제 9장. 분산분석

분산분석(analysis of variance, ANOVA) :

특성값의 변동을 나타내는 제곱합을 요인별 제곱합과 오차에 의한 제곱합으로 분해하고, 이들의 비를 통계량으로 하여 요인의 유의성을 검증하는 통계적 기법

- 인자(factor) 또는 요인 : 관측값에 영향을 주는 속성
- 인자수준 (factor level) : 인자의 여러 조건
- 처리(treatment) : 인자 수준의 조합
- 일원배치법(one-way ANOVA) : 특성값에 영향을 주는 요인이 1개만 있는 모형
- 이원배치법(two-way ANOVA) : 특성값에 영향을 주는 요인이 2개 있는 모형

9.1 일원배치법

일원배치법의 자료 구조 :

	처리1	처리2	⋯	처리k	
	y_{11}	y_{21}	⋯	y_{k1}	
	y_{12}	y_{22}	⋯	y_{k2}	
	⋮	⋮	⋯	⋮	
	y_{1n_1}	y_{2n_2}	⋯	y_{kn_k}	
평균	$\bar{y}_{1\cdot}$	$\bar{y}_{2\cdot}$	⋯	$\bar{y}_{k\cdot}$	총평균 $\bar{y}_{\cdot\cdot}$

일원배치법의 모형 :

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad (i = 1, \ldots, k)(j = 1, \ldots, n_i)$$

다만,

μ: 총평균

α_i: 처리i의 효과로서 $\sum n_i \alpha_i = 0$ 을 가정

ϵ_{ij}: 오차항. 서로 독립인 $N(0, \sigma^2)$ 확률변수

제곱합의 분해

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

$$SST = SSE + SSR$$

- SST (총제곱합) = SSE(잔차제곱합) + SS_{tr} (처리제곱합)
- SST (총제곱합): 자유도 $N - 1$
- SSE(잔차제곱합): 자유도 $N - k$
- SS_{tr} (처리제곱합): 자유도 $k - 1$

(단, $N = \sum n_i$: 관측값의 총 개수)

평균제곱(mean square) :

$$MS_{tr} = \frac{SS_{tr}}{k-1}, \quad MSE = \frac{SSE}{N-k}$$

처리효과의 유의성 검정 :

$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ 또는 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = 0$

- 검정통계량

$$F = \frac{MS_{tr}}{MSE}$$

- 분산분석표 (ANOVA table)

요인	제곱합	자유도	평균제곱	F 값	유의확률
처리	SS_{tr}	$k-1$	MS_{tr}	$f = MS_{tr}/MSE$	$P(F \geq f)$
잔차	SSE	$N-k$	MSE		
계	SST	$N-1$			

예 1) 어느 농장에서 서식하고 있는 딱정투구벌레에 대한 연구를 하던 도중, 벌레들이 선호하 는 색상이 있는가를 알아보기 위해 다음과 같은 실험을 실시하였다. 재질과 크기가 같은 네 가지 색상의 판자를 각각 여섯 개씩 준비하여, 그 위에 끈끈이를 바르고 여섯 지점에 각 네 가지 판자를 일주일 동안 설치하여 잡힌 벌레의 수를 관측하였다. 그 결과가 다음과 같다.

판자의 색	잡힌 벌레 수
레몬색(1)	45 59 48 46 38 47
흰색(2)	21 12 14 17 13 17
녹색(3)	37 32 15 25 39 41
파란(4)	16 11 20 21 14 7

벌레들이 선호하는 색상에 차이가 있는지를 유의수준 5%에서 검정해보자.

[풀이] 주어진 자료는 일 원배치모형을 적용할 수 있으며 검정하고자 하는 가설은 다음과 같다.

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0, \quad H_1$: 적어도 한 α_i는 0이 아니다.

분산분석의 시행은 statmodels패키지의 ols 함수를 사용한다. 주어진 자료는 그룹을 나타내는 변수(color)와 관측값(num)을 나타내는 두 개의 열로 나타낼 수 있다. 자료입력의 편의를 위해 색상을 나타내는 변수를 숫자형 변수로 입력할 수 있는데, 이러한 경우 분산분석을 시행하기 전에 수치변수를 요인(factor)으로 변화하는 과정이 먼저 필요하다. 이를 위해서는 ols함수 안에 'C'명령어를 사용할 수 있다.

In [8]:

```
import numpy as np
import pandas

num = np.array([45, 59, 48, 46, 38, 47, 21, 12, 14, 17, 13, 17, 37, 32, 15, 25, 39, 41, 16, 11, 20,
col = np.repeat(range(1,5), np.repeat(6,4))
bugs = pandas.DataFrame({'num':num, 'col':col})
```

In [2]:

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

model = ols('num ~ C(col)', data = bugs).fit()
table = sm.stats.anova_lm(model, typ=2) # Type 2 ANOVA DataFrame
print(table)
```

```
               sum_sq    df          F        PR(>F)
C(col)    4218.458333   3.0  30.551934  1.151046e-07
Residual   920.500000  20.0        NaN           NaN
```

분산분석 결과, 검정통계량의 값은 30.55이고 유의확률은 0.001이하로 매우 작은 것으로 나타났다. 따라서 유의수준 5%에서 모평균이 모두 동일하다는 귀무가설을 기각할 수 있다. 즉, 딱정벌레들이 선호하는 색상에는 차이가 있다는 결론을 얻을 수 있다

9.2 반복이 없는 이원배치법

반복이 없는 이원배치법의 자료구조 :

인자A \ 인자B	B_1	B_2	$\cdots$	B_q	평균
A_1	y_{11}	y_{12}	$\cdots$	y_{1q}	$\bar{y}_{1\cdot}$
A_2	y_{21}	y_{22}	$\cdots$	y_{2q}	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
A_p	y_{p1}	y_{p2}	$\cdots$	y_{pq}	$\bar{y}_{p\cdot}$
평균	$\bar{y}_{\cdot 1}$	$\bar{y}_{\cdot 2}$	$\cdots$	$\bar{y}_{\cdot q}$	$\bar{y}_{\cdot\cdot}$

반복이 없는 이원배치법의 모형 :

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad (i = 1, \ldots, p)(j = 1, \ldots, q)$$

다만,

μ: 총평균

α_i: 인자 A의 i번째 수준의 효과로서 $\sum n_i \alpha_i = 0$ 을 가정

β_j: 인자 B의 j번째 수준의 효과로서 $\sum n_i \beta_j = 0$ 을 가정

ϵ_{ij}: 오차항. 서로 독립인 $N(0, \sigma^2)$ 확률변수

제곱합의 분해

$$\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = q\sum_{i=1}^{p}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + p\sum_{j=1}^{q}(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2$$

$$SST = SS_A + SS_B + SSE$$

- SST (총제곱합) = SS_A(인자A 제곱합) + SS_B(인자B 제곱합) + SSE(잔차제곱합)
- SST (총제곱합): 자유도 $N - 1$

- SS_A(인자A 제곱합): 자유도 $p-1$
- SS_B(인자B 제곱합): 자유도 $q-1$
- SSE(잔차제곱합): 자유도 $(p-1)(q-1)$

인자 A와 B의 효과의 검정:

$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0$

- 분산분석표 (ANOVA table)

요인	제곱합	자유도	평균제곱	F값	유의확률
인자A	SS_A	$p-1$	MS_A	$f_1 = MS_A/MSE$	$P(F \geq f_1)$
인자B	SS_B	$q-1$	MS_B	$f_2 = MS_B/MSE$	$P(F \geq f_2)$
잔차	SSE	$(p-1)(q-1)$	MSE		
계	SST	$pq-1$			

예 2) 어떤 금속 파이프의 부식방지를 위한 코팅방법을 고려하고 있다. 12개의 동일한 파이프에 네 가지 방법(A_1, A_2, A_3, A_4)으로 코팅을 하고 세 가지 토양($$)에 동일한 깊이로 묻고 일정한 시간이 지난 후에 부식정도를 측정한 결과가 다음과 같다. 코팅방법과 토양의 질에 따라 파이프의 부식(corrosion) 정도에 차이가 있는지를 유의수준 5%에서 검정해 보자.

부식 정도	A_1	A_2	A_3	A_4
B_1	64	53	47	51
B_2	49	51	45	43
B_3	50	48	50	52

[풀이] 주어진 자료는 반복이 없는 이원배치법 모형이 적용 가능하며 코팅방법에 따른 효과를 α_i 라고 하고 토양의 질에 따른 효과를 β_i 라고 하면 검정하고자 하는 가설은 다음과 같다.

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0, \quad H_1$: 적어도 한 α_i는 0이 아니다.

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad H_1$: 적어도 한 β_i는 0이 아니다.

주어진 이원배치법의 자료는 3개의 열(column)로 입력할 수 있다. 관측값인 파이프의 부식정도(y)와 두 개의 요인(A,B)로 입력하도록 한다. 단, 요인 A와 요인 B의 인자 수준을 (1,2,3) 등의 숫자로 입력했으므로 앞선 예제와 마찬가지로 분산분석 시행 전에 수치변수를 요인으로 변환하는 과정을 먼저 거쳐야 한다. 분산분석은 lm() 함수를 사용해서 시행할 수 있고, 두 개의 요인은 '+'기호로 연결한다.

In [11]:

```
import numpy as np
import pandas

y = np.array([64, 53, 47, 51, 49, 51, 45, 43, 50, 48, 50, 52])
A = np.tile(range(1,5), 3)
B = np.repeat(range(1,4), 4)
cor = pandas.DataFrame({'A':A, 'B':B, 'y':y})
```

In [4]:

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

model = ols('y ~ C(A) + C(B)', data = cor).fit()
table = sm.stats.anova_lm(model, typ=2) # Type 2 ANOVA DataFrame
print(table)
```

```
               sum_sq   df         F    PR(>F)
C(A)       83.583333  3.0  1.357240  0.342215
C(B)       91.500000  2.0  2.228687  0.188880
Residual  123.166667  6.0       NaN       NaN
```

분산분석표 확인 결과, 코팅방법(A)의 효과에 대한 유의 확률은 0.3422이고 토양의 질(B)의 효과에 대한 유의 확률은 0.1889로 모두 유의수준 0.05보다 높다. 따라서 파이프의 코팅방법이나 토양의 질에 따라서 파이프의 부식 정도에는 유의한 차이가 없다고 할 수 있다.

9.3 반복이 있는 이원배치법

반복이 있는 이원배치법의 자료구조 :

인자A \ 인자B	B_1	B_2	$\cdots$	B_q	평균
A_1	y_{111} ⋮ y_{11r}	y_{121} ⋮ y_{12r}	$\cdots$	y_{1q1} ⋮ y_{1qr}	$\bar{y}_{1..}$
	$\bar{y}_{11.}$	$\bar{y}_{12.}$	$\cdots$	$\bar{y}_{1q.}$	
⋮	⋮	⋮		⋮	⋮
A_p	y_{p11} ⋮ y_{p1r}	y_{p21} ⋮ y_{p2r}	$\cdots$	y_{pq1} ⋮ y_{pqr}	$\bar{y}_{p..}$
	$\bar{y}_{p1.}$	$\bar{y}_{p2.}$	$\cdots$	$\bar{y}_{pq.}$	
	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$	$\cdots$	$\bar{y}_{.q.}$	$\bar{y}_{...}$

반복이 있는 이원배치법의 모형 :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \quad (i = 1, \ldots, p)(j = 1, \ldots, q)(k = 1, \ldots, r)$$

다만,

μ: 총평균

α_i: 인자 A의 i번째 수준의 효과로서 $\sum n_i \alpha_i = 0$ 을 가정

β_j: 인자 B의 j번째 수준의 효과로서 $\sum n_i \beta_j = 0$ 을 가정

γ_{ij}: 인자 A의 i번째 수준과 인자 B의 j번째 수준의 교호작용

ϵ_{ij}: 오차항. 서로 독립인 $N(0, \sigma^2)$ 확률변수

제곱합의 분해

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk}-\bar{y}_{...})^2 = qr\sum_{i=1}^{p}(\bar{y}_{i..}-\bar{y}_{...})^2 + pr\sum_{j=1}^{q}(\bar{y}_{.j.}-\bar{y}_{...})^2 + r\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2 + \sum_{i=}^{p}$$

$$SST = SS_A + SS_B + SS_{A\times B} + SSE$$

- SST (총제곱합) = SS_A(인자A 제곱합) + SS_B(인자B 제곱합) + $SS_{A\times B}$(교호작용 제곱합) + SSE(잔차제곱합)
- SST (총제곱합): 자유도 $pqr-1$
- SS_A(인자A 제곱합): 자유도 $p-1$
- SS_B(인자B 제곱합): 자유도 $q-1$
- $SS_{A\times B}$(교호작용 제곱합): 자유도 $(p-1)(q-1)$
- SSE(잔차제곱합): 자유도 $pq(r-1)$

인자 A와 B의 효과의 검정:

$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0$

$H_0 : \gamma_{ij} = 0, \quad \forall i, j$

- 분산분석표 (ANOVA table)

요인	제곱합	자유도	평균제곱	F-값	유의확률
인자A	SS_A	$p-1$	MS_A	$f_1 = MS_A/MSE$	$P(F \geq f_1)$
인자B	SS_B	$q-1$	MS_B	$f_2 = MS_A/MSE$	$P(F \geq f_2)$
교호작용	$SS_{A\times B}$	$(p-1)(q-1)$	$MS_{A\times B}$	$f_3 = MS_{A\times B}/MSE$	$P(F \geq f_3)$
잔차	SSE	$pq(r-1)$	MSE		
계	SST	$pqr-1$			

예3) (alzheimer.txt) 다음은 음악 감상이 알츠하이머를 겪고 있는 환자들의 불안감 정도에 어떠한 영향을 미치는지 알아보기 위한 실험의 결과이다. 초기(early stage)와 중기(middle stage)의 알츠하이머 환자들 각각에 대해 가벼운 대중음악(easy), 모차르트의 곡(Mozart), 피아노 연주곡(piano)의 세 종류의 음악을 들려주고 환자들이 느끼는 불안감의 정도를 측정하였다. 측정된 점수가 높을수록 불안감의 정도가 높다는 것을 의미한다. 주어진 자료에 대해 유의수준 5%에서 이원배치 분산분석을 시행해보자.

불안감의 정도	piano	Mozart	easy
early stage Alzheimer's	21 24 22 18 20	9 12 10 5 9	29 26 30 24 26
middle stage Alzheimer's	22 20 25 18 20	14 18 11 9 13	15 18 20 13 19

[풀이] 검정하고자 하는 가설은 다음과 같다. (단, α_i는 음악의 종류에 따른 효과에 따른 효과이고, β_j는 알츠하이머 경과 정도에 따른 효과이고 $\gamma_i j$는 는 음악과 질병의 상호작용을 의미한다.)

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$

$H_0 : \beta_1 = \beta_2 = 0$

$H_0 : \gamma_{ij} = 0 \quad \forall i, j$

알츠하이머 병의 진행 정도(B)와 음악의 종류(A)에 대한 상호작용의 유무를 평균그림(interaction plot)을 통해 확인해보자. seaborn 패키지의 pointplot 함수를 사용하면 평균 그림(interaction plot)을 그릴 수 있고, 두 개의 요인변수와 관측값을 나타내는 변수를 차례대로 입력해주면 다음과 같은 결과를 얻을 수 있다. 평균 그림 확인 결과,

두 요인 사이에는 상호작용이 존재하는 것으로 보인다.

In [1]:

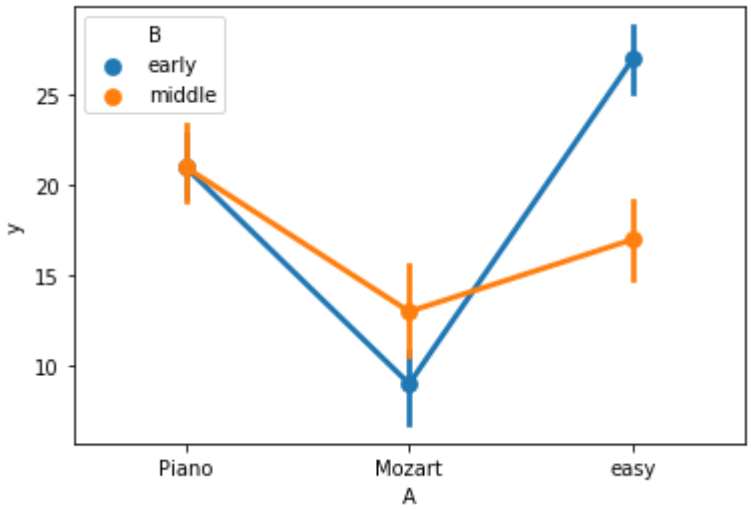
```
# %cd D:₩
```

D:₩

In [6]:

```
import pandas
import seaborn
import matplotlib.pyplot as plt

alz = pandas.read_table("dataset/ch09/alzheimer.txt", sep = " ")
seaborn.pointplot(x = "A", y = "y", hue = "B", data = alz)
plt.show()
```



상호작용을 포함하는 이원배치 분산분석은 두 개의 요인을 곱(*)으로 표현하여 다음과 같이 시행할 수 있다.

In [7]:

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

model = ols('y ~ C(A)*C(B)', data = alz).fit()
table = sm.stats.anova_lm(model, typ=2) # Type 2 ANOVA DataFrame
print(table)
```

	sum_sq	df	F	PR(>F)
C(A)	740.0	2.0	49.887640	2.824411e-09
C(B)	30.0	1.0	4.044944	5.566448e-02
C(A):C(B)	260.0	2.0	17.528090	2.029321e-05
Residual	178.0	24.0	NaN	NaN

분산분석 결과, 유의수준 5%에서 알츠하이머 병의 진행 정도(B)와 음악의 종류(A)에 따른 상호작용은 존재하는 것으로 나타났다 (F=17.5281, p-value<0.001). 이는 앞서 평균 그림을 통해서도 확인할 수 있었다. 또한 감상한 음악의 종류에 따라서 환자들의 불안감은 차이가 존재했지만 (p-value<0.001) 알츠하이머 병의 진행 정도에 따라서는 환자들이 느끼는 불안감에는 차이가 존재하지 않았다(p-value=0.056).