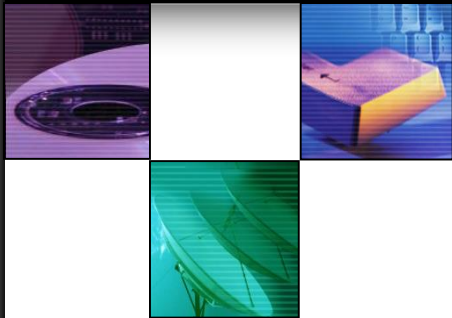
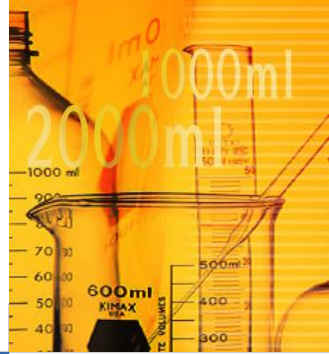


Chapter 6

가설 검정

Sejong Oh

Bio Information technology Lab.



- 개요
- 귀무가설과 대립가설
- 유의수준과 유의확률
- 카이제곱검정
- T-검정

- 통계적 가설검정 (statistical hypothesis test):
 - 통계적 추측의 하나로서, 모집단 실제의 값이 얼마가 된다는 주장과 관련해, 표본의 정보를 사용해서 가설의 합당성 여부를 판정하는 과정
 - 간단히 가설 검정(假說檢定)이라고 부르는 경우가 많음.
 - 결국은 표본을 이용해 모집단을 추정하는 작업이다

미국 성인 여성의 평균 신장은 180 cm 이상 이다
당뇨병에 A제품의 효능이 B제품보다 좋다.
민주당에 대한 지지율이 남성보다 여성이 더높다

- 가설 검정의 단계
 - 유의수준의 결정, 귀무가설과 대립가설 설정
 - 검정통계량의 설정
 - 기각역의 설정
 - 검정통계량 계산
 - 통계적인 의사결정

귀무가설과 대립가설

- 귀무가설 (null hypothesis)
 - 설정한 가설이 진실할 확률이 극히 적어 처음부터 버릴 것이 예상되는 가설. 현재의 통념
 - "차이가 없다", "관련이 없다 "
 - H_0 로 표기. 영가설이라고도 불리움
- 대립가설 (alternative hypothesis)
 - 연구를 통해 입증되기를 바라는 주장으로 연구가설이라고도 함
 - "차이가 있다", "관련이 있다 "
 - H_1 로 표기

우리가 원하는 것은 귀무가설을 기각하고 대립가설을 채택하는 것

귀무가설과 대립가설

- 예1) 상사와 부하직원의 친밀도가 업무 성과와 관련이 있는지를 알아보고 싶을 때
 - 귀무가설(H_0) : 친밀도와 업무 성과는 관련이 없다.
 - 대립가설(H_1) : 친밀도와 업무 성과는 관련이 있다.
- 예2) 치료법 A 와 치료법 B 는 효과의 차이가 있는가
 - 귀무가설(H_0) : 두 치료법은 효과의 차이가 없다.
 - 대립가설(H_1) : 두 치료법은 효과의 차이가 있다.

유의수준과 유의 확률

• 유의 수준

- 유의수준(significance level)은 통계적인 가설검정에서 사용되는 기준 값
- 표본에 의한 가설의 검정에서 귀무가설의 기각 여부를 결정하는데 사용하는 기준이 되는 확률
- 일반적으로 유의수준은 α 로 표시.
- 95%의 신뢰도를 기준으로 한다면 $(1-0.95)$ 인 0.05값이 유의수준 값 이 된다.
 - 95% 신뢰도 : 표본조사를 100번 했을 때 현재와 같은 결과가 나오는 경우가 95회라는 의미

• 유의확률(p-value)

- 귀무가설이 참임에도 이를 기각할 확률
- 우연히 해당 실험결과가 나왔을 확률
- P-value 가 낮을수록 통계적으로 유의미 하다.

유의수준과 유의 확률

- $p\text{-value} \leq \alpha$: 귀무가설 기각, 대립가설 채택
- $p\text{-value} > \alpha$: 귀무가설을 기각할 수 없다. (귀무가설이 맞다는 뜻이 아니라 대립가설을 채택할 충분한 근거가 없다는 뜻)

카이제곱 검정 (χ^2 -test)

- 관찰된 빈도가 기대되는 빈도와 의미있게 다른지의 여부를 검증하기 위해 사용되는 검증방법
- 범주형 형태의 자료에 대한 분석방법임 \Rightarrow 분할표

카이제곱 검정 (χ^2 -test)

- 비만과 당뇨병의 연관을 알아보자
- 분할표 (contingency table)
 - 독립변수와 종속변수가 모두 **범주형** 자료일 때
 - 2 x 2 분할표가 많으나 3 x 3, 2 x k 등 다양한 형태도 분석 가능
 - 당뇨환자 25명과 정상인 75명의 비만 여부 데이터 수집

	당뇨	정상	전체
비만	10 (40.0%)	10 (13.3%)	20 (20%)
정상 체중	15 (60.0%)	65 (86.7%)	80 (80%)
전체	25 (100%)	75 (100%)	100 (100%)

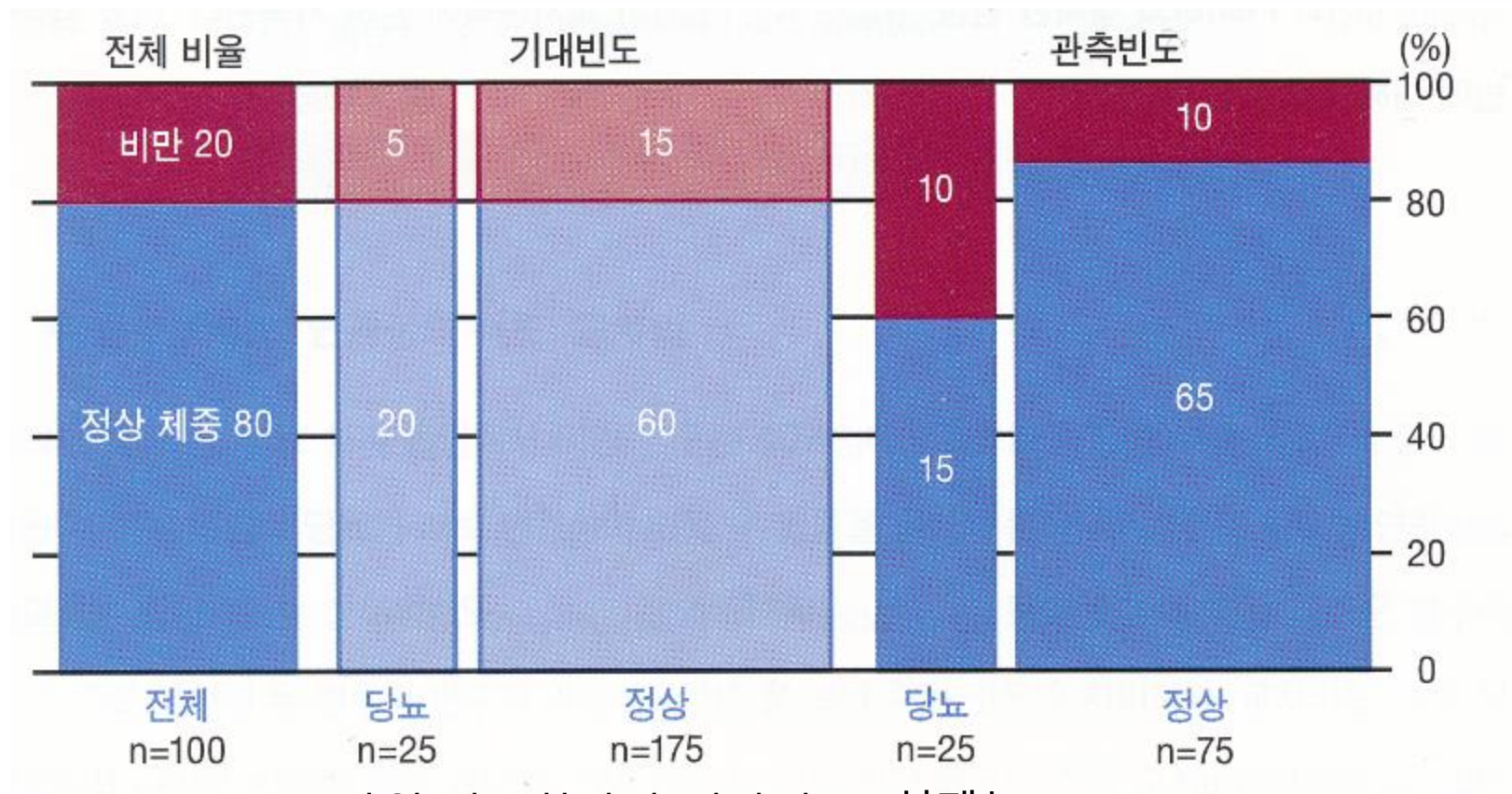


당뇨 여부	비만 여부
Y	Y
Y	N
N	N
N	Y
..	..

카이제곱 검정 (χ^2 -test)

- 아이디어

	당뇨	정상	전체
비만	10 (40.0%)	10 (13.3%)	20 (20%)
정상 체중	15 (60.0%)	65 (86.7%)	80 (80%)
전체	25 (100%)	75 (100%)	100 (100%)



(만일 당뇨병자와 정상인 사이에 차이가 없다면) (실제는 ...)

카이제곱 검정 (χ^2 -test)

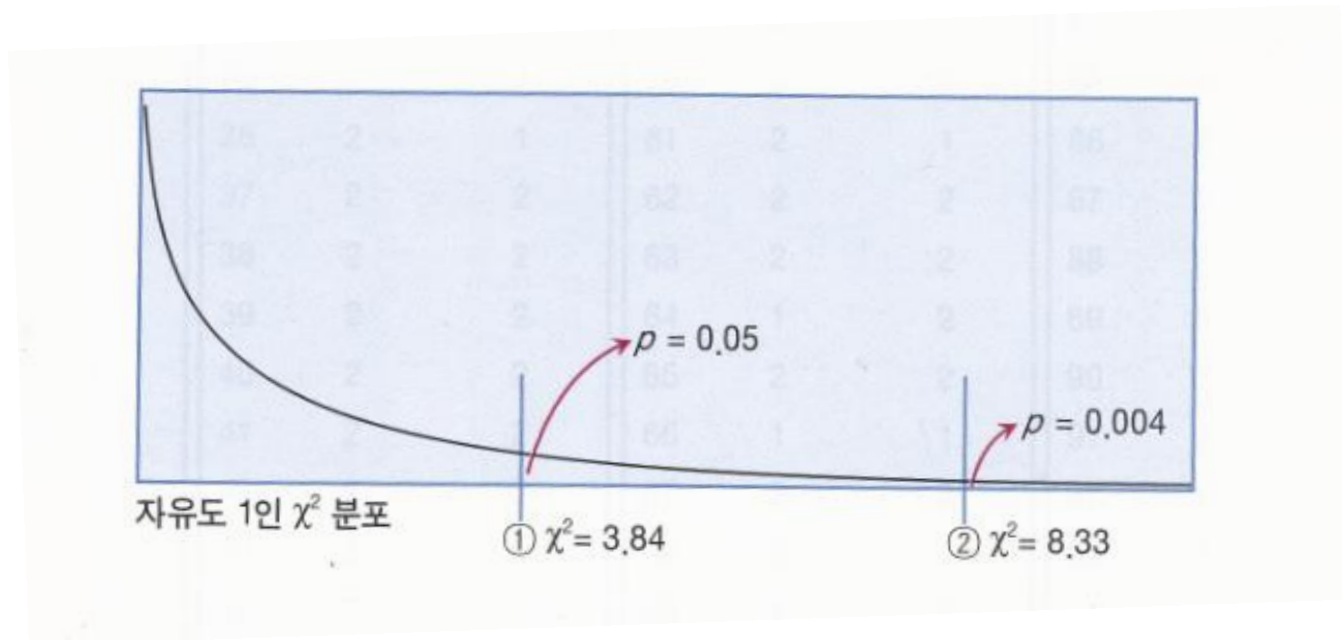
- 귀무가설(H_0) : 비만과 당뇨 여부는 관련이 없다.
- 대립가설(H_1) : 비만과 당뇨 여부는 관련이 있다.

귀무가설 H_0	두 변수는 연관성이 없다.
대립가설 H_1	두 변수는 연관성이 있다.

2×2 분할표에서 각 셀의 $\frac{(\text{관측빈도} - \text{기대빈도})^2}{\text{기대빈도}}$ 의 합은 자유도 1의 χ^2 (카이제곱) 분포를 따르는 것이 알려져 있다. 두 변수 사이의 연관성이 전혀 없다는 귀무가설하에서 검정통계량 χ^2 이 클수록 이러한 현상이 관찰될 가능성(p value)은 적어지게 된다.

$$\chi^2 = \sum \frac{(\text{관측빈도} - \text{기대빈도})^2}{\text{기대빈도}} = \frac{(+5)^2}{5} + \frac{(-5)^2}{20} + \frac{(-5)^2}{15} + \frac{(+5)^2}{60} = 8.33$$

카이제곱 검정 (χ^2 -test)



위와 같으므로 귀무가설 기각.
두 변수(당뇨,비만)는 연관성이 있다.

카이제곱 검정 (χ^2 -test)

- 카이제곱 검정의 조건
 - 기대빈도가 5 미만인 셀이 전체 셀의 20% 가 넘으면 카이제곱 분포를 따른다고 말할 수 없다
 - 2x2 분할표의 경우는 한 셀만 5 미만이어도 25%
 - 이런 경우는 카이제곱 검정 대신 Fisher's exact test 시행

카이제곱 검정 (χ^2 -test)

- [실습]

- 당뇨 환자 25명과 정상인 75명을 대상으로 비만 여부를 조사한 결과이다. (`5_chi_square_test.csv`)
- 당뇨와 비만 사이에 연관성이 있는지 검정하시오

카이제곱 검정 (χ^2 -test)

- R code

```
> # 자료 입력 및 관찰
```

```
> ds = read.csv("5_chi_square_test.csv")
```

```
> attach(ds)
```

```
> head(ds)
```

	obesity	diabetes
1	1	1
2	1	2
3	2	2
4	2	1
5	2	2
6	2	2

obesity (비만여부)

1 : 정상

2 : 비만

diabetes (당뇨여부)

1 : 정상

2 : 당뇨

카이제곱 검정 (χ^2 -test)

```
> # 분할표 작성 및 확인
```

```
> DMtable = xtabs(~obesity+diabetes)
```

```
> DMtable
```

```
      diabetes
obesity 1  2
      1 10 10
      2 15 65
```

```
> # 기대빈도 출력 뒤 chi-square test 및 Fisher's exact
test 시행
```

```
> chisq.test(DMtable)$expected
```

```
      diabetes
obesity 1  2
      1  5 15
      2 20 60
```

기대빈도 5 미만인
Cell 없음

카이제곱 검정 (χ^2 -test)

```
> chisq.test(DMtable)
```

```
      Pearson's Chi-squared test with Yates'  
continuity correction
```

```
data:  DMtable
```

```
X-squared = 6.75, df = 1, p-value = 0.009375
```

```
> fisher.test(DMtable)
```

```
      Fisher's Exact Test for Count Data
```

```
data:  DMtable
```

```
p-value = 0.007783
```

```
alternative hypothesis: true odds ratio is not equal to 1
```

```
95 percent confidence interval:
```

```
 1.331476 13.852900
```

```
sample estimates:
```

```
odds ratio
```

```
 4.255211
```

Fisher's exact test

- 기대빈도가 5 보다 작은 셀이 전체의 20% 이상인 경우 사용

	당뇨	정상	전체
비만	2 (40.0%)	2 (13.3%)	4 (20.0%)
정상 체중	3 (60.0%)	13 (86.7%)	16 (80.0%)
전체	5 (100%)	15 (100%)	20 (100%)

$$\text{Odds ratio} = \frac{2 \times 13}{2 \times 3} = 4.33$$

- 3개가 기대빈도 5 미만
- Fisher's exact test 가설

귀무가설 H_0	두 변수는 연관성이 없다.
대립가설 H_1	두 변수는 연관성이 있다.

Fisher's exact test

- 당뇨병 환자 5명과 정상인 15명을 대상으로 비만 여부를 조사한 결과이다.
- 당뇨병과 비만 사이에 연관성이 있는지 검정하시오

> # 자료 입력 및 관찰

> ds = read.csv("5_fisher_exact_test.csv")

> attach(ds)

> head(ds)

	obesity	diabetes
1	2	2
2	2	2
3	1	1
4	2	2
5	2	2
6	1	2

>

```
> # 분할표 작성 및 확인
```

```
> DMtable = xtabs(~obesity+diabetes)
```

```
> DMtable
```

	diabetes	
obesity	1	2
1	2	2
2	3	13

```
>
```

```
> # 기대빈도 출력 뒤 chi-square test 및 Fisher's exact  
test 시행
```

```
> chisq.test(DMtable)$expected
```

	diabetes	
obesity	1	2
1	1	3
2	4	12

Warning message:

```
In chisq.test(DMtable) : Chi-squared approximation may  
be incorrect
```

```
> chisq.test(DMtable)
```

```
      Pearson's Chi-squared test with Yates' continuity  
correction
```

```
data:  DMtable
```

```
X-squared = 0.4167, df = 1, p-value = 0.5186
```

```
Warning message:
```

```
In chisq.test(DMtable) : Chi-squared approximation may be  
incorrect
```

```
> fisher.test(DMtable)
```

```
      Fisher's Exact Test for Count Data
```

```
data:  DMtable
```

```
p-value = 0.2487
```

```
alternative hypothesis: true odds ratio is not equal to 1
```

```
95 percent confidence interval:
```

```
0.2088683 77.8897313
```

```
sample estimates:
```

```
odds ratio
```

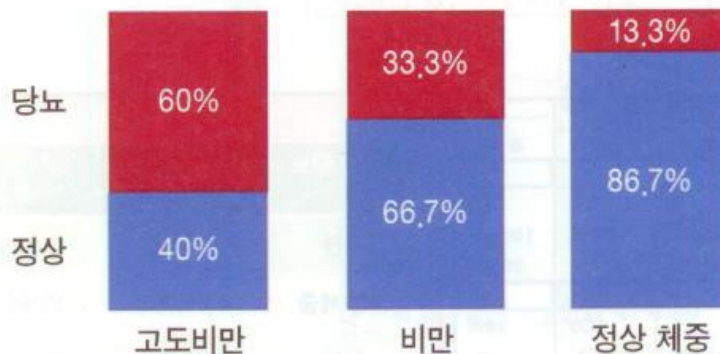
```
3.955176
```

비만 유무는 당뇨병 유무와
연관성을 보이지 못했다

Note.

- 2x k 분할표에서 독립변수가 세가지 이상의 범주로 구분되는 순위척도의 경우, 독립변수가 증가함에 따라 종속변수의 비율이 증가하는 혹은 감소하는 경향이 있는지 경향 분석이 가능하다
- 이경우 Cochran-Amitage test 가 이용된다.
(R 에서 **prop.trend.test()** 함수 이용)

	고도비만	비만	정상 체중	계
당뇨	3 (60.0%)	1 (33.3%)	2 (13.3%)	6 (26.1%)
정상	2 (40.0%)	2 (66.7%)	13 (86.7%)	17 (73.9%)
계	5 (100%)	3 (100%)	15 (100%)	23 (100%)



Note. 이 테스트는 카이제곱 분포를 이용하지 않으므로 기대빈도 5 미만 셀에 대한 제약 없음

귀무가설 H_0	비만도와 관계없이 당뇨의 비율은 일정하다.
대립가설 H_1	비만도가 증가할수록 당뇨의 비율은 증가/감소 추세가 있다.

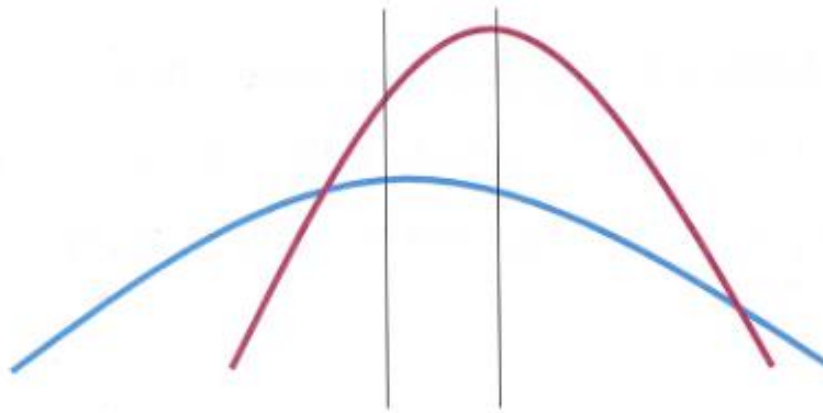
T-검정(t-test)

- 결과변수가 연속형인 독립된 두 군의 크기(평균)를 비교할 때
 - 독립표본 T 검정
 - Mann-Whitney test
- Student's t-test 라고도 불리움
- 독립표본 T 검정의 필요조건
 - Case, control 과 같이 두개의 군(group)을 비교
 - 독립성, 정규성, 등분산성 이 만족되어야 한다

독립표본 : 두 그룹의 데이터가 서로 다른 대상으로 부터 얻어짐
대응표본 : 두 그룹의 데이터가 같은 대상으로 부터 얻어짐
(예: 환자 100명의 약 복용전 혈당과 복용후 혈당)

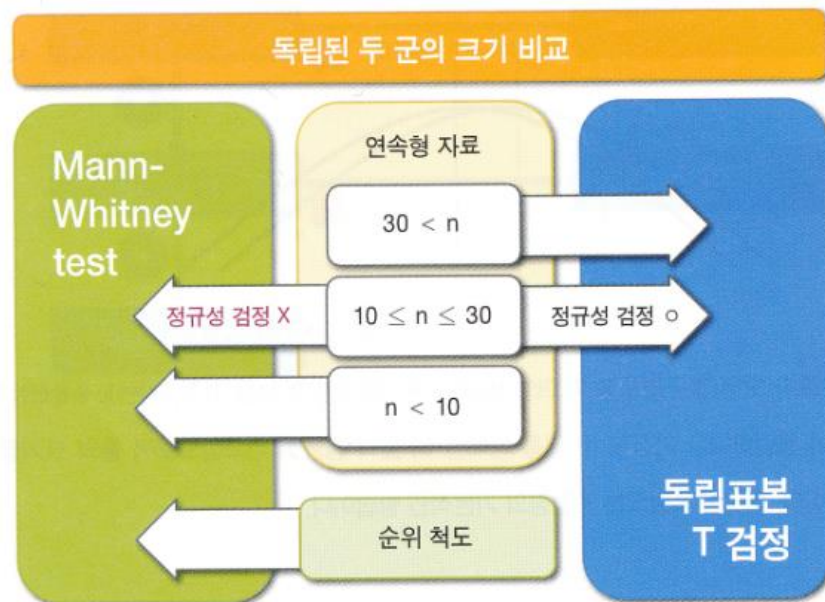
독립표본 T 검정

- 두 군의 분포가 정규분포를 따르지만 분산이 서로 다를때
 - 자유도를 수정하여 근사적인 방법으로 T 검정 실시 가능
 - 통계 프로그램에서 알아서 해 준다 (따라서 일반적으로는 신경쓰지 않는다)



독립표본 T 검정

- 정규성 검정
 - 군의 표본수가 10~30 인 경우 실시
 - Kolmogorov-Smirnov test, Shapiro-Wilk test 등



- 정규성 검정의 가설

귀무가설 H_0	자료는 정규분포를 따른다.
대립가설 H_1	자료는 정규분포를 따르지 않는다.



독립표본 T 검정

- 독립표본 T 검정의 가설 설정

귀무가설 H_0	두 군의 평균은 같다.
대립가설 H_1	두 군의 평균은 같지 않다.

독립표본 T 검정은 두 군의 평균이 같다는 귀무가설을 전제로 자료를 분석한다. 귀무가설하에서 주어진 자료가 해석될 가능성이 5% 미만인 경우, 즉 $p < 0.05$ 인 경우 귀무가설을 기각하고 '두 군의 평균은 같지 않다'는 대립가설을 채택한다. 그리고 독립표본 T 검정에서는 각 군의 평균을 통해 어느 군이 어느 군보다 얼마만큼 크기가 큰지 그 차이도 함께 제시할 수 있다.

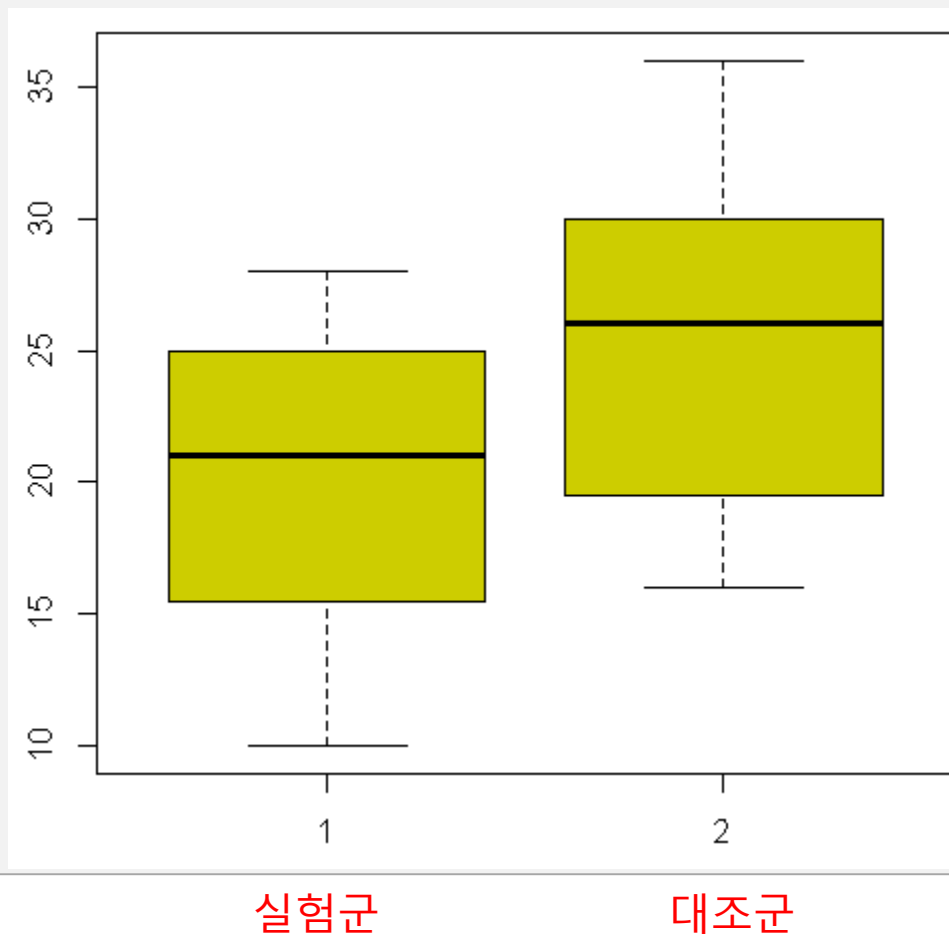
독립표본 T 검정

- 새로운 항우울제의 치료효과를 증명하기 위해 40명의 우울증 환자를 군당 20명씩 실험군과 대조군으로 배정하여 실험군에는 항우울제를, 대조군에는 위약을 투여하였다.
- 1개월뒤 측정한 우울증 점수를 통해서 두 군의 치료효과의 크기를 비교하고자 한다.
- 두군 사이에 우울증 점수의 차이가 있는지 검정하시오

데이터파일 : 2_independent_t_test.csv

독립표본 T 검정

```
> # 자료 입력 및 관찰  
> ds = read.csv("2_independent_t_test.csv")  
> head(ds)  
> boxplot(score~group, col="yellow3", data=ds)  
>
```



독립표본 T 검정

```
> # 정규성 검정
```

```
> shapiro.test(ds[ds$group==1,3])
```

Shapiro-Wilk normality test

```
data: ds[ds$group == 1, 3]
```

```
W = 0.93475, p-value = 0.1905
```

정규성 만족

```
> shapiro.test(ds[ds$group==2,3])
```

Shapiro-Wilk normality test

```
data: ds[ds$group == 2, 3]
```

```
W = 0.95115, p-value = 0.3848
```

정규성 만족

```
>
```

독립표본 T 검정

```
> # 등분산 검정 및 독립표본 T 검정 시행  
> var.test(score~group, data= ds)
```

F test to compare two variances

data: score by group

F = 0.814, num df = 19, denom df = 19, p-value = 0.6583

alternative hypothesis: true ratio of variances is not
equal to 1

95 percent confidence interval:

0.3221986 2.0565791

sample estimates:

ratio of variances

0.814019

귀무가설 기각 못함
-> 두 그룹의 분산은 같다

독립표본 T 검정

```
> t.test(score~group, var.equal=TRUE, data= ds)
```

Two Sample t-test

data: score by group

t = -2.7164, df = 38, p-value = 0.00988

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

95 percent confidence interval:

-8.726292 -1.273708

sample estimates:

mean in group 1	mean in group 2
20.25	25.25

실험군

대조군

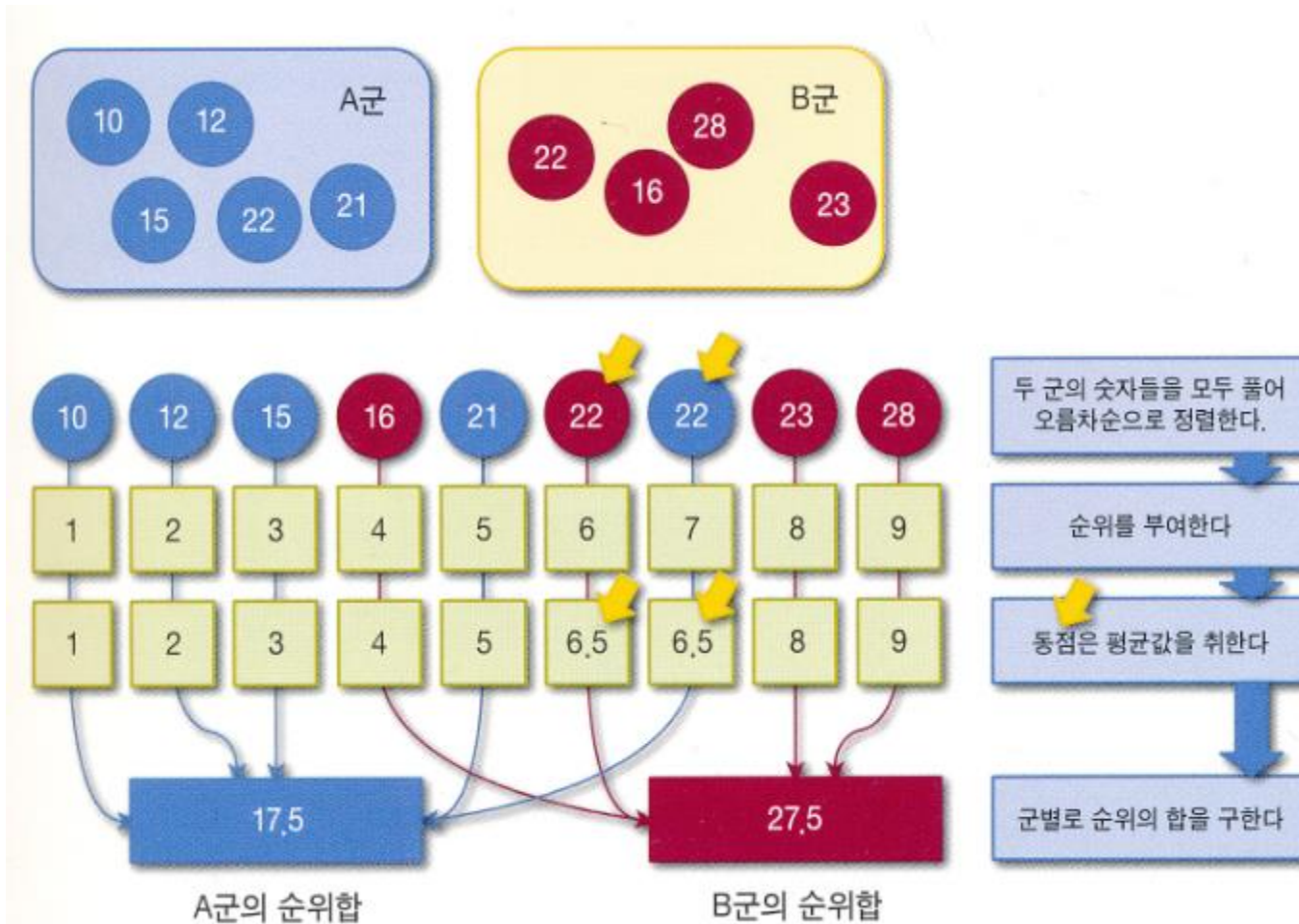
두 군의 우울증 지수에 유의한 차이가 있으며, 차이의 정도는 5 이다

Wilcoxon rank sum test

- 정규분포를 따른다고 가정할수 없는 두 군은 평균을 통해 크기를 비교할수 없다
- 비모수적 방법 필요
- Mann-Whitney test, Wilcoxon rank sum test
 - 두가지 모두 순위의 합을 기본으로 하며 동일한 검정 결과를 도출한다
 - 순위합 검정이란 두 군의 자료를 크기순으로 정렬하여 순위를 부여한 뒤, 그 순위의 합을 군별로 구하여 두군의 순위합의 크기가 통계적으로 차이가 있는지를 본다
 - **순위척도**도 연속형 변수는 아니지만 순위정보가 있으므로 순위합 검정으로 크기를 비교하는 것이 가능하다

Wilcoxon rank sum test

- 순위합 검정의 아이디어



Wilcoxon rank sum test

- 순위합 검정의 가설 설정

귀무가설 H_0	두 군의 크기가 같다.
대립가설 H_1	두 군의 크기가 같지 않다.

순위합 검정은 일반적으로 독립표본 T 검정에 비해 검정력이 낮으며, 순위만 비교한 것이기 때문에 두 군의 크기의 차이(평균의 차이)를 언급할 수 없는 단점이 있지만, 정규분포에 대한 가정을 하지 않기 때문에 크기 순서가 있는 어떤 상황에도 적용할 수 있는 장점이 있다.

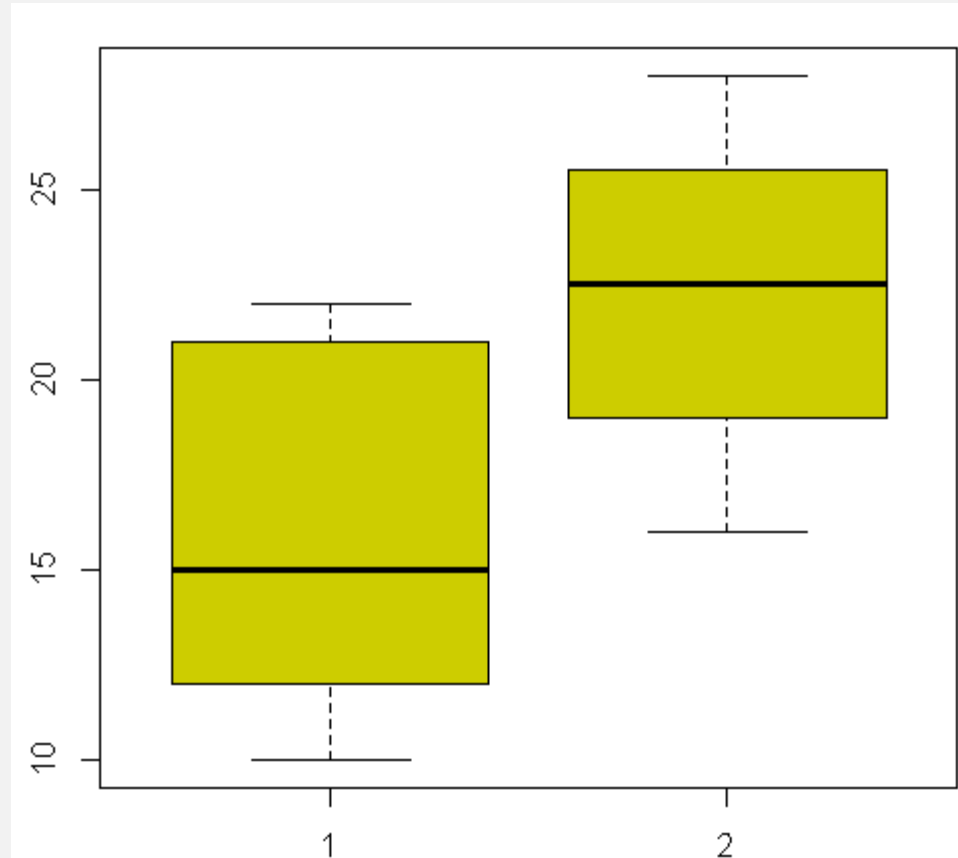
Wilcoxon rank sum test

- 새로운 고혈압 치료제의 치료 효과를 비교하기 위한 예비 연구
- 실험군(1) 5명과 대조군(2) 4명에게 각각 새 고혈압 치료제와 위약을 투여한 뒤 1개월 후 고혈압 점수를 비교하여 치료효과를 비교하고자 함
- 새 고혈압 치료제가 위약에 비해 효과가 있는지 검정하시오

두 군 모두 샘플 크기가 10 미만 이므로 정규성을 가정 할 수 없다.

Wilcoxon rank sum test

```
> # 자료 입력 및 관찰  
> ds = read.csv("2_wilcoxon_rank_sum_test.csv")  
> head(ds)  
> boxplot(score~group, col="yellow3", data=ds)
```



Wilcoxon rank sum test

```
> # Wilcoxon rank sum test 시행
```

```
> wilcox.test(score~group, exact=FALSE, data= ds)
```

```
Wilcoxon rank sum test with continuity  
correction
```

```
data: score by group
```

```
W = 2.5, p-value = 0.0851
```

```
alternative hypothesis: true location shift is not  
equal to 0
```

두 군 사이에 유의한 차이가
있다고 볼 수 없다

[과제 1]

- 8종류의 소주에 대한 소비자의 선호도 조사를 위해서 상표를 가리고 128명을 상대로 시음을 시킨다음 가장 좋아하는 소주를 선택하도록 하였다. 소주 맛에 대한 선호도의 차이가 있다고 할 수 있는지 검정하시오 (유의수준 0.05)

소주종류	관측도수(O)
1	12
2	18
3	11
4	21
5	23
6	16
7	7
8	20

[과제2]

- (1) iris 데이터셋에서 setosa 품종과 versicolor 품종 사이에 Sepal.Length의 평균의 차이가 있는지 검정하시오 (유의수준 0.05)
- (2) iris 데이터셋에서 setosa 품종과 versicolor 품종 사이에 Petal.Length의 평균의 차이가 있는지 검정하시오 (유의수준 0.05)

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4          0.2  setosa
2          4.9         3.0          1.4          0.2  setosa
3          4.7         3.2          1.3          0.2  setosa
4          4.6         3.1          1.5          0.2  setosa
5          5.0         3.6          1.4          0.2  setosa
6          5.4         3.9          1.7          0.4  setosa

> unique(iris$Species)
[1] setosa      versicolor virginica
Levels: setosa versicolor virginica
```

[과제3]

- state.x77 에는 미국50개주의 통계정보가 들어있고 state.region 은 미국의 50개주를 4개 지역으로 나눈 정보가 들어있다.

```
> head(state.x77)
      Population Income Illiteracy Life Exp Murder HS Grad Frost Area
Alabama      3615   3624        2.1   69.05   15.1   41.3    20  50708
Alaska        365   6315        1.5   69.31   11.3   66.7   152 566432
Arizona      2212   4530        1.8   70.55    7.8   58.1    15 113417
Arkansas     2110   3378        1.9   70.66   10.1   39.9    65  51945
California   21198   5114        1.1   71.71   10.3   62.6    20 156361
Colorado     2541   4884        0.7   72.06    6.8   63.9   166 103766

> state.region
 [1] South      West      West      South      West
 [6] West      Northeast South      South      South
[11] West      West      North Central North Central North Central
[16] North Central South      South      Northeast  South
[21] Northeast North Central North Central South      North Central
[26] West      North Central West      Northeast  Northeast
[31] West      Northeast  South      North Central North Central
[36] South      West      Northeast  Northeast  South
[41] North Central South      South      West      Northeast
[46] South      West      South      North Central West
Levels: Northeast South North Central West
```

[과제3]

- (1) 서부(West) 지역과 동부(East) 지역의 주들은 소득(Income) 에 있어서 차이가 있는지를 검정하시오 (유의수준 0.05)
- (2) 북부(North) 지역과 남부(South) 지역의 주들은 문맹률(Illiteracy) 에 있어서 차이가 있는지를 검정하시오 (유의수준 0.05)