HW4. for Multivariate Statistics II

November 10, 2020

201611531/Department of Statistics/Jeong Hojae

Chapter 8. Correspondence Analysis(CRA)

1. [Exercise 8.7] [Table 1] is a two-way table of the frequency and age of women's breast diagnosis

[Table 1]

Age	Frequency of Breast Self Diagnosis		
	monthly	frequently	do not
< 45	91	90	51
45 −49	150	200	155
50 +	109	198	172

(1) Make a null hypothesis for [Table 1] and apply the Chi-square test for this null hypothesis.

H0 : The two variables 'Age' and 'Frequency of Breast Self Diagnosis' are independent of each other. (not relevant)

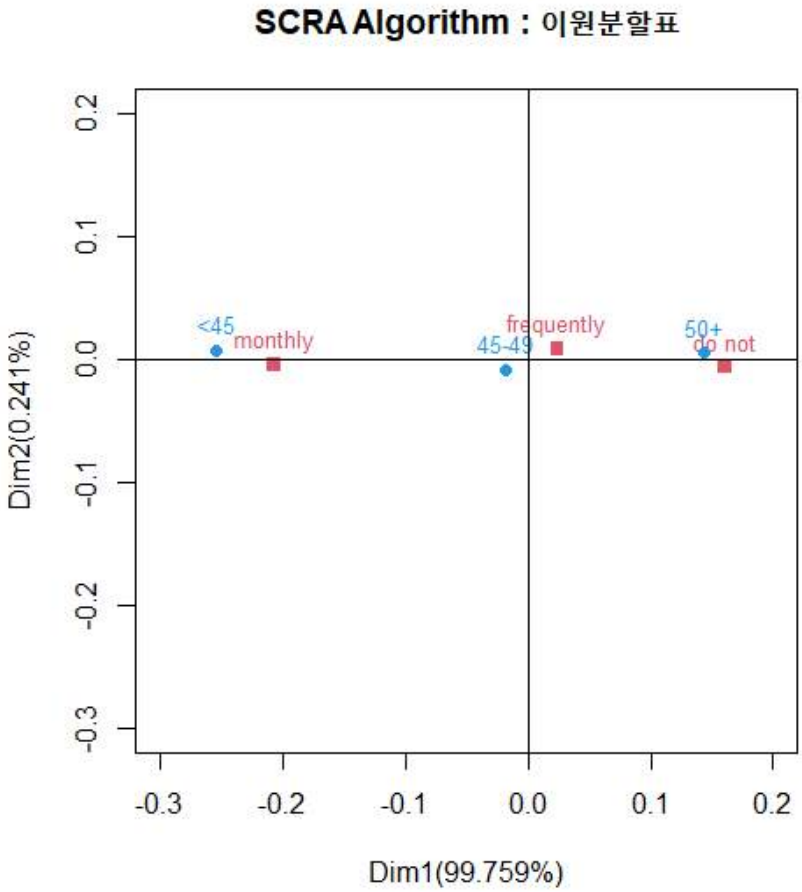
```
> chisq.test(O)

        Pearson's Chi-squared test

data:  O
X-squared = 25.086, df = 4, p-value = 4.835e-05
```

P-value is 4.835e-05. It is small than 0.05 (significance level). We can reject H0. So, the two variables 'Age' and 'Frequency of Breast Self Diagnosis' are related to each other.

(2) Apply the simple CRA according to the Simple CRA algorithm [Table 8.2.2].

```
# [Step 1] n*p two-way table data matrix:
> O
     [,1] [,2] [,3]
[1,]   91   90   51
[2,]  150  200  155
[3,]  109  198  172

# [Step 2] Corresponence Matrix:
> F
           [,1]       [,2]       [,3]
[1,] 0.07483553 0.07401316 0.04194079
[2,] 0.12335526 0.16447368 0.12746711
[3,] 0.08963816 0.16282895 0.14144737

# Row and Column centroids
> r;c;
[1] 0.1907895 0.4152961 0.3939145
[1] 0.2878289 0.4013158 0.3108553

# Centred correspondence matrix
> cF
              [,1]          [,2]          [,3]
[1,]  0.019920793 -0.002553670 -0.017367123
[2,]  0.003821037 -0.002191179 -0.001629858
[3,] -0.023741830  0.004744849  0.018996981

# [Step 3] SVD of residual matrix
# marginal sum of rows and columns f
> Dr;Dc
         [,1]     [,2]     [,3]
[1,] 2.289406 0.000000 0.000000
[2,] 0.000000 1.551748 0.000000
[3,] 0.000000 0.000000 1.593305
         [,1]     [,2]     [,3]
[1,] 1.863944 0.000000 0.000000
[2,] 0.000000 1.578545 0.000000
[3,] 0.000000 0.000000 1.793581

# [Step 4] Coordinates of rows and columns
of Simple CRA Map:
> A;B
            [,1]          [,2]
<45    -0.2548027  0.007356626
45-49  -0.0185402 -0.008323941
50+     0.1429583  0.005212637
                [,1]          [,2]
monthly     -0.2068429 -0.004437538
frequently   0.0237898  0.008539783
do not       0.1608084 -0.006916074

# [Step 5] Goodness-of fit of
s(>=2)-dimensional Simple CRA Map:

# eigenvalue and GOF
> rbind(round(eig, 3),round(per, 3))
       [,1]  [,2] [,3]
[1,]  0.021 0.000    0
[2,] 99.759 0.241    0

1-dimsion's GOF is 99.759.
2-dimsion's GOF is 0.241.
```

(3) Obtain and interpret the simple CRA Plot.



Relative positions of the points : similarities and differences among rows and columns categories.

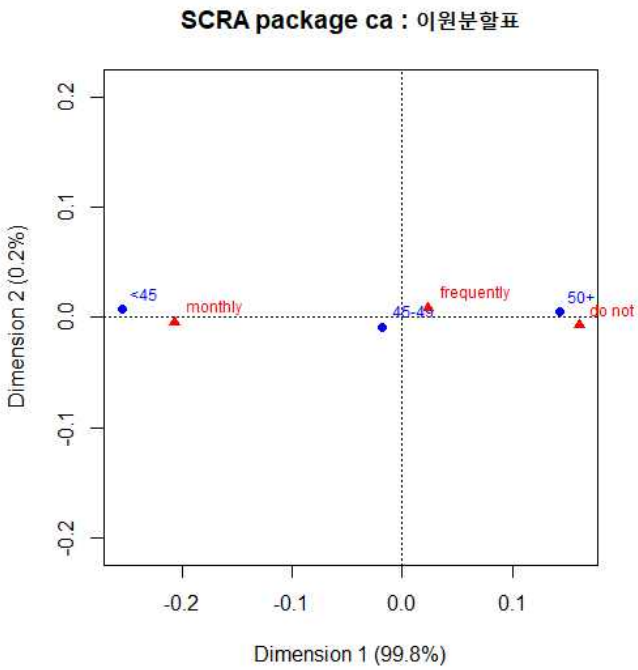
"<45" and "45-49" of Age are relatively similar group corresponding to "monthly" of frequency of breast self-diagnosis because they have same direction.

"50+" of Age is relatively similar group corresponding to "do not" and "frequently" of frequency of breast self-diagnosis because they have same direction.

We can see that frequency of breast self-diagnosis is high when their age is less than 50.

(4) Apply the simple CRA using the function ca() of R's library(ca).

```
> sca

 Principal inertias (eigenvalues):
             1         2
Value        0.02058 5e-05
Percentage 99.76%   0.24%

 Rows:
                <45       45-49        50+
Mass       0.190789   0.415296 0.393914
ChiDist    0.254909   0.020323 0.143053
Inertia    0.012397   0.000172 0.008061
Dim. 1    -1.776152  -0.129238 0.996519
Dim. 2     1.042431  -1.179499 0.738629

 Columns:
            monthly frequently     do not
Mass       0.287829   0.401316   0.310855
ChiDist    0.206890   0.025276   0.160957
Inertia    0.012320   0.000256   0.008053
Dim. 1    -1.441839   0.165831   1.120947
Dim. 2    -0.628797   1.210084  -0.980005
```



The use of R's function ca() for the CRA gives the same result as (2) and (3).

2. Consider a three-way [Table 2] for AIDS Symptoms by AZT Use and Race.

[Table 2]

Race	AZT Use	AIDS Symptoms	
		Yes	No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

(1) Make a null hypothesis for [Table 2] and test this null hypothesis.

H0 : The three variables 'Race' and 'AZT USE', 'AIDS Symptoms' are independent of each other. (not relevant)

```
> mantelhaen.test(AZT)

        Mantel-Haenszel chi-squared test with continuity correction

data:  AZT
Mantel-Haenszel X-squared = 6.0799, df = 1, p-value = 0.01367
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.2818621 0.8414208
sample estimates:
common odds ratio
        0.4869955
```

P-value is 4.835e-05. It is small than 0.05 (significance level). We can reject H0. So, the two variables 'Age' and 'Frequency of Breast Self Diagnosis' are related to each other.

(2) How can you make a 2-ways table [Table 3] from [Table 2]?

[Table 3]

Race	AZT Use	AIDS Symptoms	
		Yes	No
White + Black	Yes	25	145
	No	44	124

[Table 3] does not divide table according to race compared to [Table 2]
Ignore race variables and add values corresponding to each variable.

```
> apply(AZT,c(1,2),sum)
        AIDS_Symptoms
AZT_Use Yes  No
    Yes  25 145
    No   44 124
```

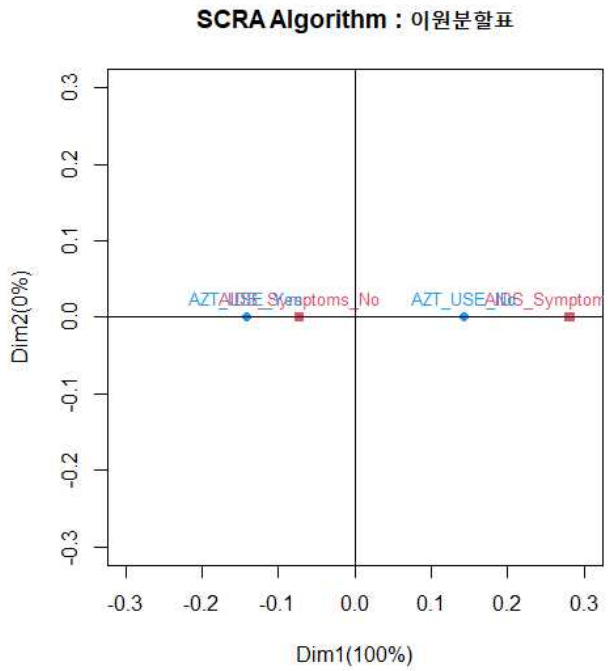
(3) Make a null hypothesis and apply the Chi-square test for [Table 3].

H0 : The two variables ‘AZT Use’ and ‘AIDS Symptoms’ are independent of each other. (not relevant)

```
> chisq.test(AZT_sum)

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

data:  AZT_sum
X-squared = 6.171, df = 1, p-value = 0.01299
```

P-value is 0.01299. It is small than 0.05 (significance level). We can reject H0. So, the two variables ‘AZT Use’ and ‘AIDS Symptoms’ are related to each other.

(4) Obtain and interpret the simple CRA Plot for [Table 3].



Relative positions of the points : similarities and differences among rows and columns categories.

"AZT_USE_No" is relatively similar group corresponding to "AIDS_Symptoms_Yes" because they have same direction.

"AZT_USE_Yes" is relatively similar group corresponding to "AIDS_Symptoms_No" because they have same direction.

We can see that AZT is effective in eliminating AIDS symptoms.

```
#####################################
#1-1
O<-matrix(c(91, 90, 51,
            150, 200, 155,
            109, 198, 172), byrow=T, nrow=3)
chisq.test(O)

#1-2
F <- O/sum(O)
r <- apply(F,1,sum)
c <- apply(F,2,sum)
r;c;

Dr<- diag(1/sqrt(r))
Dc<- diag(1/sqrt(c))
Dr;Dc
cF<- F-r%*%t(c)
cF
Y <- Dr%*%(cF)%*%Dc
svd.Y <- svd(Y)
U <- svd.Y$u
V <- svd.Y$v
D <- diag(svd.Y$d)

A <- (Dr%*%U%*%D)[,1:2]
B <- (Dc%*%V%*%D)[,1:2]
rownames(A) <- c("<45", "45-49", "50+")
rownames(B) <- c("monthly", "frequently", "do
not")
A;B

eig <- (svd.Y$d)^2
per <- eig/sum(eig)*100
gof <- sum(per[1:2])
rbind(round(eig, 3),round(per, 3))

#1-3
par(pty="s")
lim <-range(pretty(A))
plot(B[, 1:2],
xlab="Dim1(99.759%)",ylab="Dim2(0.241%)",xlim
=lim,ylim=lim,pch=15,col=2,
     main="SCRA Algorithm : 이원분할표")
text(B[, 1:2],rownames(B),cex=0.8,col=2,pos=3)
points(A[, 1:2],pch=16, col=4)
text(A[, 1:2],rownames(A),cex=0.8,pos=3,
col=4)
abline(v=0,h=0)

#1-4
O<-matrix(c(91, 90, 51,
            150, 200, 155,
            109, 198, 172), byrow=T, nrow=3)
rownames(O)<-c("<45", "45-49", "50+")
colnames(O)<-c("monthly", "frequently", "do
not")
O
install.packages("ca")
library(ca)
sca=ca(O)
sca
par(pty="s")
plot(sca, main="SCRA package ca :
이원분할표")

#####################################
#2-1
AZT<-array(c(14,32,93,81,11,12,52,43),
           dim=c(2, 2, 2),
           dimnames=list(AZT_Use=c("Yes",
"No"),
                         AIDS_Symptoms
=c("Yes", "No"),
                         response
=c("White", "Black")))

mantelhaen.test(AZT)

#2-2
apply(AZT,c(1,2),sum)
AZT_sum=apply(AZT,c(1,2),sum)

#2-3
chisq.test(AZT_sum)
```

```
#2-4
O=AZT_sum
F <- O/sum(O) #상대도수의 이원분할표로 이용
r <- apply(F,1,sum)
c <- apply(F,2,sum)

Dr<- diag(1/sqrt(r))
Dc<- diag(1/sqrt(c))
r;c;Dr;Dc
cF<- F-r%*%t(c)
Y <- Dr%*%(cF)%*%Dc
svd.Y <- svd(Y)
U <- svd.Y$u
V <- svd.Y$v
D <- diag(svd.Y$d)

A <- (Dr%*%U%*%D)[,1:2]
B <- (Dc%*%V%*%D)[,1:2]
rownames(A) <- c("AZT_USE_Yes",
"AZT_USE_No")
rownames(B) <- c("AIDS_Symptoms_Yes",
"AIDS_Symptoms_No")
A;B

eig <- (svd.Y$d)^2
per <- eig/sum(eig)*100
gof <- sum(per[1:2])
rbind(round(eig, 3),round(per, 3))

par(pty="s")
lim <-range(c(-0.3,0.3))
plot(B[, 1:2],
xlab="Dim1(100%)",ylab="Dim2(0%)",xlim=lim,yli
m=lim,pch=15,col=2,
      main="SCRA Algorithm : 이원분할표")
text(B[, 1:2],rownames(B),cex=0.8,col=2,pos=3)
points(A[, 1:2],pch=16, col=4)
text(A[, 1:2],rownames(A),cex=0.8,pos=3,
col=4)
abline(v=0,h=0)
```