第十章 集群分析(Cluster Analysis)

將比較相似的樣本聚集在一起，形成集群(cluster)。以『距離』作為分類的依據，『相對距離』愈近的，『相似程度』愈高，歸類成同一群組。此統計分析方法不需要任何的假設。集群分析可分成分層法(Hierarchical)、非分層法(Nonhierarchical)和兩階段法。

1. **分層法**有凝聚分層法(Agglomerative)和分離分層法(Divisive)。『距離』可分為『點間距離』和『群間距離』。

『點間距離』：

歐氏距離(Euclidean Distance)：

馬氏距離(Mahalanobis Distance)：

城市街距離(City Block Distance)：

(1)凝聚分層法(Agglomerative)：開始時每一個體為一群，然後最近的兩個體合成一群，一次結合使群組越變越少，最後所有個體結合成一群。依不同的『群間距離』分為，

(A)最近法(單一聯結法 Single Linkage)：

$$d_{A,B} = \underset{\substack{i\in A \\ j\in B}}{Min}\, d_{ij}$$

(B)最遠法(完全聯結法 Complete Linkage)：

$$d_{A,B} = \underset{\substack{i\in A \\ j\in B}}{Max}\, d_{ij}$$

(C)平均法(Average Linkage)：

$d_{A,B} = \Sigma\Sigma\, d_{ij} / n$ ，n 為全部距離的個數

(D)中心法(Centroid Method)：

$$d_{A,B} = d(\bar{\bar{x}}_A, \bar{\bar{x}}_B) = \| \bar{x}_A - \bar{x}_B \|^2$$

(E)華德法(Wards Method 華德最小變異法)：

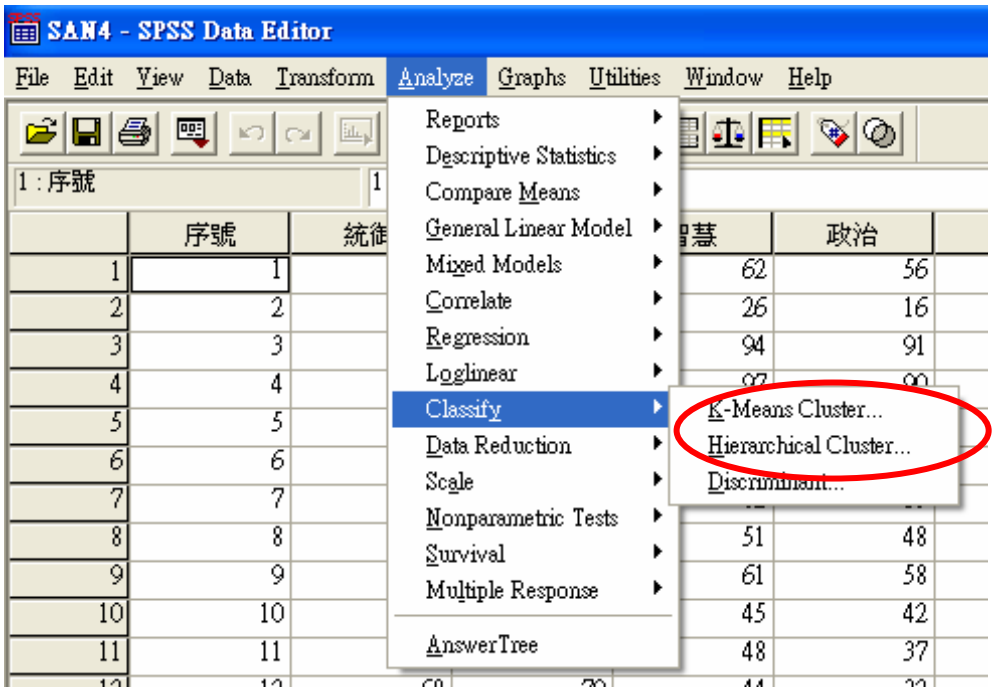
$$d_{A,B} = n_A \| \bar{x}_A - \bar{\bar{x}} \|^2 + n_B \| \bar{x}_B - \bar{\bar{x}} \|^2$$

(2)分離分層法(Divisive)：開始所有個體為一群，然後分成兩群、三群，直到每個體為一群。此法不常用。

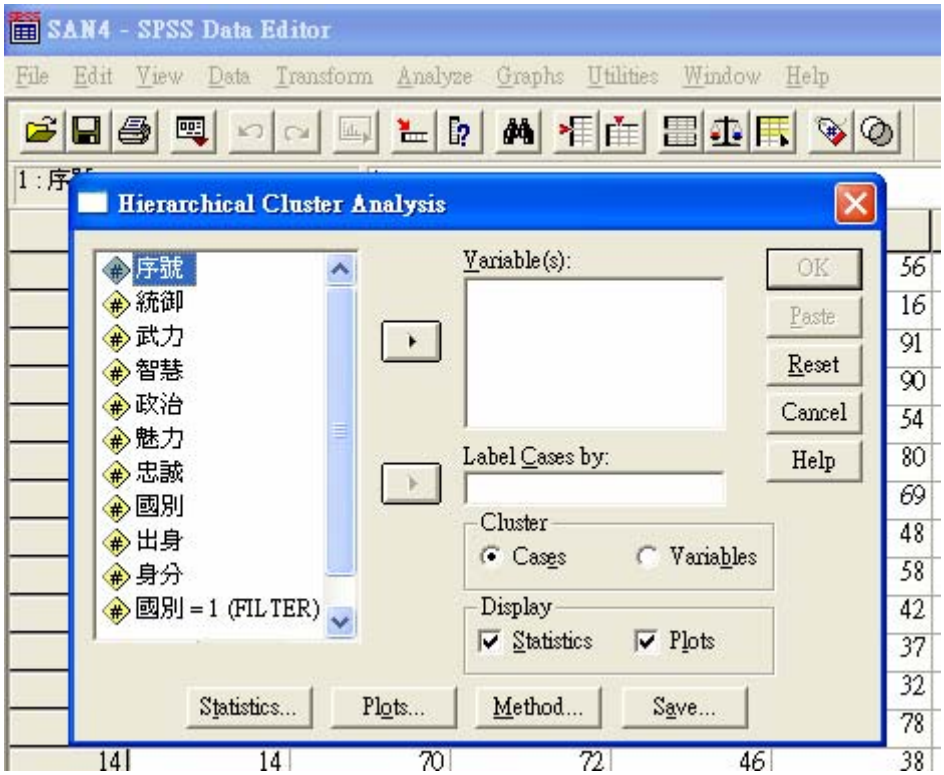
2. **非分層法**最具代表性的為 K 組平均法(K-Meanns)。開始任意將個體分成 K 組，然後將個體在個群間移動，使(1)群內變異最小；(2)群間變異最大。

3. **兩階段法**為第一階段分層法分群，決定群組個數，第二階段再以 K 組平均法進行群集，移動各群組內的個體，保持全部群組為 k 組。

SPSS 點選方式：



分層法：

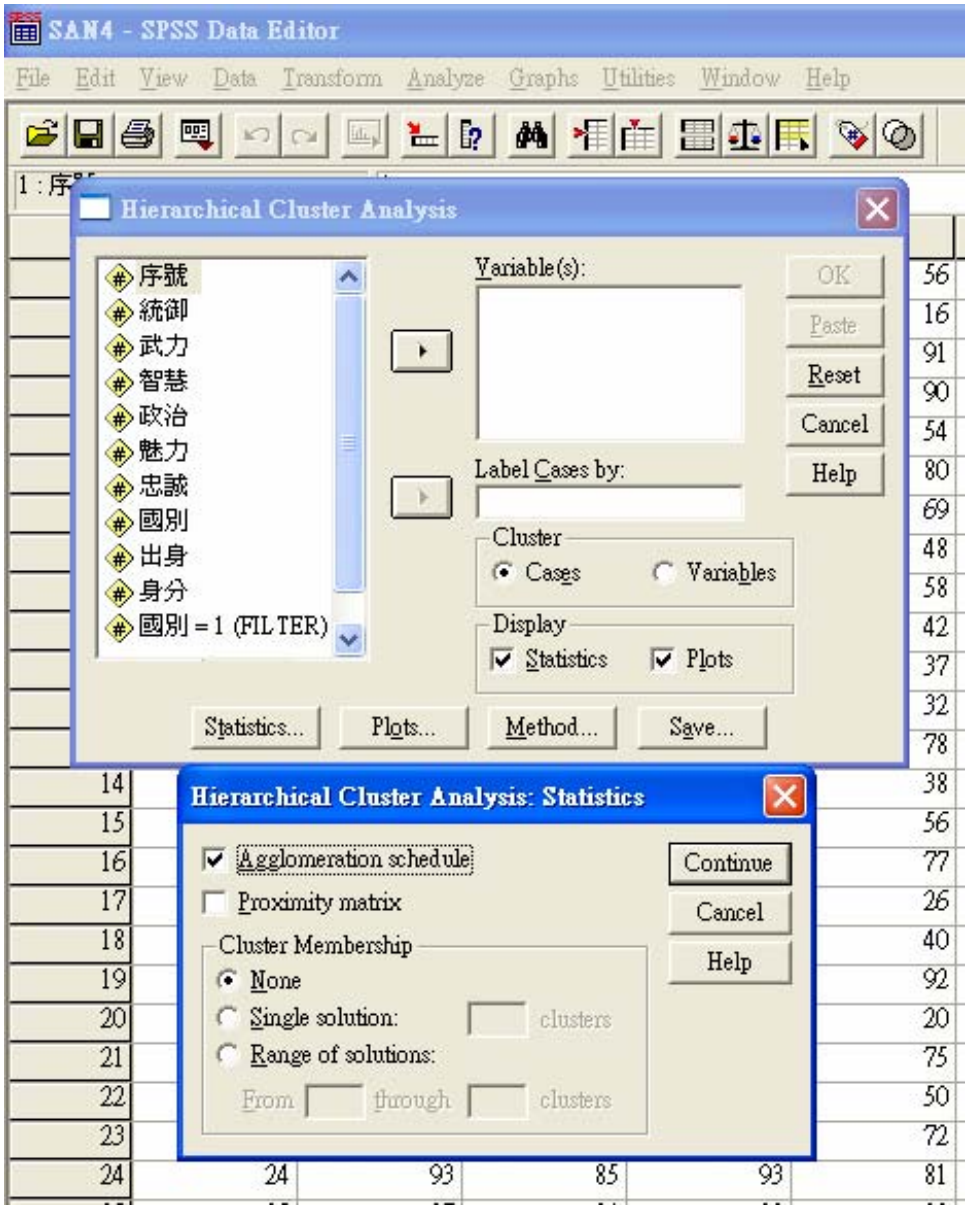


Variable(s)：放入考慮之群集變數

Label Cases by：放入顯示個體的名稱變數

Cluster：Cases(做個體的群集分析)；Variable s(做變數的群集分析)

Display：Statistics(統計量)； Plots(圖形) 。預設值通常會保留。



Statistics：

凝聚過程(Agglomerative schedule)

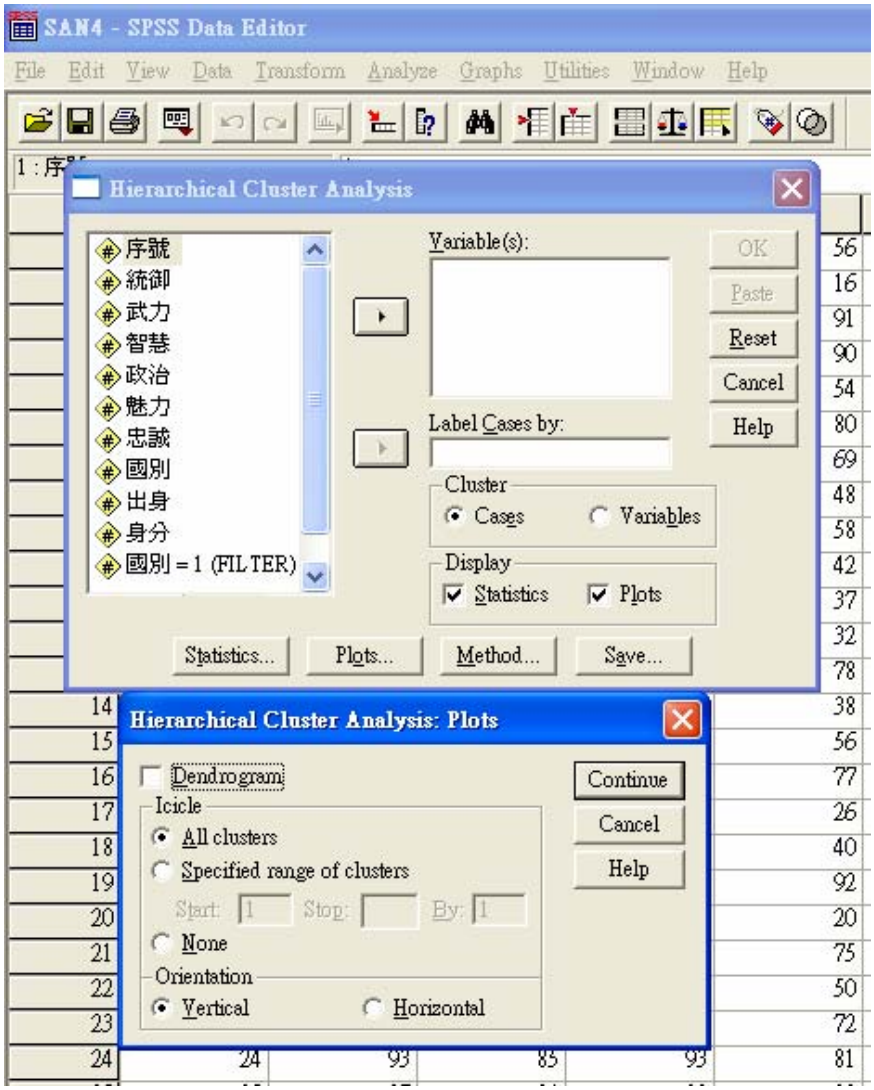
近似性矩陣(Proximity matrix)

集群組員(Cluster Membership)：(設定群集個數)

None：沒有設限制

Single solution：指定一個大於 1 的數

Range of solutions：指定一個範圍



Plots：

樹狀圖(Dendrogram)

冰柱圖(Icicle)：

All clusters(顯示所有群集)

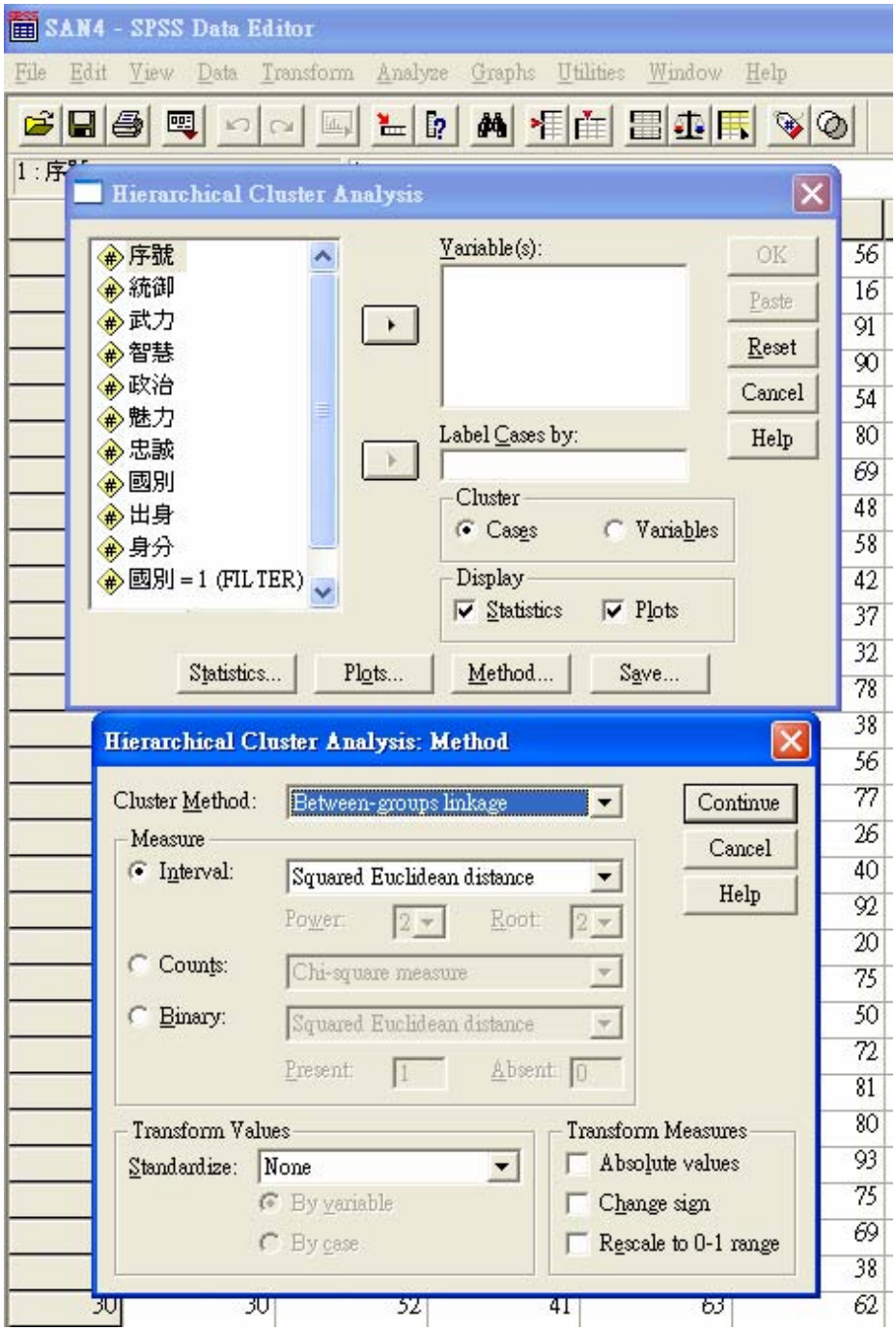
Specified range of clusters(顯示指定的群集範圍之結果)

None(不會顯示冰柱圖)

方向 (Orientation)：

Vertical(垂直)

Horizontal(水平)



Method：

凝聚分層法的方法(Cluster Method) (選取群間距離的算法)(有七種選擇)

Measure：(資料型態) (選取點間距離的算法)

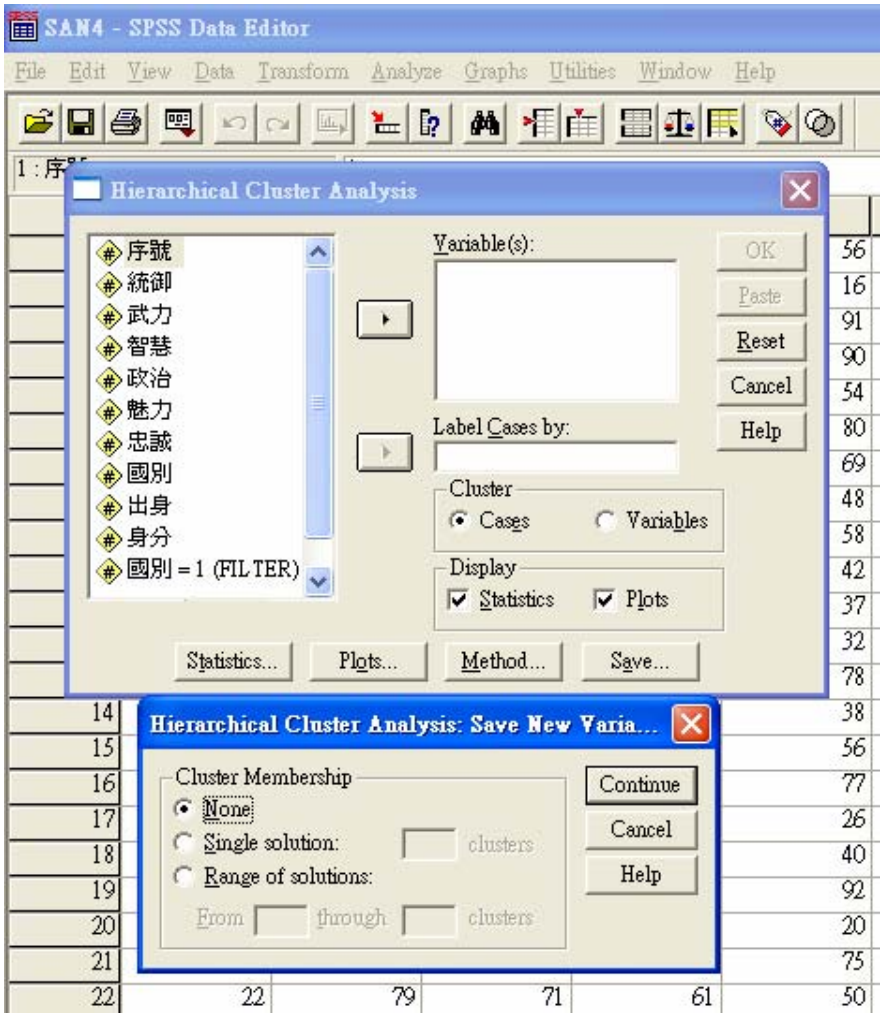
Interval(區間資料)

Counts(計數資料)

Binary(二元資料)

Transform Values：(轉換值) 各種標準化的方式

Transform Measures：(轉換衡量)



Save：

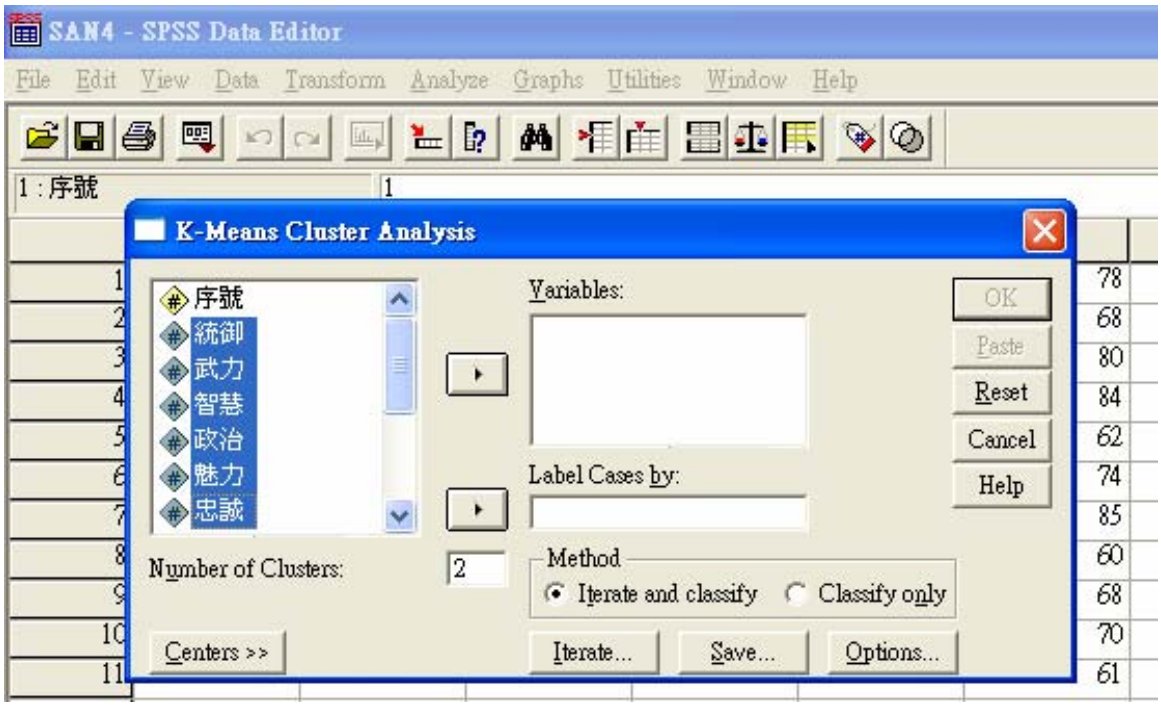
集群組員(Cluster Membership)：(儲存群集組員變數)

None：不儲存

Single solution：儲存指定一個大於 1 的數

Range of solutions：儲存指定的範圍

K- MEAN：

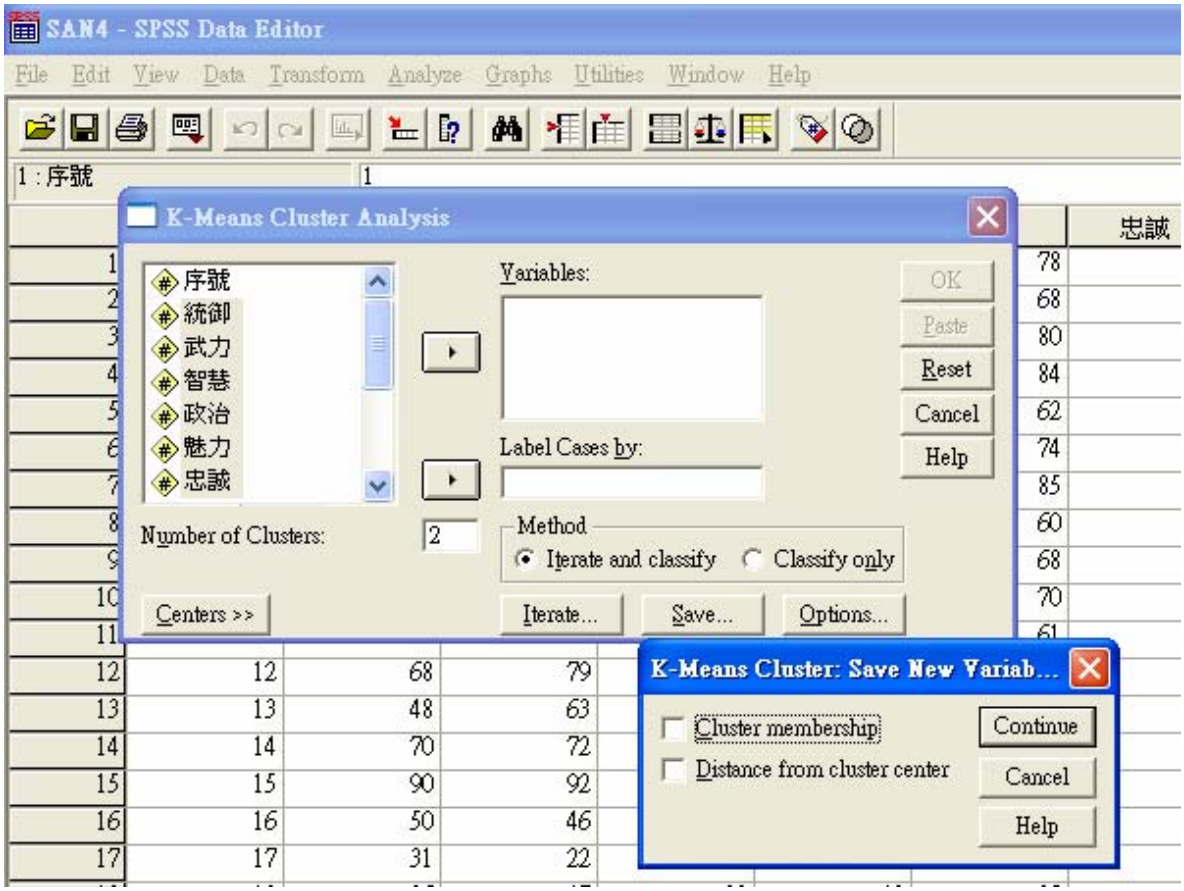


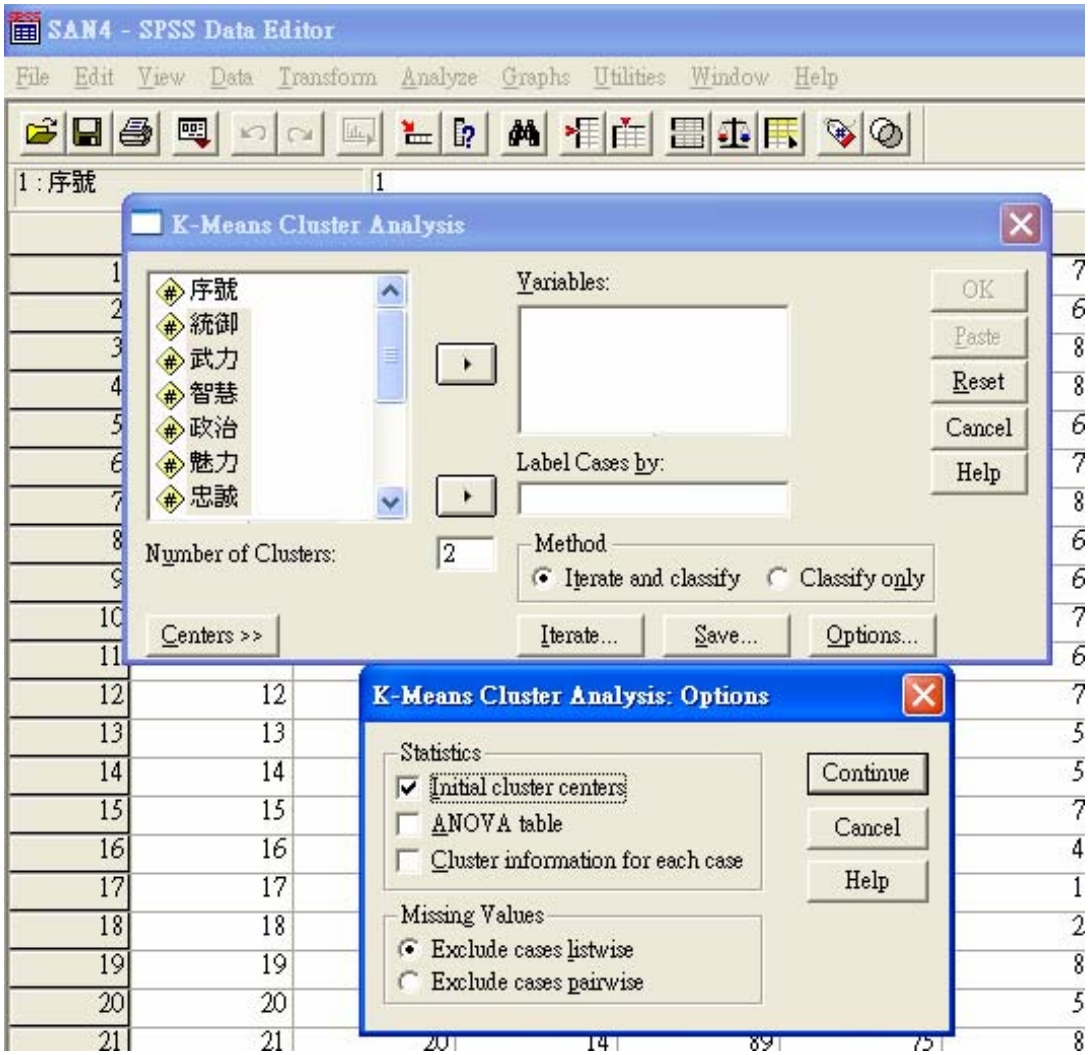
Number of Clusters: 自行指定群組個數(內設 2)

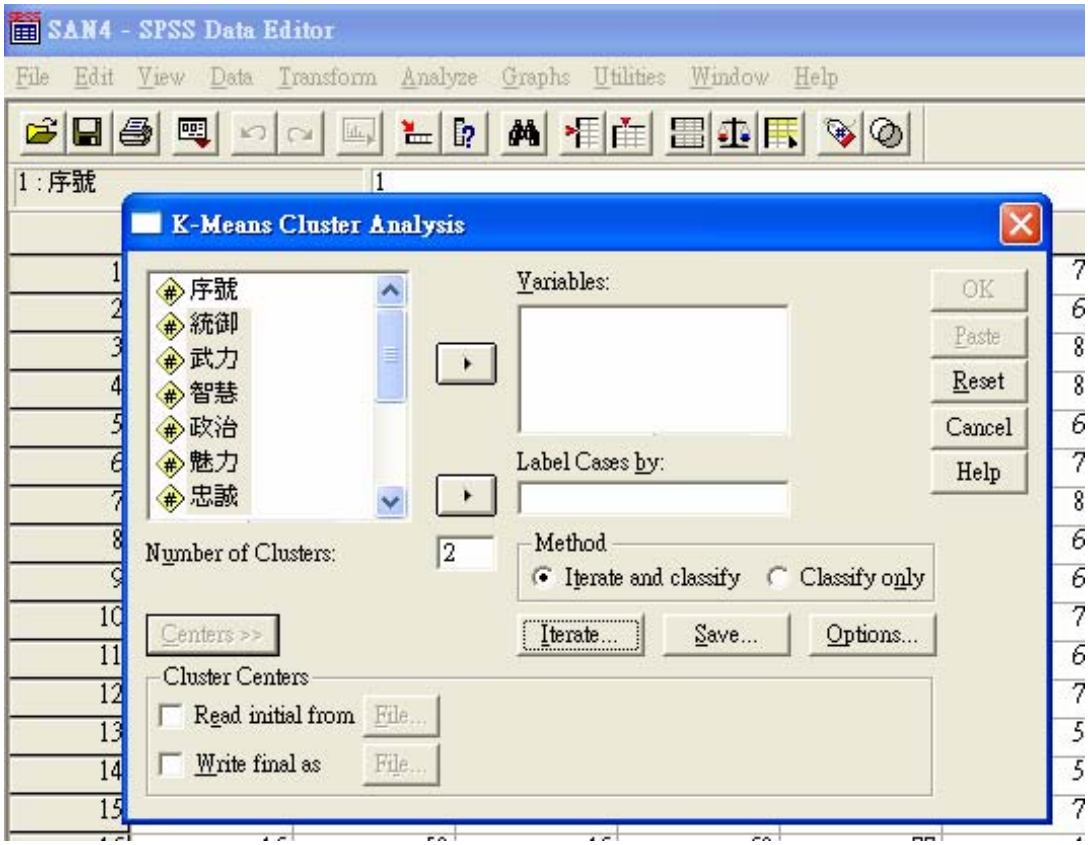
Method：

Iterate and classify(疊代與分群)：不斷疊代與更新群中心位置

Classify only(分群)：使用初始設定之群中心位置


SAN4 - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : 序號
1
忠誠
K-Means Cluster Analysis
序號
統御
武力
智慧
政治
魅力
忠誠
Variables:
Label Cases by:
OK
Paste
Reset
Cancel
Help
Number of Clusters:
2
Method
Iterate and classify
Classify only
Centers >>
Iterate...
Save...
Options...
K-Means Cluster: Save New Variab...
Cluster membership
Distance from cluster center
Continue
Cancel
Help


SAN4 - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : 序號
K-Means Cluster Analysis
序號
統御
武力
智慧
政治
魅力
忠誠
Variables:
Label Cases by:
OK
Paste
Reset
Cancel
Help
Number of Clusters:
2
Method
Iterate and classify
Classify only
Centers >>
Iterate...
Save...
Options...
K-Means Cluster Analysis: Options
Statistics
Initial cluster centers
ANOVA table
Cluster information for each case
Continue
Cancel
Help
Missing Values
Exclude cases listwise
Exclude cases pairwise


SAN4 - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : 序號
K-Means Cluster Analysis
序號
統御
武力
智慧
政治
魅力
忠誠
Variables:
Label Cases by:
OK
Paste
Reset
Cancel
Help
Number of Clusters:
2
Method
Iterate and classify
Classify only
Centers >>
Iterate...
Save...
Options...
Cluster Centers
Read initial from
File...
Write final as
File...

以 TOEFL.sav 為例， (鄧家駒著, 多變量分析)

1. 凝聚分層法(Agglomerative) 『群間距離』：華德法

『點間距離』：歐氏距離平方(Squared Euclidean Distance)

SPSS 程式：

```
CLUSTER  gpa toefl gmat work other
  /METHOD WARD
  /MEASURE= SEUCLID
  /PRINT SCHEDULE
  /PLOT DENDROGRAM VICICLE.
```

Cluster

Case Processing Summary [a,b]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
15	100.0	0	.0	15	100.0

a. Squared Euclidean Distance used

b. Ward Linkage

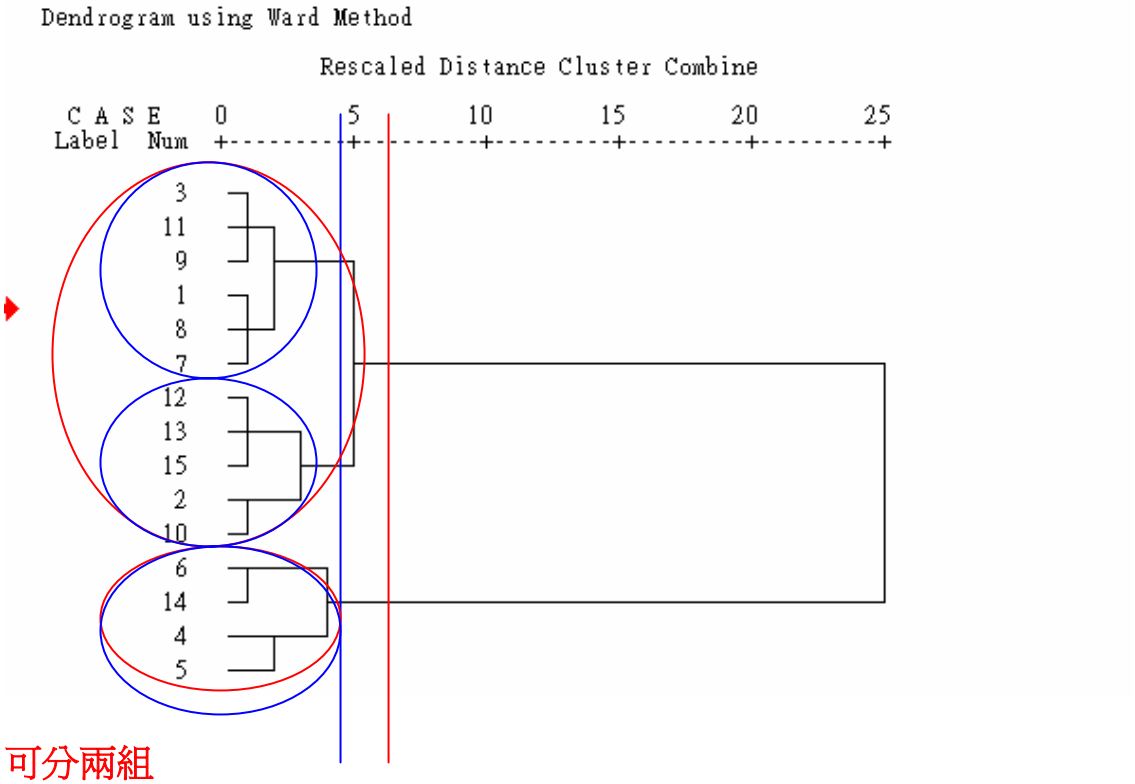
Ward Linkage

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	11	26.000	0	0	4
2	12	13	76.505	0	0	6
3	1	8	130.750	0	0	5
4	3	9	198.777	1	0	10
5	1	7	285.485	3	0	10
6	12	15	441.853	2	0	11
7	6	14	692.533	0	0	12
8	2	10	949.078	0	0	11
9	4	5	1599.583	0	0	12
10	1	3	2488.977	5	4	13
11	2	12	3842.210	8	6	13
12	4	6	6342.713	9	7	14
13	1	2	9150.495	10	11	14
14	1	4	24741.596	13	12	0

Dendrogram

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *



可分兩組

或分三組(較佳)

可接著找出各組內觀測者的特色，以了解所分成的各組之差異。

2. K組平均法(K-Meanns)：分三組之下

SPSS 程式：

```
QUICK CLUSTER
  gpa toefl gmat work other
  /MISSING=LISTWISE
  /CRITERIA= CLUSTER(3) MXITER(10) CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /PRINT INITIAL ANOVA CLUSTER DISTAN.
```

Quick Cluster

Initial Cluster Centers

	Cluster		
	1	2	3
GPA	3.70	3.20	3.40
TOEFL	580.00	530.00	580.00
GMAT	540.00	550.00	640.00
WORK	.00	.00	.00
OTHER	1.00	1.00	1.00

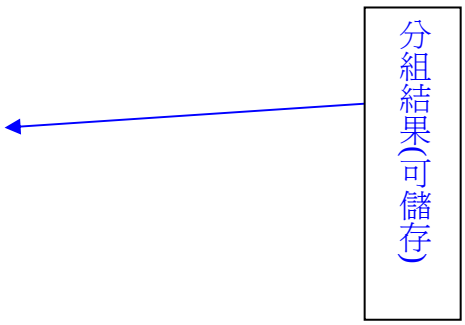
Iteration History [a]

	Change in Cluster Centers		
Iteration	1	2	3
1	15.365	12.991	28.294
2	5.901	5.716	.000
3	.000	.000	.000

a. Convergence achieved due to no or small distance change. The maximum distance by which any center has changed is .000. The current iteration is 3. The minimum distance between initial centers is 50.993.

Cluster Membership

Case Number	Cluster	Distance
1	1	9.056
2	2	16.363
3	1	14.227
4	1	31.358
5	3	36.057
6	3	10.009
7	1	19.952
8	1	18.979
9	1	7.541
10	2	28.297
11	1	13.737
12	2	22.863
13	2	13.527
14	3	28.294
15	2	6.515



Final Cluster Centers

	Cluster		
	1	2	3
GPA	3.41	3.36	3.73
TOEFL	577.14	544.00	600.00
GMAT	558.57	542.00	620.00
WORK	2.14	2.40	.00
OTHER	2.57	2.40	1.67

Distances between Final Cluster Centers

Cluster	1	2	3
1		37.056	65.585
2	37.056		96.054
3	65.585	96.054	

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
GPA	.144	2	.076	12	1.910	.191
TOEFL	3215.238	2	205.238	12	15.666	.000
GMAT	6003.810	2	313.810	12	19.132	.000
WORK	6.171	2	3.505	12	1.761	.214
OTHER	.876	2	1.465	12	.598	.565

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster	1	7.000
	2	5.000
	3	3.000
Valid		15.000
Missing		.000