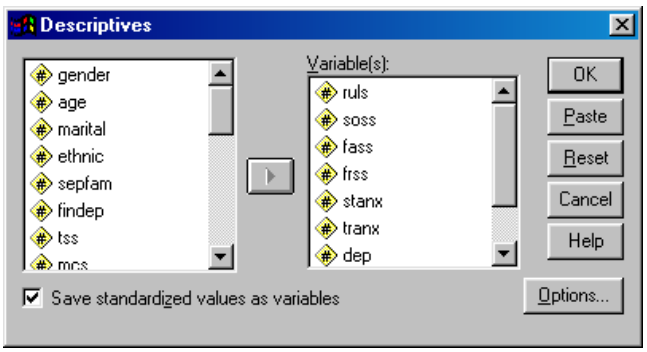
Clustering Example

The purpose of the analysis was to look for "sub-populations" of adult females, with respect to a selection of clinically relevant variables.

Converting Variables to Standardized Form (Z-scores)

It is a good idea to work with Z-scores of the variables If the variables being used differ in their variability. Otherwise, the variables with greater variability will dominate clustering.

Analyze → Descriptive Statistics → Descriptives

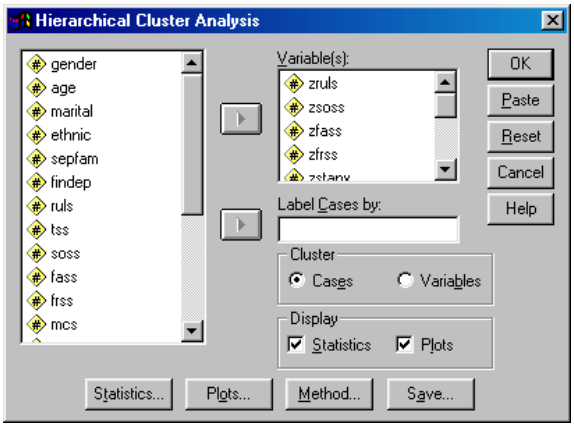


Select the variables for the analysis and click the "Save standardized values as variables" box.

The clustering will be done with the resulting Z-score variables, zruls, zsoss, etc.

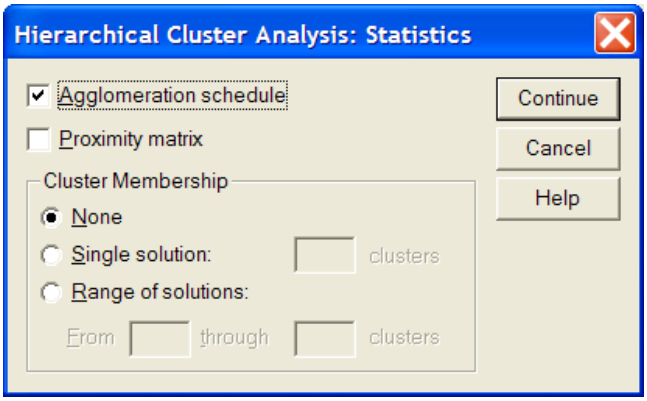
Getting Clustering Analysis

Analyze → Classify → Hierarchical Clustering



Select the variables to be clustered.

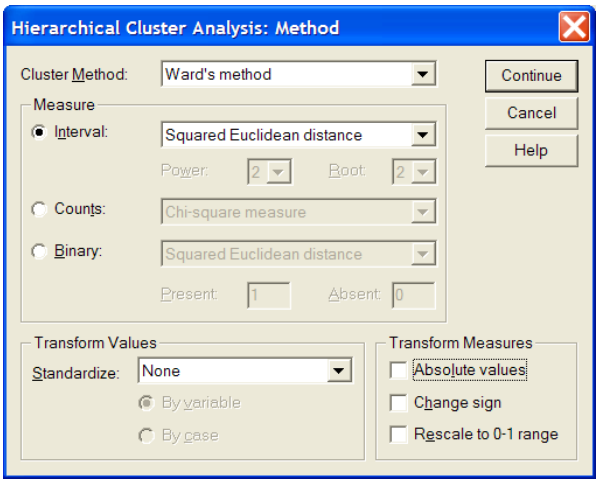
Remember to use the Z-score form of each variable



Open the Statistics window

The "agglomeration schedule" will help us decide how many clusters to include in our solution.

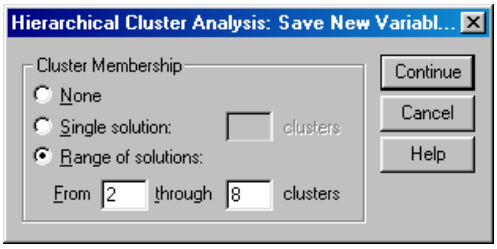
Knowing the cluster membership of each case for different # of clusters can be very useful also, but we'll use a different way of looking at this information.



Open the Method window

This is how you select the clustering method (how to decide which clusters will be combined on each step) and the dissimilarity measures (how to represent how similar the cases/clusters are to each other)

You can tell SPSS to work with transformed values. I prefer to save the transformed values separately (as above), so that they are available for additional analyses.



This allows you to save the cluster membership of each case for each clustering solution you specify.

Usually 2-12 is enough…depends upon whether groups or "strays" are being combined to form the successive clusters.

Clustering Output

Examining the Agglomeration Schecule

The agglomeration schedule shows the step-by-step clustering process.

- Which clusters were combined on that step
- The resulting total "error" in the clustering solution

We look for the "big jump" in error -- as a sign that two "different" clusters have been combined.

Pretty big jump on step 120 (from 4→ 3 clusters), suggesting that 3 is "too few" and 4 is "just right".

Have to worry about "strays"!!!!

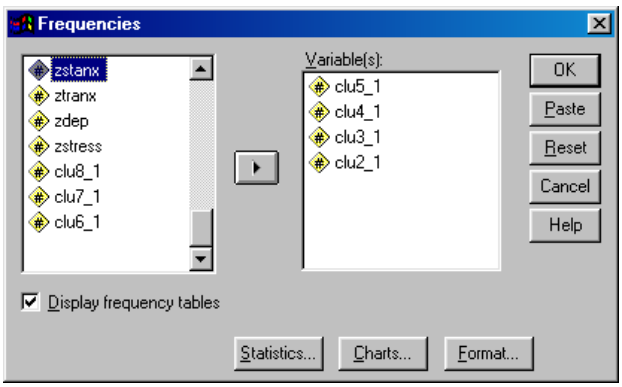
6 clusters → 5
5 clusters → 4
4 clusters → 3
3 clusters → 2
2 clusters → 1

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	235	289	.092	0	0	78
2	245	338	.223	0	0	10
3	212	387	.409	0	0	48
111	210	226	289.703	101	93	119
112	212	215	304.766	108	78	121
113	207	208	320.378	100	90	114
114	207	247	336.982	113	97	118
115	219	242	355.247	103	0	118
116	206	213	375.485	104	109	117
117	206	297	402.101	116	105	121
118	207	219	432.390	114	115	120
119	210	218	469.263	111	110	120
120	207	210	542.696	118	119	122
121	206	212	633.798	117	112	122
122	206	207	976.000	121	120	0

It can be very helpful to also consider the frequencies of the clusters for the different solutions. This can help you think about how the groups form and separate.

Analyze → Descriptive Statistics → Frequencies



The variables saved during the clustering tell the membership of each case in each number-of-clusters solution.

Use several of them to identify clustering patterns, strays, etc.

Ward Method

		Frequency	Percent
Valid	1	84	68.3
	2	39	31.7
	Total	123	100.0

Ward Method

		Frequency	Percent
Valid	1	41	33.3
	2	39	31.7
	3	43	35.0
	Total	123	100.0

Most likely solutions

Ward Method

		Frequency	Percent
Valid	1	41	33.3
	2	19	15.4
	3	20	16.3
	4	43	35.0
	Total	123	100.0

Ward Method

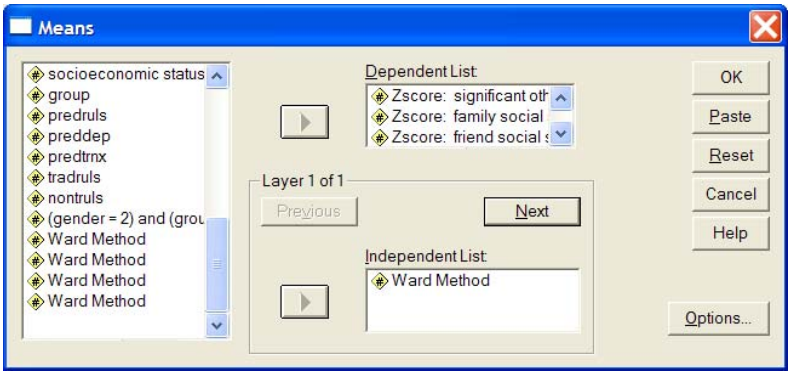
		Frequency	Percent
Valid	1	41	33.3
	2	19	15.4
	3	8	6.5
	4	43	35.0
	5	12	9.8
	Total	123	100.0

Group 1 (n=43) and Group 4 (n=41) look pretty stable. The questions is whether to keep just a 3rd group of n=39 or a 3rd and 4th group of n=19 & n=21 ???

The best way to make this decision is to look at the plots of the 4-group solutions. If the 3rd and 4th groups have "similar enough" profiles you may decide to go with the 3-group solution. If they are "sufficiently different" you may decide to keep the 4-group solution.

Getting Custer Profiles

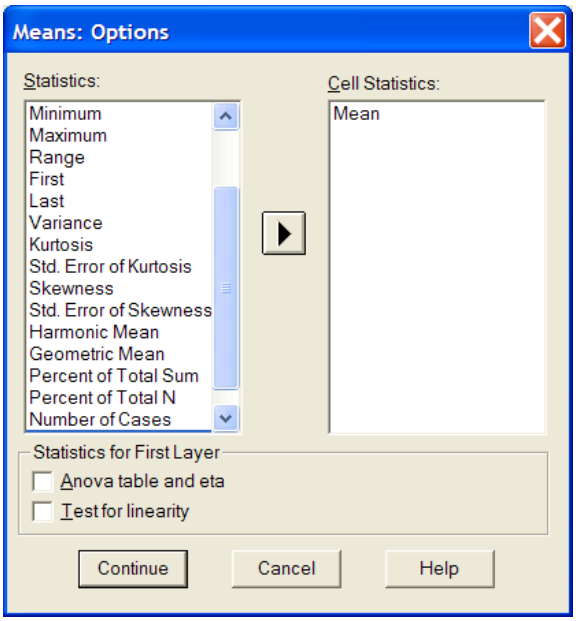
Analyze → Compare Means → Means



Use the same variables that were used to perform the cluster solution (remember to use the Z-score form of each)

Select one of the solutions for examination.

This examines the 4-cluster analysis – the variable is "clus1_4" (but doesn't show up until you highlight the variable in the listing)



Open the Options window

Remove everything from the "Cell Statistics" window except "Mean"

You get the following table as output.

Mean

Ward Method	Zscore: significant other social suppor	Zscore: family social support	Zscore: friend social support	Zscore: stait anxiety	Zscore: trait anxiety	Zscore: depression (BDI)	Zscore(STRESS)	Zscore: loneliness
1	.0253659	-.1586084	-.3073686	-.3003678	-.1480552	-.2308712	-.4889418	.1718599
2	-1.6423312	-1.4103485	-1.2006885	.9320337	.9621648	1.1836913	.9867314	1.3585691
3	-.0809595	-.0896958	.1634026	1.2053890	.9346288	.6707560	.9604135	.2841101
4	.7391507	.8161275	.7476080	-.6860777	-.7186848	-.6148729	-.4165012	-.8963086
Total	.0000000	.0000000	.0000000	.0000000	.0000000	.0000000	.0000000	.0000000

Notice that the table includes the group means for each variable for each group and for the total (overall population). You can decide whether or not you want that "overall" profile included in your graph. (They will always all be 0.00 –> average Z-scores)

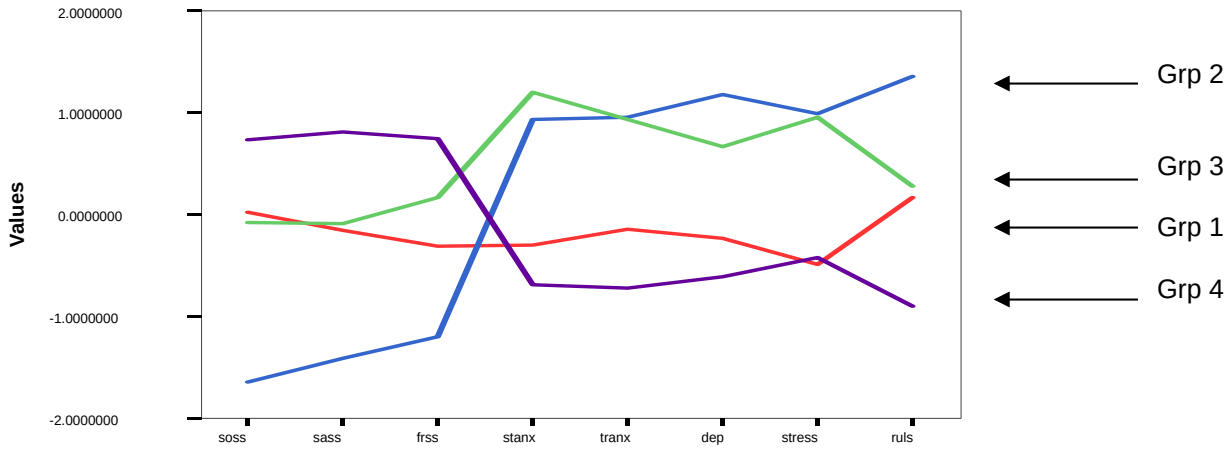
If you don't want the total data plotted you should double-click the table and then highlight and delete that row. You can also edit the various names, etc. Here's the table as I edited before graphing.

To obtain the graph → Double-click the table (to put it in "edit mode"). Then right-click the table and a menu appears that includes "Create Graph". Move the cursor to that phrase and another menu appears. Click on "Line" .

Mean

Ward Method	soss	sass	frss	stanx	tranx	dep	stress	ruls
Grp 1 N=41	.025366	-.1586084	-.3073686	-.3003678	-.1480552	-.2308712	-.48894	.171860
Grp 2 N=19	-1.6423	-1.4103485	-1.20069	.9320337	.9621648	1.1836913	.986731	1.35857
Grp 3 N=20	-.08096	-.0896958	.1634026	1.2053890	.9346288	.6707560	.960414	.284110
Grp 4 N=43	.739151	.8161275	.7476080	-.6860777	-.7186848	-.6148729	-.41650	-.89631

Here's the 4-group plot



Deciding between the 3- and 4-group models → separate or combine Grp 2 & Grp 3 ???

Group 4 – "Healthy cluster" – above average social support, below average for lonely, anxious, dep & stress

Group 1 – "Average custer" – pretty flat

Group 2 – "Unsupported, Lonely & Unhappy" -- low support, high on lonely, anxiety, dep, stress and loneliness

Group 3 – "Semi-supported, Not Lonely, but Unhappy " – average support, low on lonely, high on anx, dep & stress

I'd keep 2 & 3 separate, because of the differences on social support and loneliness. Combining tem really hides their considerable difference on these variables