CHAPTER 9.

다변량 분산 분석

우리 지역에 호수가 3 개 있다. 산소량의 차이는 있을까? 이는 일원 분산분석에 의해 호수별 산소량의 평균 차이가 있는지 알아 보면 된다. 만약 각 호수 별 오염 정도를 알아보기 위하여 수은 함유량을 측정하였다고 하자. 이런 경우 호수 별 산소량 차이와 수은 함유량 차이 각각에 대해 일원 분산 분석을 실시하면 된다.

그러나 호수의 산소량과 수은 함유량은 서로 상관 관계가 있으므로 이를 고려한 차이를 보는 것이 필요할 것이다. 이처럼 분산 분석에서 종속 변수가 2 개 이상인 경우 이를 다변량 분산 분석이라 한다. 요인이 하나인 일원 다변량 분산 분석이다. 종속 변수인 근심 지수, 불면 지수, 불쾌 지수 역시 상관 관계가 존재하기 때문에 성별, 재산 정도에 따른 근심 지수, 불면 지수, 불쾌 지수의 차이가 있는지 알아보기 위하여 다변량 분산 분석을 이용하는 것이 바람직하다. (물론 1 장에서 언급한 것처럼 유의 수준 조정의 의미도 있지만, 1 장 참고) 이런 경우 이를 이원 다변량 분석이라 한다.

일원 다변량 분산 분석은 집단간 다변량 평균 차이 검정과 동일하다. 즉 산소량과 수은 평균 벡터가 호수 별 집단간 차이가 있는지 분석하는 것과 동일하다. 이 책에서는 예제 위주로 일원 다변량 분산 분석만을 다루기로 한다. 이원 다변량 분산 분석도 유사한 방법으로 이원 분산 분석을 확장할 수 있다.

9.1. 일원 분산 분석

다음은 3 개 호수의 산소량의 차이가 있는지 알아보기 위하여 각 호수의 중앙에서 깊이 1m 의 물로부터 산소량(ppm)을 측정한 자료이다. 호수에서 위치에 따라 산소량의 차이가 있을 것이므로 10 곳을 선택하여 각 산소량을 측정한 것이다.

Lake	Observation
1	0 2 1 3 1 2 3 4 1 5
2	1 3 4 6 8 7 5 3 4 5
3	14 26 25 18 19 22 21 16 20 30

등분산 가정 검정: 분산분석에서 등분산 가정은 Hartley's test 를 이용하면 된다. 그러나 일반적으로 분산 분석에서는 등분산 가정을 검정하지 않는다.

(1)귀무가설: $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_t^2$

(2)검정 통계량: $F_{\max} = \frac{\max(s_i^2)}{\min(s_i^2)} \sim F(\text{자유도} : \text{분자표본크기} - 1, \text{분모표본크기} - 1)$

```
data one;
    input lake oxygen @@;
    cards;
1 0 1 2 1 1 1 3 1 1 1 2 1 3 1 4 1 1 1 5
2 1 2 3 2 4 2 6 2 8 2 7 2 5 2 3 2 4 2 5
3 14 3 26 3 25 3 18 3 19 3 22 3 21 3 16 3 20 3 30
run;

proc glm data=one;
    class lake;
    model oxygen=lake;
    means lake/scheffe lines;
run;
```

means 문장은 다중 비교를 위한 것이다. LINES 옵션은 집단간 평균의 차이가 유의한 것을 구별하기 위해 사용되었다.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2117.400000	1058.700000	105.52	<.0001
Error	27	270.900000	10.033333		
Corrected Total	29	2388.300000			

변동	자유도	자승합 SS	평균자승합 MS	F-값	유의 확률 p-값
처리 효과	2	2117.4	1058.7	105.52	<.0001
오차	27	270.9	10.033333		
총합	29	2388.3			

유의 확률이 0.0001 미만이므로 호수에 따른 산소량의 차이는 매우 유의하다. 이제 호수 별 차이를 보기 위하여 Scheffe 다중 비교 결과를 보자. 다중 비교는 분산 분석표의 F-검정 결과가 유의하지 않더라도 실시한다.

Means with the same letter are not significantly different.

Scheffe Grouping	Mean	N	lake
A	21.100	10	3
B	4.600	10	2
B			
B	2.200	10	1

알파벳이 같은 범주(집단은) 차이가 없음을 의미한다. 호수 3 의 산소량 평균은 21, 호수 2 는 4.6, 호수 1 은 2.2 이다. 호수 1 과 2 의 산소량 차이는 없으나(알파벳이 A 로 동일) 호수 3 과 호수 2, 호수 3 과 호수 1 은 산소량의 차이가 존재한다.

9.2. 일원 다변량 분산 분석

9.2.1. 예제 자료

SAS 의 예제 자료(POTTERY.txt)를 이용하여 일원 다변량 분산 분석을 설명하기로 한다. 도자기 회사별 5 가지 금속의 산화물 함유량의 차이가 있는지 알아보고자 실험 조사하였다. 도자기 회사는 4 개이고 회사별로 14, 5, 2, 5 반복하여 26 개 관측치를 얻었다.

```
DATA POTTERY;
      INPUT Site $ 1-15 Al Fe Mg Ca Na;
      DATALINES;
   Llanederyn    14.4 7.00 4.30 0.15 0.51
   Llanederyn    13.8 7.08 3.43 0.12 0.17
   Llanederyn    14.6 7.09 3.88 0.13 0.20
   Llanederyn    11.5 6.37 5.64 0.16 0.14
```

우선 회사별 금속 산화물 함유량의 평균과 분산을 보기 위하여 다음 프로그램을 실행하자.

```
PROC TABULATE DATA=POTTERY;
    CLASS SITE;
    VAR Al Fe Mg Ca Na;
    TABLE (SITE ALL), (Al Fe Mg Ca Na)*(MEAN STD);
RUN;
```

(1)TABLE 문은 표의 형태를 정해 주는 명령문이다. 콤마(,) 앞에는 행을 뒤는 열을 지정한다.

(2)ALL 은 전체 기초 통계량을 구하는 것이다. 그러므로 (SITE ALL) 대신 SITE 만 사용하면 마지막 열은 출력되지 않는다.

(3)*의 의미는 교차를 의미한다. 그러므로 열에 각 금속 산화물에 대해 평균, 표준편차가 출력된다.

	Al		Fe		Mg		Ca		Na	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Site										
AshleyRails	17.32	1.66	1.51	0.74	0.61	0.06	0.05	0.03	0.05	0.01
Caldicot	11.70	0.14	5.42	0.04	3.86	0.12	0.30	0.01	0.05	0.01
IslandThorns	18.18	1.78	1.71	0.44	0.67	0.03	0.03	0.03	0.05	0.03
Llanederyn	12.56	1.38	6.37	0.79	4.83	1.09	0.20	0.06	0.25	0.12
All	14.49	2.99	4.47	2.41	3.14	2.18	0.15	0.10	0.16	0.14

각 도자기 회사별 각 금속 함유량의 평균을 한 눈에 볼 수 있어 다변량 분석 결과를 어느 정도 예측할 수 있다. 만약 다변량 분산 분석 대신 일원 분산 분석을 5 번 시행한다면 각 열의 평균 차이를 분석하면 된다.

9.2.2. 모형

집단의 수를 K 라 하고 각 모집단으로부터 확률 표본을 다음과 같이 정의하자.

모집단 1: $y_{11}, y_{12}, \ldots, y_{1n_1}$, 모집단 2: $y_{21}, y_{22}, \ldots, y_{1n_2}$, … 모집단 K : $y_{11}, y_{12}, \ldots, y_{1n_k}$

이를 모형화 하면 다음과 같다.

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ip} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} + \begin{pmatrix} \alpha_{i1} \\ \alpha_{i2} \\ \vdots \\ \alpha_{ip} \end{pmatrix} + \begin{pmatrix} e_{ij1} \\ e_{ij2} \\ \vdots \\ e_{ijp} \end{pmatrix}, \quad \begin{pmatrix} e_{ij1} \\ e_{ij2} \\ \vdots \\ e_{ijp} \end{pmatrix} \underset{\sim}{iid} N_p(0, \Sigma)$$

p 는 측정 변수의 개수이고 $i = 1,2,...,k$ 는 모집단의 수, $j = 1,2,...,n_i$ 는 반복 수이다. 귀무가설은 다음과 같다. (모든 변수의 집단간 차이는 유의하지 않다)

$$H_0 : \begin{pmatrix} \alpha_{11} \\ \alpha_{12} \\ \vdots \\ \alpha_{1p} \end{pmatrix} = \begin{pmatrix} \alpha_{21} \\ \alpha_{22} \\ \vdots \\ \alpha_{2p} \end{pmatrix} = ... = \begin{pmatrix} \alpha_{k1} \\ \alpha_{k2} \\ \vdots \\ \alpha_{kp} \end{pmatrix}$$

위를 행렬로 표현하면 다음과 같다.

$$Y_{n\times p} = X_{n\times k}\beta_{k\times p} + e_{n\times p}, \quad V(e) \sim N(0, I_n \otimes \Sigma_{p\times p}), \quad \otimes \text{는 Kronecker 곱}$$

귀무가설은

$$L\beta M_{p\times p} = 0$$

이다.

9.2.3. 검정

일변량 분산 분석을 확장하여 위의 귀무가설을 검정할 수 있다. 총 변동을 within 변동과 between 변동으로 나눌 수 있다.

$$B = M'(Lb)'(L(X'X)^{-1}L')^{-1}(Lb)M$$

$$W = M'(Y'Y - b'(X'X)b)M$$

다변량 검정 통계량은 모두 $W^{-1}B$ 고유치의 함수이다.

(1)Wilks' lambda = $\frac{|W|}{|W+B|}$, $|W|$는 행렬 W 의 행렬식이다.

(2)Pillai's trace = $tr(B(B+W)^{-1})$

(3)Hotelling-Lawley trace = $tr(W^{-1}B)$

(4)Roy's maximum root = $W^{-1}B$ $B^{-1}W$ 의 최대 고유치

9.2.4. SAS 결과

```
PROC GLM DATA=POTTERY;
      CLASS Site;
      MODEL Al Fe Mg Ca Na = Site;
      CONTRAST 'Llanederyn vs. the rest' Site 1 1 1 -3;
      MEANS SITE/TUKEY LINES;
      MANOVA H=_ALL_ / PRINTE PRINTH;
RUN;;
```

(1)MODEL 을 설정할 때는 측정 변수를 왼쪽에 집단 변수(요인)를 오른쪽에 사용하면 된다. 만약 요인이 두 개 이상인 경우에는 오른쪽에 변수 명을 적어 주면 된다. 만약 블록 효과가 있다면 MODEL AL--NA=SITE BLOCK; 이런 식으로 적어 주면 된다.

(2)MEANS 는 개별 분산 분석의 다중 비교(사후 검정) 결과를 보고 싶을 때 사용한다. 다변량 일원 분석의 사후 검정은 아니다.

(3)CONTRAST 는 집단간 비교를 위하여 사용되는 대비이다. ''에는 적절한 검정 이름을 적어 주면 된다.''은 사용하지 않아도 된다.'' 다음에는 집단 변수 명을 적어주면 된다. 그 다음은 합이 0 이 되게 숫자를 적어 주면 된다. 숫자의 개수는 반드시 집단의 개수와 동일해야 한다. 본 예제에서는 도자기 회사가 4 개 이므로 숫자가 4 개이다. 만약 IslandThons 과 AshleyRails 와 다른 두 도자기 회사를 비교하려면 다음과 같이 쓰면 된다. 물론 2 2 -2 -2 를 사용해도 된다. 결과는 동일하다.

```
CONTRAST 'IslandThons+AshleyRails vs. the rest' Site 1 1 -1 -1;
```

(5)마지막 열을 사용해야 검정 결과가 출력된다.

먼저 각 측정 변수에 대한 일원 분산 분석 결과가 출력된다. MEANS 옵션 때문에 각 변수에 대한 TUKEY 다중 비교 결과가 출력된다.

Dependent Variable: Al

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	175.6103187	58.5367729	26.67	<.0001
Error	22	48.2881429	2.1949156		
Corrected Total	25	223.8984615			

도자기 회사별 알루미늄(AL) 함유량의 차이는 존재한다. 그리고 TUKEY 다중 비교 결과 IslandThons 과 AshleyRails 도자기의 알루미늄 함유량이 다른 두 회사에 비해 높음을 알 수 있다. 그러나 이런 해석은 단순 일원 분산 분석 결과이지 다변량 분산 분석 결과는 아니다.

Tukey's Studentized Range (HSD) Test for Al

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	Site
A	18.180	5	IslandThorns
A			
A	17.320	5	AshleyRails
B	12.564	14	Llanederyn
B			
B	11.700	2	Caldicot

유의 확률이 0.0001 미만이므로 5 가지 금속 함유량은 도자기 회사별 차이가 존재한다.

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall Site Effect
H = Type III SSCP Matrix for Site
E = Error SSCP Matrix

S=3 M=0.5 N=8

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.01230091	13.09	15	50.091	<.0001
Pillai's Trace	1.55393619	4.30	15	60	<.0001
Hotelling-Lawley Trace	35.43875302	40.59	15	29.13	<.0001
Roy's Greatest Root	34.16111399	136.64	5	20	<.0001

대비 검정 결과 유의 확률이 0.0001 미만이므로 IslandThons 도자기 금속 함유량은 다른 도자기와 차이가 있음을 알 수 있다. 금속 별 차이는 앞에서 출력한 도자기별 평균을 참고하여 해석하면 된다.

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Llanederyn vs. the rest Effect
H = Contrast SSCP Matrix for Llanederyn vs. the rest
E = Error SSCP Matrix

S=1 M=1.5 N=8

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.05839360	58.05	5	18	<.0001
Pillai's Trace	0.94160640	58.05	5	18	<.0001
Hotelling-Lawley Trace	16.12516462	58.05	5	18	<.0001
Roy's Greatest Root	16.12516462	58.05	5	18	<.0001

다음은 두 회사별(IslandThons+AshleyRails vs. Caldicot+IslandThorns) 차이를 보기 위한 대비 검정 결과이다.

```
CONTRAST 'IslandThons+AshleyRails vs. the rest' Site 1 1 -1 -1;
```

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall IslandThons+AshleyRails vs. the rest Effect
H = Contrast SSCP Matrix for IslandThons+AshleyRails vs. the rest
E = Error SSCP Matrix

S=1 M=1.5 N=8

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.50217801	3.57	5	18	0.0204
Pillai's Trace	0.49782199	3.57	5	18	0.0204
Hotelling-Lawley Trace	0.99132576	3.57	5	18	0.0204
Roy's Greatest Root	0.99132576	3.57	5	18	0.0204