

## 회귀분석/분산분석- 추가내용

서울대학교 통계연구소

# 상관분석

## ▶ Anscombe's quartet

- four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed.

[https://en.wikipedia.org/wiki/Anscombe%27s\\_quartet](https://en.wikipedia.org/wiki/Anscombe%27s_quartet)

```
head(anscombe)
```

	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04

**summary**(anscombe)

x1		x2		x3	
Min.	: 4.0	Min.	: 4.0	Min.	: 4.0
1st Qu.:	6.5	1st Qu.:	6.5	1st Qu.:	6.5
Median	: 9.0	Median	: 9.0	Median	: 9.0
Mean	: 9.0	Mean	: 9.0	Mean	: 9.0
3rd Qu.:	11.5	3rd Qu.:	11.5	3rd Qu.:	11.5
Max.	:14.0	Max.	:14.0	Max.	:14.0

x4		y1		y2	
Min.	: 8	Min.	: 4.26	Min.	:3.10
1st Qu.:	8	1st Qu.:	6.32	1st Qu.:	6.70
Median	: 8	Median	: 7.58	Median	:8.14
Mean	: 9	Mean	: 7.50	Mean	:7.50
3rd Qu.:	8	3rd Qu.:	8.57	3rd Qu.:	8.95
Max.	:19	Max.	:10.84	Max.	:9.26

y3		y4	
Min.	: 5.39	Min.	: 5.25
1st Qu.:	6.25	1st Qu.:	6.17
Median	: 7.11	Median	: 7.04
Mean	: 7.50	Mean	: 7.50
3rd Qu.:	7.98	3rd Qu.:	8.19
Max.	:12.74	Max.	:12.50

```
cor(anscombe$x1, anscombe$y1)
```

```
[1] 0.8164
```

```
cor(anscombe$x2, anscombe$y2)
```

```
[1] 0.8162
```

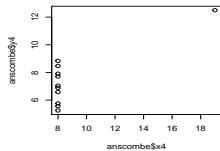
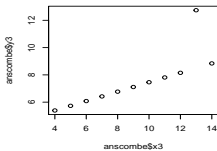
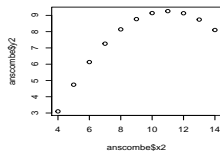
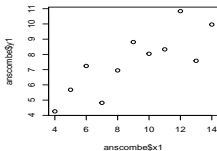
```
cor(anscombe$x3, anscombe$y3)
```

```
[1] 0.8163
```

```
cor(anscombe$x4, anscombe$y4)
```

```
[1] 0.8165
```

```
par(mfrow=c(2,2))
plot(anscombe$x1, anscombe$y1)
plot(anscombe$x2, anscombe$y2)
plot(anscombe$x3, anscombe$y3)
plot(anscombe$x4, anscombe$y4)
```



## 분산분석-이원배치법

- ▶ 반응변수에 대해서 두 종류의 요인의 영향을 조사하고자 할 때 사용하는 방법을 이원배치법(Two-way ANOVA)라 한다.
- ▶ 이원배치법은, 두 인자의 각 수준의 조합에 대해 반복이 있는 경우와 반복이 없는 두 가지 경우로 크게 나눌 수 있다.
- ▶ 이원배치법에서 반복의 여부는, 모형에 교호작용(Interaction)을 추가할 수 있는지 여부로 이어진다.

## 반복이 없는 이원배치법

- ▶ 2개의 인자를 A, B로 표시하고, 각각의 인자의 수준을 다음과 같이 표시하자.  
인자 A의 수준 :  $A_1, A_2, \dots, A_p$   
인자 B의 수준 :  $B_1, B_2, \dots, B_q$
- ▶ 이원배치법에서는 인자 A의 한 수준과 인자 B의 한 수준이 처리가 된다. 따라서, 처리의 총 개수는  $p \times q$ 개이다.
- ▶ 이원배치법에서 완전랜덤화계획은, 총  $pq$ 개의 실험의 순서를 랜덤하게 선택하여 시행하는 것이다.
- ▶ 인자 A의 수준을 고정시켜놓고, B의 수준을 변화시켜가며 차례대로 실험을 하면, 확률화의 원칙에 벗어나므로 이원배치법에 의한 실험이 아니다.



## 반복이 없는 이원배치법의 자료구조

인자A \ 인자B	$B_1$	...	$B_j$	...	$B_q$	평균
$A_1$	$y_{11}$	...	$y_{1j}$	...	$y_{1q}$	$\bar{y}_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$A_i$	$y_{i1}$	...	$y_{ij}$	...	$y_{iq}$	$\bar{y}_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$A_p$	$y_{p1}$	...	$y_{pj}$	...	$y_{pq}$	$\bar{y}_{p\cdot}$
평균	$\bar{y}_{\cdot 1}$	...	$\bar{y}_{\cdot j}$	...	$\bar{y}_{\cdot q}$	$\bar{y}_{\cdot\cdot}$

$$\bar{y}_{i\cdot} = \frac{1}{q} \sum_{j=1}^q y_{ij}, \quad \bar{y}_{\cdot j} = \frac{1}{p} \sum_{i=1}^p y_{ij}, \quad \bar{y}_{\cdot\cdot} = \frac{1}{pq} \sum_{i=1}^p \sum_{j=1}^q y_{ij}$$

반복이 없는 이원배치법의 자료구조

- ▶  $\bar{y}_{i\cdot}$  ( $i = 1, 2, \dots, p$ )는 각 인자 A의 각 수준 별 관측값의 평균이고,  $\bar{y}_{\cdot j}$  ( $j = 1, 2, \dots, q$ )는 각 인자 B의 각 수준 별 관측값의 평균이다. 그리고  $\bar{y}_{\cdot\cdot}$ 는 전체 관측값의 평균이다.

## 반복이 없는 이원배치법의 모집단 모형

- ▶ 반복이 없는 이원배치법에서는 인자 A와 B의 효과가 관측값에 영향을 주게 된다.
- ▶ 처리효과 전체의 평균을  $\mu$ 라 하고, 인자 A의  $i$ 번째 처리효과를  $\alpha_i$ , 인자 B의  $j$ 번째 처리효과를  $\beta_j$ 라 하면, 모형은 다음과 같이 나타난다. ( $i = 1, 2, \dots, p, j = 1, 2, \dots, q$ )

$$\begin{cases} Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \\ \epsilon_{ij} \sim iid \mathbf{N}(0, \sigma^2) \\ \sum_{i=1}^p \alpha_i = 0, \sum_{j=1}^q \beta_j = 0 \end{cases}$$

# 반복이 없는 이원배치법의 제곱합 분해와 처리효과의 유의성 검정

- ▶ 이원배치법에서 관심이 있는 가설은, 인자의 각 수준에 따라 처리효과의 차이가 있는가 하는 것으로서, 그 인자의 유의성을 검정하는 것이다.
- ▶ 각 인자의 유의성을 검정하기 위한 가설은 다음과 같이 나타난다.

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0, H_1 : \text{not } H_0$$

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0, H_1 : \text{not } H_0$$

- ▶ 가설의 검정법을 유도하기 위해, 총제곱합을 인자 A의 효과에 의한 제곱합과 인자 B의 효과에 의한 제곱합, 그리고 잔차제곱합으로 분해할 수 있다.

# 반복이 없는 이원배치법의 제곱합 분해와 처리효과의 유의성 검정

요인	제곱합	자유도	평균제곱	F값	유의확률
인자A	$SSA$	$p - 1$	$M SA$	$f_1 = M SA / M SE$	$P(F_1 > f_1)$
인자B	$SSB$	$q - 1$	$M SB$	$f_2 = M SB / M SE$	$P(F_2 > f_2)$
잔차	$SSE$	$(p - 1)(q - 1)$	$M SE$		
계	$SST$	$pq - 1$			

반복이 없는 이원배치법의 분산분석표

## 반복이 없는 이원배치법 : 예

- ▶ 무연탄에서 코크스를 제조하는 데 첨가하는 역청탄(A)를 총 5 종류( $A_1, \dots, A_5$ ) 선택하고, 타르피치의 첨가량(B)를 총 4수준 ( $B_1 : 4\%, B_2 : 6\%, B_3 : 8\%, B_4 : 10\%$ ) 선택하여 첨가한 후에 가열 성형하고, 코크스의 내압강도( $kg/cm^2$ )를 측정한 결과가 다음과 같다. 각 요인의 효과가 유의한지에 대해 이원배치법을 활용하여 유의수준 5%에서 검정해 보자.

	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$	평균
$B_1$	79	72	51	58	68	65.6
$B_2$	75	66	48	56	65	62
$B_3$	69	64	44	51	61	57.8
$B_4$	65	62	41	45	58	54.2
평균	72	66	46	52.5	63	총평균 59.9

코크스의 내압강도 자료

# 반복이 없는 이원배치법 : 예

## R Code

```
pres <- c(79, 72, 51, 58, 68, 75, 66, 48, 56, 65, 69, 64, 44, 51, 61, 65, 62, 41, 45, 58)
coal <- factor(rep(1:5, 4))
tar <- factor(rep(1:4, each=5))
cokes <- data.frame(coal, tar, pres)
anova(lm(pres~coal+tar, data=cokes))
```

Analysis of Variance Table

Response: pres

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
coal	4	1765	441	264.7	1.4e-11 ***
tar	3	369	123	73.8	5.3e-08 ***
Residuals	12	20	2		

---

Signif. codes:

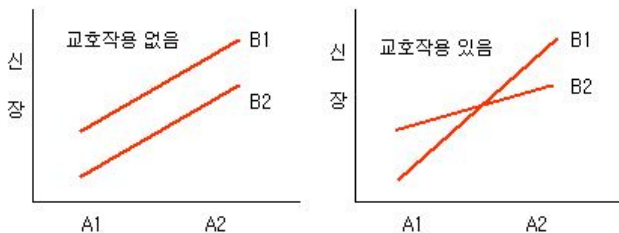
0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## 반복이 있는 이원배치법

- ▶ 이원배치법에서 2개의 인자를  $A(A_1, \dots, A_p)$ ,  $B(B_1, \dots, B_q)$ 로 표시하도록 하자.
- ▶ 반복이 없는 이원배치법에서는, 총  $pq$ 개의 처리에 대해 각각 1회씩의 실험을 하게 된다. 이 경우, 각 처리에 대한 변동이 존재하지 않게 된다.
- ▶ 반복이 있는 이원배치법이란, 각 처리에 대해 각각 반복수  $r$ 의 실험을 하는 방법이다. 이 경우, 각각의 처리에 대해서도 변동이 발생하며, 이를 통해 인자 수준의 조합에서 발생하는 효과를 분리하여 구할 수 있다.
- ▶ 인자수준의 조합에서 발생하는 효과를 교호작용(interaction)이라 하며, 인자  $A$ 의 효과가 인자  $B$ 의 수준에 따라 달라지는 모형에서 존재한다.

## 교호작용

- 서로 다른 보충제 A와 B를 각 인자의 수준에 따라 먹여 성장시킨 기니피그들의 평균 신장을 표시한 그래프를 그려본다고 하자. 교호작용 유무에 따라 그래프는 다음과 같이 나타날 것이다.



왼쪽 : 교호작용이 없는 경우, 오른쪽 : 교호작용이 있는 경우



## 반복이 있는 이원배치법의 모집단 모형

- ▶ 처리효과 전체의 평균을  $\mu$ 라 하고, 인자 A의  $i$ 번째 처리효과를  $\alpha_i$ , 인자 B의  $j$ 번째 처리효과를  $\beta_j$ , 그리고 인자 A의  $i$ 번째 수준과 인자 B의  $j$ 번째 수준의 교호작용을  $\gamma_{ij}$  라 하면, 모형은 다음과 같이 나타난다. ( $i = 1, 2, \dots, p, j = 1, 2, \dots, q, k = 1, 2, \dots, r$ )

$$\begin{cases} Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \\ \epsilon_{ijk} \sim iid \mathbf{N}(0, \sigma^2) \\ \sum_{i=1}^p \alpha_i = 0, \sum_{j=1}^q \beta_j = 0 \\ \sum_{i=1}^p \gamma_{ij} = 0, \sum_{j=1}^q \gamma_{ij} = 0 \end{cases}$$

# 반복이 있는 이원배치법의 제곱합 분해와 처리효과의 유의성 검정

- ▶ 반복이 있는 이원배치법에서 관심이 있는 가설은, 인자의 각 수준에 따른 처리효과가 유의한지와 함께, 인자의 조합으로 인해 발생하는 교호작용이 유의한지에 대한 것이다.
- ▶ 각 인자 및 교호작용의 유의성을 검정하기 위한 가설은 다음과 같이 나타난다.

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0, H_1 : \text{not } H_0$$

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0, H_1 : \text{not } H_0$$

$$H_0 : \gamma_{ij} = 0, (i = 1, 2, \cdots, p, j = 1, 2, \cdots, q) H_1 : \text{not } H_0$$

- ▶ 가설의 검정법을 유도하기 위해, 총제곱합을 인자 A의 효과에 의한 제곱합과 인자 B의 효과에 의한 제곱합, 교호작용에 의한 제곱합, 그리고 잔차제곱합으로 분해할 수 있다.

# 반복이 있는 이원배치법의 제곱합 분해와 처리효과의 유의성 검정

요인	제곱합	자유도	평균제곱	F값	유의확률
인자A	$SSA$	$p - 1$	$M SA$	$f_1 = M SA / M SE$	$P(F_1 > f_1)$
인자B	$SSB$	$q - 1$	$M SB$	$f_2 = M SB / M SE$	$P(F_2 > f_2)$
교호작용	$SSA \times B$	$(p - 1)(q - 1)$	$M SA \times B$	$f_3 = M SA \times B / M SE$	$P(F_3 > f_3)$
잔차	$SSE$	$pq(r - 1)$	$M SE$		
계	$SST$	$pqr - 1$			

반복이 있는 이원배치법의 분산분석표

## 반복이 있는 이원배치법 : 예

- ▶ 세 종류의 기계( $A_1, A_2, A_3$ )와 세 사람의 기능공( $B_1, B_2, B_3$ )이 제품 품질에 미치는 영향을 조사하고자 하여 2회 반복이 있는 이원배치법에 의해 생산성을 측정하였다. 이원배치법을 활용하여 인자에 의한 처리효과와 교호작용을 유의수준 5%에서 검정해보자.

	$B_1$	$B_2$	$B_3$	평균
$A_1$	9 14	14 16	19 22	15.67
$A_2$	13 16	18 26	14 18	17.5
$A_3$	11 12	11 17	15 16	13.67
평균	12.5	17	17.33	총평균 15.61

기계와 기능공에 따른 제품의 생산성 자료

## 반복이 있는 이원배치법 : 예

- ▶ 요인의 수준의 개수  $p = 3, q = 3$ , 반복수  $r = 2$
- ▶ 각 효과의 유의성을 검증하기 위한 가설은 다음과 같다.

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0, H_1 : \text{not } H_0$$

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0, H_1 : \text{not } H_0$$

$$H_0 : \gamma_{ij} = 0, (i = 1, 2, 3, j = 1, 2, 3), H_1 : \text{not } H_0$$

# 반복이 있는 이원배치법 : 예

## R Code

```
machine <- factor(rep(1:3, each=6))
technician <- factor(rep(1:3, each=2, 3))
quality <- c(9, 14, 14, 16, 19, 22, 13, 16, 18, 26, 14, 18, 11, 12, 11, 17, 15, 16)
product <- data.frame(machine, technician, quality)
anova(lm(quality~machine*technician, data=product))
```

Analysis of Variance Table

Response: quality

	Df	Sum Sq	Mean Sq	F value
machine	2	44.1	22.1	2.41
technician	2	87.4	43.7	4.77
machine:technician	4	74.2	18.6	2.02
Residuals	9	82.5	9.2	

Pr(>F)

machine	0.146
technician	0.039 *
machine:technician	0.174
Residuals	

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1