

# 계층적 군집분석

계층적 군집분석(cluster analysis)은 처음 N개의 모든 샘플에서 시작하여, 점차 군집의 개수를 줄여나가는 계층적 군집방법입니다. 처음에는 모든 관측 값들이 모든 군집으로 할당 된 후, 거리측정 방식에 따라 각 그룹 간의 유사성을 측정하여, 최종 두 개의 군집만 남을 때까지 수행됩니다. 범주형 변수, 연속형 변수 모두 포함 가능하나 범주형 변수의 경우, 관측벡터 간들의 거리측정시 'gower' 알고리즘이 사용됩니다. 비지도 분류분석에서 가장 많이 사용되는 K-평균 군집분석(cluster analysis)과 가장 큰 차이점은, 범주형 변수도 선택 가능하다는 것과, 최종군집 수가 필요하지 않다는 것입니다. 출력옵션 탭에 있는 '군집의 수' 는 계층적분석의 결과에는 아무런 영향을 주지 않으며, 최종 보고싶은 군집들을 시각화해 주기 위한 것입니다.

## 메뉴 호출하기

- 고급분석 > 분류분석 > 비지도 학습 > 계층적군집



## • 변수설정 탭

계층적 군집분석

변수설정 분석옵션 출력옵션

① 입력 데이터 형식 (데이터 외의 경우 작업기록 기능에서 제외)

☒ 데이터 ☐ 거리행렬

데이터

전체변수

id  
lowbw  
preterm  
matage  
hyp  
sex

>

<

양적변수(선택-1개이상가능)

bweight  
gestwks

③ ☒ 양적변수 표준화

④ 질적변수(선택-1개이상가능)

>

<

도움말 재설정 확인 취소

메뉴 요소	설명
① 입력 데이터 형식	<p>데이터와 거리행렬 두 가지 중 하나를 택합니다.</p> <ul style="list-style-type: none"> <li>데이터 (Default) : 엑셀 스프레드 시트에 있는 데이터에서 변수를 택하여 분석하고자 할 때 선택합니다</li> <li>거리행렬 : 데이터가 거리행렬인 경우에 사용합니다. 거리행렬을 선택할 경우, [데이터]-'양적변수', '질적변수'가 비활성화 됩니다.</li> </ul>
② 양적변수	<p>군집분석에 사용할 변수를 지정합니다. 질적변수는 선택할 수 없습니다. 적어도 하나 이상의 양적변수를 지정해야 합니다.</p>
③ 양적변수 표준화	<p>양적변수가 1개 이상 선택된 경우 활성화됩니다. 군집분석 시, 표준화된 데이터 값을 사용합니다.</p>
④ 질적변수	<p>질적변수를 지정합니다. 질적변수로 선택한 변수들은 문자로 인식되어 분석에 사용됩니다. 질적변수로 양적변수를 선택할 수 없으며, 선택된 경우 분석에서 제외됩니다. 질적변수를 선택한 경우 [거리행렬 계산 방법]에서 'Gower'의 거리만 활성화됩니다.</p>

## • 분석옵션 탭

계층적 군집분석

변수설정 [분석옵션] 출력옵션

① 거리행렬 계산방법

☒ Euclidean      ☐ Manhattan  
☐ Maximum      ☐ Minkowski  
☐ Gower      Minkowski power

② 군집의 수

③ 군집 기준

☒ 케이스      ☐ 변수

④ 군집 알고리즘

☒ Complete      ☐ Average  
☐ Single      ☐ Median  
☐ Ward.D      ☐ Mcquitty  
☐ Ward.D2      ☐ Centroid

메뉴 요소	설명
① 거리행렬 계산방법	<p>관측값 간의 거리계산 방법을 지정합니다.</p> <ul style="list-style-type: none"> <li>Euclidean (Default) : 두 점 사이의 거리를 구할 때 가장 많이 쓰는 방식입니다. <math>d = \sqrt{\sum  P_i - Q_i ^2}</math></li> <li>Manhattan : 두 점 사이의 절대적 거리를 이용한 거리 계산 방식입니다. <math>d = \sum  P_i - Q_i </math></li> <li>Maximum : 두 점 사이의 거리가 좌표 차원에서의 가장 큰 벡터공간에서 정의됩니다.</li> <li>Minkowski : power에 입력된 값은 수식 상에 <math>p</math>로 반영됩니다. <math>d = (\sum  P_i - Q_i ^p)^{\frac{1}{p}}</math></li> <li>Minkowski power : Minkowski를 선택할 경우 활성화됩니다. 1 이상의 정수를 입력할 수 있으며, Default는 3입니다.</li> <li>Gower : 질적변수가 포함되어 있을 때 사용 가능한 방법입니다. 양적변수만 존재할 때에도 사용이 가능합니다. 선택된 변수들을 [0, 1] 사이의 값으로 표준화 시킨 후, 모든 변수들 간의 거리를 가중평균하여 합한 값을 사용합니다.</li> </ul>
② 군집의 수	<p>최종적으로 보고싶은 군집의 수를 지정해줍니다. 결과 출력물 창에 선택된 군집의 수만큼 묶일 수 있도록 시각화 합니다. 2 이상의 정수만 입력 가능하며, Default는 2입니다.</p>
③ 군집 기준	<p>[변수설정]-'데이터' 선택 시 활성화됩니다. 군집 기준을 선택합니다.</p> <ul style="list-style-type: none"> <li>케이스 (Default) : 군집 기준으로 각 케이스를 사용합니다.</li> <li>변수 : 군집 기준으로 각 변수를 사용합니다.</li> </ul>

- 분석옵션 탭

계층적 군집분석

변수설정 | **분석옵션** | 출력옵션

① 거리행렬 계산방법

☒ Euclidean      ☐ Manhattan  
☐ Maximum      ☐ Minkowski  
☐ Gower      Minkowski power

② 군집의 수

③ 군집 기준

☒ 케이스      ☐ 변수

④ 군집 알고리즘

☒ Complete      ☐ Average  
☐ Single      ☐ Median  
☐ Ward.D      ☐ Mcquitty  
☐ Ward.D2      ☐ Centroid

## ④ 군집 알고리즘

메뉴 요소	설명
	<p>군집 간의 거리 계산에 사용할 알고리즘을 선택합니다.</p> <ul style="list-style-type: none"> <li>Complete (Default) : 최장연결법으로 두 군집 간의 최장 거리를 군집 간 거리로 정의합니다.</li> <li>Single : 최단연결법으로 두 군집 간의 최단 거리를 군집 간 거리로 정의합니다.</li> <li>Ward.D : 군집 내의 편차제곱합에 근거를 두고 군집을 병합합니다. 군집을 병합하는 과정에서 생기는 정보의 손실이 최소가 되도록 정의합니다.</li> <li>Ward.D2 : Ward.D 방법에 표준화 수치를 사용한 것으로 절대값 대신 거듭제곱값을 사용합니다.</li> <li>Average : 평균연결법으로 각 군집에 속한 모든 개체들 간의 거리의 평균으로 정의합니다.</li> <li>Median : 중앙연결법으로 군집 간의 거리를 군집의 모든 샘플의 중앙값으로 정의합니다.</li> <li>Mcquitty : 산술평균을 이용한 가중 쌍그룹 방법으로 가장 가까운 두 군집이 합쳐져 하나의 그룹을 형성한 후, 다른 군집과의 거리는 산술평균으로 구합니다..</li> <li>Centroid : 중심연결법으로 두 군집 간의 거리가 중심간 거리로 정의됩니다.</li> </ul>

## 출력옵션 탭

계층적 군집분석

변수설정 분석옵션 **출력옵션**

① 출력

☐ 기술통계량 ☒ 수평덴드로그램

☐ 분산분석표 ☒ 수직덴드로그램

☒ Silhouette plot

② 최적 군집수 탐색

☐ Within cluster sum of squares

☒ Silhouette

☐ Dunn Index

③ 최대 군집의 수

④ 저장

☐ 최종군집

☐ 최종군집중심으로부터의 거리

도움말 재설정 **확인** 취소

메뉴 요소	설명
① 출력	<p>선택한 내용을 출력합니다.</p> <ul style="list-style-type: none"> <li>기술통계량 : 최종적으로 선정된 군집의 기술통계량을 출력합니다.</li> <li>분산분석표 : 분산분석표를 출력합니다.</li> <li>Silhouette plot : 실루엣도표를 출력합니다.</li> <li>수평덴드로그램 : 수평적으로 그린 덴드로그램을 출력합니다.</li> <li>수직덴드로그램 : 수직으로 그린 덴드로그램을 출력합니다.</li> </ul>
② 최적 군집수 탐색	<ul style="list-style-type: none"> <li>Within cluster sum of squares : 군집내 제곱합을 계산하여 최적의 군집 수를 탐색합니다.</li> <li>Silhouette : 실루엣 지표를 계산해 최적의 군집 수를 탐색합니다.</li> <li>Dunn Index : 군집 간 거리의 최소값을 분자, 군집 내 요소 간 거리의 최대값을 분모로 하는 지표입니다. 군집화 결과가 좋을수록 Dunn Index는 커지게 됩니다.</li> </ul>
③ 최대 군집의 수	<p>최적 군집 수 탐색에 사용할 최대 군집 수를 지정합니다. 2 이상의 정수만 입력 가능하며, Default는 10입니다.</p>
④ 저장	<p>선택한 결과를 괄호 안의 변수명으로 저장합니다.</p> <ul style="list-style-type: none"> <li>최종군집 : 각 관측값이 최종적으로 할당된 군집을 출력한 후 저장합니다. (HCluster)</li> <li>최종군집중심으로부터의 거리 : 각 관측값과 해당 관측값이 최종적으로 할당된 군집의 중심 사이의 거리를 출력한 후 저장합니다. (HC_dist)</li> </ul>