Chapter 8. 분산분석

일원배치 분산분석(One-way ANOVA)

❖ 일원분산분석의 개요

- 개념
  - 두 개 이상의 집단들의 평균값을 비교하는 데 사용하는 통계기법
  - 검증통계량: F
  - 처치변수가 한 개인 가장 간단한 분산분석
- 자료
  - 독립변수: 범주를 나타내는 명목척도
  - 종속변수: 간격척도 혹은 비율척도
- 가정
  - 각 모집단은 정규분포를 이루며, 분산이 동일하다는 가정이 필요
  - 분산의 동질성 검증을 위해서 Levene's test가 사용

일원배치 분산분석(One-way ANOVA)

❖ 자료의 구조

	관 측 치	평 균	제 곱 합
처리 1	$y_{11}, y_{12}, \cdots, y_{1n_1}$	$\overline{y_1}$	$\sum_{j=1}^{n_1}(y_{1j}-\overline{y_1})^2$
처리 2	$y_{21}, y_{22}, \cdots, y_{2n_2}$	$\overline{y_2}$	$\sum_{j=1}^{n_2}(y_{2j}-\overline{y_2})^2$
.			
.			
처리 k	$y_{k1}, y_{k2}, \cdots, y_{kn_k}$	$\overline{y_k}$	$\sum_{j=1}^{n_k}(y_{kj}-\overline{y_k})^2$

$$\text{총 평균} : \overline{y} = \frac{\text{모든 관측값의 합}}{n_1+n_2+\cdots+n_k} = \frac{n_1\overline{y_1}+\cdots+n_k\overline{y_k}}{n_1+n_2+\cdots+n_k}$$

일원배치 분산분석(One-way ANOVA)

❖ 총 편차의 분해식

- 관측값 = 총평균 + 처리로인한편차 + 잔차(residual)

$$y_{ij} = \overline{y} + (\overline{y_i} - \overline{y}) + (y_{ij} - \overline{y_i})$$

- 총 편차 = 처리 효과 + 잔차

$$\therefore y_{ij} - \overline{y} = (\overline{y_i} - \overline{y}) + (y_{ij} - \overline{y_i})$$

❖ 제곱합의 분할

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 = \sum_{i=1}^{k} n_i(\overline{y_i} - \overline{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2$$

↑ 총 제곱합 (total sum of Squares : SST)

↑ 처리 제곱합 (treatment sum of Squares : SS_T)

↑ 잔차 제곱합 (residual sum of Squares : SSE)

일원배치 분산분석(One-way ANOVA)

- 제곱합의 자유도
  - SST의 자유도 = SS_{tr} 의 자유도 + SSE 의 자유도

    $(n-1) = (n-k) + (k-1)$

- 평균 제곱
  - 평균제곱 = $\frac{\text{제곱합}}{\text{자유도}}$

일원배치 분산분석(One-way ANOVA)

❖ 분산분석표 (ANOVA table)

요 인	제 곱 합	자 유 도	평 균 제 곱
처 리	$SS_T = \sum_{i=1}^{k} n_i(\overline{y_i} - \overline{y})^2$	k − 1	$MST = \frac{SS_T}{k-1}$
잔 차	$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2$	n − k	$MSE = \frac{SS_E}{n-k}$
합 계	$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2$	n − 1	

일원배치 분산분석(One-way ANOVA)

- K개의 처리를 비교하기 위한 통계적 모형
  - 통계적 모형

$$Y_{ij} = \mu + \alpha_i + e_{ij}\ ,\ i = 1, \cdots, k\ ,\ \ j = 1, \cdots, n$$

  - 선 형 성 : $E(e_{ij}) = 0$, 즉 $E(Y_{ij}) = \mu + \alpha_i$
  - 등분산성 : $Var(e_{11}) = \cdots = Var(e_{kn}) = \sigma^2 > 0$
  - 독 립 성 : $e_{11}, \cdots, e_{kn}$는 서로 독립
  - 정 규 성 : $e_{ij} \sim N(0, \sigma^2)$

일원배치 분산분석(One-way ANOVA)

- 평균들의 동일성에 대한 F-검정
  - 처리 효과의 유의성 검정
    - 가설

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \quad \text{vs} \quad H_1 : not\, H_0$$
$$(\mu_1 = \mu_2 = \cdots = \mu_k)$$

    - 검정통계량

$$F = \frac{SS_T/(k-1)}{SSE/(n-k)} = \frac{MST}{MSE}$$

    - 기각역

$$R : F \geq F_\alpha(k-1, n-k)$$

일원배치 분산분석(One-way ANOVA)

- 예제(수확량)
  - 분류변수(요인): FERTIL
  - 종속변수: YIELD

수확량					
FERTIL	YIELD				
F1	148	76	134	98	
F2	166	153	255		
F3	264	214	327	304	
F4	335	436	423	380	465

일원배치 분산분석(One-way ANOVA)

- ❖ 분산 동일성 검정
  - HOVTEST = 옵션
    - BARTLETT(Bartlett)
    - BF(Brown & Forsythe)
    - LEVENE(Levene)
    - OBRIEN(O'Brien)
- ❖ 사후분석 – 다중비교
  - LSD, DUNCAN, TUKEY, SHEFFE
  - 다중비교 출력형식
    - CLDIFF: 모평균 차이에 대한 신뢰구간
    - LINES: 그룹화

일원배치 분산분석(One-way ANOVA)

일원배치 분산분석(One-way ANOVA)

```
DATA harvest;
INPUT fertil$ yield @@;
CARDS;
F1 148 F1   76 F1 134 F1   98
F2 166 F2 153 F2 255
F3 264 F3 214 F3 327 F3 304
F4 335 F4 436 F4 423 F4 380 F4 465
;RUN;
```

```
PROC ANOVA DATA=harvest;
CLASS fertil;
MODEL yield = fertil;
MEANS fertil / HOVTEST=BARTLETT;
MEANS fertil / TUKEY CLDIFF ALPHA = 0.10;
MEANS fertil / TUKEY LINES ALPHA = 0.10;
RUN;
```

이원배치 분산분석(Two-way ANOVA)

❖ 이원분산분석의 개요

- 개념
  - 팩토리얼 디자인(factorial design)
    - 두 개 이상의 독립처치변수의 수준변화에 따른 결과변수값의 변화를 조사하기 위한 실험디자인
    - 각 처치변수를 factor라고 부름
    - 예) 처치수준a(factor A), 처치수준b(factor B) => a×b factorial design
  - 처치변수가 두 개인 경우 처치효과를 조사하기 위해 이원분산분석을 적용
  - 주효과(main effect)
    - 한 처치변수의 변화가 결과변수에 미치는 영향에 관한 것
  - 상호작용효과(interaction effect)
    - 다른 처치변수의 변화에 따라 한 처치변수가 결과변수에 미치는 영향에 관한 것
- 자료
  - 일원분산분석의 경우와 동일
- 가정
  - 일원분산분석의 경우와 동일

이원배치 분산분석(Two-way ANOVA)

❖ 이원배치 분산분석(Two-way ANOVA) – 반복이 없는 경우

- 자료의 구조

	처리 1	처리 2		처리 k	평균
블록1	y_{11}	y_{12}	…	y_{1k}	$\overline{y_{1.}}$
블록2	y_{21}	y_{22}	…	y_{2k}	$\overline{y_{2.}}$
⋮			…		
블록b	y_{b1}	y_{b2}	…	y_{bk}	$\overline{y_{b.}}$
처리평균	$\overline{y_{.1}}$	$\overline{y_{.2}}$	…	$y_{.k}$	$\overline{y_{..}}$

- 제곱합

관측값 = 전체평균 + 블록편차 + 처리편차 + 잔차

$$y_{ij} = \overline{y_{..}} + (\overline{y_{i.}} - \overline{y_{..}}) + (\overline{y_{.j}} - \overline{y_{..}}) + (y_{ij} - \overline{y_{i.}} - \overline{y_{.j}} + \overline{y_{..}})$$

이원배치 분산분석(Two-way ANOVA)

- ANOVA 표

요인	제곱합	자유도	평균제곱	F-비
블 록	$SS_B = k\sum_{i=1}^{b}(\overline{y_{i.}} - \overline{y_{..}})^2$	b-1	$MS_B = \frac{SS_B}{b-1}$	$\frac{MS_B}{MSE}$
처 리	$SS_T = b\sum_{j=1}^{k}(\overline{y_{.j}} - \overline{y_{..}})^2$	K-1	$MS_T = \frac{SS_T}{k-1}$	$\frac{MS_T}{MSE}$
잔 차	$SSE = \sum_{i=1}^{b}\sum_{j=1}^{k}(y_{ij} - \overline{y_{i.}} - \overline{y_{.j}} - \overline{y_{..}})^2$	(b-1)(k-1)	$MSE = \frac{SSE}{(b-1)(k-1)}$	
합 계	$SST = \sum_{i=1}^{b}\sum_{j=1}^{k}(y_{ij} - \overline{y_{..}})^2$			

이원배치 분산분석(Two-way ANOVA)

- 통계적 모형

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}\ ,\ i = 1, \cdots, b\ ,\ \ j = 1, \cdots, k$$

  - $\sum_{i=1}^{b} \alpha_i = 0\ ,\ \sum_{j=1}^{k} \beta_i = 0$
  - 독립성 : $e_{11}, \cdots, e_{kn}$는 서로 독립
  - 정규성 : $e_{ij} \sim N(0, \sigma^2)$

- 유의성 검정
  - 가설

$$H_0 : \alpha_1 = 0 = \cdots = \alpha_b = 0 \quad \text{vs} \quad H_1 : not\ H_0$$

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{vs} \quad H_1 : not\ H_0$$

  - 검정 통계량

$$F \equiv \frac{MS_B}{MSE} \quad (F \equiv \frac{MS_T}{MSE})$$

  - 기각역

$$R : F \geq F_\alpha(b-1, (b-1)(k-1))(F \geq F_\alpha(k-1, (b-1)(k-1)))$$

이원배치 분산분석(Two-way ANOVA)

❖ 예제(수확량)

- 분류변수(요인): product, customer
- 종속변수: prefer

주행거리					
제품\소비자	1	2	3	4	5
A1	5	7	9	10	8
A2	2	3	4	5	2
A3	4	7	6	5	7
A4	6	4	2	2	1

이원배치 분산분석(Two-way ANOVA)

이원배치 분산분석(Two-way ANOVA)

```
DATA prefer;
DO product =“A1”, “A2”, “A3”, “A4”;
DO customer = 1 TO 5 BY 1;
INPUT prefer @@;
OUTPUT;END;END;
CARDS;
5 7 9 10 8
2 3 4 5 2
4 7 6 5 7
6 4 2 2 1
;RUN;
```

```
PROC ANOVA DATA=prefer;
CLASS product customer;
MODEL prefer = product customer;
MEANS product / DUNCAN TUKEY ALPHA = 0.10;
RUN;
```

이원배치 분산분석(Two-way ANOVA)

- 이원배치 분산분석(Two-way ANOVA) – 반복이 있는 경우
  - 통계적 모형

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}\ ,\ i = 1, \cdots, a\ ,\ \ j = 1, \cdots, b,\ \ k = 1, \cdots, r$$

  - 유의성 검정
    - 가설

$$H_0 : \alpha_1 = 0 = \cdots = \alpha_b = 0 \quad \text{vs} \quad H_1 : not\ H_0$$
$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{vs} \quad H_1 : not\ H_0$$
$$H_0 : \gamma_{ij} = 0,\ \ i = 1, \cdots, a;\ j = 1, \cdots, b$$

이원배치 분산분석(Two-way ANOVA)

- 예제(주행거리)
  - 분류변수(요인): city, design
  - 종속변수: sales

주행거리			
도시구분\디자인	A	B	C
대	23 20 21	22 19 20	19 18 21
중	22 20 19	24 25 22	20 19 22
소	18 18 16	21 23 20	20 22 24

이원배치 분산분석(Two-way ANOVA)

이원배치 분산분석(Two-way ANOVA)

```
DATA sales;
DO city =“LARGE”, “MIDDLE”, “SMALL”;
DO design = “A”, “B”, “C”;
DO rep = 1, 2, 3;
INPUT sales @@;
OUTPUT;END;END; END;
CARDS;
23 20 21 22 19 20 19 18 21
22 20 19 24 25 22 20 19 22
18 18 16 21 23 20 20 22 24
;RUN;
```

```
PROC ANOVA DATA=sales;
CLASS city design;
MODEL sales = city design city*design;
MEANS city design city*design;
RUN;
```

이원배치 분산분석(Two-way ANOVA)

이원배치 분산분석(Two-way ANOVA) – 평균 프로파일 도표의 작성

```
PROC SUMMARY DATA=sales NWAY;
CLASS city design;
VAR sales;
OUTPUT OUT = MEANOUT MEAN(sales)=MEAN;
RUN;
PROC PRINT DATA=MEANOUT;
RUN;
SYMBOL1 I=JOIN W=1 V=DOT CV=BLACK H=2;
SYMBOL2 I=JOIN W=1 V=CIRCLE CV=BLACK H=2;
SYMBOL3 I=JOIN W=1 V=SQUARE CV=BLACK H=2;
PROC GPLOT DATA=MEANOUT;
PLOT mean*city = design;
RUN;
```

연습문제 1

❖ 어느 공장에서 서식하고 있는 딱정투구벌레에 대한 연구를 하던 도중 벌레들이 선호하는 색상이 있는지를 알아보기 위해 다음과 같은 실험을 실시하였다. 네 가지 색상의 판자를 각각 여섯개씩 준비하여, 그 위에 끈끈이를 바르고 여섯 지점에 각 네 가지 판자를 일 주일 동안 설치하여 잡힌 벌레의 수를 관측하였다. 딱정벌레가 색상을 선호하는 정도가 다르다고 말할 수 있는지를 검정하여 보아라.

벌레 수						
레몬색	45	59	48	46	38	47
흰색	21	12	14	17	13	17
녹색	37	32	15	25	39	41
파란색	16	11	20	21	14	7

연습문제 2

- 다음의 데이터는 다이어트 음식 3종류의 콜레스테롤 함량을 4개의 다른 실험실에서 측정한 결과이다. 분산분석표를 작성하고 다이어트 음식과 실험실에 따라 콜레스테롤 함량에 차이가 있는지를 검정하여 보아라.

	음식1	음식2	음식3
A실험실	3.4	2.6	2.8
B실험실	3.0	2.7	3.1
C실험실	3.3	3.0	3.4
D실험실	3.5	3.1	3.7

연습문제 3

❖ 여섯 종류의 사료를 쥐에게 먹였을 때 쥐의 체중증가에 어떤 영향을 주는지를 알아보기 위해 단백질의 요소에 대한 세 수준(A:쇠고기,곡류,돼지고기)과 단백질의 수준에 대한 두 수준(B:높음,낮음)을 고려한 후 실험을 하였다. 다음의 데이터는 실험을 통해서 얻어진 쥐의 체중증가량을 기록한 결과이다. 요인 A,B및 상호작용의 효과가 있는지를 분석하여라.

단백질수준 / 단백질요소	B1	B2
A1	73 102 118 104 81 107 100 87 117 111	90 76 90 64 86 51 72 90 95 78
A2	98 74 56 111 95 88 82 77 86 92	107 95 97 80 98 74 74 67 89 58
A3	94 79 96 98 102 102 108 91 120 105	49 82 73 86 81 97 106 70 61 82