

# 컴퓨터응용통계

## 10. 분산분석법

최경미

# 실험계획법

- 실험방법을 미리 설계하고, 계획에 따라서 얻어진 자료를 분석하는 통계방법
- 일원배치분산분석법  
다중비교법  
독립표본 t-검정의 확장

# 일원배치 분산분석법

## InsectSprays

6종류의 살충제를 뿌려서 죽은 벌레수가 동일한가?

6종류의 살충제 효과는 동일한가?

어떤 살충제의 효과가 다른가?

독립변수 요인 (factor)

수준 (level)

반응변수 (response variable)

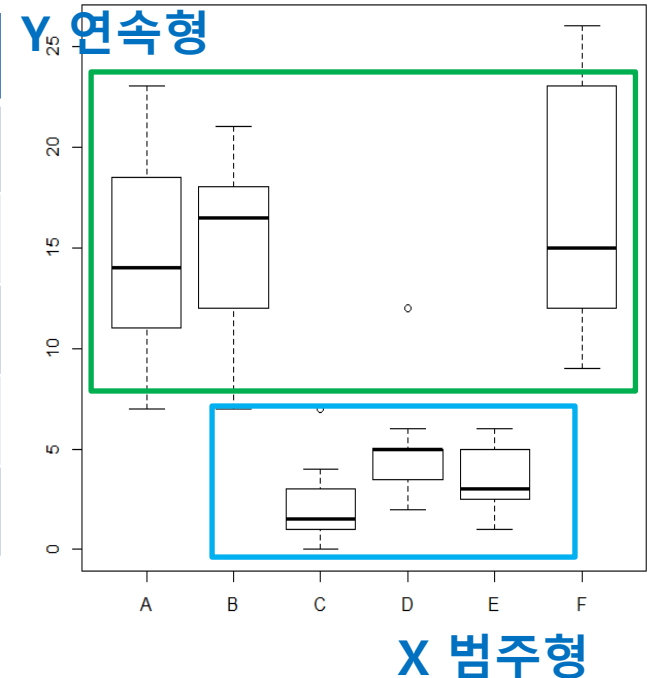
X = 살충제 종류

X의 값 A, B, C, D, E, F

Y = 죽은 벌레 수

A	10	7	20	14	14	12	10	23	17	20	14	13
B	11	17	21	11	16	14	17	17	19	21	7	13
C	0	1	7	2	3	1	2	1	3	0	1	4
D	3	5	12	6	4	3	5	5	5	5	2	4
E	3	5	3	5	3	6	1	1	3	2	6	4
F	11	9	15	22	15	16	13	10	26	26	24	13

$$6 \times 12 = 72$$



# 일원배치분산분석의 자료

72번의 실험을 랜덤한 순서로 실시함

반복 수준	1	2	...	$n_i = r$ 12
1	$y_{11}$	$y_{12}$	...	$y_{1n_1}$
2	$y_{21}$	$y_{22}$	...	$y_{2n_2}$
⋮	⋮	⋮	...	⋮
k	$y_{k1}$	$y_{k2}$	...	$y_{kn_k}$

데이터 엑셀 형식  
변수 한 개당 한 열

X	Y
1	$y_{11}$
1	$y_{12}$
⋮	⋮
1	$y_{1n_1}$
⋮	⋮
k	$y_{k1}$
k	$y_{k2}$
⋮	⋮
k	$y_{kn_k}$

# 모형

- $y_{ij} = \mu_i + \varepsilon_{ij}, i = 1, 2, \dots, k, j = 1, 2, \dots, n_i$

i: 처리집단  $i = 1, 2, \dots, k$

k: 집단수

j: 집단 내에서의 반복

$n_i$ : i번째 집단에서의 반복 수

n: 전체 관측수

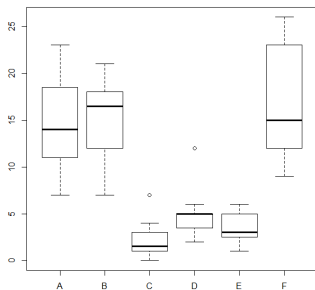
$y_{ij}$ : i번째 처리집단에서 j번째 관측값

$\mu_i$ : i번째 집단평균

- 가정: 오차  $\varepsilon_{ij} \sim iid N(0, \sigma^2)$

- 가설

$H_0: \mu_1 = \mu_2 = \dots = \mu_k$      $H_1$ : 모든  $\mu_i$ 가 같지는 않다.



$$y_{ij} = \mu_i + \varepsilon_{ij}$$

$$i = 1, \dots, 6, j = 1, \dots, 12$$

A

$$y_{11} = \mu_1 + \varepsilon_{11}$$

$$y_{12} = \mu_1 + \varepsilon_{12}$$

.....

$$y_{1,12} = \mu_1 + \varepsilon_{1,12}$$

B

$$y_{21} = \mu_2 + \varepsilon_{21}$$

$$y_{22} = \mu_2 + \varepsilon_{22}$$

.....

$$y_{2,12} = \mu_2 + \varepsilon_{2,12}$$

.....

F

$$y_{61} = \mu_6 + \varepsilon_{61}$$

$$y_{62} = \mu_6 + \varepsilon_{62}$$

.....

$$y_{6,12} = \mu_6 + \varepsilon_{6,12}$$

# 효과

총평균

$$\mu = \frac{1}{k} \sum_{i=1}^k \mu_i = \frac{1}{6} (\mu_1 + \mu_2 + \dots + \mu_6)$$

고정효과 (fixed effect)

$$\alpha_i = \mu_i - \mu, i = 1, 2, \dots, k$$

$$\sum_{i=1}^k \alpha_i = 0$$

모형

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$i = 1, 2, \dots, k, j = 1, 2, \dots, n_i$$

$$\varepsilon_{ij} \sim iid N(0, \sigma^2)$$

가설

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$$

$H_1$ : 적어도 한 개의  $\alpha_i$ 가 0이 아니다.

$$\begin{aligned} \sum_{i=1}^k \alpha_i &= \alpha_1 + \alpha_2 + \dots + \alpha_k \\ &= (\mu_1 - \mu) + (\mu_2 - \mu) + \dots + (\mu_k - \mu) \\ &= (\mu_1 + \mu_2 + \dots + \mu_k) - k\mu \\ &= k\mu - k\mu = 0 \end{aligned}$$

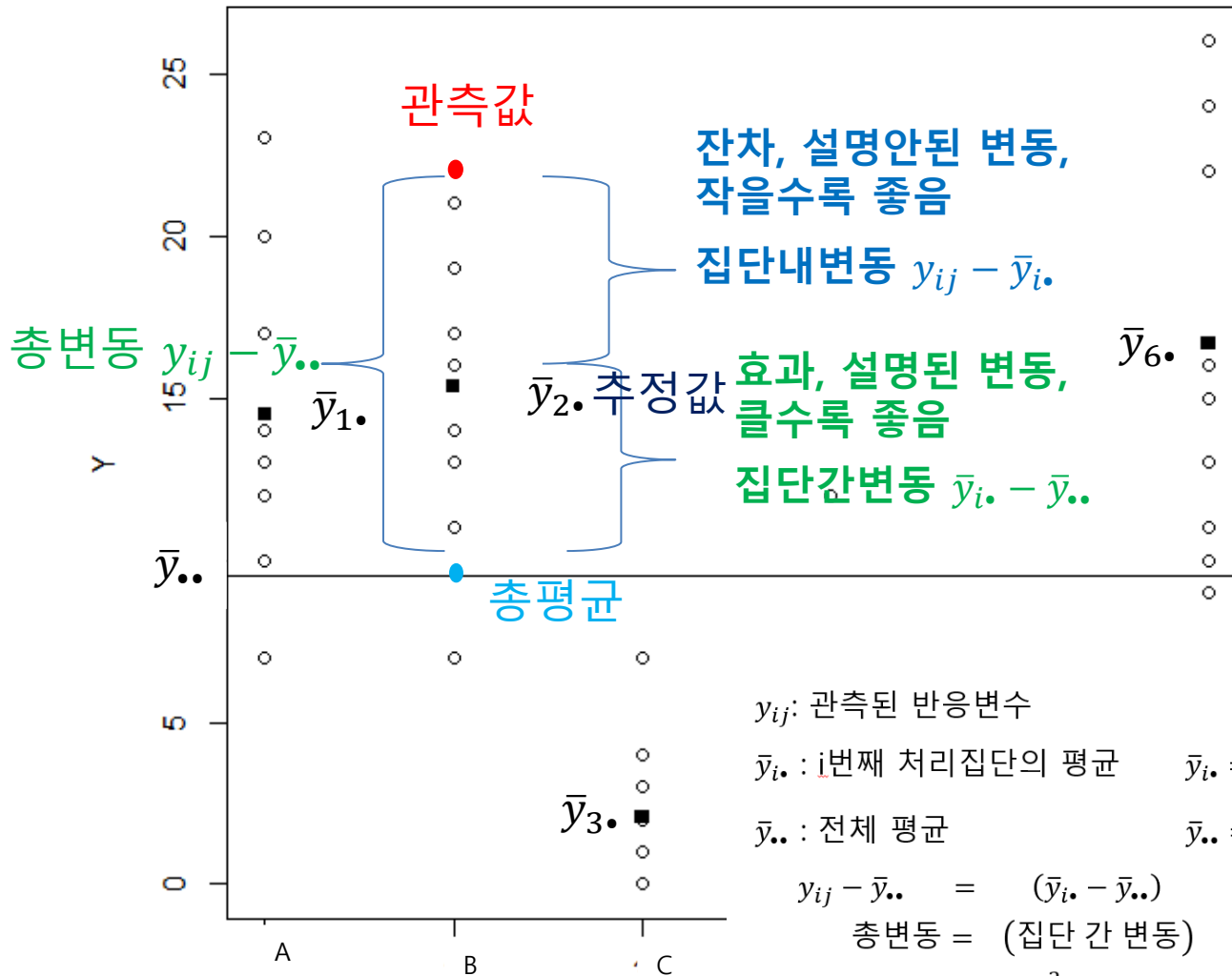
$$\begin{aligned} y_{ij} &= \mu_i + \varepsilon_{ij} \\ &= \mu + \alpha_i + \varepsilon_{ij} \end{aligned}$$

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$$

실험자가 미리 설계한  $k$  처리수준을 비교 분석한 결과를 실험에 포함된  $k$  수준에만 적용하는 경우를 고정효과모형(fixed effect model)이라고 부른다.

# 변동의 분해



$y_{ij}$ : 관측된 반응변수

$\bar{y}_{i\cdot}$ :  $i$ 번째 처리집단의 평균

$$\bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

$\bar{y}_{\cdot\cdot}$ : 전체 평균

$$\bar{y}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}$$

$$y_{ij} - \bar{y}_{\cdot\cdot} = (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot})$$

$$\text{총변동} = (\text{집단 간 변동}) + (\text{집단 내 변동})$$

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

$$SST = SStreatment + SSEerror$$

$$SST = SSbetween + SSwithin$$

# 다중비교법 (Multiple comparison)

- 모든 가능한  $\binom{k}{2}$  조합에 대한 이표본평균비교
- 각 쌍에 대해서 유의수준  $\alpha$ 를 적용하면, 전체 유의수준은  $\binom{k}{2}\alpha$ 가 됨.
- 총 유의수준 (familywise error rate) =  $\binom{6}{2}(0.05) = 0.75 >> 0.05$
- 총 유의수준 (familywise error rate) 0.05가 되도록,  
조정된 개별 유의수준  $\frac{0.05}{\binom{6}{2}} = 0.003333333$

- $\binom{k}{2}$  가설들  
 $H_0$ : 두 집단의 평균이 동일하다. ( $\mu_i = \mu_j, i \neq j$ )  
 $H_1$ :  $H_0$ 이 아니다. ( $\mu_i \neq \mu_j, i \neq j$ )

$$H_0: \mu_A = \mu_B \quad H_1: \mu_A \neq \mu_B$$

$$H_0: \mu_A = \mu_C \quad H_1: \mu_A \neq \mu_C$$

$$H_0: \mu_A = \mu_D \quad H_1: \mu_A \neq \mu_D$$

...

$$H_0: \mu_E = \mu_F \quad H_1: \mu_E \neq \mu_F$$

- 암세포 찾기 등에 응용됨

0.05	0.003333333
0.05	0.003333333
0.05	0.003333333
...	...
0.05	0.003333333
<b>0.75</b>	<b>0.05</b>



# 예제

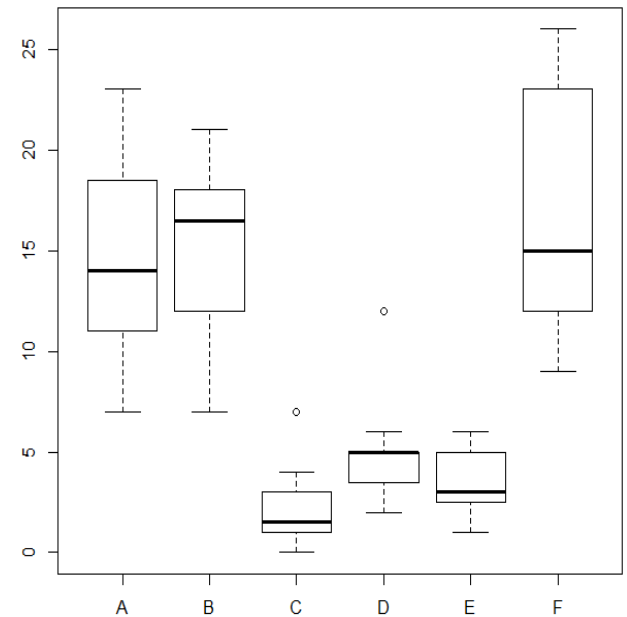
## 기술통계표

Spray 수준	평균(means)	표준편차(std)	반복수(r)	최소(Min)	최대(Max)
A	14.500000	4.719399	12	7	23
B	15.333333	4.271115	12	7	21
C	2.083333	1.975225	12	0	7
D	4.916667	2.503028	12	2	12
E	3.500000	1.732051	12	1	6
F	16.666667	6.213378	12	9	26

# 던컨(Duncan)의 다중비교법

\* Means with the same letter are not significantly different.

Groups *	Treatments	means
a	F	16.67
a	B	15.33
a	A	14.5
b	D	4.917
b	E	3.5
b	C	2.083



$$\mu_A = \mu_B = \mu_F$$

$$\mu_C = \mu_D = \mu_E$$

- `aggregate(count~spray,  
data=InsectSprays,mean)`

spray	count
F	16.666667
B	15.333333
A	14.500000
D	4.916667
E	3.500000
C	2.083333

> `TukeyHSD(aov.out)`  
 Tukey multiple comparisons of means  
 95% family-wise confidence level

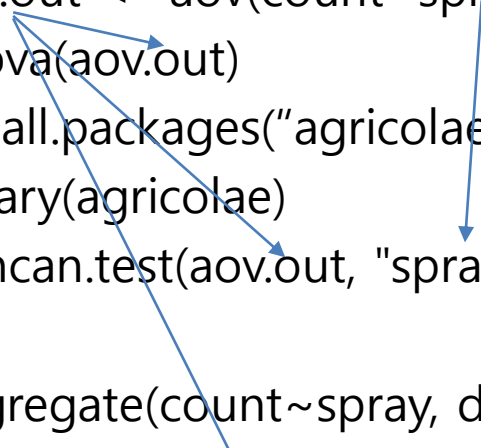
Fit: `aov(formula = count ~ spray, data = InsectSprays)`

		diff	lwr	upr	p adj
B-A	0.8333333	-3.866075	5.532742	0.9951810	> 0.05
C-A	-12.4166667	-17.116075	-7.717258	0.0000000	
D-A	-9.5833333	-14.282742	-4.883925	0.0000014	< 0.05*
E-A	-11.0000000	-15.699409	-6.300591	0.0000000	
F-A	2.1666667	-2.532742	6.866075	0.7542147	
C-B	-13.2500000	-17.949409	-8.550591	0.0000000	
D-B	-10.4166667	-15.116075	-5.717258	0.0000002	
E-B	-11.8333333	-16.532742	-7.133925	0.0000000	
F-B	1.3333333	-3.366075	6.032742	0.9603075	> 0.05
D-C	2.8333333	-1.866075	7.532742	0.4920707	
E-C	1.4166667	-3.282742	6.116075	0.9488669	> 0.05
F-C	14.5833333	9.883925	19.282742	0.0000000	
E-D	-1.4166667	-6.116075	3.282742	0.9488669	> 0.05
F-D	11.7500000	7.050591	16.449409	0.0000000	
F-E	13.1666667	8.467258	17.866075	0.0000000	

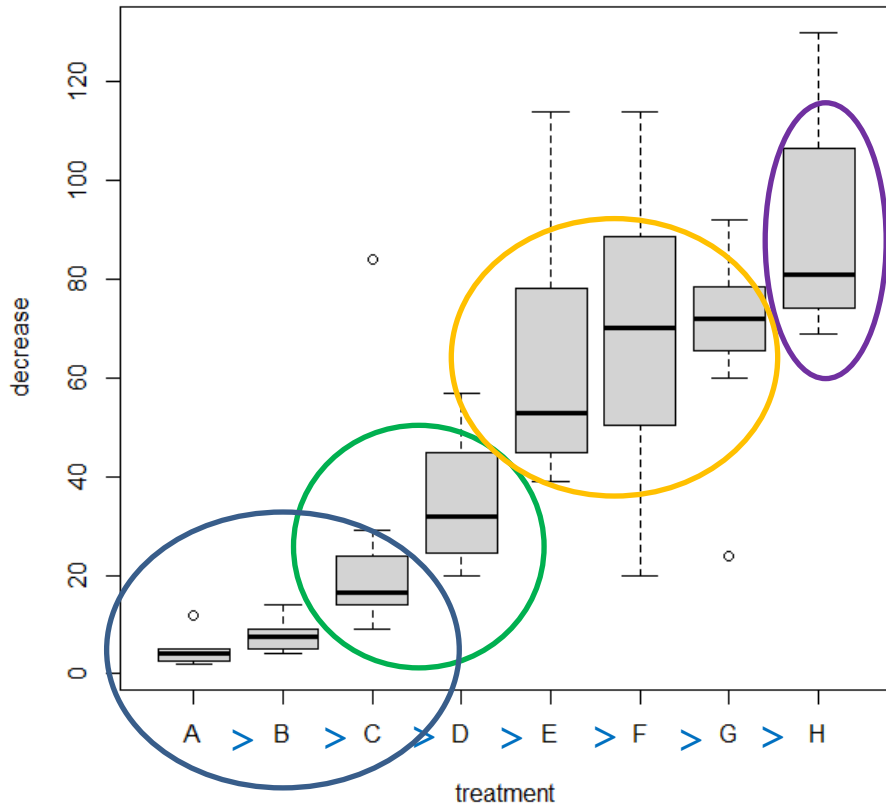
# R

```
aov.out <- aov(count~spray, data=InsectSprays)
anova(aov.out)
install.packages("agricolae")
library(agricolae)
duncan.test(aov.out, "spray", alpha=0.05, console=TRUE)

aggregate(count~spray, data=InsectSprays, mean)
TukeyHSD(aov.out)
```



# 혼자 풀기



```
> fit <- lm( decrease ~ treatment, data= OrchardSprays)
```

```
> anova(fit)
```

Analysis of Variance Table

Response: decrease

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatment	7	56160	8022.9	19.062	9.499e-13 ***
Residuals	56	23570	420.9		

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0

Pr(>F)

9.499e-13 \*\*\*

<0.05 농도별 자당  
감소 효과가 다르다.

```
> Install.packages("agricolae")
```

```
> library(agricolae)
```

```
> duncan.test(fit, "treatment", alpha=0.05, console=TRUE)
```

Study: fit ~ "treatment"

Duncan's new multiple range test  
for decrease

Mean Square Error: 420.8862

treatment, means

	decrease	std	r	Min	Max
A	4.625	3.204350	8	2	12
B	7.625	3.292307	8	4	14
C	25.250	24.429198	8	9	84
D	35.000	13.437687	8	20	57
E	63.125	26.909571	8	39	114
F	69.000	29.189039	8	20	114
G	68.500	20.142351	8	24	92
H	90.250	24.223660	8	69	130

Alpha: 0.05 ; DF Error: 56

Means with the same letter are not significantly different.

decrease groups

H	90.250	a
F	69.000	b
G	68.500	b
E	63.125	b
D	35.000	c
C	25.250	cd
B	7.625	d
A	4.625	d

$\alpha_B = \alpha_C$

$\alpha_C = \alpha_D$

$\alpha_B \neq \alpha_D$

## 석회황이 벌에게 미치는 영향 연구

R의 OrchardSprays는 8가지 농도(A>B>...>H)의 석회황유화액(lime sulphur emulsion)을 자당 용액(sucrose solution)에 섞은 후, 농도별로 8개의 벌 방에 발랐다. 여기에 100 마리 벌을 넣은 후 2시간 뒤에, 각 벌 방에서 줄어든 자당 용액이 얼마인지 측정하였다. 그림1은 농도별로 자당 감소량의 상자도표이고, 표1은 농도별 자당 감소량의 평균, 표준편차를 나타낸다. 던컨의 다중비교법을 이용하여, 어떤 농도에서 줄어든 자당 용액이 다른지 살펴보자. 유의수준 0.0.5를 사용한다.

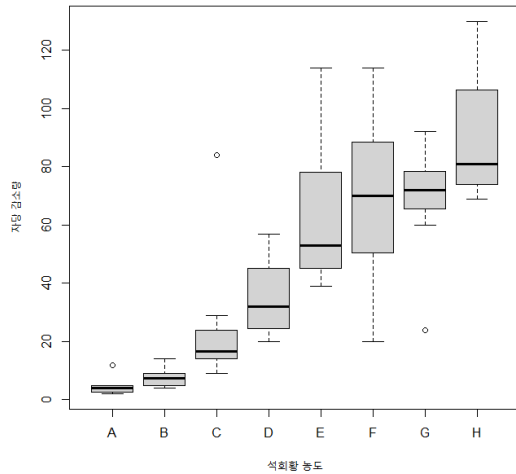


그림 1. 석회황 농도에 따른 자당 감소량

표1. 석회황 농도별 자당 감소액의 기술통계

농도	평균	표준편차	반복수	최소	최대
A		3.204			
B		3.292			
C		24.429			
D		13.438			
E		26.910			
F		29.189			
G		20.142			
H		24.224			

## 석회황이 벌에게 미치는 영향 연구

표2는 던컨의 다중비교법을 실시한 결과이다. 농도 ( )의 자당 감소량은 동일하다. ( )의 자당 감소량은 동일하다. ( )의 자당 감소량은 동일하다. ( )의 자당 감소량은 다르다. ( )의 자당 감소량은 다르다. ....

표3. 던컨의 다중비교법 결과.  
던컨 집단이 같은 글자이면,  
평균이 유의하게 다르지 않다.

농도	평균	던컨집단
H	90.250	a
F		
G		
E		
D		
C		
B		
A		

부록

### R 코드와 결과