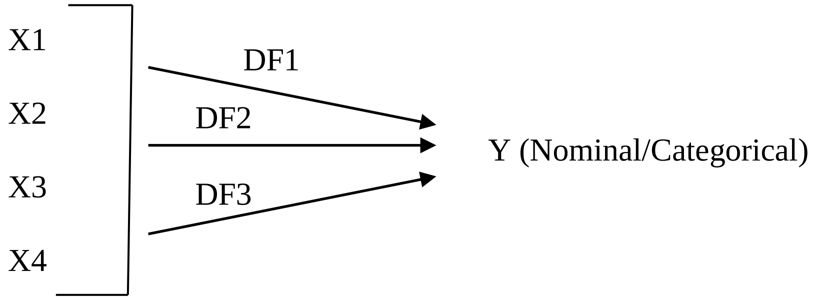
*Neuendorf*

**Discriminant Analysis**

**The Model**



**Assumptions:**

1. Metric (interval/ratio) data for 2+ IVs, and categorical (nominal) data for a single DV

2. Linearity--in relationships among variables--discriminant functions (DFs) are linear constructions (variates) of the IVs that best differentiate among the DV groups.

   The number of DFs that may be derived is:

   c-1 (where c=# of categories on the DV)<br>
   OR<br>
   k (where k=# of IVs),

   whichever is smaller.

3. Univariate and multivariate normal distributions for the IVs

4. Little or no multicollinearity. However, SPSS will not assess this in the Discriminant procedure; we can run a "fake" Multiple Regression to at least get the tolerances. . .

5. Homogeneity of variances/covariances (for the different DV groups). . . Box's M tests the assumption of homogeneity of variances/covariances of the DV groups. Based on the determinants of the group variance/covariance matrices, Box's M uses an F transformation. A significant F indicates substantial group differences, showing heterogeneity of variances/covariances, a type of *heteroscedasticity* (which we do not want).

**Decisions to make:**

1. Simultaneous/Forced entry ("Enter independents together," in SPSS-ese) OR Stepwise entry of IVs; Hierarchical models are not available

2. Use (or not) of a hold-out sample for validation of the discriminant function. This is a split halves test, where a portion of the cases are randomly assigned to an *analysis sample* for purposes of deriving the discriminant function(s), and then the function(s) are validated by assessing their performance with the remaining cases in the *hold-out sample*.

**Statistics:**

**Part #1**: Derivation and Interpretation of Discriminant Functions:

1. Box's M—as noted above, this tests the assumption of homogeneity of variances/covariances across the DV groups. We hope for non-significance.

2. Standardized canonical discriminant coefficients/weights--like betas in multiple regression, they indicate the relative, unique contribution of each IV to each DF (discriminant function) (in "Standardized Canonical Discriminant Function Coefficients" table in SPSS).

   Each "ß" below:

   $DF1 = ß_1X1_z + ß_2X2_z + ß_3X3_z + \ldots$
   $DF2 = ß_4X1_z + ß_5X2_z + ß_6X3_z + \ldots$
   etc.

   Strangely, Hair et al. call the calculated DF1, DF2, etc., scores "Discriminant Z scores," which seems to invite confusion with simple standardized scores (z-scores).

3. Structure coefficients/discriminant "loadings" (in SPSS's "Structure Matrix")--simple r's between each IV and a DF. These loadings are viewed by many as a better way to interpret a DF, since the discriminant coefficients are partials and these loadings are not. A common cutoff value is the absolute value of .4. [NOTE: The term "loading" may have a slightly different meaning across different statistical procedures and also across stat books. In general, it refers to some coefficient that we are supposed to use to make sense of a variate (factor, discriminant function, etc.).]

4. An eigenvalue for each DF--the eigenvalue has no absolute meaning (much like eigenvalues in factor analysis, they have only "comparative" meaning). As Klecka says, "they cannot be interpreted directly." Each eigenvalue is a *relative* measure of how much of the total discriminating power a DF has. Examining the eigenvalues tells us the relative strength of each DF. For example, from Klecka:

| DF | Eigenvalue | Relative % |
|---|---|---|
| 1 | 9.66 | 85.5% |
| 2 | 1.58 | 14.0 |
| 3 | .05 | 0.5 |

5. Wilks' lambda (Λ)--assesses the statistical significance of each DF, based on eigenvalues. It is a multivariate measure of group differences over several IVs. Rather than testing a DF itself, lambda examines the residual discrimination in the system *prior to* deriving that function (Klecka). Λ is interpretable as an *inverse* measure of how much discrimination there is among the groups (i.e., how much the groups differ on the pool of IVs). As DFs are derived, the lambda typically starts small and gets bigger, and ranges from 0 to 1:

0 -------------------------------------------------------------- 1

| | |
|---|---|
| No discrimination (by DF(s))<br>among groups | Great discrimination (by DF(s))<br>among groups |

One formula for lambda:

$$\Lambda = \prod_{i=k+1}^{q} \frac{1}{1 + \text{eigen}_i}$$

where Π is like Σ, only with multiplication instead of addition, and q = # of DFs total, k = # of DFs derived at that point

So, from the e.g., in #3 above:

$$\Lambda = \frac{1}{1+9.66} \quad x \quad \frac{1}{1+1.58} \quad x \quad \frac{1}{1+.05}$$

= .035 for NO DFs DERIVED YET

Thus, with Λ = .035, there's a lot of discrimination left to capture by the deriving of DF(s). Each Λ is tested with a chi-square statistic. To follow through on our example:

| | | Wilks' | Chi-square | | |
|---|---|---|---|---|---|
| DFs derived | "Test of Function(s)" | lambda (Λ) | ($\chi^2$) | df | Sig. (*p*) |
| 0 | 1 through 3 | .035 | 43.76 | 18 | .001 |
| 1 | 2 through 3 | .368 | 13.00 | 10 | .224 |
| 2 | 3 | .949 | 0.68 | 4 | .954 |

How many DFs are significant? [Answer: 1] NOTE: SPSS uses the column titled "Test of Function(s)." Interpret this as "Test of the significance of aggregate group differences for the pool of IVs *prior to* deriving this/these DF(s)."

The calculation of df (from Klecka, p. 40) is:

df = (p-k) (g-k-1)

where p = # of IVs

g = # of categories on DV

k = # of DFs derived at that point

6. Canonical correlation coefficient (CC)—This is another way to judge the substantive utility of a DF. Each DF has a CC with the DV (treated here as a collection of c-1 dummies). The squared CC ($CC^2$ = coefficient of determination) is shared variance, as always. Here, the shared variance is between the individual DF and a set of dummies representing the DV groups.

7. Group centroids--the means of the DFs are reported for each of the DV groups. This is central to discriminant analysis, yet is sometimes overlooked in writeups. It tells us *how* the groups differ on the function(s) that have been derived for that very purpose. We can look at these centroids graphically in SPSS's *territorial maps*, which plot the centroids in the first two dimensions, i.e., the first two DFs.

**Part #2:** Group Classification (a rather *practical* aspect of Discriminant Analysis):

8. The territorial map--the optimal cutting scores are shown visually for two DFs at a time in a territorial map. With this SPSS output component, you can plot the position of any given individual case for the two DFs, and see which group that individual is predicted to be in.

9. Classification matrix (found in SPSS's "Classification Results")--a chart shows predicted group membership against actual group membership. We hope for large values on the diagonal, and small values on the off-diagonal. We also hope for a high "percent. . . correctly classified." The pattern shown in the matrix can be assessed for statistical significance with two different statistics--tau and Press' Q. Neither is provided in the SPSS output; each must be calculated by hand (but neither is very difficult). Here they are:

10. Tau--very much like a special form of a $\chi^2$, it tests whether a given classification analysis improves one's prediction to groups over chance.

$$\text{Tau} = \frac{n_{cor} - \Sigma p_i n_i}{n - \Sigma p_i n_i}$$

where:

$n_{cor}$ = # of cases correctly classified
$n$ = # of cases
$p_i$ = chance probability of membership in each group (e.g., .25 for each of 4 groups)
$n_i$ = # of cases in that group
$i$ = each group

This test for "classification errors" is interpreted as the proportion fewer errors obtained by the classification analysis than what would be expected by chance (see Klecka p. 51 for more info.)

11. Press' Q—an alternative to tau, its calculation is shown on p. 266 of Hair, and below. Using a chi-square table, with 1 degree of freedom, one can actually get a *significance test* for the difference from chance.

$$\text{Press' Q} = \frac{[N - (nK)]^2}{N(K - 1)}$$

where

N = total sample size
n = number of observations correctly classified
K = number of groups on the dependent variable

12. Fisher's linear discriminant functions (i.e., classification functions)--*not* to be confused with the DFs. These are contained in the "Classification Function Coefficients" table in SPSS, and provide a handy-dandy method of placing a "new" case in its predicted DV group without running data through SPSS. That is, a new case's values for the IVs (raw, unstandardized) may be inserted in the functions and a score is calculated for each function for that case. The case is then classified into the group for which it has the highest classification score. This is a practical application, rather than informative of relationships among variables. This is sometimes used in clinical situations.

Additional references:

Klecka, W. R. (1991). *Discriminant analysis*. Newbury Park, CA: Sage.

Discriminant Analysis
COM 631, 2008 “Classic” Class Example:

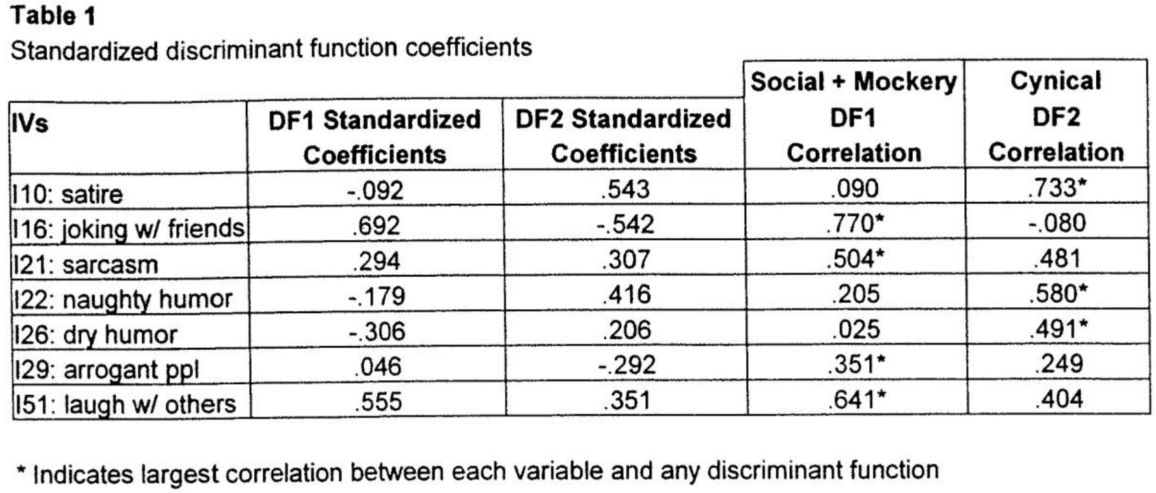


**Table 1**
Standardized discriminant function coefficients

| IVs | DF1 Standardized Coefficients | DF2 Standardized Coefficients | Social + Mockery DF1 Correlation | Cynical DF2 Correlation |
|---|---|---|---|---|
| I10: satire | -.092 | .543 | .090 | .733* |
| I16: joking w/ friends | .692 | -.542 | .770* | -.080 |
| I21: sarcasm | .294 | .307 | .504* | .481 |
| I22: naughty humor | -.179 | .416 | .205 | .580* |
| I26: dry humor | -.306 | .206 | .025 | .491* |
| I29: arrogant ppl | .046 | -.292 | .351* | .249 |
| I51: laugh w/ others | .555 | .351 | .641* | .404 |

* Indicates largest correlation between each variable and any discriminant function

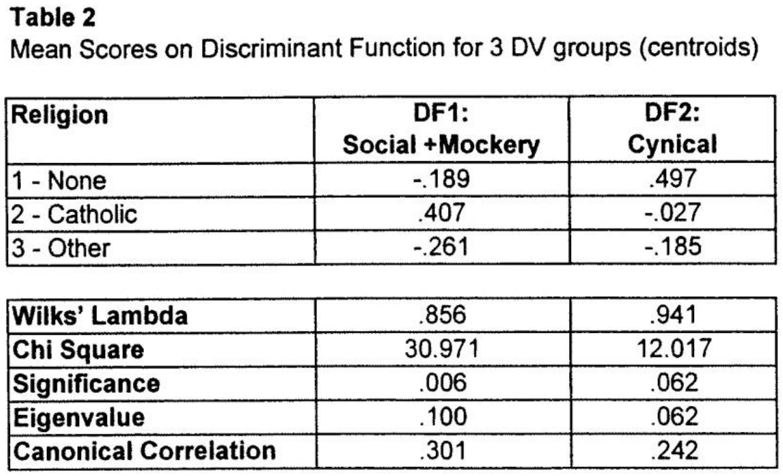


**Table 2**
Mean Scores on Discriminant Function for 3 DV groups (centroids)

| Religion | DF1: Social +Mockery | DF2: Cynical |
|---|---|---|
| 1 - None | -.189 | .497 |
| 2 - Catholic | .407 | -.027 |
| 3 - Other | -.261 | -.185 |

| | | |
|---|---|---|
| **Wilks’ Lambda** | .856 | .941 |
| **Chi Square** | 30.971 | 12.017 |
| **Significance** | .006 | .062 |
| **Eigenvalue** | .100 | .062 |
| **Canonical Correlation** | .301 | .242 |

**Table 3**
Classification Matrix results for 3 group discriminant analysis

| **Actual Group:** | | **Predicted Group:** | | |
|---|---|---|---|---|
| **Group** | **Actual Group Size** | **None** | **Catholic** | **Other** |
| None | 38 | **19** | 10 | 9 |
| Catholic | 76 | 16 | **41** | 19 |
| Other | 91 | 30 | 22 | **39** |
| Total | 205 | | | |

48.3% of original grouped cases were correctly classified.

**Press' Q:**

$$\frac{[N-(nK)]^2}{N(K-1)}$$

N=205
n=99
K=3

$$\frac{[205-(99*3)]^2}{205(3-1)} =$$

$$\frac{[205-(297)]^2}{410} =$$

$$\frac{8464}{410} =$$

20.64

df=1 on chi square table
Significant at less than .001