

Data Analysis with R



: PCT Proportions Comparison test

Visualizations for CDA Categorical Data Analysis

최 용 석 (부산대학교 통계학과)

yschoi.pusan.ac.kr

2017.09.6.

부산대학교 피부과학교실 워크숍

1. 독립 표본의 모비율 비교

Comparison of Population Proportion of Independent Two Samples

- 최용석(2014). R과 함께하는 통계학의 이해, 교보문고 POD

독립표본의 모비율 비교 검정 단계 : 독립표본의 크기가 n_1 과 n_2 가 대표본인 경우

[1단계] H_o : 두 모집단의 비율(모비율)은 차이가 없다. $\Leftrightarrow H_o : p_1 - p_2 = 0$.

[2단계] $H_o : p_1 - p_2 = 0$ 하에 합동 추정치 계산 :

$$\hat{p} = \frac{X + Y}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

[3단계] 검정통계량 : $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$.

[4단계] 유의수준 α 에서 Z -검정

귀무가설	대립가설	기각역	p -값
	$H_1 : p_1 - p_2 > 0$	$z \geq z_\alpha$	$P(Z \geq z)$
$H_o : p_1 - p_2 = 0$	$H_1 : p_1 - p_2 < 0$	$z \leq -z_\alpha$	$P(Z \leq z)$
	$H_1 : p_1 - p_2 \neq 0$	$ z \geq z_{\alpha/2}$	$P(Z \geq z)$

참고 :

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \simeq N(0,1) \Rightarrow Z^2 \simeq X_{df=1}^2$$

[보기 1] 독감 예방 백신에 의한 남녀 항체율 비교

60세에서 65세 노인: 독감 예방 백신 접종 항체의 생성 비율이 남성 노인보다 여성 노인 쪽이 더 높다고 할 수 있는지를 유의수준 5%에서 검정하고 p -값에 의한 결과와 비교하라.

모집단	표본의 크기	항체 생성
남자	113	$X = 34$
여자	139	$Y = 54$

1) 남성 노인과 여성 노인의 항체 생성 모비율을 각각 p_1 과 p_2 :

$$H_0 : p_1 - p_2 = 0, \quad H_1 : p_1 - p_2 < 0.$$

2) $n_1 = 113, n_2 = 139$: 대표본 $\Rightarrow$ 합동 추정치: $\hat{p} = \frac{34 + 54}{113 + 139} = 0.349$

3) 검정통계량의 관측값 :

$$z = \frac{0.301 - 0.388}{\sqrt{0.349 \times (1 - 0.349)}} \sqrt{\frac{1}{113} + \frac{1}{139}} = -1.45$$

4) 유의수준 5% = 0.05에서 기각역은 $z \leq -1.64$ 이다. $z = -1.45$ 는 기각역을 벗어나므로 귀무가설 $H_0 : p_1 - p_2 = 0$ 을 기각하지 못하게 된다.

5) 여성의 항체 생성률이 남성의 항체 생성률 보다 높다고 할 수 없다.

[R-코드: 독감백신.R]

```
## 독감 예방 백신에 의한 항체율 비교
## 방법 1
z.prop = function(x1,x2,n1,n2){
  numerator = (x1/n1) - (x2/n2)
  p.common = (x1+x2) / (n1+n2)
  denominator = sqrt(p.common * (1-p.common) * (1/n1 + 1/n2))
  z.prop.ris = numerator / denominator
return(z.prop.ris)}
n1 <- 113
n2 <- 139
x <- 34
y <- 54
z=z.prop(x, y, n1, n2)
z
pvalue1=(1-pnorm(z))/2
pvalue1

## 방법 2
# Chisquare test
chi=z^2
pvalue2=(1-pchisq(chi, df=1))/2
list(chi, pvalue2)

## 방법 3
# prop.test()
prop.test(x=c(34, 54), n=c(113, 139),alternative="less", correct=F)

참고: <alt(ernative) = "two.sided", "less", "greater">
```

[결 과]

```
> list(z, pvalue1)
[[1]]
[1] -1.450804

[[2]]
[1] 0.4632914

> list(chi, pvalue2)
[[1]]
[1] 2.104833

[[2]]
[1] 0.07341719

> prop.test(x=c(34, 54), n=c(113, 139),alternative="less", correct=F)

      2-sample test for equality of proportions without continuity
      correction

data:  c(34, 54) out of c(113, 139)
X-squared = 2.1048, df = 1, p-value = 0.07342
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000  0.01068364
sample estimates:
   prop 1   prop 2 
0.3008850 0.3884892
```

[응용 1] 특이 혈관이 관찰될 질환별 비율 비교(김문범 등, 2017)

Ann Dermatol Vol. 29, No. 3, 2017

pISSN 1013-9087 • eISSN 2005-3894
https://doi.org/10.5021/ad.2017.29.3.288



ORIGINAL ARTICLE

Arborizing Vessels on Dermoscopy in Various Skin Diseases Other Than Basal Cell Carcinoma

Hyunju Jin^{1,2}, Min-Young Yang^{1,3}, Jeong-Min Kim^{1,2}, Gun-Wook Kim^{1,3}, Hoon-Soo Kim^{1,3},
Hyun-Chang Ko^{1,2}, Byung-Soo Kim^{1,3,4}, Moon-Bum Kim^{1,3,4}

Table 3. Comparison of AV characteristics in BCC and non-BCC

	BCCs (n=67)	Non-BCCs (n=57)	<i>p</i>	Benign cystic (n=14)	<i>p</i>	Benign non-cystic (n=28)	<i>p</i>	Premalignant & malignant (n=11)	<i>p</i>	Non- tumor (n=4)	<i>p</i>
Focused AV	43 (64.2)	30 (52.6)	0.19	11 (78.6)	0.30	15 (53.6)	0.33	3 (27.3)*	0.02	1 (25.0)	0.12
Unfocused AV	34 (50.7)	36 (63.2)	0.17	6 (42.9)	0.59	17 (60.7)	0.37	10 (90.9)*	0.01	3 (75.0)	0.35
Mixed AV	10 (14.9)	9 (15.8)	0.89	3 (21.4)	0.55	4 (14.3)	0.94	2 (18.2)	0.68	0 (0.0)	-
Widest diameter of the stem vessel (mm)	0.14	0.15	0.54	0.12	0.58	0.17	0.28	0.15	0.68	0.15	0.56
Widest diameter ratio of stem vessel to the first branch	1.52*	1.78*	0.03	1.84	0.25	1.75	0.10	1.79	0.14	1.76	0.68
Number of ramification	3.63	3.40	0.09	3.21*	0.03	3.50	0.49	3.36	0.25	3.32	0.70

Values are presented as number (%) or number only. AV: arborizing vessel, BCC: basal cell carcinoma. *Significant findings ($p < 0.05$).

[R-코드: 특이혈관.R]

```
##Arborizing Vessels ... 특이 혈관이 관찰될 질환별 비율
z.prop = function(x1,x2,n1,n2){
  numerator = (x1/n1) - (x2/n2)
  p.common = (x1+x2) / (n1+n2)
  denominator = sqrt(p.common * (1-p.common) * (1/n1 +
1/n2))
  z.prop.ris = numerator / denominator
return(z.prop.ris)}

## Focused AV : BCCs vs. Non-BCCs
n1 <- 67
n2 <- 57
x <- 43
y <- 30

# Z-test
z=z.prop(x, y, n1, n2)
pvalue1=1-pnorm(z)
list(z, pvalue1)

# Chisquare test
chi=z^2
pvalue2=1-pchisq(chi, df=1)
list(chi, pvalue2)
# prop.test()
prop.test(x=c(43, 30), n=c(67, 57),alternative="two.sided",
correct=F)

# Unfocused AV : BCCs vs. Non-BCCs
prop.test(x=c(34, 36), n=c(67, 57),alternative="two.sided",
correct=F)

# Mixed AV : BCCs vs. Non-BCCs
prop.test(x=c(10, 9), n=c(67, 57),alternative="two.sided",
correct=F)
```

[결 과]

```
> list(z, pvalue1)
[[1]]
[1] 1.302351

[[2]]
[1] 0.09639819

>
> # Chisquare test
> chi=z^2
> pvalue2=1-pchisq(chi, df=1)
> list(chi, pvalue2)
[[1]]
[1] 1.696118

[[2]]
[1] 0.1927964

> # prop.test()
> prop.test(x=c(43, 30), n=c(67, 57),alternative="two.sided", correct=F)

      2-sample test for equality of proportions without continuity
      correction

data:  c(43, 30) out of c(67, 57)
X-squared = 1.6961, df = 1, p-value = 0.1928
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.05768053  0.28863104
sample estimates:
   prop 1    prop 2 
0.6417910 0.5263158
```

[응용 2] 피부암의 조직학적 침윤도 평가(김문범 등, 2015)

Ann Dermatol Vol. 27, No. 3, 2015

<http://dx.doi.org/10.5021/ad.2015.27.3.1>

ORIGINAL ARTICLE

Usefulness of Dermatoscopy for the Preoperative Assessment of the Histopathologic Aggressiveness of Basal Cell Carcinoma

Hoon-Soo Kim¹, Jung-Min Park¹, Je-Ho Mun¹, Margaret Song¹, Hyun-Chang Ko^{1,2},
Byung-Soo Kim^{1,2}, Moon-Bum Kim^{1,2}

Table 1. Comparison of classic, nonclassic and other basal cell carcinoma dermatoscopic patterns between non-aggressive and aggressive types*

Variable	Non-aggressive type (n = 105)	Aggressive type (n = 40)	Total (n = 145)	χ^2 statistic p-value	Kappa coefficient (κ)
Classic					
Large blue-gray ovoid nests	85 (80.9)	21 (52.5)	106 (73.1)	0.001	0.774
Multiple blue-gray globules	57 (54.3)	29 (72.5)	86 (59.3)	0.046	0.928
Maple leaf-like areas	22 (20.9)	6 (15.0)	28 (19.3)	0.417	
Spoke wheel areas	9 (8.6)	3 (7.5)	12 (8.3)	0.834	
Arborizing telangiectasia	48 (45.7)	29 (72.5)	77 (53.1)	0.004	0.834
Ulceration	65 (61.9)	25 (62.5)	90 (62.1)	0.947	
Nonclassic					
Short fine superficial telangiectasia	33 (31.4)	9 (22.5)	42 (29.0)	0.289	
Multiple small erosions	13 (12.4)	3 (7.5)	16 (11.0)	0.557	
Concentric structure	12 (11.4)	11 (27.5)	23 (15.9)	0.018	0.868
Multiple blue-gray dots	28 (26.7)	13 (32.5)	41 (28.3)	0.486	
Other					
Non-arborizing vessels	2 (1.9)	3 (7.5)	5 (3.4)	0.129	
Brown-black dots	3 (2.8)	4 (10.0)	7 (4.8)	0.092	
Blue-white veil	43 (40.9)	10 (25.0)	53 (36.6)	0.075	
Pigment network	2 (1.9)	1 (2.5)	3 (2.1)	0.822	

Values are presented as number (%). * χ^2 test, $p < 0.05$.

[R-코드: 피부암침윤도.R]

```
## Usefulness of Dermatoscopy... 피부암의 조직학적  
침윤도 평가
```

```
# Non-aggressive vs. Aggressive  
n1 <- 105  
n2 <- 40
```

```
## Classic  
# Large blue-gray ovoid nests  
prop.test(x=c(85, 21), n=c(105,  
40),alternative="two.sided", correct=F)
```

```
# Ulceration  
prop.test(x=c(65, 25), n=c(105,  
40),alternative="two.sided", correct=F)
```

```
## Nonclassic  
# Short fine superficial telangiectasia  
prop.test(x=c(33, 9), n=c(105,  
40),alternative="two.sided", correct=F)
```

```
# Multiple blue-gray dots  
prop.test(x=c(28, 13), n=c(105,  
40),alternative="two.sided", correct=F)
```

[결 과]

data: c(85, 21) out of c(105, 40)
X-squared = 11.9257, df = 1, p-value = 0.0005536

data: c(65, 25) out of c(105, 40)
X-squared = 0.0044, df = 1, p-value = 0.9474

data: c(33, 9) out of c(105, 40)
X-squared = 1.1223, df = 1, p-value = 0.2894

data: c(28, 13) out of c(105, 40)
X-squared = 0.486, df = 1, p-value = 0.4857

2. 범주형자료의 시각화

Visualizations for CDA

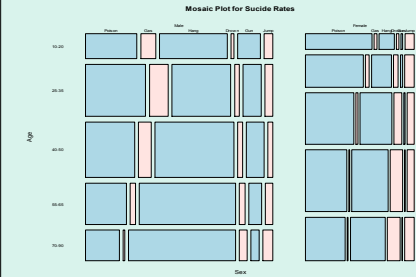
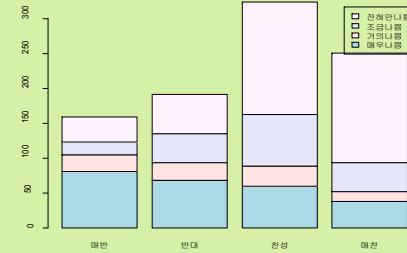
- 최용석(2017). R과 함께하는 다변량 자료분석, 경문사

막대 그림

Bar Plot

모자이크그림

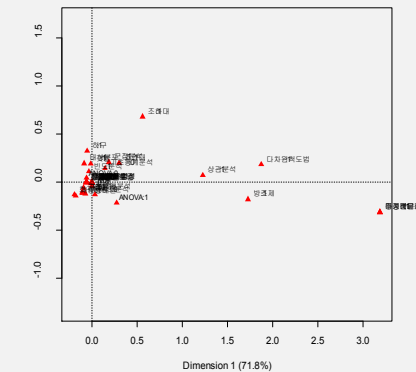
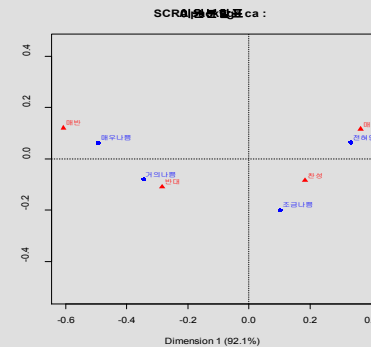
Mosaic Plot



단순 대응분석 그림

Simple Correspondence Analysis Map

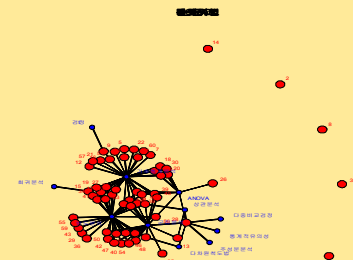
이원 분할표



다중 대응분석 그림

Multiple Correspondence Analysis Map

삼원, 오원, 다원 분할표, 버트행렬 자료, 비정형화 자료



사회네트워크분석

Social Network Analysis

워드 클라우드

Word Cloud

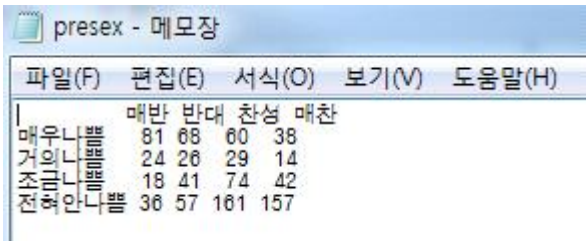
2.1 단순 대응분석 그림 Simple Correspondence Analysis Map

이원분할표 Tow-Way Contingency Table

presex.txt

[자료 2.1] 결혼 전 성관계에 대한 반응과 피임허용의 이원분할표

혼전 성관계	10대 임신조절			
	매우 반대	반	대	찬성
매우 나쁘다	81	68	60	38
거의 나쁘다	24	26	29	14
조금 나쁘다	18	41	74	42
전혀 나쁘지 않다	36	57	161	157



	매반	반대	찬성	매찬
매우나쁨	81	68	60	38
거의나쁨	24	26	29	14
조금나쁨	18	41	74	42
전혀안나쁨	36	57	161	157

presex-SCRA.R

Simple CRA ca() : Matrix for Two-Way Table

```
O<-matrix(c(81, 68, 60, 38,
            24, 26, 29, 14,
            18, 41, 74, 42,
            36, 57, 161, 157), byrow=T, nrow=4)
rownames(O)<-c("매우나쁨", "거의나쁨", "조금나쁨", "전혀안나쁨")
colnames(O)<-c("매반", "반대", "찬성", "매찬")
library(ca)
sca=ca(O)
par(pty="s")
plot(sca, main="SCRA package ca : 이원분할표")
```

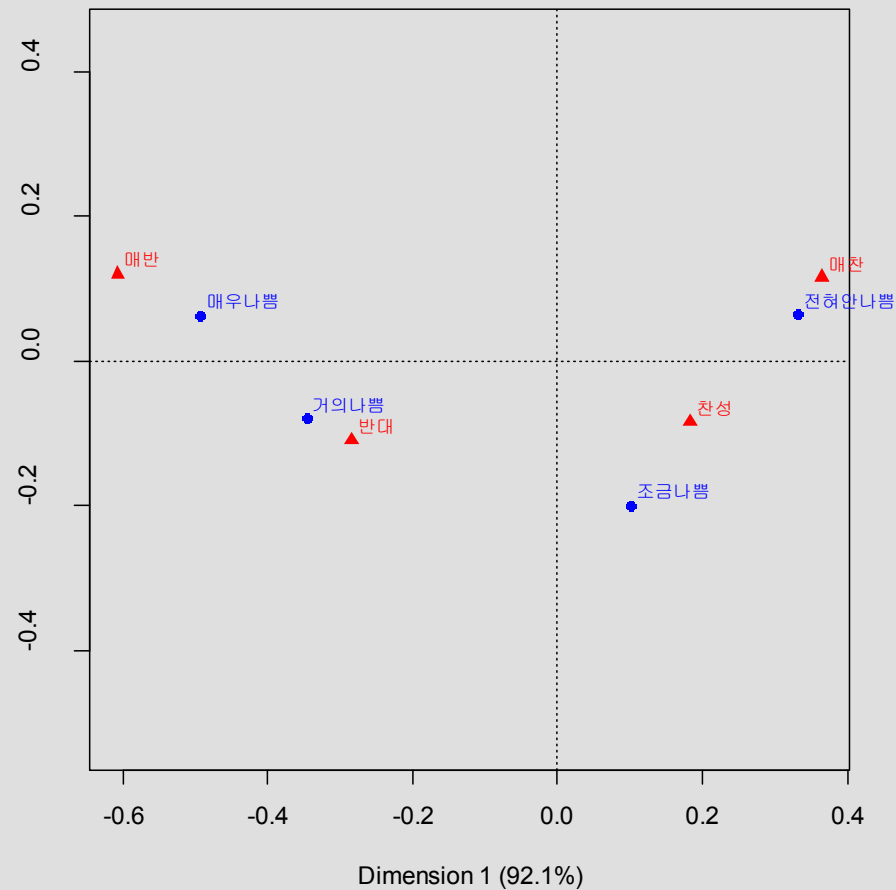
Simple CRA ca() : Text for Two-Way Table

```
setwd("C:/2017빅데이터/R_data")
O<-read.table("presex.txt", header=T)
library(ca)
sca=ca(O)
sca
win.graph()
par(pty="s")
plot(sca, main="SCRA package ca : 이원분할표")
```

presex - 메모장

파일(F)	편집(E)	서식(O)	보기(V)	도움말(H)
매우나쁨	81	68	60	38
거의나쁨	24	26	29	14
조금나쁨	18	41	74	42
전혀안나쁨	36	57	161	157

SCRA package ca :



- 수평축(Dim1)의 왼편 : 행범주 (매우나쁨, 거의나쁨)은 열범주 (매반, 반대)방향으로 놓여져 있다. 이는 혼전 성관계가 나쁘다고 여기는 사람들은 10대에 피임을 반대하는 입장에 있음을 나타낸다.

- 오른편 : 행범주 (조금나쁨, 전혀안나쁨)은 열범주 (찬성, 매찬) 방향으로 놓여져 있어 이를 서로 대응한다고 한다. 이는 혼전 성관계에 대해 부정적이지 않은 사람들은 10대에 피임을 찬성하고 있다.

- 혼전 성경험에 대해서도 관대한 사람들은 10대에게 피임을 허용해야 한다고 좀 더 강하게 느끼는 입장을 보이고 있다.

2.2 다원 분할표 Multi-Way Contingency Table

ptsd.txt

[자료 2.1] 외상 후 스트레스증후군에 대한 다시점자료

subject	time	ptsd	control	problems	sevent	cohes
15	1	0	3.22222	5.625	1	8
15	2	0	3.16667	5.375	0	8
15	3	0	3.27778	3.750	1	8
18	1	1	2.55556	9.250	0	8
18	2	0	3.44444	4.375	0	8
18	3	0	3.33333	2.375	0	8

----- 중 간 생 략 -----

571	1	0	3.55556	3.000	0	7
571	2	0	2.94444	1.875	0	7
571	3	0	3.50000	2.750	0	7

ptsd - 메모장

파일(F)	편집(E)	서식(O)	보기(V)	도움말(H)		
개체	시점	스트레스증후	자아통제력	인생문제수	스트레스수	가족결속력
15	1	0	3.22222	5.625	1	8
15	2	0	3.16667	5.375	0	8
15	3	0	3.27778	3.750	1	8
18	1	1	2.55556	9.250	0	8
18	2	0	3.44444	4.375	0	8
18	3	0	3.33333	2.375	0	8
19	1	1	2.72222	7.750	1	7
19	2	1	2.77778	7.750	1	7
19	3	0	2.77778	7.500	1	7

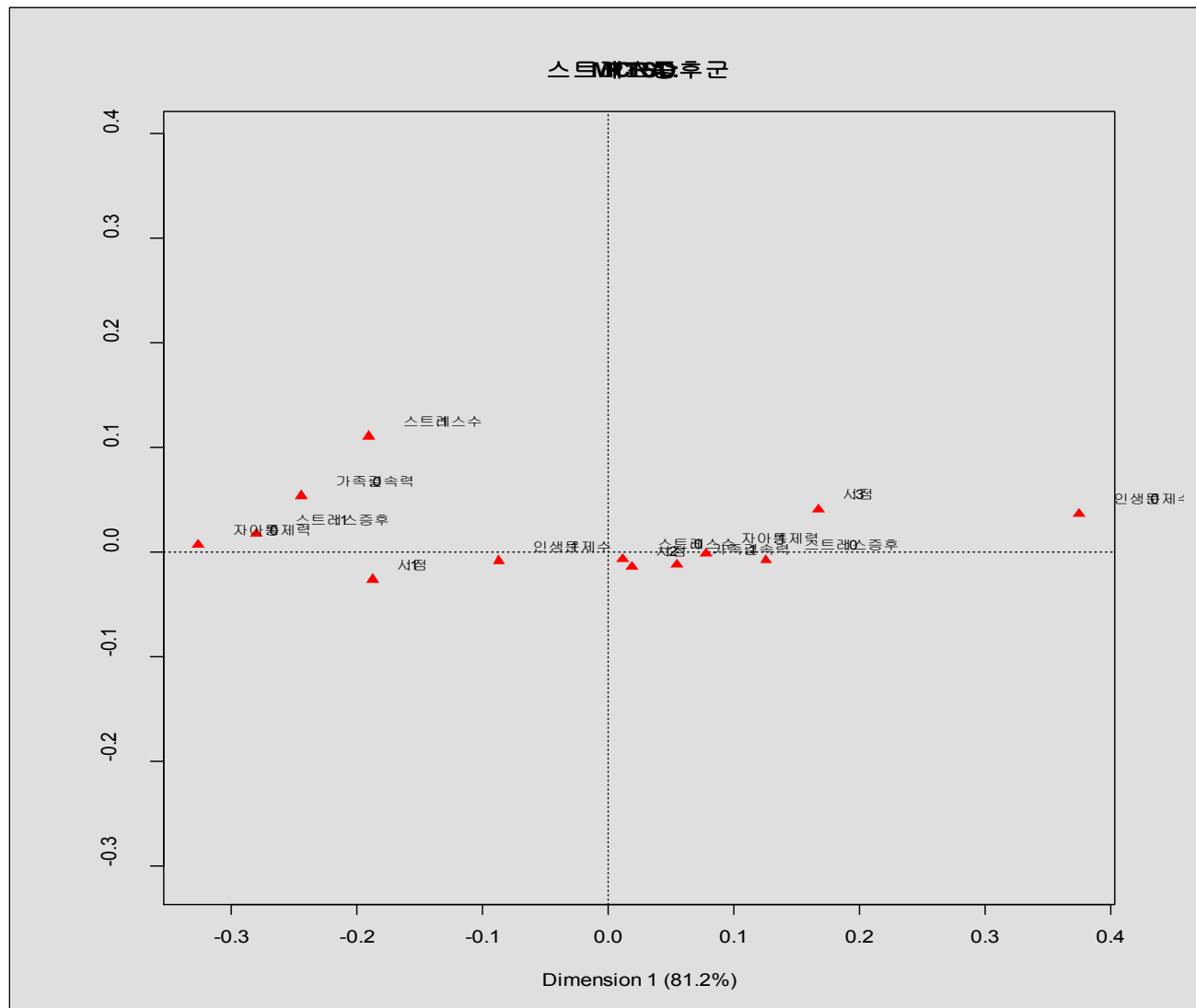
pstd-MCRA.R

setwd("C:/2017빅데이터/R_data")
ptsd<-read.table("ptsd.txt", header=T)
v1 <- as.numeric(ptsd\$자아통제력>=3)
v2 <- as.numeric(ptsd\$인생문제수>=3)
v3<- as.numeric(ptsd\$스트레스수>=3)
v4 <- as.numeric(ptsd\$가족결속력>=6)
nptsd <- data.frame(ptsd[,1:3],
 자아통제력=v1,인생문제수=v2, 스트레스수=v3, 가족결속
 력=v4)
nptsd<-nptsd[, -1]
library("ca")
mjca(nptsd)
par(pty="s")
plot(mjca(nptsd), main="MCRA : 스트레스증후군 PTSD")

```

setwd("C:/2017빅데이터/R_data")
ptsd<-read.table("ptsd.txt", header=T)
v1 <- as.numeric(ptsd$자아통제력>=3)
v2 <- as.numeric(ptsd$인생문제수>=3)
v3<- as.numeric(ptsd$스트레스수>=3)
v4 <- as.numeric(ptsd$가족결속력>=6)
nptsd <- data.frame(ptsd[,1:3],
  자아통제력=v1,인생문제수=v2, 스트레스수=v3, 가족결속
  력=v4)
nptsd<-nptsd[, -1]
library("ca")
mjca(nptsd)
par(pty="s")
plot(mjca(nptsd), main="MCRA : 스트레스증후군 PTSD")

```



- 제1축 원편 : 시점:1(3개월 후), 자아통제력:0, 가족결속력:0, 인생문제수:1, 스트레스수:1 서로 대응한다.
- 환자가 외상 3개월 후 자아통제력과 가족결속력이 낮으며 인생문제와 스트레스가 높음을 보여주고 있다.
- 제2축의 오른편 : 외상 후 환자가 시점2-3(3-12개월 후)에서 자아통제력(1), 가족결속력(1)이 높고, 인생문제수(0)와 스트레스(0)가 낮아졌음을 보여주고 있다.