Neuendorf
Cluster Analysis

<u>Model</u>:

| Cluster Analysis | Profiling the Clusters |

X1
X2
X3 → NEW VARIABLE → Y1
"Clusters" (Nominal) → Y2
X4
X5 → Y3

Clustering/Internal Variables   External Variables
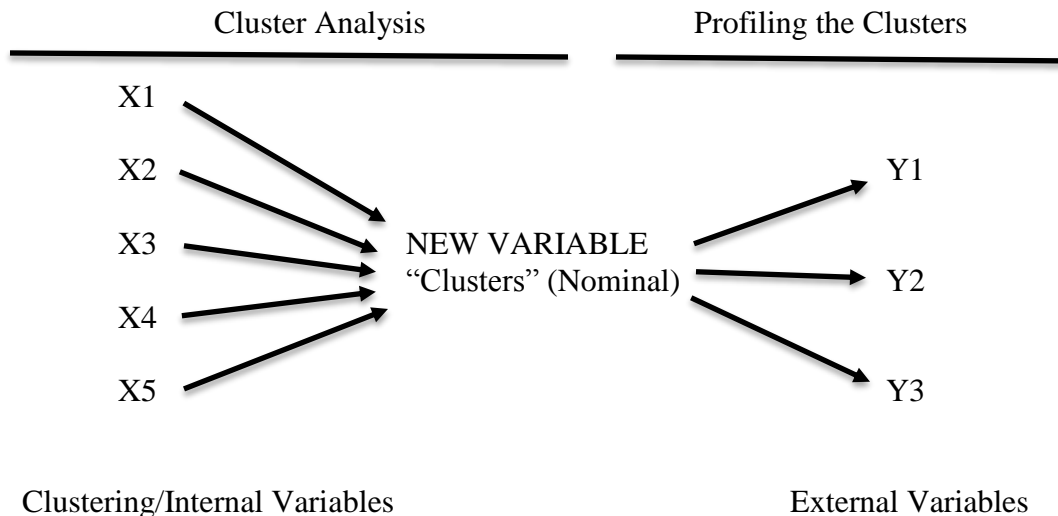
<u>Assumes</u>:

1.   Actually, <u>any</u> level of measurement (nominal, ordinal, interval/ratio) is acceptable for certain types of clustering. The typical methods, though, require metric (I/R) data. The most basic assumption is that there are two or more (clustering or "internal" or "independent") variables, and that cases (often people, sometimes companies or products or other identifiable units; Hair et al. call them "objects") will be placed into groups on the basis of how similar they are on the selected variables. [Then, usually, these clusters are used to predict status on one or more "dependent" or "external" variables.]

2.   Sample representativeness. That means some type of probability/random sampling.

3.   Not very much else is assumed! As Hair et al. note, even multicollinearity is not technically prohibited (although it's hard to imagine *why* you would want to cluster on a set of variables that could better be represented by just one of the variables, or a scale that combines them). As Aldenderfer and Blashfield point out in their Sage QASS (little green) book:
-"Most cluster analysis methods are relatively simple procedures that. . . are not supported by an extensive body of statistical reasoning."
-"Different clustering methods can and do generate different solutions to the same data set."

<u>Decisions to Make</u>:

1.   [Whether you need to run PermuCLUSTER or not. This add-on to SPSS will help determine the "optimal" ordering for your data set to avoid problems that can occur when there are distance "ties" in your data. If used, this step must be conducted *before* running Cluster in SPSS. See the references below for more info.]

2.   Standardized variables vs. original ("raw") data. This choice is very similar to the choice of whether to standardize variables before putting them together in a scale. If the variables are measured on

different response scales, we typically standardize. If the variables are measured on the same response scale, we typically do not standardize. But, even when data are measured on the same measurement scale, if they are unstandardized the variable(s) with the largest variance will dominate.

3. Similarity measure

    A. Correlation coefficients (Pearson or Spearman)--the unit of analysis is now the variable, and the "variable" is now a person or object.   For example, there will be a correlation between person A and person B, for their values across, say, ten selected variables. This is quite similar to a factor analysis of persons or objects! This method of analysis is sensitive to patterns of relationships, rather than absolute values on variables (i.e., the magnitude of the difference).

    B. Association coefficients--for categorical/nominal data; there are more than 30 of these stats. available; e.g., the simple matching coefficient and Jaccard's coefficient (see Aldenderfer and Blashfield, p. 29).

    C. Distance coefficients--assumes I/R data; except for #4, can select either original variables or standardized variables. See the attached two-variable example of these distance measures:

        1. Euclidean distance

        $$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

        2. Squared Euclidean distance

        $$d_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

        3. Manhattan or city-block distance ("Block" in SPSS)

        $$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

        4. Mahalanobis D^2 or generalized distance

        $$d_{ij} = (X_i - X_j)' \sum{}^{-1} (X_i - X_j)$$

where Σ is the pooled within-groups variance-covariance matrix, and Xi and Xj are vectors of the values for the variables for cases i and j. When the correlations between variables are all zero, Mah. D^2 = squared Euclidean distance. The noteworthiness of Mah. D^2 is that it takes intercorrelations among the variables into account. Also, it "automatically" standardizes the variables. Mah. D^2 does not currently seem to be available in SPSS.

4. Methods of clustering

    A. Hierarchical
        1. Agglomerative (the most common; *see below*)
        2. Divisive

    B. Nonhierarchical--includes several options; see Hair p. 511. Such procedures have grown in popularity, but are <u>not</u> a good choice without theory. That's because nonhierarchical procedures require the choice of "cluster seeds" (sets of values on all variables) to begin the clustering process. They also assume that the researcher can intelligently select the number of clusters to be derived. . . prior to the clustering.

    A. 1. Hierarchical Agglomerative, cont'd.: Assuming that Hierarchical Agglomerative is chosen, there are at least 7 options from which to choose, in terms of <u>how</u> cases are chosen for inclusion in clusters:
        a. Average linkage ("between-groups linkage")--criterion is average distance from individuals in one cluster to individuals in another; probably the most common method
        b. Within-groups linkage--this method combines clusters so that the average distance between all cases in the resulting cluster is as small as possible
        c. Single linkage ("nearest neighbor")--can be a problem in that it may produce snake-like clusters
        d. Complete linkage ("furthest neighbor")
        e. Centroid method--uses distances between cluster centroids (multivariate means); for every new grouping, centroids are recomputed, thus the disadvantage is that "reversals" may occur (where centroid distance between two unlinked clusters can be smaller than that of two clusters linked earlier)
        f. Median clustering—the distance between two clusters is represented by the distance between the SPSS-determined median for the cases in cluster A and the median for the cases in cluster B
        g. **Ward's method--uses squared Euclidean distances from each case to the mean (centroid) of its cluster, summed across all cases; at each step, the two clusters that merge are those that result in the smallest increase in the overall sum of the squared within-cluster distances; useful in creating clusters of relatively equal size; probably the second most common method

5. Number of clusters

    A. A priori–A decision based on theory, past work, etc. (With the "K-Means" non-hierarchical procedure in SPSS, this is the only option.) With the "Hierarchical Cluster" procedure, there are two additional options:

B.   Intuitive, data-based procedures--(1) look at the dendrogram or icicle plot, and/or (2) look at *n*'s for clusters (need to save cluster groupings and then run frequencies!).

C.   Formal tests--SPSS provides an agglomeration coefficient. Small values indicate that fairly homogeneous clusters are being merged at that step. We may look for large increases as a "cut-off," similar to the scree test in factor analysis. See Hair et al. p. 527 for an example of the application of this "stopping rule."

6.   Treatment of outliers--Single-case clusters may be candidates to be dropped from the analysis, particularly in a solution with a small number of clusters.

7.   Validation and interpretation techniques (Profiling)--Hair et al. indicate that one may validate by splitting the sample and comparing cluster analyses. Another procedure more commonly used to interpret or "make sense" of the clusters is to look at cluster means on (1) the "internal" (clustering) variables used to produce the clusters, and (2) "external" variables of interest, not used to produce the clusters. Hair et al. call this "profiling." We can simply profile each cluster as an heuristic in order to get a clearer picture of that group, or we can use the cluster groupings as an independent variable for an ANOVA or MANOVA procedure, or as a dependent variable in a discriminant analysis procedure. To use the cluster groupings for further analyses, use the "save" function in cluster analysis, and cluster membership variables will be added to the data set.

Statistics:

Not much!   Just what's already been mentioned:

1.   Agglomerative coefficient--the within-cluster sum of squares (Ward's method) or the squared Euclidean distance between the two cases of clusters being combined (other methods), this coefficient indicates whether the clusters being joined are very homogenous (small coefficient) or different (large coefficient). Again, a scree-type inspection or "stopping rule" may be used.

2.   ANOVA/MANOVA analyses or Discriminant analysis--to "profile" the cluster groupings with both "internal" and "external" variables. Must be run *after* clustering, with cluster solution(s) saved. I tend to use "MEANS" with the ANOVA stats clicked under Options. For this application of ANOVA, the cluster variable is the IV.

Other notes:

1.   SPSS provides "K-Means" Cluster Analysis, which is faster (but this doesn't make a lot of difference these days); but it limits the user to simple Euclidean distances only. This procedure asks for you to indicate how many clusters you want, and to provide "centers" (i.e., cluster "seeds," or values on the clustering variables to serve as the comparison points). If you do not specify the centers, SPSS selects some for you. (SPSS says, "a number of well-spaced cases equal to the number of clusters is selected from the data.")

2.   When using regular "Hierarchical Cluster" in SPSS, the user might need to wait a few minutes--it's much slower to run than other procedures we've used. It is found under "Analyze → Classify → Hierarchical Cluster."

3.   Cluster analysis has been a favorite among marketing professionals, for applied use in consumer or audience "segmentation" research. An old classic is the famed "VALS" system (now owned by Strategic Business Insights; see also http://en.wikipedia.org/wiki/VALS), which has segmented consumers into eight groups. Another proprietary system is "PersonicX" by Acxiom, which has an online function that will categorize you into one of 70 consumer groups (https://isapps.acxiom.com/personicx/personicx.aspx).

References:

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Beverly Hills, CA: Sage.

Spaans, A., & van der Kloot, W. (2004). *PermuCLUSTER 1.0 user's guide*. The Netherlands: Department of Psychology, University of Leiden. (Available at: http://www.leidenuniv.nl/fsw/mtlab/software/PermuCluster%201.0%20User%20Guide.pdf)

van der Kloot, W. A., Spaans, A. M. J., & Heiser, W. J. (2005). Instability of hierarchical cluster analysis due to input order of the data: The PermuCLUSTER solution. *Psychological Methods*, *10*, 468-476.

Attachment A: Examples of distances with two variables (K1 and K2):
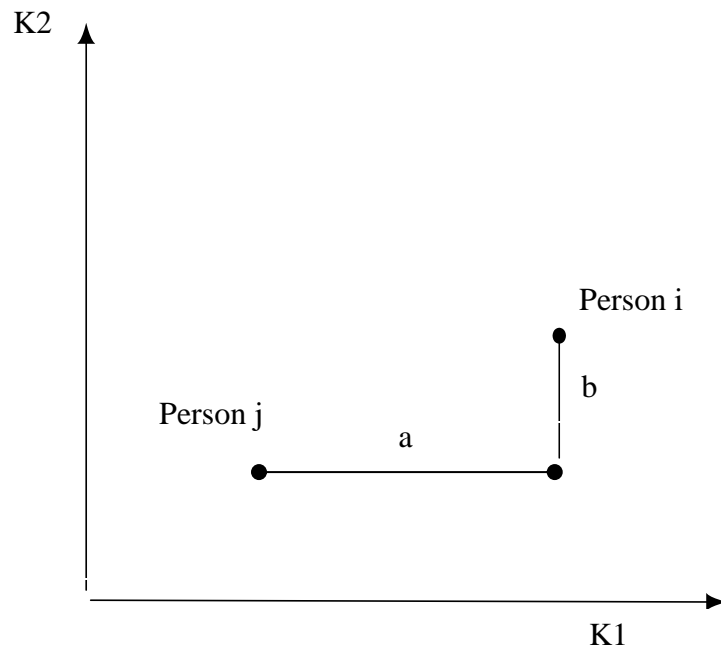
$$a = x_{iK1} - x_{jK1}$$

$$b = x_{iK2} - x_{jK2}$$

1. Euclidean distance $\qquad\qquad\qquad\qquad\qquad d = \sqrt{a^2 + b^2}$

2. Squared Euclidean distance $\qquad\qquad\qquad d = a^2 + b^2$

3. Manhattan or city-block distance $\qquad\qquad d = a + b$

4. Mahalanobis D^2 or generalized distance $\qquad d = a^2 + b^2$
       (IFF K1 and K2 are orthogonal)

Attachment B: General Population Humor Cluster Profiling

Cluster Name: Variable:	1 (20)	2 (44)	3 (71)	4 (66)	5 (20)	6 (52)	7 (16)	8 (32)	Total (312)	F	Sig.
Internal Variables: Mean-spirited humor	$-.14^c$	$-.57^c$	1.10^a	$.61^b$	$-.72^c$	$-.94^d$	$-.54^{c,d}$	$-.38^c$	0.00	$55.61_{(7,313)}$	<.001
Visual/Verbal humor	-1.69^a	$-.04^d$	$-.10^d$	$.24^d$	1.01^b	$-.65^c$	$.44^{b,c,d}$	1.04^b	0.00	$36.45_{(7,313)}$	<.001
Absurd/Stupid humor	$-.63^{a,b}$	-1.11^a	$-.30^b$	$.41^c$	$.86^c$	$.75^c$	$.94^c$	$-.38^b$	0.00	$33.46_{(7,313)}$	<.001
Social humor	$-.78^{a,c}$	$.29^b$	$-.21^{c,d}$	$.75^b$	$.35^{b,d}$	$.31^b$	-1.20^a	-1.12^a	0.00	$29.46_{(7,313)}$	<.001
Satire/Death humor	$-.71^{a,b}$	$-.37^{b,d}$	$.69^{c,e}$	$-.79^{a,d}$	1.05^c	$.45^e$	-1.26^a	$.30^e$	0.00	$40.92_{(7,313)}$	<.001
External Variables: Age (in years)	$45.0^{a,b}$	46.0^a	$34.5^{b,c,d}$	32.0^d	48.5^a	48.4^a	$48.4^{a,c}$	49.4^a	41.6	$8.50_{(7,297)}$	<.001
Gender (% female)	$75\%^a$	$68\%^a$	$32\%^b$	$62\%^a$	$45\%^{a,b}$	$77\%^a$	$88\%^a$	$66\%^a$	60%	$6.25_{(7,313)}$	<.001
Education (1-6 scale; 5=college grad)	$3.7^{a,b}$	$3.9^{a,b}$	4.1^a	3.5^b	$4.2^{a,b}$	4.2^a	$3.9^{a,b}$	$4.1^{a,b}$	4.0	$2.47_{(7,302)}$	.018
Race (% non-white)	$42\%^a$	$9\%^b$	$14\%^{a,b}$	$23\%^{a,b}$	$15\%^{a,b}$	$19\%^{a,b}$	$38\%^{a,b}$	$16\%^{a,b}$	19%	$2.20_{(7,312)}$	.034
Marital status (% married)	$45\%^{a,b}$	$48\%^{a,b}$	$41\%^{a,b}$	$21\%^a$	$45\%^{a,b}$	$48\%^{a,b}$	$38\%^{a,b}$	$53\%^b$	41%	$2.18_{(7,313)}$	.036
Days/week read newspaper	$3.6^{a,b}$	$4.6^{a,b}$	$3.9^{a,b}$	3.2^a	5.5^b	$4.1^{a,b}$	$3.6^{a,b}$	$4.1^{a,b}$	4.0	$2.11_{(7,307)}$	.042
Videos watched in past month	$3.5^{a,b}$	3.9^a	$6.1^{a,b}$	9.5^b	$7.7^{a,b}$	$5.0^{a,b}$	$3.5^{a,b}$	$3.6^{a,b}$	5.8	$2.39_{(7,307)}$	.021
Have DVD player	$15\%^{a,b}$	$7\%^{a,b}$	$13\%^{a,b}$	$25\%^a$	$11\%^{a,b}$	$0\%^b$	$25\%^{a,b}$	$6\%^{a,b}$	12%	$3.20_{(7,306)}$	.003
Have camcorder	$15\%^a$	$36\%^{a,b}$	$56\%^b$	$58\%^b$	$37\%^{a,b}$	$40\%^{a,b}$	$63\%^{a,b}$	$34\%^{a,b}$	46%	$3.12_{(7,306)}$	.003
Have satellite dish	$10\%^{a,b}$	$0\%^a$	$4\%^{a,b}$	$17\%^b$	$5\%^{a,b}$	$6\%^{a,b}$	$13\%^{a,b}$	$0\%^a$	7%	$2.51_{(7,306)}$	.016
How eager for DTV (0-10 scale)	0.8^a	$2.1^{a,b}$	$3.0^{a,b}$	3.3^b	$1.5^{a,b}$	$1.7^{a,b}$	$2.3^{a,b}$	$2.1^{a,b}$	2.4	$2.43_{(7,289)}$	.020
Depression scale (0-140 scale)	$34.6^{a,b}$	$27.7^{a,b}$	$31.5^{a,b}$	39.0^a	$22.4^{a,b}$	$27.2^{a,b}$	$26.2^{a,b}$	19.4^b	29.9	$3.28_{(7,304)}$	.001
Wallowing scale (0-100 scale)	29.7^a	27.3^a	33.1^a	33.4^a	29.3^a	34.9^a	26.9^a	26.9^a	31.3	$2.25_{(7,312)}$	.030
Favorite TV show is comedy	$27\%^a$	$44\%^{a,b}$	$63\%^{a,b}$	$63\%^{a,b}$	$83\%^b$	$46\%^{a,b}$	$64\%^{a,b}$	$57\%^{a,b}$	57%	$2.60_{(7,265)}$	.013
Favorite TV show is weepy/melodrama	$47\%^a$	$31\%^a$	$15\%^a$	$11\%^a$	$11\%^a$	$11\%^a$	$7\%^a$	$19\%^a$	17%	$2.13_{(7,265)}$	.041
Favorite movie is weepy (2), melodrama (1), or not (0)	$1.2^{a,b}$	$0.8^{a,b}$	0.4^a	$0.7^{a,b}$	$0.9^{a,b}$	$0.9^{a,b}$	1.5^b	$0.9^{a,b}$	0.8	$3.41_{(7,272)}$	.002
Violence in favorite movie (graphic=2, lite=1; no=0)	$.46^{a,b}$	$.63^{a,b}$	1.07^a	1.00^a	$.68^{a,b}$	$.36^b$	$.58^{a,b}$	$.74^{a,b}$	.78	$3.69_{(7,272)}$	.001

Means that do not share a superscript are significantly different at p<.05 using Tukey's HSD post hoc test.

Source: Neuendorf, K. A., & Skalski, P. (2000, June). *Senses of humor: The development of a multi-factor scale in relationship to moving image utility*. Paper presented to the Mass Communication Division of the International Communication Association, Acapulco, Mexico.