

통계자료 분석방법 소개

실제문제를 다루면서

hmkang98@naver.com

와이즈인컴퍼니

데이터분석

모수적 방법 – 평균비교

모수적 검정(parametric test)

모수적 검정 방법

- 한 집단 평균비교
 - 집단의 크기가 30 이상인 경우 : 정규 검정
 - 집단의 크기가 30 미만인 경우 : t – 검정
- 두 집단 평균비교
 - 대응표본인 경우 : 두 집단의 차이를 구한 후 한 집단의 평균비교 방법에 동일하게 적용
 - 독립인 두 표본인 경우
 - 두 집단의 크기가 모두 30 이상인 경우 : 정규 검정
 - 적어도 한 집단의 크기가 30 미만인 경우 : t – 검정, 두 집단의 분산이 같인지 다른지에 따라 검정통계량의 차이가 있음.
- 셋 이상의 집단 평균비교
 - 모든 집단의 분산이 같은 경우 : F – 검정
 - 적어도 한 집단의 분산이 다른 경우 : Welch 검정

한 집단, 두 집단의 평균비교에서 집단의 크기가 커질수록 t 분포와 표준정규분포는 차이가 작아지므로 통계프로그램에서 t 검정을 사용함

한 집단 평균 비교

단일 집단 t-test

* 단일 집단의 모평균 μ 에 대한 검정법

□ 어떤 알약의 부작용으로 혈압강하의 효과가 있는지를 알아보기 위하여 10명의 환자를 대상으로 알약의 복용 전후의 혈압을 측정한 후 복용전 혈압과 복용후 혈압 차가 다음과 같았다.

환자	1	2	3	4	5	6	7	8	9	10
복용전 - 복용후	2	8	10	6	18	10	4	26	18	-8

이 자료로부터 알약 복용전-복용후의 평균이 0이라고 주장할 수 있는가?

한 집단 평균 비교

단일 집단 t-test

① 귀무가설 : 복용전과 복용후 차이의 평균은 0이다.

대립가설 : 복용전과 복용후 차이의 평균은 0이 아니다.

② 검정통계량 :

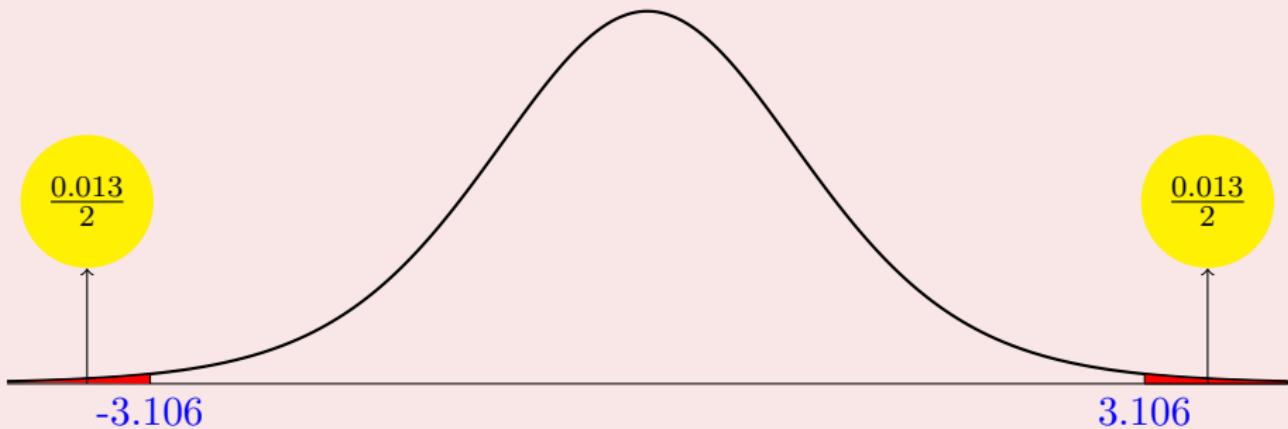
$$t = \frac{\text{표본평균} - \text{귀무가설에서 설정한 평균}}{\text{표준편차}/\sqrt{\text{표본수}}} = \frac{9.4 - 0}{9.571/\sqrt{10}} = 3.106$$

유의확률 : $2 \times P[t \geq 3.106] = 0.013$

③ 연구자가 유의수준을 5%로 설정하였다면 귀무가설을 기각한다.
따라서 평균은 0이 아니라고 할만한 통계적 유의성이 있다.

한 집단 평균 비교

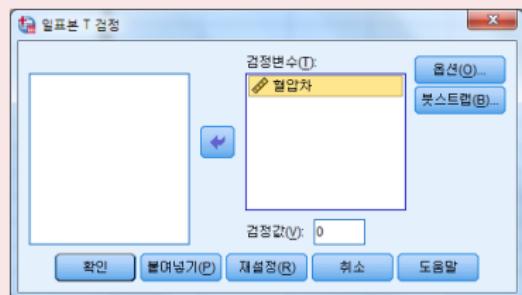
한 집단 평균 비교 - 그래프로



한 집단 평균 비교 – SPSS

한 집단 t-test

SPSS 실행은 분석→평균 비교→일표본 T 검정 메뉴 선택



(a) 단일표본 설정

일표본 통계량					
	N	평균	표준편차	표준오차	
혈압차	10	9.40	9.571	3.027	

일표본 검정					
	검정값 = 0				
	t	자유도	유의확률 (양쪽)	평균차	처미의 95% 신뢰구간
혈압차	3.106	9	.013	9.400	2.55 - 16.25

(b) 단일표본 결과

유의확률이 0.05보다 작으므로 통계적으로 복용전과 복용후 혈압차이는 0이 아님

한 집단, 두 집단의 평균비교 – R

t-test 사용방법

```
t.test(formula, data, subset, na.action, ...)
```

- 한 집단, 대응 집단, 독립은 두 집단의 검정에 사용
- alternative 대립가설의 방향을 설정.
- mu 가설을 검정하는 평균값 또는 차이.
- paired 짹검정(paired samples t-test) 인지 설정하는 논리값.
- var.equal 독립인 두 표본 검정에서 두 그룹의 분산이 같은지 설정
- conf.level 신뢰구간의 계산에서 신뢰도.
- formula data ~ group 와 같은 형태. formula는 두표본 검정인 경우에만 설정

한 집단 평균비교 – R

한 집단 t-test

```
> library(foreign)
> dsn <- read.spss(file="http://pluto.hallym.ac.kr/data/onesam")
> t.test(dsn$diff)
```

One Sample t-test

```
data: dsn$diff
t = 3.1058, df = 9, p-value = 0.0126
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.55347 16.24653
sample estimates:
mean of x
9.4
```

두 집단 평균 비교

독립인 두 집단 t-test

* 두 집단의 모평균 μ_1 과 μ_2 차이에 대한 검정법

□ 나병환자의 치료법을 연구하기 위하여 항생제의 효과를 비교하려고 한다. 실험에 참여하는 나병환자를 랜덤하게 20명을 뽑은 후 10명에게는 항생제 A, 나머지 10명 대조군에게는 생리식염수를 투여하고 나병균 세균수가 다른지 그 경과를 관찰하였다.

항생제	세균수									
	항생제 A	6	0	2	8	11	4	13	1	8
대조군 F	13	10	18	5	23	12	5	16	1	20

이 자료로부터 항생제를 사용한 환자와 사용하지 않은 환자의 세균수는 차이가 있다고 할 수 있는가?

두 집단 평균 비교

독립인 두 집단의 평균비교 과정

- 집단의 크기가 커질수록 정규분포와 t 분포는 차이가 작아지므로 두 집단의 평균비교는 t – 검정을 사용
- 우선 두 집단의 분산이 같은지 검정
- 두 집단의 분산이 같으면 두 집단의 크기수 – 2 인 t – 분포로 검정하고 분산이 같지 않으면 자유도를 보정한 t – 분포로 검정한다.

두 집단 평균 비교

독립인 두 집단의 등분산성 검정 F-test

- ① 귀무가설 : 항생제 A와 대조군 F의 분산은 같다.
대립가설 : 항생제 A와 대조군 F의 분산은 같지 않다.
- ② 검정통계량 : $F = 1.646$, 유의확률 : $P[F \geq 1.646] = 0.216$
- ③ 연구자가 유의수준을 5%로 설정하였다면 귀무가설을 기각할 수 없다. 따라서 두 집단의 분산은 통계적으로 같다.

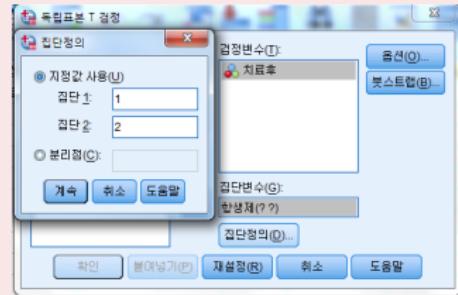
독립인 두 집단 t-test

- ① 귀무가설 : 항생제 A와 대조군 F의 평균은 같다.
대립가설 : 항생제 A와 대조군 F의 평균은 같지 않다.
- ② 검정통계량 : $t = -2.596$, 유의확률 : $2 \times P[t \leq -2.596] = 0.018$
- ③ 연구자가 유의수준을 5%로 설정하였다면 귀무가설을 기각한다.
따라서 두 집단의 평균은 같지 않다는 통계적 유의성이 있다.

두 집단 평균 비교 – SPSS

독립인 두 집단 t-test

SPSS 실행은 분석→평균 비교→독립표본 T 검정 메뉴 선택



(c) 독립인 두 표본 설정

집단통계량										
	N	평균	표준偏差	표준오차	표준오차					
평균 A	10	5.30	4.644	1.469						
평균 B	10	12.30	7.150	2.281						
독립표본 검정										
Levene의 등분산 검증										
	F	유의율	t	자유도	유의율 (양쪽)	평균차	표준오차	차이의 95% 신뢰구간	하한	상한
평균 A	1.646	.216	-2.598	18	.018	-7.000	2.696	-12.864	-1.336	
평균 B			-2.598	15.446	.020	-7.000	2.696	-12.732	-1.268	

(d) 독립인 두 표본 결과

두 집단의 분산이 같은지 검정한 결과 유의확률이 0.215로 두 집단은 통계적으로 같음. 따라서 두 집단의 평균차에 대한 유의확률은 등분산이 같을 때 유의확률 0.018이므로 통계적으로 항생제를 사용한 경우와 사용하지 않은 경우의 세균수는 차이가 있음.

두 집단 평균 비교 – R

독립인 두 변량 F-test : 두 집단의 분산의 동질성 검정

```
> library(foreign)
> dsn <- read.spss(file="http://pluto.hallym.ac.kr/data/twosam")
> var.test(dsn$after ~ dsn$antibiotics)
```

F test to compare two variances

data: dsn\$after by dsn\$antibiotics

F = 0.42186, num df = 9, denom df = 9, p-value = 0.2146

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

95 percent confidence interval:

0.1047853 1.6984253

sample estimates:

ratio of variances

0.4218648

두 집단 평균 비교 – R

독립인 두 변량 t-test

```
> t.test(dsn$after ~ dsn$antibiotics, var.equal=T)
```

Two Sample t-test

```
data: dsn$after by dsn$antibiotics
t = -2.5964, df = 18, p-value = 0.01824
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
-12.66427 -1.33573
sample estimates:
mean in group A mean in group F
      5.3          12.3
```

두 집단 평균 비교

대응 집단 t-test

* 대응 집단의 차이에 대한 검정법

□ 어떤 알약의 부작용으로 혈압강하의 효과가 있는지를 알아보기 위하여 10명의 환자를 대상으로 알약의 복용 전후의 혈압을 측정한 결과 다음과 같았다.

환자	1	2	3	4	5	6	7	8	9	10
복용전	70	80	72	76	76	76	72	78	82	64
복용후	68	72	62	70	58	66	68	52	64	72

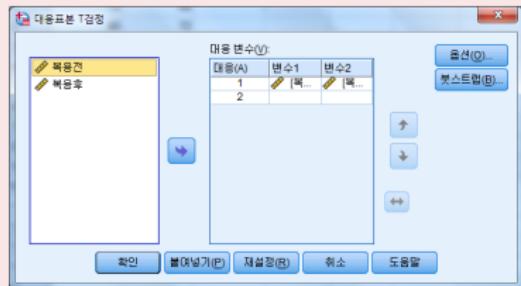
이 자료로부터 알약이 혈압을 내린다는 주장을 할 수 있는가?

주의 대응 표본에 대한 검정은 동일한 개체에서 두 개의 측정값을 관측한 경우이고, 독립인 두 표본은 서로 다른 개체에서 측정값을 관측한 경우이다.

두 집단 평균 비교 – SPSS

대응 집단 t-test

SPSS 실행은 분석→평균 비교→대응비교 T 검정 메뉴 선택



(e) 대응비교 설정

대응표본 통계량				
	평균	N	표준판차	표준오차
대응 1 복용전	74.60	10	5.254	1.661
복용후	65.20	10	6.408	2.026

대응표본 상관계수			
	N	상관계수	유의확률
대응 1 복용전 & 복용후	10	-0.341	0.336

대응표본 검정								
대응자	평균			표준판차			t	
	평균	N	표준판차	표준오차	자마의 95% 신뢰구간	하한	상한	
대응 1 복용전 & 복용후	9.400	9.571	3.027	2.553	16.247	3.108	9	0.013

(f) 대응비교 결과

유의확률이 0.05보다 작으므로 통계적으로 약 복용전과 복용후 효능 차이는 있음

두 집단 평균 비교 – R

대응집단 t-test

```
> library(foreign)
> dsn <- read.spss(file="http://pluto.hallym.ac.kr/data/paired")
> t.test(dsn$before, dsn$after, paired=T)
```

Paired t-test

data: dsn\$before and dsn\$after

t = 3.1058, df = 9, p-value = 0.0126

alternative hypothesis: true difference in means is not equal

95 percent confidence interval:

2.55347 16.24653

sample estimates:

mean of the differences

9.4

셋 이상의 집단 평균 비교

분산분석(ANOVA, ANalysis Of VAriance)

* 서로 독립인 세 가지 이상 집단의 모평균 비교($\mu_1 = \mu_2 = \dots = \mu_n$)

□ 나병환자의 치료법을 연구하기 위하여 항생제의 효과를 비교하려고 한다. 실험에 참여하는 나병환자를 랜덤하게 30명을 뽑은 후 10명에게는 항생제 A, 10명에게는 항생제 D, 나머지 10명 대조군에게는 생리식염수를 투여하고 나병균 세균수가 다른지 그 경과를 관찰하였다.

항생제	세균수									
	항생제 A	항생제 D	대조군 F	항생제 A	항생제 D	대조군 F	항생제 A	항생제 D	대조군 F	항생제 A
항생제 A	6	0	13	2	0	10	8	2	18	5
항생제 D	11	18	23	4	4	12	13	4	14	5
대조군 F	4	9	16	13	14	12	1	9	1	20

이 자료로부터 항생제별 세균수는 차이가 있다고 할 수 있는가?

셋 이상의 집단 평균 비교

관측값의 분해

$$\begin{array}{c} \text{각 집단} \\ \left(\begin{array}{l} \text{항생제 A} \\ \text{항생제 D} \\ \text{대조군 F} \end{array} \right) = \left(\begin{array}{cccccccccc} 6 & 0 & 2 & 8 & 11 & 4 & 13 & 1 & 8 & 0 \\ 0 & 2 & 3 & 1 & 18 & 4 & 14 & 9 & 1 & 9 \\ 13 & 10 & 18 & 5 & 23 & 12 & 5 & 16 & 1 & 20 \end{array} \right) = \\ \text{전체 평균 } (\bar{y}) \\ \left(\begin{array}{cccccccccc} 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 \\ 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 \\ 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 & 7.9 \end{array} \right) + \\ \text{처리효과 } (\bar{y}_i - \bar{y}) \\ \left(\begin{array}{cccccccccc} -2.6 & -2.6 & -2.6 & -2.6 & -2.6 & -2.6 & -2.6 & -2.6 & -2.6 & -2.6 \\ -1.8 & -1.8 & -1.8 & -1.8 & -1.8 & -1.8 & -1.8 & -1.8 & -1.8 & -1.8 \\ 4.4 & 4.4 & 4.4 & 4.4 & 4.4 & 4.4 & 4.4 & 4.4 & 4.4 & 4.4 \end{array} \right) + \\ \text{잔차 } (y_{ij} - \bar{y}_i) \\ \left(\begin{array}{cccccccccc} 0.7 & -5.3 & -3.3 & 2.7 & 5.7 & -1.3 & 7.7 & -4.3 & 2.7 & -5.3 \\ -6.1 & -4.1 & -3.1 & -5.1 & 11.9 & -2.1 & 7.9 & 2.9 & -5.1 & 2.9 \\ 0.7 & -2.3 & 5.7 & -7.3 & 10.7 & -0.3 & -7.3 & 3.7 & -11.3 & 7.7 \end{array} \right) \end{array}$$

셋 이상의 집단 평균 비교

k개의 처리집단에 대한 완전 랜덤화 설계의 구조

	관측자료				평균	제곱합
처리 1	Y_{11}	Y_{12}	\cdots	Y_{1n_1}	\bar{Y}_1	$\sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)^2$
처리 2	Y_{21}	Y_{22}	\cdots	Y_{2n_2}	\bar{Y}_2	$\sum_{j=1}^{n_2} (Y_{2j} - \bar{Y}_2)^2$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
처리 k	Y_{k1}	Y_{k2}	\cdots	Y_{kn_k}	\bar{Y}_k	$\sum_{j=1}^{n_k} (Y_{kj} - \bar{Y}_k)^2$

전체 자료수 $n = n_1 + n_2 + \cdots + n_k$

전체평균 $\bar{Y} = \frac{n_1\bar{Y}_1 + n_2\bar{Y}_2 + \cdots + n_k\bar{Y}_k}{n_1 + n_2 + \cdots + n_k}$

표: 네 종류의 코팅 자료와 통계량

$$\begin{aligned}\text{관측값} &= (\text{전체평균}) + (\text{처리의 편차}) + (\text{잔차}) \\ Y_{ij} &= \bar{Y} + (\bar{Y}_i - \bar{Y}) + (Y_{ij} - \bar{Y}_i)\end{aligned}$$

셋 이상의 집단 평균 비교

제곱합의 분해

$$\begin{aligned} & (\text{관측값} - \text{전체평균}) = (\text{처리의 편차}) + (\text{잔차}) \\ & \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} [(\bar{Y}_i - \bar{Y}) + (Y_{ij} - \bar{Y}_i)]^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} [(\bar{Y}_i - \bar{Y})^2 + 2(\bar{Y}_i - \bar{Y})(Y_{ij} - \bar{Y}_i) + (Y_{ij} - \bar{Y}_i)^2] \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} 2(\bar{Y}_i - \bar{Y})(Y_{ij} - \bar{Y}_i) + \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_i - \bar{Y})^2 + 2 \underbrace{\sum_{i=1}^k (\bar{Y}_i - \bar{Y})}_{0} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i) + \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \\ &= \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \\ & \text{전체제곱합} = \text{처리제곱합} + \text{잔차제곱합} \end{aligned}$$

셋 이상의 집단 평균 비교

제곱합의 성질

$$\begin{array}{ccl} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2 & = & \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \\ \text{전체제곱합} & = & \text{처리제곱합} + \text{잔차제곱합} \\ \sum_{i=1}^k n_i - 1(\text{자유도}) & & \sum_{i=1}^k n_i - k(\text{자유도}) \end{array}$$

- SST(total sum of square) : 전체제곱합
- SSt(treatment sum of square) : 처리 제곱합(집단간 제곱합)
- SSE(error sum of square) : 잔차 제곱합(오차제곱합, 집단내 제곱합)
- MS(mean square) : 평균제곱 = $\frac{\text{제곱합}}{\text{자유도}}$

셋 이상의 집단 평균 비교

분산분석표

요인	제곱합	자유도	평균제곱합
처리	$SSt = \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2$	$k - 1$	$MSt = \frac{SSt}{k - 1}$
잔차	$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$	$\sum_{i=1}^k n_i - k$	$MSE = \frac{SSE}{\sum_{i=1}^k n_i - k}$
합계	$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$	$\sum_{i=1}^k n_i - 1$	

표: 분산분석표

셋 이상의 집단 평균 비교

분산분석 과정

- ① $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ vs $H_1 : \text{not } H_0$ (적어도 한 집단의 평균은 다른)
- ② 검정통계량 :
$$F_0 = \frac{\text{평균처리오차(MSt)}}{\text{평균오차제곱(MSE)}} = \frac{SSt/(k-1)}{SSE/(n-k)} \sim F(k-1, n-k),$$
유의확률 : $P[F \geq F_0]$
- ③ 유의확률로 가설에 대한 결론

셋 이상의 집단 평균 비교

분산분석표 작성

요인	제곱합	자유도	평균제곱합	F_0	유의확률
처리	293.6	2	146.8	$3.983(\frac{146.8}{36.856})$	0.030
잔차	995.1	27	36.856		
합계	1288.7	29			

- ① $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_1 : \text{not } H_0$
- ② 검정통계량 : $F_0 = 3.983$, 유의확률 : $P[F \geq 3.983] = 0.030$
- ③ 유의수준 5%에서 귀무가설을 기각한다. 따라서 적어도 한 집단의 평균은 통계적으로 다르다고 할 수 있다.

셋 이상의 집단 평균 비교

서로 독립인 세 집단의 등분산성 검정

- ① 귀무가설 : 세 집단의 분산은 모두 같다. 대립가설 : 적어도 한 집단의 분산은 다르다.
- ② Levene W 통계량 : 0.890, 유의확률 : $P[W \geq 0.89] = 0.422$
- ③ 유의수준 5%에서 귀무가설을 기각하지 않으므로 세 집단의 분산은 모두 같다.

서로 독립인 세 집단의 평균 비교

- ① 귀무가설 : 세 집단의 평균은 모두 같다. 대립가설 : 적어도 한 집단의 평균은 다르다.
- ② F 통계량 : 3.983, 유의확률 : $P[F \geq 3.983] = 0.03$
- ③ 유의수준 5%에서 귀무가설을 기각하므로 적어도 한 집단의 평균은 다르다.

셋 이상의 집단 평균 비교

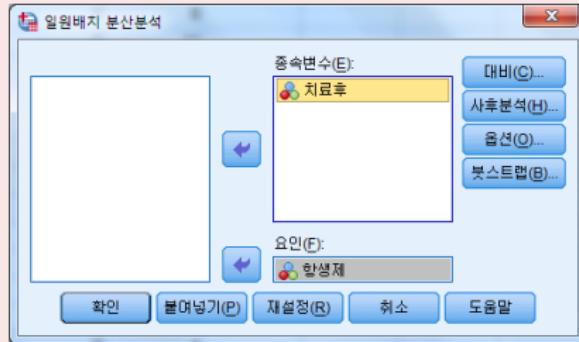
셋 이상의 집단 평균 비교 과정

- ① 비교하려는 셋 이상의 집단의 분산이 모두 같은지 등분산성 검정
- ② 셋 이상 집단의 평균이 모두 같은지 검정
 - 집단의 분산이 모두 같으면 F - 검정,
 - 적어도 하나 이상 집단의 분산 같지 않으면 Welch 검정
- ③ 셋 이상 집단의 평균이 모두 같지 않다면 두 집단씩 평균이 같은지 비교하며 그 갯수는 $\binom{\text{집단의 수}}{2}$ 이며 사후분석이라고 함

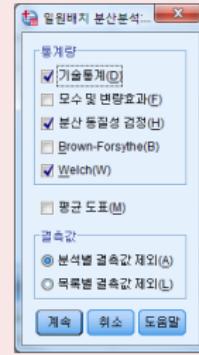
셋 이상의 집단 평균 비교 – SPSS

분산분석(ANOVA, ANalysis Of VAriance)

SPSS 실행은 분석→평균 비교→일원배치 분산분석 메뉴 선택



(g) 분산분석 설정



(h) 옵션

설정은 집단변수는 요인, 측정변수는 종속변수에 입력하고, 옵션에 집단의 요약정보인 기술통계, 분산분석이 가능한지 분산 동질성 검정, 분산이 다를 때 검정할 Welch를 선택

셋 이상의 집단 평균 비교 – SPSS

분산분석(ANOVA, ANalysis Of VAriance)

일원배치분산분석(one-way ANOVA)으로 분석한 SPSS 출력결과

기술통계								
치료후	N	평균	표준편차	표준오차	평균에 대한 95% 신뢰구간		최소값	최대값
					하한값	상한값		
항생제 A	10	5.30	4.644	1.469	1.98	8.62	0	13
항생제 D	10	6.10	6.154	1.946	1.70	10.50	0	18
대조군 F	10	12.30	7.150	2.261	7.19	17.41	1	23
합계	30	7.90	6.866	1.217	5.41	10.39	0	23

분산의 동질성 검정				
치료후				
Levene 통계량	df1	df2	유의확률	
.890	2	27	.422	

분산분석					
치료후					
	제곱합	df	평균 제곱	F	유의확률
집단-간	293.800	2	146.800	3.983	.030
집단-내	995.100	27	36.856		
합계	1288.700	29			

분산의 동질성 검정결과 유의확률이 0.422로 세 집단의 분산이 같으므로 분산분석을 사용할 수 있으며 세 집단의 평균비교 결과 유의확률이 0.030으로 세 집단의 평균은 적어도 한 집단이 다르다.

셋 이상의 집단 평균 비교 – R

leveneTest 사용방법

```
leveneTest(x, ...)
```

```
leveneTest(x, g, ...)
```

```
leveneTest(formula, data, subset, na.action, ...)
```

- 세 집단 이상의 분산이 같은지 검정
- 모형식은 formula에 response ~ group로 설정

셋 이상의 집단 평균 비교 – R

분산분석(ANOVA); Levene 검정 - 분산의 동질성 검정

```
> library(foreign)
> dsn <- read.spss(file="http://pluto.hallym.ac.kr/data/oneway")
> install.packages("Rcmdr") # 패키지 설치
> library(Rcmdr) # 패키지 불러오기
> leveneTest(dsn$after ~ dsn$antibiotics, center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
    Df F value Pr(>F)
group  2  0.8904 0.4222
      27
```

셋 이상의 집단 평균 비교 – R

oneway.test 사용방법

```
oneway.test(formula, data, subset, na.action,  
           var.equal = FALSE)
```

- 분산분석 결과 F 통계량과 유의확률의 계산결과를 제공하는 함수
- formula response ~ group 와 같은 형태의 모형식을 지정
- var.equal 그룹별로 분산이 같은지 설정. 값이 TRUE이면 분산분석표와 F 분포를 이용한 가설검정을 실행하며, FALSE이면 Welch 검정(Welch(1951))을 시행.

셋 이상의 집단 평균 비교 – R

분산분석(ANOVA); 분산분석 결과

```
> oneway.test(dsn$after ~ dsn$antibiotics, var.equal=T)
```

One-way analysis of means

```
data: dsn$after and dsn$antibiotics  
F = 3.9831, num df = 2, denom df = 27, p-value = 0.03049
```

셋 이상의 집단 평균 비교 – R

aov 사용방법

```
aov(formula, data = NULL, projections = FALSE, qr = TRUE,  
contrasts = NULL, ...)
```

- 분산분석 모형을 적합시키는 함수
- formula : response ~ group 와 같은 형태의 모형식을 지정
- data : 데이터 이름 설정

셋 이상의 집단 평균 비교 – R

분산분석(ANOVA); 분산분석표

```
> aov(dsn$after ~ dsn$antibiotics)
```

Call:

```
  aov(formula = dsn$after ~ dsn$antibiotics)
```

Terms:

	dsn\$antibiotics	Residuals
Sum of Squares	293.6	995.1
Deg. of Freedom	2	27

Residual standard error: 6.070878

Estimated effects may be unbalanced

셋 이상의 집단 평균 비교 – R

anova 사용방법

`anova(object, ...)`

- 분산분석표 모형 계산
- `object` : 모형을 적합시키는 `aov` 또는 `lm` 함수에서 결과가 저장된 객체

셋 이상의 집단 평균 비교 – R

분산분석(ANOVA); 분산분석표

```
> anova(aov(dsn$after ~ dsn$antibiotics))
```

```
Analysis of Variance Table
```

```
Response: dsn$after
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dsn\$antibiotics	2	293.6	146.800	3.9831	0.03049 *
Residuals	27	995.1	36.856		

```
---
```

```
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1  
‘ ’ 1
```

분산분석(ANOVA)

다중비교(multiple comparison)

- 일원배치법의 분산분석에서 귀무가설 H_0 를 기각한 경우, 그룹간의 평균차가 있는 것을 추론하는 통계기법
- 쳐리 수가 k 인 것에 대한 평균차의 비교 개수는 $\binom{k}{2}$ 이며, 이와 같은 비교방법을 다중비교, 사후 분석이라고 부름
- LSD, Scheffe, Bonferroni, Duncan, Tukey 등이 있음

셋 이상의 집단 평균 비교

분산분석(ANOVA, ANalysis Of VAriance)

평균이 다른 집단이 적어도 하나 존재하므로 SPSS에서 분석→평균 비교→일원배치 분산분석 메뉴 선택하고 사후분석을 선택한 후 분산이 같을 때 다중비교의 분석 종류를 선택



셋 이상의 집단 평균 비교

분산분석(ANOVA, ANalysis Of VAriance)

일원배치분산분석(one-way ANOVA)으로 분석한 SPSS 사후분석 출력결과

차료후 Duncan ^a			
항생제	N	유의수준 = 0.05에 대한 부집단	
		1	2
항생제 A	10	5.30	
항생제 D	10	6.10	
대조군 F	10		12.30
유의확률		.771	1.000

둘일 집단군에 있는 집단에 대한 평균이 표시됩니다.

a. 조화평균 표본 크기 10.000을(를) 사용합니다.

사후분석 분석종류를 Duncan으로 설정하고 분석한 결과 항생제 A, 항생제 D 두 집단의 평균은 같고 대조군 F 집단은 항생제 집단과 평균이 다른 집단으로 검정됨

셋 이상의 집단 평균 비교 – R

duncan.test 사용방법

```
duncan.test(y, trt, DFerror, MSerror, alpha = 0.05,  
            group=TRUE, main = NULL,console=FALSE)
```

- Duncan 사후분석방법
- y : aov나 lm으로 적합된 모형.
- trt : 그룹을 표시하는 변수.
- DFerror : 오차 자유도.
- MSerror : 평균제곱오차.
- alpha : 유의수준.
- group : 그룹을 표시하는 변수.
- main : 제목 설정.

셋 이상의 집단 평균 비교 – R

분산분석(ANOVA); Duncan 사후분석

```
> install.packages("agricolae")
> library(agricolae)
> model <- aov(dsn$after ~ dsn$antibiotics)
> out <- duncan.test(model, "dsn$antibiotics")
> out
$statistics
  MSerror Df Mean      CV
  36.85556 27  7.9 76.84655
$parameters
  test      name.t ntr alpha
  Duncan dsn$antibiotics   3  0.05
$duncan
  Table CriticalRange
  2 2.901727      5.570677
  3 3.048662      5.852760
```

셋 이상의 집단 평균 비교 – R

분산분석(ANOVA); Duncan 사후분석 – 전 화면에 이어서

\$means

	dsn\$after	std	r	Min	Max	Q25	Q50	Q75
A	5.3	4.643993	10	0	13	1.25	5.0	8.0
D	6.1	6.154492	10	0	18	1.25	3.5	9.0
F	12.3	7.149981	10	1	23	6.25	12.5	17.5

\$comparison

NULL

\$groups

dsn\$after groups

F	12.3	a
D	6.1	b
A	5.3	b

비모수적 방법

비모수적 검정(nonparametric test)

비모수적 검정 방법

- 한 집단 비교

- 부호 검정(sign test)
- Wilcoxon 부호 순위 검정(Wilcoxon signed rank test)

- 두 집단 비교

- 대응표본인 경우 : Wilcoxon 부호 순위 검정(Wilcoxon signed rank test)
- 독립인 두 표본인 경우 : Wilcoxon 순위합 검정(Wilcoxon rank sum test, Mann-Whitney U test, Wilcoxon-Mann-Whitney test)

- 셋 이상의 집단 평균비교

- 모든 집단이 서로 독립인 경우 : Kruscal-Wallis 검정
- 반복 측정인 경우 : Friedman 검정

한 집단 중앙값 비교 – 비모수 방법

단일 집단 Wilcoxon signed rank test

* 단일 집단의 중앙값에 대한 검정법

□ 어떤 알약의 부작용으로 혈압강하의 효과가 있는지를 알아보기 위하여 10명의 환자를 대상으로 알약의 복용 전후의 혈압을 측정한 후 복용전 혈압과 복용후 혈압 차가 다음과 같았다.

환자	1	2	3	4	5	6	7	8	9	10
복용전 – 복용후	2	8	10	6	18	10	4	26	18	-8

이 자료로부터 알약 복용전–복용후의 중앙값이 0이라고 주장을 할 수 있는가?

한 집단 중앙값 비교 – 비모수 방법

단일 집단 Wilcoxon signed rank test

① 귀무가설 : 복용전과 복용후 차이의 중앙값은 0이다.

대립가설 : 복용전과 복용후 차이의 중앙값은 0이 아니다.

② 검정통계량 : $s(X_i) = I_{[x_i > 0]}$, $W = \sum_{i=1}^n R_i s(X_i)$, $Z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$,

$s(X_i)$ 는 양수는 1이고 음수는 0인 함수, R_i 는 순위, n 은 자료수.

$$z = \frac{4.5 - \frac{10 \cdot (10+1)}{4}}{\sqrt{\frac{10(10+1)(2 \cdot 10+1)}{24}}} = -2.344$$

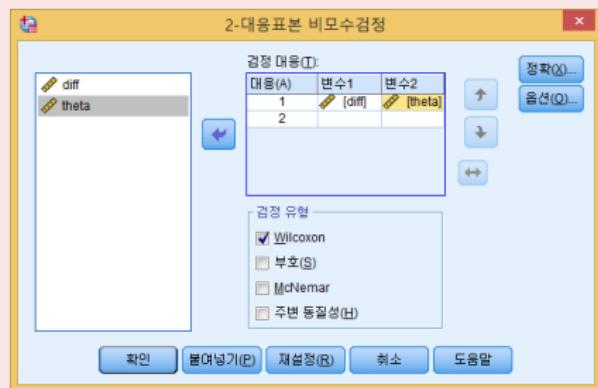
유의확률 : $2 \times P[t \leq -2.344] = 0.019$

③ 연구자가 유의수준을 5%로 설정하였다면 귀무가설을 기각한다.
따라서 중앙값은 0이 아니라고 할만한 통계적 유의성이 있다.

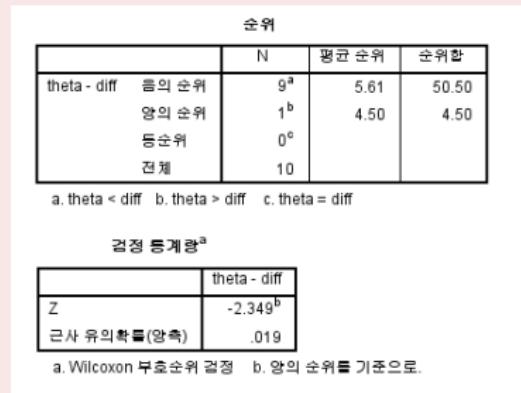
한 집단 중앙값 비교 – SPSS

한 집단 Wilcoxon signed rank test

SPSS 실행은 분석→비모수 검정→레거시 대화상자→2-대응표본 메뉴 선택



(i) 단일표본 설정



(j) 단일표본 결과

유의확률이 0.05보다 작으므로 통계적으로 복용전과 복용후 혈압차이는 0이 아님

한 집단, 두 집단의 비교 – R

wilcox.test 사용방법

```
wilcox.test(x, y=null, alternative, mu, paired, conf.level, ...)
```

- 한 집단, 대응 집단, 독립은 두 집단의 검정에 사용
- **alternative** : 대립가설의 방향을 설정.
- **mu** : 대응표본에서 두 집단 차이를 설정하는 값으로 기본은 0
- **paired** : 짹검정(paired samples) 인지 설정하는 논리값.
- **var.equal** : 독립인 두 표본 검정에서 두 그룹의 분산이 같은지 설정
- **conf.level** : 신뢰구간에서 신뢰도.

한 집단 평균비교 – R

한 집단 Wilcoxon signed rank test

```
> library(foreign)
> dsn <- read.spss(file="http://pluto.hallym.ac.kr/data/onesam")
> wilcox.test(dsn$diff)
```

Wilcoxon signed rank test with continuity correction

```
data: dsn$diff
V = 50.5, p-value = 0.02157
alternative hypothesis: true location is not equal to 0
```

Warning message:

In wilcox.test.default(dsn\$diff) :
tie가 있어 정확한 p값을 계산할 수 없습니다

두 집단 분포 비교 – 비모수 방법

독립인 두 집단 분포 비교 Wilcoxon rank sum test

* 두 집단의 모평균 μ_1 과 μ_2 차이에 대한 검정법

□ 나병환자의 치료법을 연구하기 위하여 항생제의 효과를 비교하려고 한다. 실험에 참여하는 나병환자를 랜덤하게 20명을 뽑은 후 10명에게는 항생제 A, 나머지 10명 대조군에게는 생리식염수를 투여하고 나병균 세균수가 다른지 그 경과를 관찰하였다.

항생제	세균수									
	항생제 A	6	0	2	8	11	4	13	1	8
대조군 F	13	10	18	5	23	12	5	16	1	20

이 자료로부터 항생제를 사용한 환자와 사용하지 않은 환자의 세균수는 분포가 같은가?

두 집단 분포 비교 – 비모수 방법

독립인 두 집단 분포 비교 Wilcoxon rank sum test

- ① 귀무가설 : 항생제 A와 대조군 F의 분포는 같다.
대립가설 : 항생제 A가 대조군 F보다 세균수가 작다.

- ② 검정통계량 : $W = \text{항생제 A 순위합}, U = W - \frac{m(m+1)}{2}, z = \frac{U - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}$, m 은 항생제 A n 은 대조군 F 집단의 자료수,

$$W = 76, U = 76 - \frac{10(11)}{2} = 21, Z = \frac{21 - \frac{10 \cdot 10}{2}}{\sqrt{\frac{10 \cdot 10(10+10+1)}{12}}} = -2.192$$

유의확률 : $P[t \leq -2.192] = 0.014$

- ③ 연구자가 유의수준을 5%로 설정하였다면 귀무가설을 기각한다.
따라서 항생제 A의 세균수는 대조군 F의 세균수보다 통계적으로 작다.

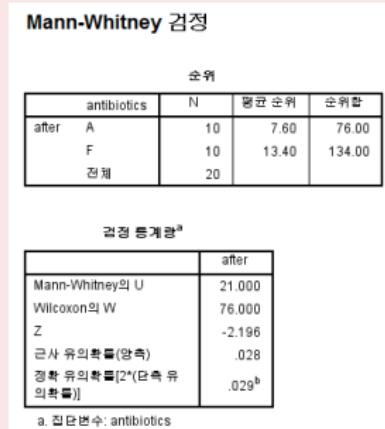
두 집단 분포 비교 – SPSS

독립인 두 집단 비교 Wilcoxon rank sum test

SPSS 실행은 분석→비모수 검정→레거시 대화상자→2-독립표본 선택



(k) 독립인 두 표본 설정



(l) 독립인 두 표본 결과

검정결과 유의확률은 $0.028/2=0.014$ (단측검정)이므로 통계적으로 항생제를 사용이 사용하지 않은 환자보다 세균수는 작음.

두 집단 분포 비교 – R

독립인 두 집단 비교 Wilcoxon rank sum test

```
> dsn <- read.spss(file="http://pluto.hallym.ac.kr/data/twosam")
> groupA <- subset(dsn$after, dsn$antibiotics=='A')
> groupF <- subset(dsn$after, dsn$antibiotics=='F')
> wilcox.test(groupA, groupF)
```

Wilcoxon rank sum test with continuity correction

```
data: groupA and groupF
W = 21, p-value = 0.03089
alternative hypothesis: true location shift is not equal to 0
```

Warning message:

In wilcox.test.default(groupA, groupF) :
tie가 있어 정확한 p값을 계산할 수 없습니다

두 집단 비교 – 비모수 방법

대응 집단 Wilcoxon signed rank test

* 대응 집단의 차이에 대한 검정법

□ 어떤 알약의 부작용으로 혈압강하의 효과가 있는지를 알아보기 위하여 10명의 환자를 대상으로 알약의 복용 전후의 혈압을 측정한 결과 다음과 같았다.

환자	1	2	3	4	5	6	7	8	9	10
복용전	70	80	72	76	76	76	72	78	82	64
복용후	68	72	62	70	58	66	68	52	64	72

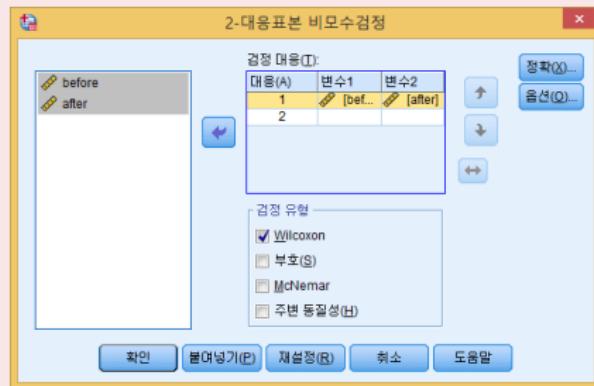
이 자료로부터 알약이 혈압을 내린다는 주장을 할 수 있는가?

주의 대응 표본에 대한 검정은 동일한 개체에서 두 개의 측정값을 관측한 경우이고, 독립인 두 표본은 서로 다른 개체에서 측정값을 관측한 경우이다.

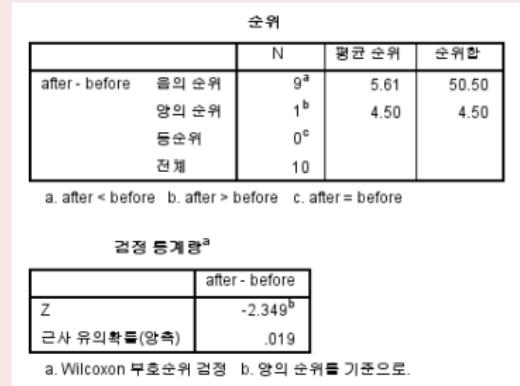
두 집단 비교 – SPSS

대응 집단 Wilcoxon signed rank test

SPSS 실행은 분석→비모수 검정→레거시 대화상자→2-대응표본 메뉴 선택



(m) 대응비교 설정



(n) 대응비교 결과

유의확률이 0.05보다 작으므로 통계적으로 약 복용전과 복용후 효능 차이는 있음

두 집단 비교 – R

대응집단 Wilcoxon signed rank test

```
> library(foreign)
> dsn <- read.spss(file="http://pluto.hallym.ac.kr/data/paired")
> wilcox.test(dsn$before, dsn$after, paired=T)
```

Wilcoxon signed rank test with continuity correction

data: dsn\$before and dsn\$after

V = 50.5, p-value = 0.02157

alternative hypothesis: true location shift is not equal to 0

Warning message:

In wilcox.test.default(dsn\$before, dsn\$after, paired = T) :
tie가 있어 정확한 p값을 계산할 수 없습니다

셋 이상의 집단 비교 – 비모수 방법

셋 이상의 집단 비교 Kruskal-Wallis test

- ＊ 서로 독립인 세 가지 이상 집단의 중앙값 비교($\mu_1 = \mu_2 = \dots = \mu_n$)
- 나병환자의 치료법을 연구하기 위하여 항생제의 효과를 비교하려고 한다. 실험에 참여하는 나병환자를 랜덤하게 30명을 뽑은 후 10명에게는 항생제 A, 10명에게는 항생제 D, 나머지 10명 대조군에게는 생리식염수를 투여하고 나병균 세균수가 다른지 그 경과를 관찰하였다.

항생제	세균수									
	항생제 A	0	2	8	11	4	13	1	8	0
항생제 D	0	2	3	1	18	4	14	9	1	9
대조군 F	13	10	18	5	23	12	5	16	1	20

이 자료로부터 항생제별 세균수는 차이가 있다고 할 수 있는가?

셋 이상의 집단 분포 비교

분산분석 과정

- ① H_0 : 집단의 분포는 모두 같다. vs H_1 : *not* H_0 (적어도 한 집단의 분포는 다름)

- ② 검정통계량 : $H = \frac{12}{N(N+1)} \sum_{i=1}^g \frac{1}{n_i} \left(\sum_{j=1}^{n_i} R_{ij} - n_i \frac{N+1}{2} \right)^2$ 이고

정리하면 $H = \frac{12}{N(N+1)} \sum_{i=1}^g n_i \bar{R}_{i\cdot}^2 - 3(N+1)$, N 은 전체 자료수,

R_{ij} 는 전체 순위, n_i 는 각 집단 자료수, $\bar{R}_{i\cdot}$ 은 각 집단 순위 평균.

유의확률 : $P[\chi^2(df = g-1) \geq H]$

- ③ 유의확률로 가설에 대한 결론

셋 이상의 집단 분포 비교

검정 과정

- ① H_0 : 집단의 분포는 모두 같다. vs H_1 : *not* H_0 (적어도 한 집단의 분포는 다름)

- ② 검정통계량 :

$$H = \frac{12}{30(30+1)} (10 \cdot 12.2^2 + 10 \cdot 13.25^2 + 10 \cdot 21.05^2) - 3(30+1) = 6.033,$$

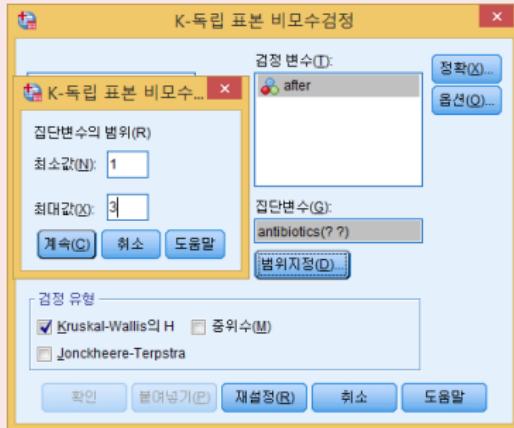
유의확률 : $P[\chi^2(df=2) \geq 6.033] = 0.049$

- ③ 유의수준 5%에서 귀무가설을 기각한다. 따라서 적어도 한 집단의 분포는 통계적으로 다르다고 할 수 있다.

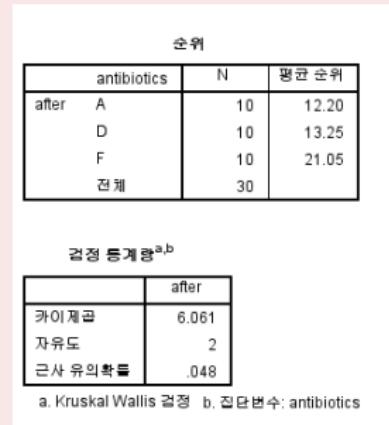
셋 이상의 집단 평균 비교 – SPSS

검정과정 – SPSS

SPSS 실행은 분석→비모수 검정→레거시 대화상자→k-독립표본 선택



(o) 설정



(p) 옵션

유의수준 5%에서 귀무가설을 기각한다. 따라서 적어도 한 집단의 분포는 통계적으로 다르다고 할 수 있다.

셋 이상의 집단 분포 비교 – R

kruskal.test 사용방법

```
kruskal.test(formula, data, subset, na.action, ...)
```

- g 집단의 분포가 모두 동일한지 검정하는 함수
- formula response ~ group 와 같은 형태의 모형식을 지정
- data : 데이터 이름 설정

posthoc.kruskal.nemenyi.test 사용방법

```
posthoc.kruskal.nemenyi.test(x, g, formula, data, na.action,
```

- 독립인 k 집단의 비모수 사후분석 방법
- x : 반응 변수, g : 집단 변수, data : 데이터 이름 설정
- formula : response ~ group 와 같은 형태의 모형식을 지정

셋 이상의 집단 분포 비교 – R

분석 과정 및 결과

```
> library(foreign)
> dsn <- read.spss(file="http://pluto.hallym.ac.kr/data/oneway")
> kruskal.test(dsn$after~dsn$antibiotics)
> install.packages("PMCMRplus")
> library(PMCMRplus)
> kwAllPairsNemenyiTest(x=dsn$after,g=dsn$antibiotics,method="Nemenyi")
```

Pairwise comparisons using Tukey-Kramer-Nemenyi all-pairs test

data: dsn\$after and dsn\$antibiotics

	A	D
D	0.962	-
F	0.063	0.117

범주형 자료분석

범주형 자료분석

빈도분석(frequency analysis)

- * 빈도분석은 한 변량의 각 범주에 속한 사례(응답수)가 얼마인지 빈도로 표현하며 상대 듯수와 함께 나타낸다.
- 현혈한 사람의 혈액형을 요약한 자료이다.

혈액형	O	A	B	AB
인원수	38	43	10	5

- 혈액형의 비율이 같은지 알아보자.

범주형 자료분석

빈도분석(frequency analysis)

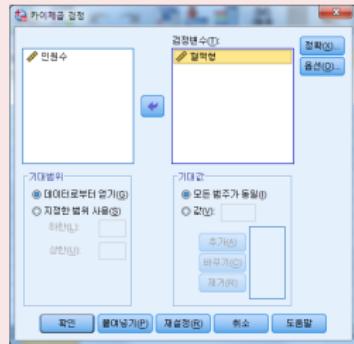
적합도 검정(goodness of fit test) : 각 범주마다 비율이 $p_1 : p_2 : \dots : p_k$ 인지 검정하는 분석방법이다. 모든 비율은 같지 않아도 된다.

- ① 귀무가설 : 각 혈액형의 비율은 모두 같다. 대립가설 : 혈액형 비율이 같지 않는 것이 있다.
- ② χ^2 검정통계량 : 46.417, 유의확률 : $P[\chi^2 \geq 46.417] \leq 0.0001$
- ③ 유의수준 5%에서 귀무가설을 기각한다. 따라서 각 혈액형 비율은 모두 같다고 할 수 없다.

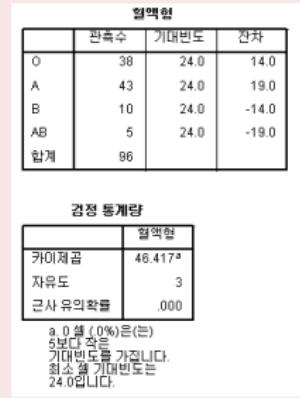
범주형 자료분석 – SPSS

빈도분석(frequency analysis)

SPSS에서 분석→비모수 검정→레거시 대화 상자→카이제곱 검정 실행



(q) 적합도 검정 설정



(r) 적합도 검정 결과

기대값에 각 범주의 비율이 같으면 모든 범주가 동일, 다르면 값을 선택하고 각 비율을 입력한다. 유의확률이 0.05보다 작으므로 통계적으로 혈액형 비율은 모두 같지 않다.

범주형 자료분석 – R

chisq.test 사용방법

```
chisq.test(x, y = NULL, correct = TRUE,  
           p = rep(1/length(x), length(x)), rescale.p = FALSE,  
           simulate.p.value = FALSE, B = 2000)
```

- 범주형 자료 분석에 사용할 카이제곱 검정
- x 분석할 행 변수
- y 분석할 열 변수

xtabs 사용방법

```
xtabs(formula = ~., data = parent.frame(), subset, sparse = FA  
na.action, exclude = c(NA, NaN), drop.unused.levels = FA
```

- 어떤 벡터에 대하여 분할표를 만드는 함수
- formula : 종속변수 ~ 설명변수 형태로 분석할 식을 설정
- data 분석할 자료의 data.frame

범주형 자료분석 – R

빈도분석(frequency analysis)

```
> blood <- c("O", "A", "B", "AB")
> freq <- c(38,43,10,5)
> cat <- xtabs(freq~blood)
> chisq.test(cat)
```

Chi-squared test for given probabilities

```
data: cat
X-squared = 46.417, df = 3, p-value = 4.625e-10
```

범주형 자료분석

교차분석(cross tabulation analysis)

- * 교차분석은 두 변량의 각 범주가 교차하는 빈도를 표현하며, 행별, 열별, 전체별 상대 뜻수를 함께 나타낸다.
- ▣ 노인에 대한 뼈의 광물질 손실을 막기 위하여 물리치료나 운동을 실시한 후 뼈속의 광물질 변화를 측정한 자료이다.

치료 방법 \ 뼈속 변화	상당한 손실	작은 변화	상당한 증가
아무 것도 안함	38	15	7
물리 치료 시행	22	32	16
운동 요법 시행	15	30	25

- ▣ 치료방법과 광물질의 변화는 서로 연관성, 상관성이 있는지 검토가 필요

범주형 자료분석

교차분석(cross tabulation analysis)

이 분석은 자료 수집 방법에 따라 두 가지로 나눈다.

- 동질성 검정(hemogeneity test) : 모집단에서 이미 정한 수 만큼의 표본을 뽑아 범주별로 분류
- 독립성 검정(independence test) : 표본의 수를 정하지 않고 자료를 수집하는 경우

교차분석(cross tabulation analysis)

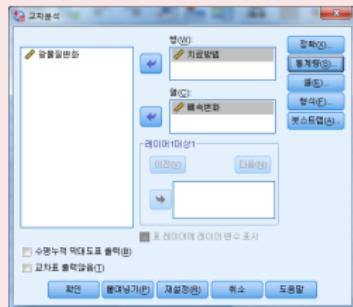
독립성 검정(independence test) : 가로 범주와 세로 범주는 서로 영향을 주는지 검정한다. 즉 두 변수는 서로 독립인지 검정한다.

- ① 귀무가설 : 치료방법과 뼈속변화는 서로 관련이 없다. 대립가설 : 치료방법과 뼈속변화는 서로 관련이 있다.
- ② χ^2 검정통계량 : 28.162, 유의확률 : $P[\chi^2 \geq 28.162] \leq 0.0001$
- ③ 유의수준 5%에서 귀무가설을 기각한다. 따라서 치료방법과 뼈속변화는 서로 관련이 있다.

범주형 자료분석 – SPSS

교차분석(cross tabulation analysis)

SPSS에서 분석→기술통계량→교차분석 실행



(s) 교차분석 설정



(t) 카이제곱 선택

통계량 버튼을 누르고 카이제곱을 선택한 후 분석실행.

범주형 자료분석 – SPSS

교차분석(cross tabulation analysis)

SPSS에서 교차분석 실행 결과

치료방법 × 빠속변화 교차표					
반도	치료방법	빠속변화			전체
		상당한 흔들	작은 변화	상당한 증가	
	아무 것도 안함	38	15	7	60
	풀리 치료 시행	22	32	16	70
	운동 요법 시행	15	30	25	70
전체		75	77	48	200

카이제곱 검정			
	값	자유도	첨부 유의확률 (상호간접정)
Pearson 카이제곱	28.162*	4	.000
우도비	27.956	4	.000
선형 대 선형결합	23.091	1	.000
유호 케이스 수	200		

a. 0. 설 (0%) 을 (는) 5보다 작은 기대 반도를 가지는 샘입니다. 최소 기대반도는 14.40입니다.

Pearson 카이제곱에 대한 유의확률이 0.05보다 작으므로 통계적으로 치료방법에 따른 빠속의 광물질 변화하는 비율은 차이가 있다.

범주형 자료분석 - R

교차분석(cross tabulation analysis)

```
> dsn <- read.spss(file="http://pluto.hallym.ac.kr/data/categ")
> cat <- xtabs(amount ~ treatment+bone, data=dsn)
> chisq.test(cat)
```

Pearson's Chi-squared test

data: cat

X-squared = 28.162, df = 4, p-value = 1.156e-05

범주형 자료분석

교차분석(cross tabulation analysis)

- * 동질성 검정(homogeneity test)은 독립된 그룹이 각 범주별로 동일한 반응형태를 보이는가 검정하는 통계분석기법이다.
- 직업별로 알콜중독자의 비율이 다른지 조사하여 보았다. 사무원, 교육자, 기업가, 상인들의 집단에서 임의로 표본을 추출하여 면담을 통하여 알콜중독자의 수를 조사한 결과가 아래와 같았다.

직업 \ 환자유무	알콜중독	정상	표본크기
사무원	32	268	300
교육자	51	199	250
기업인	67	233	300
상인	83	267	350
전체	233	967	1200

알콜중독자의 발생비율이 모든 4개의 직업별로 같은지 검정하자.

범주형 자료분석

교차분석(cross tabulation analysis)

동질성 검정(homogeneity test)은 독립된 그룹이 각 범주별로 동일한 반응형태를 보이는지 검정한다.

- ① 귀무가설 : 직업별 알콜중독자 발생비율은 모두 같다. 대립가설 : 직업별 알콜중독자 발생비율은 모두 같지 않다.
- ② χ^2 검정통계량 : 20.59, 유의확률 : $P[\chi^2 \geq 20.59] \leq 0.0001$
- ③ 유의수준 5%에서 귀무가설을 기각한다. 따라서 직업별 알콜중독자 비율은 모두 같지 않다.

범주형 자료분석

교차분석(cross tabulation analysis)

* 교차분석은 두 변량의 각 범주가 교차하는 빈도를 표현하며, 행별, 열별, 전체별 상대 둑수를 함께 나타낸다.

▣ 독소노출과 천식증상의 관계

		독소 노출	없음	보통	강함
		천식 증상	없음	보통	강함
천식 증상	없음	9	3	3	
	있음	1	2	7	

▣ 독소 노출과 천식 증상이 서로 연관성, 상관성이 있는지 검토가 필요

교차분석(cross tabulation analysis)

독립성 검정(independence test) : 가로 범주와 세로 범주는 서로 영향을 주는지 검정한다. 즉 두 변수는 서로 독립인지 검정한다.

- ① 귀무가설 : 독소노출과 천식증상은 서로 관련이 없다. 대립가설 : 독소노출과 천식증상은 서로 관련이 있다.
- ② χ^2 검정통계량 : 8.202, 유의확률 : $P[\chi^2 \geq 8.202] = 0.017$
- ③ 유의수준 5%에서 귀무가설을 기각한다. 따라서 독소노출과 천식증상은 서로 관련이 있다.

범주형 자료분석 – SPSS

교차분석(cross tabulation analysis)

SPSS에서 교차분석 실행 결과

천식증상 * 독소노출 교차표					
빈도		독소노출			전체
		없음	보통	강합	
천식증상	없음	9	3	3	15
	있음	1	2	7	10
	전체	10	5	10	25

카이제곱 검정			
	값	자유도	점근 유의확률 (양복검정)
Pearson 카이제곱	7.500 ^a	2	.024
우도비	8.202	2	.017
선형 대 선형결합	7.200	1	.007
유효 케이스 수	25		

a. 4 셀 (66.7%)은(는) 5보다 작은 기대 빈도를 가지는 셀입니다. 최소 기대빈도는 2.00입니다.

우도비의 유의확률이 0.05보다 작으므로 통계적으로 독소노출과 천식증상은 서로 연관성이 있다.

범주형 자료분석 - R

교차분석(cross tabulation analysis) – chisquare test

```
> dsn <- read.spss(file="http://pluto.hallym.ac.kr/data/catego  
> chisq.test(dsn$asthma, dsn$poison)
```

Pearson's Chi-squared test

```
data: dsn$asthma and dsn$poison  
X-squared = 7.5, df = 2, p-value = 0.02352
```

경고메시지(들) :

In chisq.test(dsn\$asthma, dsn\$poison) :

카이제곱 approximation은 정확하지 않을수도 있습니다

범주형 자료분석 – R

likelihood.test 함수 사용방법

```
likelihood.test(x, y=NULL, conservative=FALSE)
```

- 분할표에 대한 likelihood ratio test
- Deducer 패키지
- x : 분석할 범주형 자료를 벡터나 행렬로 입력
- y : 분석할 범주형 자료를 벡터로 입력

범주형 자료분석 - R

교차분석(cross tabulation analysis) – likelihood ratio test

```
> install.packages("Deducer")
> library(Deducer)
> likelihood.test(dsn$asthma, dsn$poison)
```

Log likelihood ratio (G-test) test of independence without correction

data: dsn\$asthma and dsn\$poison

Log likelihood ratio statistic (G) = 8.2015, X-squared df = 2,
p-value = 0.01656

표본수 계산

표본수 계산

단일 집단 표본수 계산식

$$n = \frac{(t_{n-1;1-\alpha/2} + t_{n-1;1-\beta})^2 S^2}{(\mu_0 - \mu_a)^2}$$

- $t_{n-1;1-\alpha/2}$ 는 t-분포 분위수, α 는 제 1종의 오류
- $t_{n-1;1-\beta}$ 는 t-분포 분위수, β 는 제 2종의 오류
- S 는 표준편차
- μ_0 는 귀무가설 H_0 에서 설정한 평균, μ_a 는 대립가설 H_1 에서 설정한 평균

표본수 계산 – R

power.t.test 함수 사용방법

```
power.t.test(n, power, delta, sd, sig.level, type, alternative)
```

- 한 집단, 독립인 두 집단, 대응 집단에 대한 표본수, 검정력 계산
- n : 표본수, power : 검정력. 이 두 값 중 한 값만 입력.
- delta : 평균차이 허용 값. sd : 표준편차
- sig.level : 유의수준(α)
- type : 집단은 two.sample, one.sample, paired 중 선택
- alternative : 단측(one sided)과 양측(two sided) 중 선택

표본수 계산

단일 집단 표본수 계산

- * 환자를 대상으로 한 연구에서 처음 MMSE의 복용 전과 복용 후 차이의 평균이 3.3점으로 알려져 있다. 이 때 복용 전 점수와 복용 후 점수 차이에 대한 표준편차는 5.1, 유의수준 $\alpha = 0.05$, 검정력 $1 - \beta = 0.8$ 일 때 표본수는 얼마인가?

표본수 계산

단일 집단 표본수 계산 – R

```
> power.t.test(sd=5.1,delta=3.3,type="one.sample",
+ alternative="two.side",power=.8,sig.level=0.05)
```

One-sample t test power calculation

```
n = 20.74904
delta = 3.3
sd = 5.1
sig.level = 0.05
power = 0.8
alternative = two.sided
```

표본수 계산

독립인 두 집단 표본수 계산식

$$n = \frac{2S_p^2(t_{2n-2;1-\alpha/2} + t_{2n-2;1-\beta})^2}{(\mu_{1a} - \mu_{2a})^2}$$

- $t_{2n-2;1-\alpha/2}$ 는 t-분포 분위수, α 는 제 1종의 오류
- $t_{2n-2;1-\beta}$ 는 t-분포 분위수, β 는 제 2종의 오류
- S_p 두 그룹의 합동 표준편차
- μ_{1a} 는 첫 번째 집단의 평균, μ_{2a} 는 두 번째 집단의 평균

표본수 계산

독립인 두 집단 표본수 계산

- * 임상적 유의성에 대한 기준은 MADRS 점수의 평균이 3점, 표준편차가 10점 이하로 차이가 나면 임상적으로 유의하지 않다고 알려져 있다. 대조군과 실험군을 고려하여 유의수준 $\alpha = 0.05$, 검정력 $1 - \beta = 0.8$ 일 때 표본수는 얼마인가?

표본수 계산

독립인 두 집단 표본수 계산 – R

```
> power.t.test(sd=10,delta=3,type="two.sample",
+ alternative="two.side",power=.8,sig.level=0.05)
```

Two-sample t test power calculation

```
n = 175.3851
delta = 3
sd = 10
sig.level = 0.05
power = 0.8
alternative = two.sided
```

NOTE: n is number in *each* group

표본수 계산

독립인 두 집단 표본수 계산식 – 두 집단 표본수가 다를 때
두 집단 표본수가 $n_1 = rn_2$ 일 때.

$$n_2 = \frac{(1/r + 1)S_p^2 \left(t_{(1/r+1)n_2-2;1-\alpha/2} + t_{(1/r+1)n_2-2;1-\beta} \right)^2}{(\mu_{1a} - \mu_{2a})^2}$$

- r 은 표본 배율
- $t_{(1/r+1)n_2-2;1-\alpha/2}$ 는 t-분포 분위수, α 는 제 1종의 오류
- $t_{(1/r+1)n_2-2;1-\beta}$ 는 t-분포 분위수, β 는 제 2종의 오류
- S_p 두 그룹의 합동 표준편차
- μ_{1a} 는 첫 번째 집단의 평균, μ_{2a} 는 두 번째 집단의 평균

표본수 계산

독립인 두 집단 표본수 계산 – 두 집단 표본수가 다를 때

- * 임상적 유의성에 대한 기준은 MADRS 점수의 평균이 3점, 표준편차가 10점 이하로 차이가 나면 임상적으로 유의하지 않다고 알려져 있다. 유의수준 $\alpha = 0.05$, 검정력 $1 - \beta = 0.8$ 일 때 처리군을 대조군보다 3배 더 많게 뽑으려면 각 군당 표본수는 얼마인가?

$$n_2 = \frac{(1/3 + 1) \times 10^2 (t_{(1/3+1)n_2-2;1-0.05/2} + t_{(1/3+1)n_2-2;0.8})^2}{3^2} = 117.7645$$

계산결과 대조군 118명, 처리군 $118 \times 3 = 354$
계산은 엑셀에서 해찾기 사용한다.

표본수 계산

단일 집단 비율에 대한 표본수 계산식

$$n = \frac{\left(z_{1-\alpha/2} \sqrt{p_0(1-p_0)} + z_{1-\beta} \sqrt{p_a(1-p_a)} \right)^2}{(p_0 - p_a)^2}$$

- $z_{1-\alpha/2}$ 는 정규분포 분위수, α 는 제 1종의 오류
- $z_{1-\beta}$ 는 정규분포 분위수, β 는 제 2종의 오류
- p_0 는 귀무가설 H_0 에서 비율, p_a 는 대립가설 H_1 에서 비율

표본수 계산 – R

pwr.p.test 함수 사용방법

```
pwr.p.test(h, n, sig.level, alternative)
```

- 한 집단, 독립인 두 집단, 대응 집단 비율에 대한 표본수, 검정력 계산
- h : 효과크기(effect size; Cohen, J)
- n : 표본수, power : 검정력. 이 두 값 중 한 값만 입력.
- sig.level : 유의수준(α)
- alternative : 단측(one sided)과 양측(two sided) 중 선택

표본수 계산

단일 집단 비율에 대한 표본수 계산

- * 이미 판매되는 위의 미란 치료약은 투여한 2주 후 완치율이 40%로 알려져 있다. 어느 제약회사에서 신약을 개발한 후 2주 후 완치율이 65%라고 주장하려고 한다. 유의수준 $\alpha = 0.05$, 검정력 $1 - \beta = 0.8$ 일 때 표본수는 얼마인가?

표본수 계산

단일 집단 비율에 대한 표본수 계산 – R

```
> install.packages("pwr")
> library(pwr)
> pwr.p.test(h=ES.h(0.4,0.65),sig.level=0.05,power=0.8,alternative="two.sided")
```

proportion power calculation for binomial distribution (arcsine)

```
h = 0.5060506
n = 30.64917
sig.level = 0.05
power = 0.8
alternative = two.sided
```

표본수 계산

서로 독립인 두 집단 비율에 대한 표본수 계산식

$$n = \frac{\left(z_{1-\alpha/2} \sqrt{2p_{12}(1-p_{12})} + z_{1-\beta} \sqrt{p_{1a}(1-p_{1a}) + p_{2a}(1-p_{2a})} \right)^2}{(p_{1a} - p_{2a})^2}$$

- $z_{1-\alpha/2}$ 는 정규분포 분위수, α 는 제 1종의 오류
- $z_{1-\beta}$ 는 정규분포 분위수, β 는 제 2종의 오류
- p_{12} 는 각 집단의 공통확률로 알고 있지 않은 경우에는 $p_{12} = \frac{p_1+p_2}{2}$ 로 추정하여 사용할 수 있다. 여기서 p_1 은 첫 번째 집단에서 성공 확률, p_2 는 두 번째 집단에서 성공 확률이다.
- 대립가설 가정하에 p_{1a} 는 첫 번째 집단의 성공확률, p_{2a} 는 두 번째 집단의 성공확률

표본수 계산 – R

power.prop.test 함수 사용방법

```
power.prop.test(n, p1, p2, sd, sig.level, alternative)
```

- 서로 독립인 두 집단 비율에 대한 표본수, 검정력 계산
- n : 표본수, power : 검정력. 이 두 값 중 한 값만 입력.
- p1 : 첫 번째 집단 비율. p2 : 두 번째 집단 비율
- sig.level : 유의수준(α)
- alternative : 단측(one sided)과 양측(two sided) 중 선택

표본수 계산

서로 독립인 두 집단 비율에 대한 표본수 계산

- * 앞 임상시험에서 미란 완치율은 약을 투여하고 2주 후 56%로 알려져 있다. 또한 약을 사용하지 않더라도 자연적으로 완치율이 30%이라고 알려져 있다. 대조군과 실험군을 고려하여 유의수준 $\alpha = 0.05$, 검정력 $1 - \beta = 0.8$ 일 때 표본수는 얼마인가?

표본수 계산

서로 독립인 두 집단 비율에 대한 표본수 계산 – R

```
> power.prop.test(p1=0.56, p2=0.30, sig.level=0.05, power=0.8)
```

Two-sample comparison of proportions power calculation

n = 55.72232

p1 = 0.56

p2 = 0.3

sig.level = 0.05

power = 0.8

alternative = two.sided

NOTE: n is number in *each* group

표본수 계산

독립인 두 집단 비율에 대한 표본수 계산식 – 표본수가 다를 때
두 집단 표본수가 $n_1 = rn_2$ 일 때.

$$n_2 = \frac{\left(z_{1-\alpha/2} \sqrt{(1+r)p_{12}(1-p_{12})} + z_{1-\beta} \sqrt{p_{1a}(1-p_{1a}) + rp_{2a}(1-p_{2a})} \right)^2}{r(p_{1a} - p_{2a})^2}$$

- r 은 표본 배율
- $z_{1-\alpha/2}$ 는 정규분포 분위수, α 는 제 1종의 오류
- $z_{1-\beta}$ 는 정규분포 분위수, β 는 제 2종의 오류
- p_{12} 는 각 집단의 공통확률로 알고 있지 않은 경우에는 $p_{12} = \frac{p_1+p_2}{2}$ 로 추정하여 사용할 수 있다. 여기서 p_1 은 첫 번째 집단에서 성공 확률, p_2 는 두 번째 집단에서 성공일 확률이다.
- 대립가설 가정하에 p_{1a} 는 첫 번째 집단의 성공확률, p_{2a} 는 두 번째 집단의 성공확률

표본수 계산

독립인 두 집단 비율에 대한 표본수 계산 – 표본수가 다를 때

* 앞 임상시험에서 미란 완치율은 약을 투여하고 2주 후 56%로 알려져 있다. 또한 약을 사용하지 않더라도 자연적으로 완치율이 30%이라고 알려져 있다. 조건이 유의수준 $\alpha = 0.05$, 검정력 $1 - \beta = 0.8$ 일 때 처리군을 대조군보다 3배 더 많이 뽑으려면 각 집단마다 몇 명씩 조사해야 하는가?

$$n_2 = \frac{\left(1.96\sqrt{(1+3) \times 0.43(1-0.43)} + 0.84\sqrt{0.56(1-0.56) + 3 \times 0.3(1-0.3)} \right)^2}{3 \times (0.56 - 0.3)^2} = 36.67125$$

대조군은 37명 처리군은 $37 \times 3 = 111$