



# R을 이용한 기초통계학

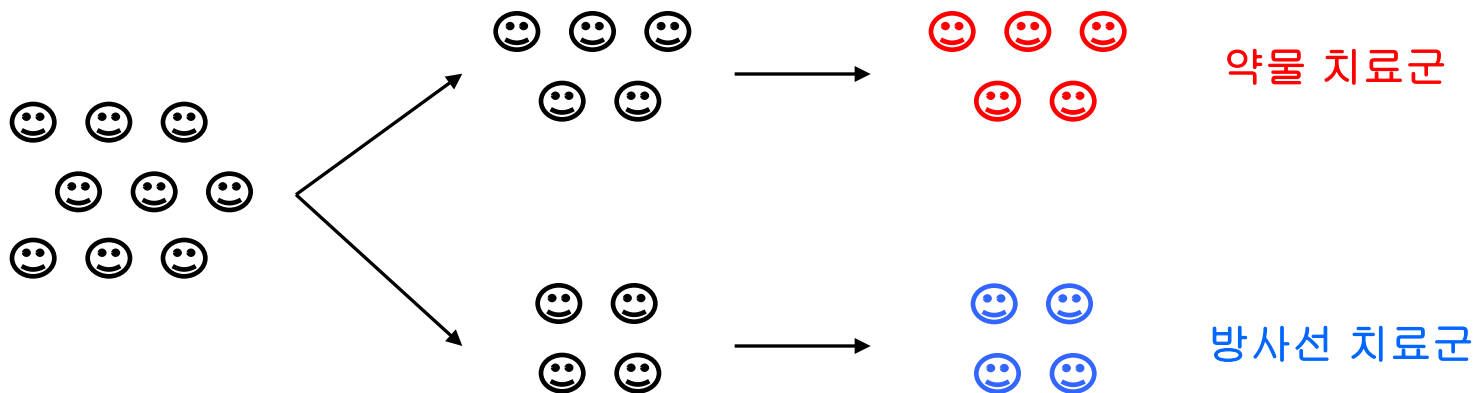
## 6강: 두 모집단의 비교와 세 개 이상 모집단의 비교를 위한 분산분석

수원대학교 데이터과학부 김진흠		서울대학교 보건환경연구소 이보라
------------------------	--	-------------------------

# 자료 추출 방법

## 1. 두 치료군의 자료가 독립적으로 추출된 경우

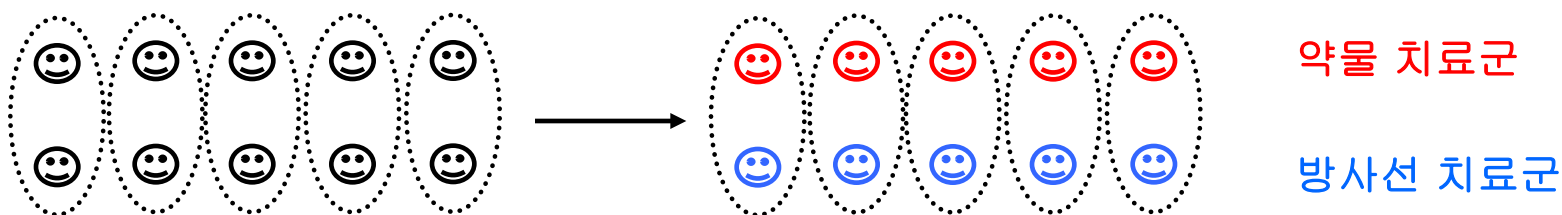
예: 암 환자에 대해 **약물 치료**와 **방사선 치료**을 비교할 때, 서로 독립적인 환자를 두 치료군으로 나누어 치료 방법을 적용시키는 경우



# 자료 추출 방법

## 2. 두 치료군의 자료가 짝을 지어 추출된 경우

예: 쌍둥이와 같이 서로 비슷한 체질의 환자를 짝을 지어 서로 다른 두 치료 방법을 비교하는 경우



# 독립 표본에서 평균 차이에 대한 구간추정

- Data

- $X_1, X_2, \dots, X_{n_x} \sim iid N(\mu_x, \sigma_x^2)$

- $Y_1, Y_2, \dots, Y_{n_y} \sim iid N(\mu_y, \sigma_y^2)$

- 가정:  $\sigma_x^2 = \sigma_y^2$

- 관심 모수:  $\mu_x - \mu_y$

- 점추정량:  $\bar{X} - \bar{Y}$

- 표준오차:  $SE(\bar{X} - \bar{Y}) = S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$

- $100(1 - \alpha)\%$  신뢰구간

- $(\bar{X} - \bar{Y}) \pm t_{\frac{\alpha}{2}}(n_x + n_y - 2) \times SE(\bar{X} - \bar{Y})$

# 합동(pooled)분산 추정량

- $$S_p^2 = \frac{(n_x-1)S_x^2 + (n_y-1)S_y^2}{n_x + n_y - 2}$$

- $$S_x^2 = \frac{1}{n_x-1} \sum_{i=1}^{n_x} (X_i - \bar{X})^2 : x \text{ 표본의 표본분산}$$

- $$S_y^2 = \frac{1}{n_y-1} \sum_{i=1}^{n_y} (Y_i - \bar{Y})^2 : y \text{ 표본의 표본분산}$$

# Is a weight-loss drug effective?

→ R code로 이동

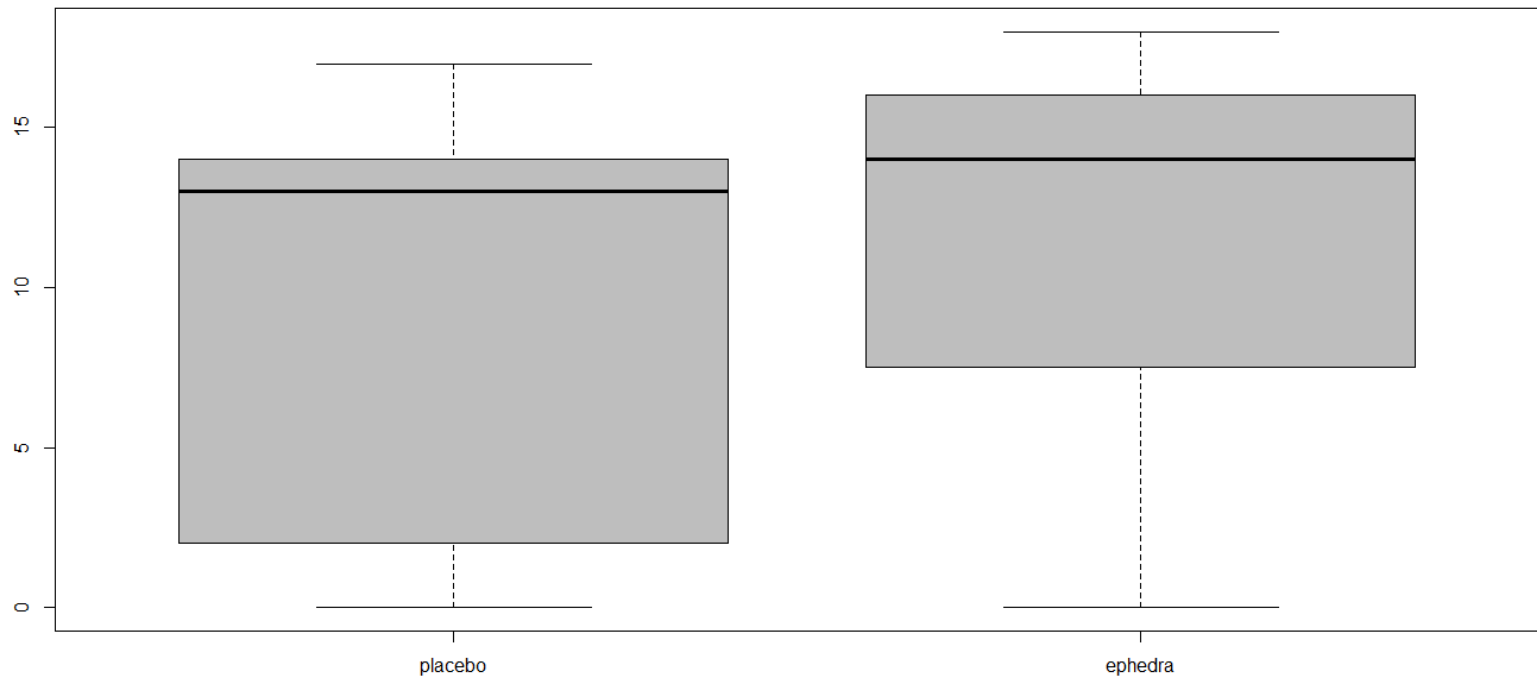
```
> x <- c(0,0,0,2,4,5,13,14,14,14,15,17,17)
> y <- c(0,6,7,8,11,13,15,16,16,16,17,18)
> t.test(x,y,var.equal = TRUE)
```

Two Sample t-test

```
data: x and y
t = -1.2071, df = 23, p-value = 0.2397
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.332489  2.191463
sample estimates:
mean of x mean of y
 8.846154 11.916667
```

```
> boxplot(list(placebo = x,ephedra = y),col = "grey")
```

# Boxplots for checking equal variance



# Is a weight-loss drug effective?

```
> var.test(x,y)
```

F test to compare two variances

```
data: x and y
F = 1.5802, num df = 12, denom df = 11, p-value = 0.4568
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4607529 5.2486187
sample estimates:
ratio of variances
 1.580204
```

```
> t.test(x,y)
```

Welch Two Sample t-test

```
data: x and y
t = -1.2185, df = 22.538, p-value = 0.2356
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.289271  2.148245
sample estimates:
mean of x mean of y
 8.846154 11.916667
```



# 짝표본에서 평균 차이에 대한 구간추정

- Data:  $(X_1, Y_1), \dots, (X_n, Y_n)$
- 관심 모수:  $\mu_D = \mu_x - \mu_y$
- 변환된 data
  - $D_1 = X_1 - Y_1, D_2 = X_2 - Y_2, \dots, D_n = X_n - Y_n$
  - 가정:  $D_i \sim iid N(\mu_D, \sigma^2)$
- $100(1 - \alpha)\%$  신뢰구간
  - $\bar{D} \pm t_{\frac{\alpha}{2}}(n - 1) \times \frac{S_D}{\sqrt{n}}$
  - $\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i, S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$

# Are the wear amounts of two types of shoes different? → R code로 이동

```
> library(MASS)
> data("shoes")
> names(shoes)
[1] "A" "B"
> with(shoes, t.test(A-B, conf.level = 0.9))
```

One Sample t-test

```
data:  A - B
t = -3.3489, df = 9, p-value = 0.008539
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 -0.6344264 -0.1855736
sample estimates:
mean of x
 -0.41
```

# Are the wear amounts of two types of shoes different?

```
> with(shoes,t.test(A,B,paired = TRUE,conf.level = 0.9))
```

Paired t-test

data: A and B

t = -3.3489, df = 9, p-value = 0.008539

alternative hypothesis: true difference in means is not equal to 0

90 percent confidence interval:

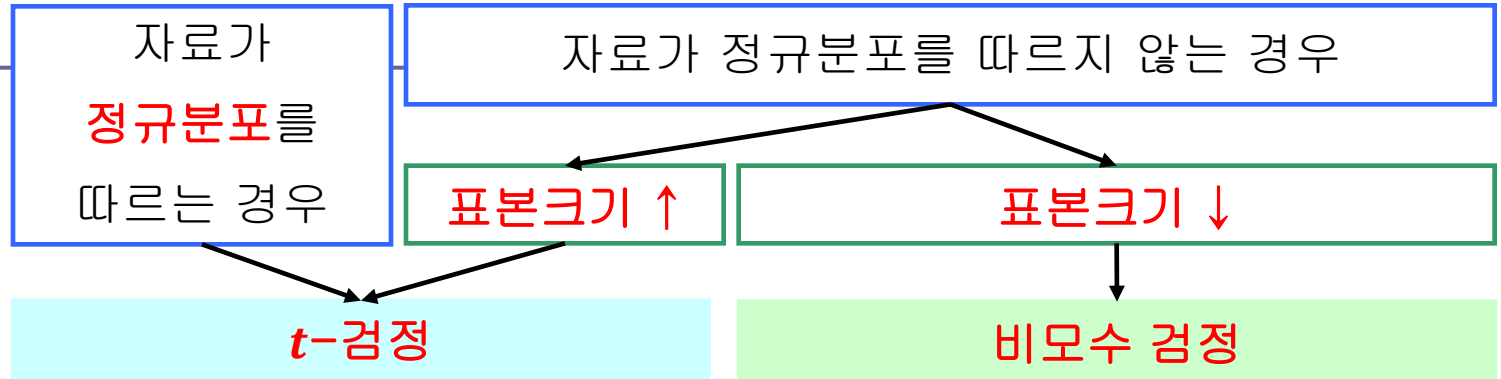
-0.6344264 -0.1855736

sample estimates:

mean of the differences

-0.41

# 검정방법의 선택



독립적(이)로 추출

<ul style="list-style-type: none"> <li>■ 이표본 <math>t</math>-검정</li> </ul>	<ul style="list-style-type: none"> <li>■ 월콕슨 순위합 검정</li> <li>■ 맨-위트니 검정 (Mann-Whitney test)</li> <li>■ Permutation test</li> </ul>
---	--

짝을 지어 추출

<ul style="list-style-type: none"> <li>■ 쌍체 <math>t</math>-검정</li> </ul>	<ul style="list-style-type: none"> <li>■ 월콕슨 부호 순위 검정</li> <li>■ 부호 검정 (Sign test)</li> <li>■ Permutation test</li> </ul>
--	---

# 이표본 $t$ -검정: 평균 차이에 대한 유의성 검정

- Hypothesis:  $H_0: \mu_x = \mu_y$  대  $H_1: \mu_x \neq \mu_y$
- Test statistics:  $T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$
- Null distribution
  - When **normal population**,  $T \sim t(n_x + n_y - 2)$
  - When **large sample**,  $T \sim N(0,1)$
- $P$ -값:  $P(|T| \geq |t_0| | H_0)$

# 실습: 이표본 $t$ -검정

- Data: Levels of **p24** in mg for two treatment groups

Amount		p24 level									
<b>300</b> mg		284	279	289	292	287	295	285	279	306	298
<b>600</b> mg		298	307	297	279	291	335	299	300	306	291

# 실습: 이표본 $t$ -검정

→ R code로 이동

```
> x <- c(284,279,289,292,287,295,285,279,306,298)
> y <- c(298,307,297,279,291,335,299,300,306,291)
> var.test(x,y)
```

F test to compare two variances

data: x and y

F = 0.34183, num df = 9, denom df = 9, p-value = 0.1256

alternative hypothesis: true ratio of variances is not equal to 1  
95 percent confidence interval:

0.0849059 1.3762082

sample estimates:

ratio of variances

0.3418306

# 실습: 이표본 $t$ -검정

```
> t.test(x,y,var.equal = T)
```

Two Sample t-test

```
data: x and y  
t = -2.034, df = 18, p-value = 0.05696  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -22.1584072  0.3584072  
sample estimates:  
mean of x mean of y  
   289.4    300.3
```



# 이표본 비모수 방법: 순위합 검정

- Hypothesis:  $H_0: m_x = m_y$  대  $H_1: m_x \neq m_y$
- Test statistic:  $W = \sum_{j=1}^{n_y} R_j$ 
  - $R_j$ :  $n_x$ 개의  $x$  표본과  $n_y$ 개의  $y$  표본을 섞은 표본에서  $y_j$ 의 순위
- Null distribution: Follows a **discrete** distribution
- $P$ -값:  $2 \times P(W \geq \max(w_0, 2M - w_0) | H_0)$ 
  - $M = \frac{n_y(n_x + n_y + 1)}{2}$

# 실습: 순위합 검정

- Data: Ten **checkout times** for two grocery checkers

Checker		Times									
Checker A		5.8	1.0	1.1	2.1	2.5	1.1	1.0	1.2	3.2	2.7
Checker B		1.5	2.7	6.6	4.6	1.1	1.2	5.7	3.2	1.2	1.3

# 실습: 순위합 검정

→ R code로 이동

```
> A <- c(5.8,1.0,1.1,2.1,2.5,1.1,1.0,1.2,3.2,2.7)
> B <- c(1.5,2.7,6.6,4.6,1.1,1.2,5.7,3.2,1.2,1.3)
> wilcox.test(A,B)
```

wilcoxon rank sum test with continuity correction

data: A and B

W = 34, p-value = 0.2394

alternative hypothesis: true location shift is not equal to 0

# 쌍체 $t$ -검정

- Hypothesis:  $H_0: \mu_x = \mu_y (\Leftrightarrow \mu_D = 0)$  대  
 $H_1: \mu_x \neq \mu_y (\Leftrightarrow \mu_D \neq 0)$
- Test statistic:  $T = \frac{\bar{D}}{\frac{s_D}{\sqrt{n}}}$
- $P$ -값:  $P(|T| \geq |t_0| | H_0), T \sim t(n - 1)$

# 실습: 쌍체 $t$ – 검정

- Data: Pre- and post-test **scores**

Test	score									
Pre-test	77	56	64	60	57	53	72	62	65	66
Post-test	88	74	83	68	58	50	67	64	74	60

# 실습: 쌍체 $t$ - 검정

→ R code로 이동

```
> x <- c(77,56,64,60,57,53,72,62,65,66)
> y <- c(88,74,83,68,58,50,67,64,74,60)
> t.test(x,y,paired = T)
```

Paired t-test

```
data: x and y
t = -1.8904, df = 9, p-value = 0.09128
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.862013  1.062013
sample estimates:
mean of the differences
          -5.4
```

# 실습: 부호순위검정

→ R code로 이동

```
> x <- c(77, 56, 64, 60, 57, 53, 72, 62, 65, 66)
> y <- c(88, 74, 83, 68, 58, 50, 67, 64, 74, 60)
> wilcox.test(x, y, paired = T)
```

wilcoxon signed rank exact test

data: x and y

V = 12, p-value = 0.1309

alternative hypothesis: true location shift is not equal to 0

```
> library(UsingR)
```

```
> confint(wilcox.test(x, y, paired = T, conf.int = 0.95))
```

(-13.00, 2.00) with 95 percent confidence

# Leukemia 예제

- 자료: **3가지 종류**의 leukemia환자 72명을 대상으로 수집한 Affymetrix microarray 자료
  - Acute myeloid leukemia(AML): 25명
  - B-cell acute lymphoblastic leukemia(B-cell ALL): 38명
  - T-cell acute lymphoblastic leukemia(T-cell ALL): 9명
- 목적: **특정 유전자의 발현 정도**가 leukemia의 종류에 따라 차이가 있는가?

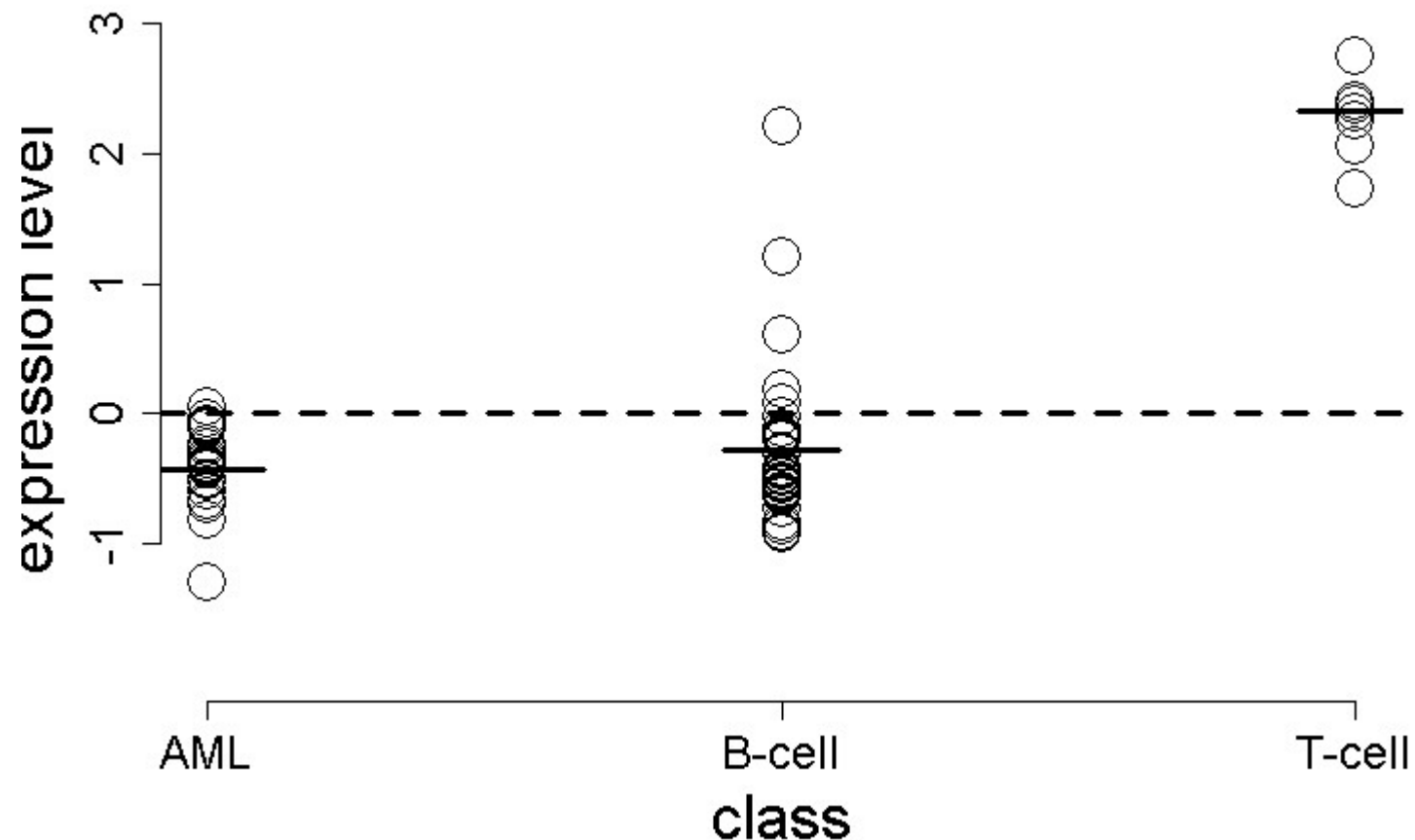


# 자료의 형태

- $y_{ij}$ :  $i$ 번째 그룹의  $j$ 번째 환자에서 얻은 특정 유전자 발현 수준의 관측값

	발현 수준	그룹평균	전체평균
AML	$y_{11}, y_{12}, \dots, y_{1n_1}$	$\bar{y}_{1.}$	$\bar{y}_{..}$
B-cell ALL	$y_{21}, y_{22}, \dots, y_{2n_2}$	$\bar{y}_{2.}$	
T-cell ALL	$y_{31}, y_{32}, \dots, y_{3n_3}$	$\bar{y}_{3.}$	

# Leukemia 예제: 자료의 요약



# 분산분석(ANOVA)

- 자료:  $y_{ij}$  =  $i$ 번째 그룹에서 얻은  $j$ 번째 관측값
- 모형:  $y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \tau_i + \varepsilon_{ij}$
- 가정:  $\varepsilon_{ij} \sim iid N(0, \sigma^2)$
- 귀무가설:  $H_0: \mu_1 = \mu_2 = \mu_3 \Leftrightarrow H_0: \tau_1 = \tau_2 = \tau_3 = 0$

# 분산분석표와 $F$ -검정

## ■ 분산분석표(ANOVA table)

요인 (source)	자유도 (df)	제곱합 (SS)	평균제곱 (MS)	$F$ -값
처리 (treatment)	$I - 1$	$SS_t = \sum_i n_i (\bar{y}_{i.} - \bar{y}_{..})^2$	$MS_t = SS_t / (I - 1)$	$F = \frac{MS_t}{MSE}$
잔차 (error)	$N - I$	$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$	$MSE = SSE / (N - I)$	
Total	$N - 1$	$SST = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$		

■  $P$ -값:  $P(F \geq f_0 | H_0), F \sim F(I - 1, N - I)$

# Leukemia 예제: 분산분석표와 $F$ -검정

요인 (source)	자유도 (df)	제곱합 (SS)	평균제곱 (MS)	$F$ -값
처리 (treatment)	2	55.684	27.842	125.43
잔차 (error)	69	15.316	0.222	
Total	71	71.000		

$P$ -값 =  $1.04514E-23 \ll 0.05 \rightarrow H_0$  기각

➔ ‘유전자 X03934의 발현 정도가 leukemia의 종류에 따라 차이가 있다’고 결론

# 실습: ANOVA

→ R code로 이동

- Data: Number of **calories** consumed by month

```
> may <- c(2166, 1568, 2233, 1882, 2019)
> sep <- c(2279, 2075, 2131, 2009, 1793)
> dec <- c(2226, 2154, 2583, 2010, 2190)
> ex5 <- stack(list(may=may, sep=sep, dec=dec))
```

# 실습: ANOVA

```
> ex5
      values ind
1      2166 may
2      1568 may
3      2233 may
4      1882 may
5      2019 may
6      2279 sep
7      2075 sep
8      2131 sep
9      2009 sep
10     1793 sep
11     2226 dec
12     2154 dec
13     2583 dec
14     2010 dec
15     2190 dec
```

# 실습: ANOVA

```
> oneway.test(values ~ ind, data = ex5, var.equal = T)
```

One-way analysis of means

data: values and ind

F = 1.7862, num df = 2, denom df = 12, p-value = 0.2094



# 실습: ANOVA table

요인 (source)	자유도 ( $df$ )	제 곱 합 (SS)	평균제 곱 (MS)	$F$ -값
처리 (treatment)	2	174664	87332	1.7862
잔차 (error)	12	586720	48893	
Total	14	761384		

# 실습: ANOVA

```
> res <- aov(values ~ ind, data = ex5)
```

```
> res
```

Call:

```
aov(formula = values ~ ind, data = ex5)
```

Terms:

	ind	Residuals
Sum of Squares	174664.1	586719.6
Deg. of Freedom	2	12

Residual standard error: 221.1183

Estimated effects may be unbalanced

```
> summary(res)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ind	2	174664	87332	1.786	0.209
Residuals	12	586720	48893		

# F-검정 후 분석

- 일단 ‘그룹 간의 차이’가 있다는 결론을 내리게 되면 추가적으로 ‘어떤 처리가 가장 효과가 있는가?’ 등과 같은 추가적인 의문 발생
- 예를 들면 ‘3가지 종류의 leukemia 간에 차이가 있다’는 결론을 내렸을 때,
  - ➔ 세 종류 leukemia가 모두 다른지  
또는 AML과 T-cell ALL이,  
또는 AML과 B-cell ALL이,  
또는 T-cell ALL과 B-cell ALL이 서로 다른가?

# 다중비교

- 모든 처리쌍에 대해 이표본 검정을 시행할 수 있으나, 유의수준  $\alpha$ 의 관리가 어려움
- 예를 들어 3개의 처리군을 비교할 때, 제1종 오류를 범할 확률(family-wise error rate, FWER) 즉, 3개 가설을 동시에 검정할 때 하나라도 제1종 오류가 발생할 확률은,

$$\begin{aligned} &P(\text{at least one Type I error}) \\ &= 1 - P(\text{no Type I error}) \\ &= 1 - (1 - \alpha)^3 \\ &\geq 1 - (1 - \alpha) = \alpha \end{aligned}$$

# 다중비교

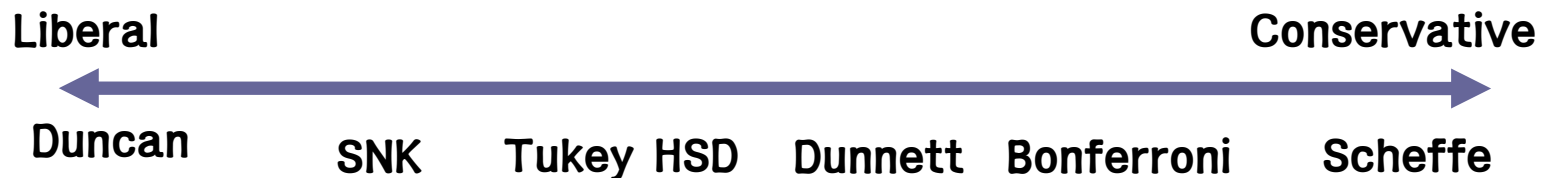
- 다중비교는 전체 제1종 오류 확률  $\alpha$  를 유지하면서 여러 처리군을 비교하는 방법

# 다중비교 방법들




- 쌍별 비교(pairwise comparison)
  - 한 번에 한 쌍의 모평균들이 같은지 다른지 검정
  - $\mu_1 = \mu_2, \mu_1 = \mu_3, \mu_2 = \mu_3$
  - Bonferroni, Scheffe, Tukey's HSD 방법
- 대조군(control)과 비교
  - 대조군의 모평균과 나머지 처리군의 모평균들 간의 비교
  - $\mu_1 = \mu_2, \mu_1 = \mu_3$
  - Dunnett 방법

# 다중비교 방법들

- 다단계 검정 (stepdown procedure)
  - **쌍별** 비교에서 5평균 차이, 4평균 차이 등과 같이 단계별로 검정
  - $\bar{X}_1 < \bar{X}_3 < \bar{X}_2$  일 때  $\mu_1 = \mu_2$ 를 먼저 검정 한 후 차이가 있으면 다음 단계로  $\mu_1 = \mu_3, \mu_2 = \mu_3$ 를 검정
  - **Duncan, SNK** (Student-Newman-Keuls) 방법



# Tukey의 다중비교: 모의실험 자료:

```
>  
> #  1) seed 고정 (재현성)  
> set.seed(123)  
> #  2) 각 그룹 데이터 생성  
> A <- rnorm(15, mean = 50, sd = 5)  
> B <- rnorm(15, mean = 60, sd = 5)  
> C <- rnorm(15, mean = 65, sd = 5)  
> #  3) stack() 사용하여 long-format 데이터프레임으로 변환  
> exABC <- stack(list(A = A, B = B, C = C))  
> exABC
```

	values	ind
1	47.19762	A
2	48.84911	A
3	57.79354	A
4	50.35254	A



# Tukey의 다중비교: 모의실험 자료:

```
> aggregate(values ~ ind, data = exABC, mean)
```

```
  ind  values
1   A 50.76192
2   B 58.76704
3   C 66.47645
```

```
>
```

```
> oneway.test(values ~ ind, data = exABC, var.equal = TRUE)
```

One-way analysis of means

data: values and ind

F = 41.783, num df = 2, denom df = 42, p-value = 1.028e-10

```
>
```

```
> res <- aov(values ~ ind, data = exABC)
```

```
> summary(res)
```

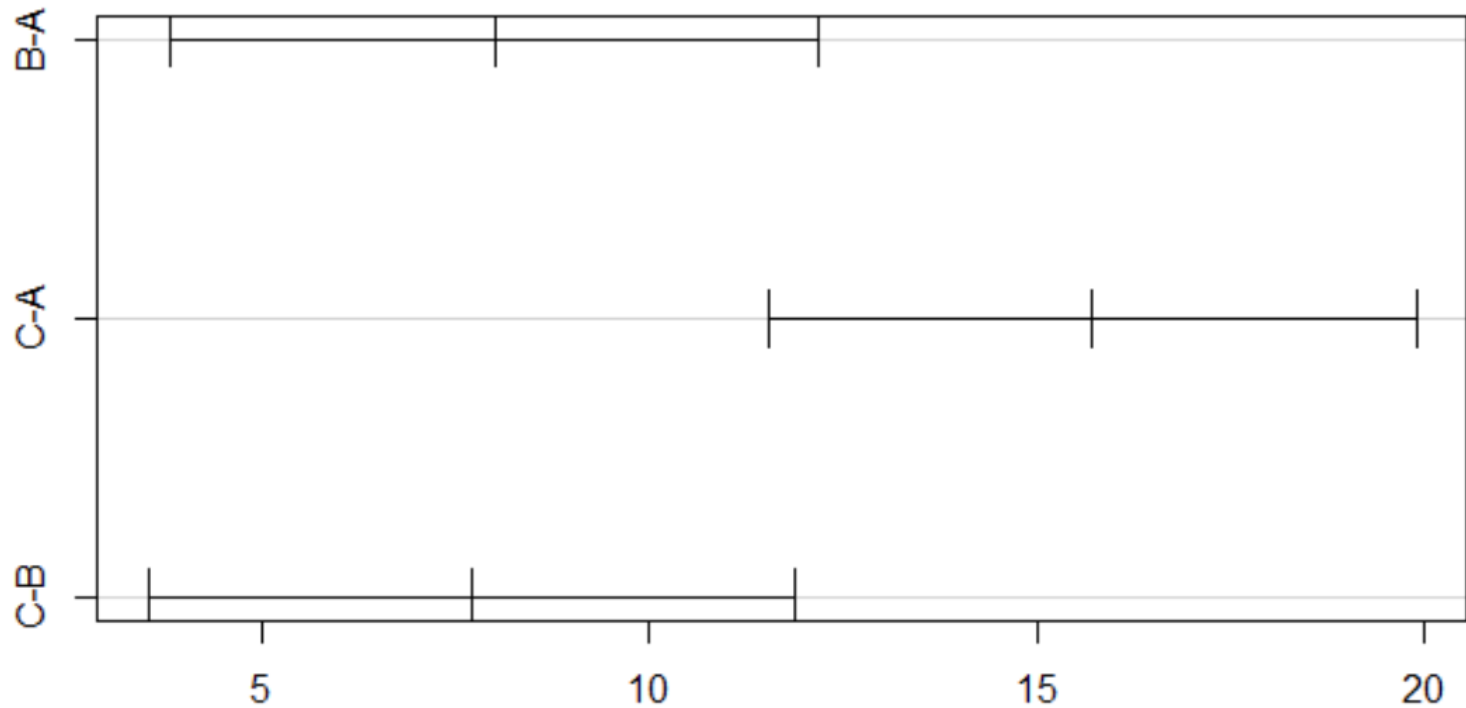
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
ind	2	1852	926.2	41.78	1.03e-10	***
Residuals	42	931	22.2			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Tukey의 다중비교: 모의실험 자료:

95% family-wise confidence level



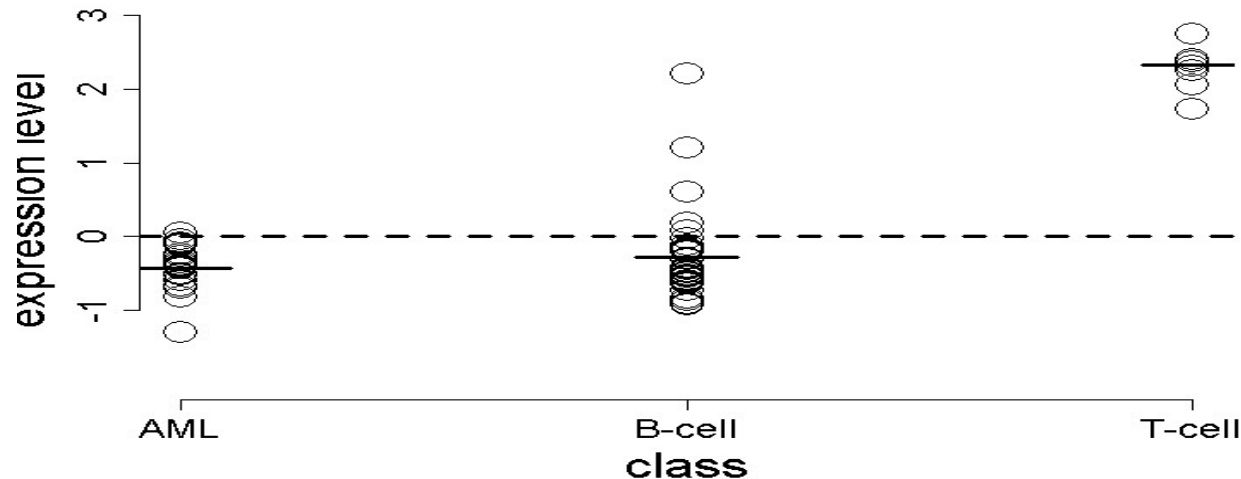
Differences in mean levels of ind

# Leukemia 예제: 다중비교 결과

AML  
-0.423

B-cell ALL  
-0.271

T-cell ALL  
2.319



# 실습: ANOVA

→ R code로 이동

```
> TukeyHSD(res)
```

Tukey multiple comparisons of means  
95% family-wise confidence level

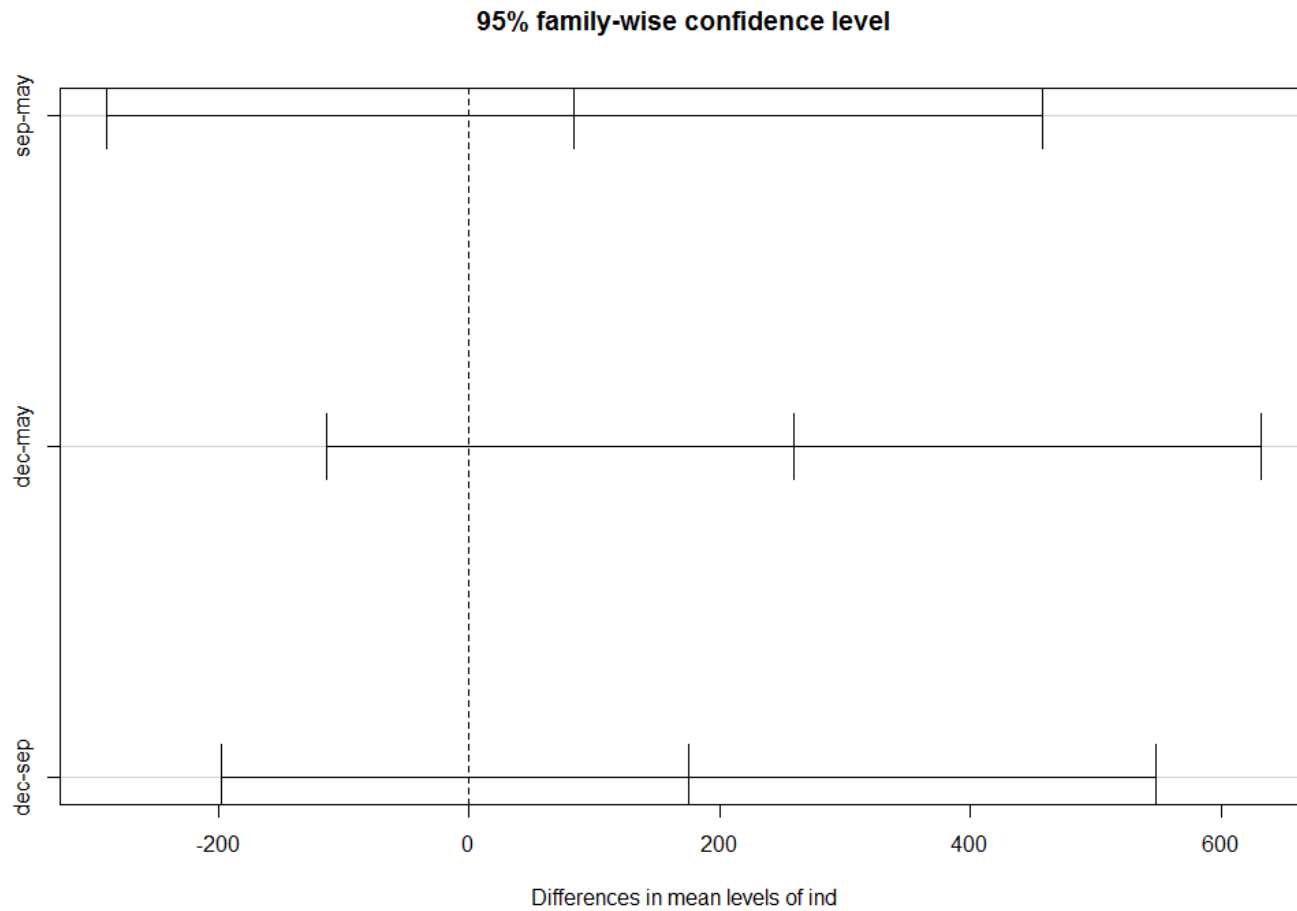
Fit: aov(formula = values ~ ind, data = ex5)

\$`ind`

	diff	lwr	upr	p adj
sep-may	83.8	-289.294	456.894	0.8231586
dec-may	259.0	-114.094	632.094	0.1949625
dec-sep	175.2	-197.894	548.294	0.4467189

```
> plot(TukeyHSD(res))
```

# Tukey의 다중비교



# 비모수 방법: 크루스칼-왈리스 검정

- 정규성 가정이 **만족되지 않는** 경우
- 방법:  $y_{ij}$  대신에 전체 관측값들 중에서 계산한  **$y_{ij}$** 의 순위  **$R_{ij}$** 를 사용하여 검정

	발현 수준	그룹별 평균순위	평균순위
AML	$R_{11}, R_{12}, \dots, R_{1n_1}$	$\bar{R}_{1\cdot} = \sum_j R_{1j} / n_1$	$\bar{R}_{..} = \frac{N+1}{2}$
B-cell ALL	$R_{21}, R_{22}, \dots, R_{2n_2}$	$\bar{R}_{2\cdot} = \sum_j R_{2j} / n_2$	
T-cell ALL	$R_{31}, R_{32}, \dots, R_{3n_3}$	$\bar{R}_{3\cdot} = \sum_j R_{3j} / n_3$	

# 크루스칼-왈리스 검정

- 귀무가설:  $H_0: \mu_1 = \mu_2 = \mu_3 \Leftrightarrow H_0: \tau_1 = \tau_2 = \tau_3 = 0$
- 검정통계량:  $H = \frac{12}{N(N+1)} \sum_{i=1}^3 n_i \left( \bar{R}_{i.} - \frac{N+1}{2} \right)^2$
- P-값:  $P(H \geq h_0 | H_0), H \sim \chi^2(2)$

# Leukemia 예제: 크루스칼-왈리스 검정

Kruskal-Wallis chi-squared = 23.0721

df = 2

$p$ -value = **9.771e-06** << 0.05

➔ ‘유전자 X03934의 발현 정도가 leukemia의 종류에 따라 차이가 있다’고 결론



# 실습: 크루스칼-왈리스 검정

## → R code로 이동

- Data: Test **score**s for three separate exams

```
> x <- c(63,64,95,64,60,85)
> y <- c(58,56,51,84,77)
> z <- c(85,79,59,89,80,71,43)
> ex6 <- stack(list(test1 = x, test2 = y, test3 = z))
> kruskal.test(values ~ ind, data = ex6)
```

Kruskal-wallis rank sum test

data: values by ind

Kruskal-wallis chi-squared = 1.7753, df = 2, p-value =  
0.4116

# 요약

## ■ 두 처리의 비교

- 독립적으로 추출된 경우: 이표본  $t$ -검정, 순위합 검정
- 짝을 지어 추출된 경우: 쌍체  $t$ -검정, 부호순위 검정

## ■ 세 개 이상 처리의 비교

- $F$ -검정
- 크루스칼-왈리스 검정