{X}_1 - \overline{X}_2$의 기대값과 공분산행렬을 구하면

$$E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2 = 0 \qquad Cov(\overline{X}_1, \overline{X}_2) = 0$$

$$Cov(\overline{X}_1 - \overline{X}_2) = Cov(\overline{X}_1) + Cov(\overline{X}_2)$$

$$= \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2$$

공분산행렬에 대한 추정량: $\widehat{Cov}(\overline{X}_1 - \overline{X}_2) = \frac{1}{n_1} S_1 + \frac{1}{n_2} S_2$

여기서 $S_1 = \hat{\Sigma}_1, \quad S_2 = \hat{\Sigma}_2$

(1) 통계적 가설

$H_0 : \mu_1 - \mu_2 = 0$ 에 대해 $H_1 : \mu_1 - \mu_2 \neq 0$

(2) 검정법

$$T^2 = (\overline{X}_1 - \overline{X}_2)' \left(\frac{S_1}{n_1} + \frac{S_2}{n_2} \right)^{-1} (\overline{X}_1 - \overline{X}_2)$$

유의수준 α에서 근사적인 검정법: $n_1 - p,\ n_2 - p$ 가 충분히 클 때,

$T^2 \geqq \chi_p^2(\alpha)$이면 H_0를 기각한다.

■ 모평균벡터 차 $\mu_1 - \mu_2$에 대한 구간 추정을 고려하고자 한다.

$\mu_1 - \mu_2$에 대해 근사적인 $100(1-\alpha)\%$ 신뢰영역은

$$(\overline{X}_1 - \overline{X}_2)' \left(\frac{S_1}{n_1} + \frac{S_2}{n_2} \right)^{-1} (\overline{X}_1 - \overline{X}_2) \leq \chi_p^2(\alpha)$$

이 된다. 또한 선형조합$l'(\mu_1 - \mu_2)$에 대한 $100(1-\alpha)\%$ 동시 신뢰영역은

$$l'(\overline{X}_1 - \overline{X}_2) \pm \sqrt{\chi_p^2(\alpha)} \sqrt{l' \left(\frac{S_1}{n_1} + \frac{S_2}{n_2} \right) l}$$

와 같이 주어진다.

≪예제 5.3≫ Azotobacter라는 유기체를 포함한 토양과 포함하지 않은 토양을 비교하기위해 다음의 변수들을 측정하여 [표 5.4]의 결과를 얻었다. 두 토양의 성분이 같다고 할 수 있는지 유의수준 5%에서 두 집단에 대한 모평균벡터에 대한 검정을 하고자 한다. 단 두 집단은 공분산행렬이 다른 다변량 정규분포를 따른다고 할 수 있다고 가정한다. 관측된 변수는 다음과 같다:

$X_1 =$ pH 산성도, $X_2 =$ 인산(phosphate) 양, $X_3 =$ 질소(nitrogen) 양.

관심있는 통계적 가설은 $H_0 : \mu_1 = \mu_2$에 대해 $H_1 : \mu_1 \neq \mu_2$이므로 Hotelling의 검정을 이용하고자 한다. 우선

$$\overline{X}_1 = \begin{pmatrix} 7.81 \\ 108.77 \\ 44.92 \end{pmatrix}, \quad \overline{X}_2 = \begin{pmatrix} 5.89 \\ 41.90 \\ 20.80 \end{pmatrix}$$

$$S_1 = \begin{pmatrix} 0.461 & 1.18 & 4.49 \\ 1.18 & 3776.4 & -17.35 \\ 4.49 & -17.35 & 147.24 \end{pmatrix}, \quad S_2 = \begin{pmatrix} 0.148 & -0.679 & 0.209 \\ -0.679 & 96.10 & 20.20 \\ 0.209 & 20.20 & 24.18 \end{pmatrix}.$$

Hotelling의 T^2 검정통계량을 계산하면

$$T^2 = (\overline{X}_1 - \overline{X}_2)' \left(\frac{S_1}{n_1} + \frac{S_2}{n_2} \right)^{-1} (\overline{X}_1 - \overline{X}_2) = 96.818$$

근사적인 기각값 $\chi_3^2(0.05) = 7.81$를 얻고 $T^2 = 96.818 \geqq 7.81$이므로 H_0를 기각한다.

[표 5.4] Azotobacter 자료

Azotobacter 있는 경우			Azotobacter 없는 경우		
X_1	X_2	X_3	X_1	X_2	X_3
8.0	60	58	6.2	49	30
8.0	156	68	5.6	31	23
8.0	90	37	5.8	42	22
6.1	44	27	5.7	42	14
7.4	207	31	6.2	40	23
7.4	120	32	6.4	49	18
8.4	65	43	5.8	31	17
8.1	237	45	6.4	31	19
8.3	57	60	5.4	62	26
7.0	94	43	5.4	42	16
8.5	86	40			
8.4	52	48			
7.9	146	52			

5.5 짝지어진(Paired) 두 개 집단에 대한 검정

5.5.1 소표본($n \leq 30$)이며 모분산 σ^2을 모르고 있는 경우

각 개체에 대해 정규분포를 따르는 모집단으로부터 쌍(pair)으로 측정하여 확률표본 $(X_1, Y_1), ..., (X_n, Y_n)$을 얻은 경우 두 집단(처리)의 모평균이 같은지 검정하고자한다.

관심있는 가설

$$H_0 : \mu_1 = \mu_2 \text{ 에 대해 } H_1 : \mu_1 \neq \mu_2$$

마찬가지로

$$H_0 : \mu_1 - \mu_2 = 0 \text{ 에 대해 } H_1 : \mu_1 - \mu_2 \neq 0$$

또는

$$H_0 : \delta = 0 \text{ 에 대해 } H_1 : \delta \neq 0$$

[표 5.5]는 짝지어진 표본의 자료구조를 보여주며 각 개체에 따른 효과를 제거하기 위하여 각 개체에 대해 관측값의 차이를 구하여 이용한다.

[표 5.5] 짝지어진 표본의 자료구조

쌍	처리 1	처리 2	차이
1	Y_{11}	Y_{21}	$d_1 = Y_{11} - Y_{21}$
2	Y_{12}	Y_{22}	$d_2 = Y_{12} - Y_{22}$
⋮	⋮	⋮	⋮
n	Y_{1n}	Y_{2n}	$d_n = Y_{1n} - Y_{2n}$

여기서 $d_1, d_2, \cdots, d_n \sim iid \quad N(\delta, \sigma_\delta^2)$

■ 차이에 대한 평균과 분산의 추정량:

$$\overline{d} = \frac{\sum_{i=1}^{n} d_i}{n}, \qquad s_d^2 = \frac{\sum_{i=1}^{n} (d_i - \overline{d})^2}{n-1}$$

■ 검정통계량

$$t = \frac{\bar{d} - \delta_0}{s_d / \sqrt{n}}$$

은 H_0 하에서 t_{n-1} 분포를 따르며

모평균 차이에 관한 다음의 가설들에 대해 유의수준 α에서의 검정법:

(i) $H_0 : \delta = \delta_0$ 에 대해 $H_1 : \delta \neq \delta_0$

유의수준 α에서 검정법은 $| t | \geq t_{n-1}(\alpha/2)$이면 H_0를 기각한다.

(ii) $H_0 : \delta \leq \delta_0$ 에 대해 $H_1 : \delta > \delta_0$

유의수준 α에서 검정법은 $t \geq t_{n-1}(\alpha)$이면 H_0를 기각한다.

(iii) $H_0 : \delta \geq \delta_0$ 에 대해 $H_1 : \delta < \delta_0$

유의수준 α에서 검정법은 $t \leq -t_{n-1}(\alpha)$이면 H_0를 기각한다.

5.5.2 짝지어진 두 집단에서 추출된 다변량인 표본의 경우

관측값간에 짝지어진 관계로 인해 두 개의 표본이 서로 독립이 아닐 때

두 집단에 대한 T^2 검정은 적당하지 않다.

■ 짝지어진 관측 예

(1) 한 학생에게 특별한 교육방법 시행 전(before)과 후(after) 2회 3과목 시험이 치러진 경우,

(2) 한 환자에게 약을 투여한 뒤 1시간 후, 3시간 후에 효과와 혈압을 측정한 경우.

=> 각 측정치간의 차이를 구해, 한 개 모집단 문제로 전환하여 보는 것이 타당하다.

[표 5.6] 짝지어진 표본의 다변량 자료구조

짝 번호	처리 1	처리 2	차이
1	X_1	Y_1	$d_1 = Y_1 - X_1$
2	X_2	Y_2	$d_2 = Y_2 - X_2$
⋮	⋮	⋮	⋮
n	X_n	Y_n	$d_n = Y_n - X_n$

각 개체에 대해 p-변량 정규분포를 따르는 $p\times 1$ 확률벡터를 쌍(pair)으로 측정하여 $(X_1, Y_1),...,(X_n, Y_n)$로 확률표본을 얻은 경우

■ 두 집단의 모평균벡터가 같은지 검정하고자한다.

(1) 통계적 가설

$H_0 : d = \mu_1 - \mu_2 = 0$에 대해 $H_1 : d = \mu_1 - \mu_2 \neq 0$

(2) 검정법 : 유의수준 α에서

$$T^2 = \bar{d}'\left(\frac{S_d}{n}\right)^{-1}\bar{d} = n\bar{d}'S_d^{-1}\bar{d} \geq T^2_{\alpha,p,n-1} = \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$$

이면 H_0를 기각한다. 여기서

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad S_d = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})(d_i - \bar{d})'$$

≪예제 5.4≫ 부식을 방지해주는 두 가지 코팅 방법 A, B에 대한 비교를 하고자 15개의 파이프 양쪽에 각각의 방법으로 코팅을 한 후, 얼마의 시간이 지난 후 부식정도에 대해

$X_1 =$ 부식된 구멍의 최대 깊이, $X_2 =$부식한 구멍 수

를 측정하여 [표 5.7]의 결과를 얻었다. 짝지어진 두 집단으로 볼 수 있으므로 각 변수의 차이를 계산하여 [표 5.7]에 첨가하였다. 두 가지 코팅 방법에 대한 부식 깊이와 부식 구멍수의 평균벡터가 같은지 유의수준 5%에서 통계적 검정을 하고자한다.

(1) 통계적 가설

$H_0 : d = \mu_A - \mu_B = 0$에 대해 $H_1 : d = \mu_A - \mu_B \neq 0$

(2) 검정통계량

차이(difference)에 대한 표본평균벡터 $\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \begin{pmatrix} 8.000 \\ 3.067 \end{pmatrix}$,

차이에 대한 표본공분산행렬

$$S_d = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})(d_i - \bar{d})'$$
$$= \begin{pmatrix} 121.571 & 17.071 \\ 17.071 & 21.781 \end{pmatrix}$$

검정통계량은

$$T^2 = \bar{d}'\left(\frac{S_d}{n}\right)^{-1}\bar{d} = n\bar{d}'S_d^{-1}\bar{d} = 10.819$$

기각값은

$$T^2_{p,n-1}(\alpha) = \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha) = \frac{14 \cdot 2}{15-2}F_{2,13}(0.05) = 2.154 \cdot 3.81 = 8.207$$

$T^2 = 10.819 > 8.207$ 이므로 H_0를 기각한다.

그러므로 코팅방법에 따라 부식깊이와 구멍개수의 평균벡터간에 통계적으로 유의한 차이가 있다고 할 수 있다.

[표 5.7] 파이프의 코팅 방법에 따른 부식 깊이와 구멍 개수

위치	코팅 A		코팅 B		차이	
	깊이	개수	깊이	개수	깊이	개수
	X_1	X_2	Y_1	Y_2	$d_1 = Y_1 - X_1$	$d_2 = Y_2 - X_2$
1	51	35	73	31	22	-4
2	41	14	43	19	2	5
3	43	19	47	22	4	3
4	41	29	53	26	12	-3
5	47	34	58	36	11	2
6	32	26	47	30	15	4
7	24	19	52	29	28	10
8	43	37	38	36	-5	-1
9	53	24	61	34	8	10
10	52	27	56	33	4	6
11	57	14	56	19	-1	5
12	44	19	34	19	-10	0
13	57	30	55	26	-2	-4
14	40	7	65	15	25	8
15	68	13	75	18	7	5

5.6 Hotelling T^2 통계량과 최대우도비 검정

확률벡터 X가 p-변량 정규분포를 따를 때 즉 $X \sim N_p(\mu, \Sigma)$일 때

- 우도함수

$$L(\mu, \Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}(X_j-\mu)'\Sigma^{-1}(X_j-\mu)\right)$$

- 최대우도추정량(MLE : maximum likelihood estimator)

$$\hat{\mu} = \overline{X}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_j (X_j - \overline{X})(X_j - \overline{X})' = \frac{(n-1)}{n} S$$

- 최대우도함수는

$$\max_{\mu,\Sigma} L(\mu, \Sigma) = \frac{1}{(2\pi)^{np/2}|\hat{\Sigma}|^{n/2}} \exp\left\{-\frac{np}{2}\right\} .$$

지수함수내의 부분은 트레이스를 이용해

$$-\frac{1}{2}\sum_{j=1}^{n}(X_j-\mu)'\Sigma^{-1}(X_j-\mu) = -\frac{1}{2}\sum_{j=1}^{n}tr\left[\Sigma^{-1}(X_j-\mu)(X_j-\mu)'\right]$$
$$= -\frac{1}{2}tr\left[\Sigma^{-1}\sum_{j=1}^{n}(X_j-\mu)(X_j-\mu)'\right]$$

로 정리되고 최대우도추정량 $\hat{\mu}=\overline{X}$ 와 $\hat{\Sigma}=\frac{(n-1)}{n}S$ 를 이용하면

$S_{p\times p}^{-1}\, nS_{p\times p} = nI_{p\times p}$ 이 되어 $tr(nI_{p\times p}) = np$가 되기 때문이다.

- 모평균벡터 μ_0가 알려진 경우에는 다음의 최대우도

$$\max_{\Sigma}\ L(\mu_0,\Sigma)=\frac{1}{(2\pi)^{np/2}|\hat{\Sigma}_0|^{n/2}}\exp\left(-\frac{np}{2}\right)$$

- 최대우도추정량은 $\hat{\Sigma}_0=\frac{1}{n}\sum_j(X_j-\mu_0)(X_j-\mu_0)'$

■ 우도비 검정법의 원리:

$$\Lambda = \frac{\max\limits_{\theta \in \Theta_0} L(\theta)}{\max\limits_{\theta \in \Theta} L(\theta)}$$

를 구하고, 유의수준 α에서 Λ가 작을수록 H_0를 기각한다.

■ 우도비(likelihood ratio)를 구하면

$$\Lambda = \frac{\max\limits_{\Sigma} L(\mu_0, \Sigma)}{\max\limits_{\mu, \Sigma} L(\mu, \Sigma)} = \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_0|} \right)^{n/2}$$

■ 검정통계량은

$$\Lambda^{\frac{2}{n}} = \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_0|} = \frac{\left|\sum_{j=1}^{n}(X_j - \overline{X})(X_j - \overline{X})'\right|}{\left|\sum_{j=1}^{n}(X_j - \mu_0)(X_j - \mu_0)'\right|} = \Lambda^*$$

■ 이와 같은 통계량은 Wilks에 의해 제안되어졌고 Wilks lambda라고 부른다.

유의수준 α에서 우도비 검정법은

$\Lambda^* \leqq C_\alpha$이면 H_0를 기각한다.

여기서 C_α는 Λ^*의 오른쪽 꼬리 부분의 확률이 α가 되는 값이다.

$$\sum_{j=1}^{n}(X_j-\mu_0)(X_j-\mu_0)' = \sum_{j=1}^{n}(X_j-\overline{X}+\overline{X}-\mu_0)(X_j-\overline{X}+\overline{X}-\mu_0)'$$
$$= \sum_{j=1}^{n}(X_j-\overline{X})(X_j-\overline{X})' + \sum_{j=1}^{n}(\overline{X}-\mu_0)(\overline{X}-\mu_0)'$$
$$= \sum_{j=1}^{n}(X_j-\overline{X})(X_j-\overline{X})' + n(\overline{X}-\mu_0)(\overline{X}-\mu_0)'$$

와 같이 정리하면 $\Lambda^{\frac{2}{n}}$은

$$\Lambda^{\frac{2}{n}} = \left[\frac{\left|\sum_j (X_j-\overline{X})(X_j-\overline{X})' + n(\overline{X}-\mu_0)(\overline{X}-\mu_0)'\right|}{\left|\sum_j (X_j-\overline{X})(X_j-\overline{X})'\right|}\right]^{-1}$$

이 되고 다음의 정리 5.1과 5.2를 얻을 수 있다.

정리 5.1은 우도비 검정통계량과 T^2 통계량과의 관계를 정의한 것이다.

정리 5.1 $X_1,\cdots,X_n \sim iid\ N_p(\mu,\Sigma)$일 때 모평균벡터에 대한 가설

$$H_0 : \mu = \mu_0 \text{에 대해 } H_1 : \mu \neq \mu_0$$

에 대한 우도비 검정통계량 Λ와 Hotelling의 T^2 검정통계량에 대해

$$\Lambda^{\frac{2}{n}} = \left(1 + \frac{T^2}{n-1}\right)^{-1}$$

이 성립한다.

정리 5.2 $n \to \infty$ 이면 우도비검정통계량은

$$-2\ln \Lambda = -2\ln \left(\frac{\max\limits_{\theta \in \Theta_0} L(\theta)}{\max\limits_{\theta \in \Theta} L(\theta)} \right) \sim \chi^2_{\nu - \nu_0}$$

분포를 따른다. 여기서 Θ_0는 H_0하에서 모수공간(parameter space)이고 Θ_1은 H_1하에서 모수공간이며 $\Theta = \Theta_0 \bigcup \Theta_1$이다. 여기서 ν는 Θ의 차원수이고 ν_0는 Θ_0의 차원수를 나타낸다.

5.7 모평균벡터의 동시신뢰영역

■ 모평균벡터의 모든 가능한 선형함수들에 관한 가설을 동시에 검정할 수 있고 제 1종 오류의 확률을 일정수준으로 유지하는 검정방법으로부터 동시 신뢰영역을 구하고자한다.

$$P[R(\boldsymbol{X})\text{가 } \boldsymbol{\theta}\text{를 포함한다}] = 1-\alpha$$

를 만족하는 영역 $R(\boldsymbol{X})$를 θ 에 대한 $100(1-\alpha)\%$ 신뢰영역(confidence region)이라고 한다.

■ p-차원 정규 모집단의 모평균벡터 μ에 대해, Hotelling T^2의 분포로부터

$$P\left[n(\overline{\boldsymbol{X}}-\boldsymbol{\mu})'\boldsymbol{S}^{-1}(\overline{\boldsymbol{X}}-\boldsymbol{\mu}) \le \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right] = 1-\alpha$$

이 된다. 즉, p-차원 공간에서

$\overline{\boldsymbol{X}}$가 $\left[\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right]^{1/2}$ 이내에 있을 확률이 $1-\alpha$ 가 된다.

정리 5.3 p-변량 확률표본 $X_1,\cdots,X_n$은 서로 독립이며 $N_p(\mu,\Sigma)$분포를 따르며 Σ는 양정치행렬일 때, p-차원 정규 모집단의 모평균벡터 μ에 대한 $100(1-\alpha)\%$ 신뢰영역은

$$\left\{\mu |\ n(\overline{X}-\mu)'S^{-1}(\overline{X}-\mu) \le \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right\}$$

(i) 이 신뢰영역은 $\overline{X}$를 중심으로 하는 p-차원 타원체가 되며 축의 길이는 S의 고유값 $\lambda_1,\lambda_2,...,\lambda_p$에 의존하며 축의 방향은 S 의 고유벡터 $e_1,e_2,...,e_p$방향이다.

(ii) 축의 방향과 길이는 다음의 식

$$n(\overline{X}-\mu)'S^{-1}(\overline{X}-\mu) \le \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha) = c^2$$

으로부터 축의 길이는

$$\frac{\sqrt{\lambda_i}\,c}{\sqrt{n}} = \sqrt{\lambda_i}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}\,,$$ 축방향은 고유벡터 e_i의 방향이 된다.

(iii) 신뢰영역 타원체의 축의 방향은 $\overline{X}$ 를 중심으로

$$\pm \sqrt{\lambda_i}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}\,e_i$$

이며, 여기서 $Se_i = \lambda_i e_i, \quad i = 1,2,...,p$ 이다.

■ $p=4$인 경우 각 모평균벡터의 신뢰수준 $1-\alpha = 0.95$에 대해 이와 같은 방법으로 동시 신뢰영역을 구하면 신뢰수준은 0.95가 된다. 동시 신뢰영역의 신뢰수준을 0.95로 하기위해서는 각 모평균에 대한 신뢰구간을 따로 구할 필요없이 동시에 구하면 된다.

정리 5.4 p-변량 확률표본 $X_1,\cdots,X_n$은 서로 독립이며 $N_p(\mu,\Sigma)$분포를 따르고 Σ는 양정치행렬일때, 0이 아닌 모든 $p\times 1$벡터 l에 대해 $l'\mu$에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$\left(l'\overline{X} - \sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)l'Sl}\,,\quad l'\overline{X} + \sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)l'Sl}\right)$$

■ 각 모평균의 개별적 신뢰구간을 이용해 동시 신뢰영역을 구해보자.

각 변수에 대해 $100(1-\alpha)\%$ 신뢰구간은

$$\overline{X}_1 - t_{n-1}(\alpha/2)\sqrt{\frac{s_{11}}{n}} \leqq \mu_1 \leq \overline{X}_1 + t_{n-1}(\alpha/2)\sqrt{\frac{s_{11}}{n}}$$

$$\vdots$$

$$\overline{X}_p - t_{n-1}(\alpha/2)\sqrt{\frac{s_{pp}}{n}} \leqq \mu_p \leq \overline{X}_p + t_{n-1}(\alpha/2)\sqrt{\frac{s_{pp}}{n}}$$

■ 동시신뢰영역을 구하기위해 다음의 확률을 계산하면

$$P[t \mid \text{모든 } \mu_i\text{들을 포함하는 영역}] = (1-\alpha)\cdots(1-\alpha) = (1-\alpha)^p$$

이므로 $p=4$인 경우 각 모평균의 신뢰수준 $1-\alpha=0.95$에 대해 이와 같은 방법으로 신뢰 영역을 구하면 신뢰수준은 0.8145가 된다. 신뢰영역의 신뢰수준을 0.95 로 하기위해서는 각 모평균에 대한 신뢰수준을 0.987 정도로 하는 신뢰구간을 구한 후 신뢰구간들의 교집합을 취해야한다.

■ 모평균벡터들에 대한 선형조합 $a\mu = a_1\mu_1 + a_2\mu_2 + \cdots + a_p\mu_p$ 의 신뢰영역을 구하고자 한다. m개의 선형조합 $a_1'\mu, a_2'\mu, \ldots, a_m'\mu$ 에 대한 신뢰영역을 구해보자.

C_i를 $a_i'\mu$ 에 대한 신뢰구문(confidence statement)으로 놓고

신뢰수준을 $P[C_i$가 옳다$] = 1-\alpha_i, \quad i = 1,2,\ldots,m$ 이라 할 때,

Bonferroni 방법에 의한 신뢰영역을 구해보면

$$P[\text{모든 } C_i\text{가 옳다}] = 1 - P[\text{적어도 하나의 } C_i\text{는 틀리다}] \qquad \blacksquare$$

$$\geq 1 - \sum_{i=1}^{m} P[C_i\text{는 틀리다}] = 1 - \sum_{i=1}^{m} (1 - P[C_i\text{는 옳다}])$$

$$= 1 - (\alpha_1 + \alpha_2 + \cdots + \alpha_m).$$

$m = 4$인 경우: 각 모평균의 신뢰수준 $1-\alpha_i = 0.95$에 대해 Bonferroni 방법으로 신뢰영역을 구하면 신뢰수준은 0.8이 된다. 신뢰영역의 신뢰수준을 0.95로 하기위해서는 각 모평균에 대한 신뢰수준을 0.9875정도로 정해야한다.

■ 일반적으로 같은 신뢰수준을 얻기 위한 타원체의 동시 신뢰영역이 다른 방법으로 구한 신뢰영역보다 더 넓다.

≪예제 5.5≫ ≪예제 5.1≫에서 사용된 [표 5.1]의 데이터를 이용하여

$$X_1 = \text{땀흘리는 비율(sweat rate)}$$
$$X_2 = \text{sodium 양}$$
$$X_3 = \text{potassium 양}$$

의 모평균벡터 $\mu=(\mu_1,\mu_2,\mu_3)'$에 대한 95% 동시신뢰영역을 구하여보자.

$$\left\{\mu \mid n(\overline{X}-\mu)'S^{-1}(\overline{X}-\mu) \le \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right\}$$
$$= \left\{\mu \mid n(\overline{X}-\mu)'S^{-1}(\overline{X}-\mu) \le 8.18\right\}$$

를 만족하는 벡터 μ들의 집합이 되며 3차원 공간에서 타원체의 형태로 나타난다.

2차원 공간에서의 신뢰영역을 구하기위해 다음의 2개 변수

$X_1 =$ 땀의 비율(sweat rate), $X_2 =$ sodium 양

의 모평균벡터 $\mu=(\mu_1,\mu_2)'$에 대한 95% 동시신뢰영역을 구하여보자.

$$\overline{X}=\begin{pmatrix}4.640\\45.400\end{pmatrix},\qquad S=\begin{pmatrix}2.879 & 10.010\\10.010 & 199.788\end{pmatrix},\qquad S^{-1}=\begin{pmatrix}0.421 & -0.021\\-0.021 & 0.005\end{pmatrix}$$

이고 $\dfrac{(n-1)p}{n-p}F_{p,n-p}(\alpha)=\dfrac{19\cdot 2}{20-2}F_{2,18}(0.05)=2.11\cdot 3.55=7.49$이므로

$$n(\overline{X}-\mu)'S^{-1}(\overline{X}-\mu)\le\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$$

$$20\cdot\left[\begin{pmatrix}4.640\\45.400\end{pmatrix}-\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix}\right]'\begin{pmatrix}0.421 & -0.021\\-0.021 & 0.005\end{pmatrix}\left[\begin{pmatrix}4.640\\45.400\end{pmatrix}-\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix}\right]\le 7.49$$

$$0.421(\mu_1-4.64)^2+0.005(\mu_2-45.4)^2-0.042(\mu_1-4.64)(\mu_2-45.4)\le 0.375$$

그러므로 모평균벡터$\mu=(\mu_1,\mu_2)'$에 대한 95% 동시 신뢰영역은

$$\left\{(\mu_1,\mu_2)'\middle|1.123(\mu_1-4.64)^2+0.013(\mu_2-45.4)^2-0.112(\mu_1-4.64)(\mu_2-45.4)\le 1.0\right\}$$

이 되어 위 타원의 부등식을 만족하는 영역.

만약 동시 신뢰영역을 고려하지 않고 각각의 모수의 신뢰구간을 따로 구하여 합한다면 95%의 신뢰수준을 유지하기위해 각각의 모수에 대해 97.5%의 신뢰구간을 구해야한다.

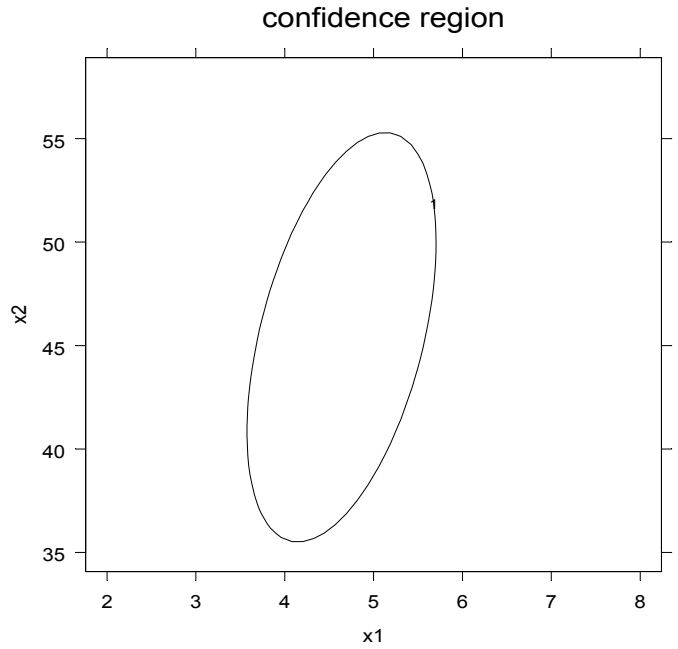
$$\overline{X}_i - t_{n-1}(\alpha/2)\ \sqrt{\frac{S_{ii}}{n}} \quad \leqq \quad \mu_i \quad \leq \quad \overline{X}_i + t_{n-1}(\alpha/2)\ \sqrt{\frac{S_{ii}}{n}}$$

를 이용해 구하면 $t_{19}(0.0125) = 2.433$이므로

$$\overline{X}_1 \pm 2.433 \cdot \sqrt{\frac{2.879}{20}} = 4.640 \pm 0.923 = (3.717,\ 5.563)$$

$$\overline{X}_2 \pm 2.433 \cdot \sqrt{\frac{199.788}{20}} = 45.4 \pm 7.690 = (37.71,\ 53.09)$$

를 얻어 2차원 공간에서 직사각형 모양의 신뢰영역으로 나타내어진다.



[그림 5.1] 타원의 중심이 (4.64, 45.4)인 $(\mu_1,\mu_2)'$ 에 대한 95% 동시 신뢰영역

5.8 R을 이용한 Hotelling 검정

● R에서는 직접 Hotelling의 통계량을 출력하지 않지만 6장에서 소개되는 manova() 함수와 summary(·, test="Wilks") 옵션을 이용하여 Wilks lambda 값을 구할 수 있으며 일집단의 경우

$$T^2 = (n-1)\frac{1-\Lambda^*}{\Lambda^*}$$

를 이용하여 통계량을 계산할 수 있다.

[프로그램 5.1] 일집단 Hotelling의 T^2 검정

```
sweat=read.csv("C:/data/s                    header=T)
sweat   ;    attach(sweat)

x=sweat[,2:4]
p = ncol(x)
xbar =apply(x,2,mean)
xbar
S=cov(x)   ;   S
n= dim(x)[[1]]
mu0=c(4,50,10)

library(MASS)   # for ginv
T2= n* t(xbar-mu0) %*% ginv(S) %*% (xbar-mu0)
T2
f= ((n-p)/(n-1)*p )*T2
pvalue= 1-pf(f,p,n-p)
pvalue
```

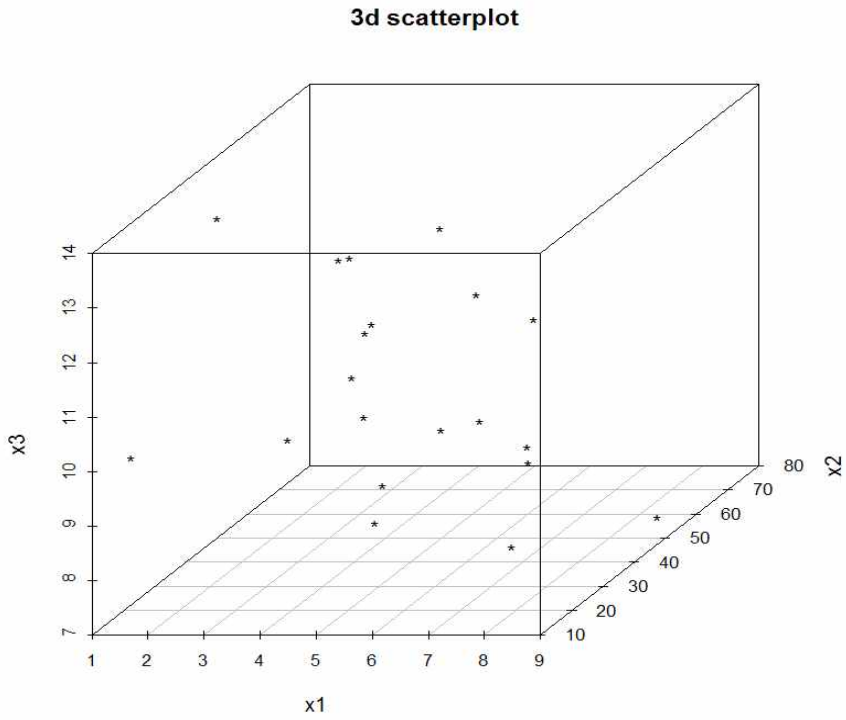
[결과 5.1] 일집단 Hotelling의 T^2 검정

```
> x=sweat[,2:4]
> p = ncol(x)
> xbar =apply(x,2,mean)
> xbar
    x1     x2     x3
 4.640 45.400  9.965
> S=cov(x)
> S
          x1       x2        x3
x1  2.879368  10.0100 -1.809053
x2 10.010000 199.7884 -5.640000
x3 -1.809053  -5.6400  3.627658
> n= dim(x)[[1]]
> mu0=c(4,50,10)
```

```
> library(MASS)   # for ginv
> T2= n* t(xbar-mu0) %*% ginv(S)
%*% (xbar-mu0)
> T2
          [,1]
[1,] 9.738773
> f= ((n-p)/(n-1)*p )*T2
> pvalue= 1-pf(f,p,n-p)
> pvalue
              [,1]
[1,] 1.347883e-06
```

[프로그램 5.2] 3차원 산점도

```
# 3D 산점도
library(scatterplot3d)
scatterplot3d(x, pch="*", main="3d scatterplot")  # 그림 5.2
```



[그림 5.2] 땀 자료에 대한 3차원 산점도