A BASIC program for classification analysis using proportional and maximum chance estimates of prior probabilities

ROBERT S. SCHLOTTMANN
Oklahoma State University, Stillwater, Oklahoma

Applied behavioral research often involves the development of procedures for assigning individuals to various groups. It is then necessary to determine whether or not the separate group hit rates and the total group hit rate are statistically better than what can be expected by chance. Assuming that sample sizes are reflections of population characteristics, the chance frequency of hits for a given group is often represented by the proportion of the group's size to the total sample. The total group chance hit rate is then the sum of the separate group chance hit rates. This is referred to as the *proportional chance criterion* (Huberty, 1984; Morrison, 1969). However, when groups differ greatly in size, the chance frequency of hits is sometimes estimated by assigning all individuals to the group with the greatest number in the population. This is the *maximum chance criterion* (Huberty, 1984; Morrison, 1969).

As an example, consider the fictitious data presented in Table 1, in which we try to classify 314 individuals into three groups: an organic brain syndrome group (OBS), a psychiatric group (PSY), and a normal group (NOR). Although it may be tempting to analyze such frequency data using the Pearson chi-squared statistic, to do so is inappropriate (Huberty, 1984), since we are only interested in the diagonal entries. In such cases, a significant chi-squared statistic would not necessarily imply that the classifications were better than chance. Huberty (1984) outlines in detail the appropriate procedures to use in analyzing data in classification tables.

A sample run of a BASIC program for an analysis using the data in Table 1 appears in the Appendix. The program runs on the IBM PC and compatibles. After first typing BASICA (or GWBASIC), the program CLASS.BAS. is loaded. (Assuming the program is on a disk in drive A, type LOAD "A:CLASS" and press the Return key.) The printer is then turned on and the Crtl key depressed while one presses PrtSc. This results in a printout of everything on the screen. To run the program, F2 is pressed. A prompt will then appear for the user to enter the dataset name, the number of groups involved, the names of the three groups, and the frequencies in each cell. After first calculating the row and column total and the grand total, the program sums the diagonal entries to determine the overall number of hits (o). In the example above, it is 200 (20+30+150). This number is then di-

vided by N to provide the percent correct classification (63.69).

Using the proportional chance criterion, the total group chance frequency of hits (e) is estimated by finding each group's proportion in the total sample and multiplying it by the group's sample size. In our example, e = (30/314 × 30) + (75/314 × 75) + (209/314 × 209). The total group chance hit rate (e/N) is then calculated. The observed number of hits is 200, the chance frequency is 159.89, and the total group chance hit rate is .51. Since the observed hit rate is greater than the chance hit rate, we can test to see if the difference is statistically significant, using the standardized normal statistic (z) as outlined by Huberty (1984, p. 166):

$$z = (o - e) \sqrt{N} / \sqrt{e(N - e)} \ .$$

Since the only interest in testing is to see if the observed frequency of hits is greater than the chance frequency of hits, and not vice versa, a one-tailed test is used. For the sample data, the z of 4.53 (rounded) yields a p value of .000.

The sample run yielded an observed hit rate (o/N) of 63.69%, using the proportional chance criterion. If instead we had predicted everyone to fall into the normal group, a 66.56% hit rate could have been achieved, since all 209 of the normals would have been correctly classified (209/314 × 100 = 66.56). This hit rate would have been obtained if the maximum chance criterion had been used to define "chance." Using 209 as the expected hit rate and recalculating z, the results are not statistically significant ($z = -1.08$).

Which definition of chance should be used? One should use the proportional chance criterion if group sizes are nearly equal, and the maximum chance criterion if they are quite diverse (Huberty, 1984, p. 170). In addition, it may be important to determine whether individuals in each group can be identified separately. In the example data, it may be very important to identify individuals with an organic brain syndrome even though normals or psychiatric cases may be misclassified. If so, we need to look at separate group hit rates. In the sample run data, individuals with organic brain syndromes and psychiatric cases can be identified, but normals cannot. Assuming we want to take into consideration the fact that three tests

---

Correspondence may be addressed to Robert S. Schlottmann, Department of Psychology, Oklahoma State University, Stillwater, OK 74078.

Table 1
Fictitious Data for Three Groups: Organic Brain Syndrome (OBS), Psychiatric (PSY), and Normal (NOR)

| Actual Group | Predicted Group | | | |
	OBS	PSY	NOR	Total
OBS	20	5	5	30
PSY	22	30	23	75
NOR	27	32	150	209
Total	69	67	178	

were conducted, in assessing the significance for each test, each obtained p value must be multiplied by three.

The program also enables the user to calculate an improvement over chance index, I:

$$I = (H_O - H_E)/(1 - H_E)$$

where H_O is the observed hit rate and H_E is the hit rate expected by chance (see Huberty, 1984, p. 168). For the sample data, the proportional chance criterion resulted in an improvement over chance of 26.02%, while the maximum chance criterion indicates that the classification procedures were not as good as could have been obtained had everyone been classified as normal. However, if separate group hit rates are important, the fictitious classification procedures resulted in an improvement over chance of 63.15% for individuals with an organic brain syndrome and much lower hit rates with the other two groups. (For examples of the use of I with real rather than fictitious data, see Sexton, McIlwraith, Barnes, & Dunn, 1987; Walters, 1985, 1986).

Program availability. A listing of the program can be obtained from Robert S. Schlottmann, Department of Psychology, Oklahoma State University, Stillwater, OK 74078. Alternatively, send a formatted disk and an addressed disk mailer, and a copy of the program will be made to your disk free of charge.

REFERENCES

HUBERTY, C. J. (1984). Issues in the use and development of discriminant analysis. *Psychological Bulletin*, **95**, 156-171.

MORRISON, D. J. (1969). On the interpretation of discriminant analysis. *Journal of Marketing Research*, **6**, 156-163.

SEXTON, D. L., McILWRAITH, R., BARNES, G., & DUNN, R. (1987). Comparison of the MCMI and MMPI-168 as psychiatric inpatient screening inventories. *Journal of Personality Assessment*, **51**, 388-398.

WALTERS, G. D. (1985). Scale 4 (Pd) of the MMPI and the diagnosis of antisocial personality. *Journal of Personality Assessment*, **49**, 474-476.

WALTERS, G. D. (1986). Screening for psychopathology in groups of black and white prison inmates by means of the MMPI. *Journal of Personality Assessment*, **50**, 257-264.

APPENDIX
Sample Run

RUN
DATASET NAME: SAMPLE DATA

NUMBER OF GROUPS INVOLVED? 3

NAME OF GROUP 1 ? OBS
NAME OF GROUP 2 ? PSY
NAME OF GROUP 3 ? NOR

NUMBER OF CASES IN GROUP OBS THAT WERE PREDICTED TO BE IN GROUP OBS? 20
NUMBER OF CASES IN GROUP OBS THAT WERE PREDICTED TO BE IN GROUP PSY? 5
NUMBER OF CASES IN GROUP OBS THAT WERE PREDICTED TO BE IN GROUP NOR? 5
NUMBER OF CASES IN GROUP PSY THAT WERE PREDICTED TO BE IN GROUP OBS? 22
NUMBER OF CASES IN GROUP PSY THAT WERE PREDICTED TO BE IN GROUP PSY? 30
NUMBER OF CASES IN GROUP PSY THAT WERE PREDICTED TO BE IN GROUP NOR? 23
NUMBER OF CASES IN GROUP NOR THAT WERE PREDICTED TO BE IN GROUP OBS? 27
NUMBER OF CASES IN GROUP NOR THAT WERE PREDICTED TO BE IN GROUP PSY? 32
NUMBER OF CASES IN GROUP NOR THAT WERE PREDICTED TO BE IN GROUP NOR? 150

ROW TOTAL FOR ACTUAL GROUP OBS IS 30
ROW TOTAL FOR ACTUAL GROUP PSY IS 75
ROW TOTAL FOR ACTUAL GROUP NOR IS 209

COLUMN TOTAL FOR PREDICTED GROUP OBS IS 69
COLUMN TOTAL FOR PREDICTED GROUP PSY IS 67
COLUMN TOTAL FOR PREDICTED GROUP NOR IS 178

TOTAL NUMBER OF CASES, N, IS 314

THE OVERALL NUMBER OF HITS (0) IS 200

THE PERCENT OF CASES CORRECTLY CLASSIFIED, O/N, IS 63.69427%

PROPORTIONAL CHANCE CRITERION: OVERALL

THE TOTAL-GROUP CHANCE FREQUENCY OF HITS, E, IS 159.8917

APPENDIX (Continued

THE TOTAL-GROUP CHANCE HIT RATE, E/N, IS .5092094

THE STANDARDIZED NORMAL STATISTIC, Z, IS 4.527651

PROPORTIONAL CHANCE CRITERION: SEPARATE GROUPS

THE CHANCE FREQUENCY OF HITS, E(G), FOR GROUP OBS IS 2.866242
THE STANDARDIZED NORMAL TEST STATISTIC, Z, FOR GROUP OBS IS 10.64148

THE CHANCE FREQUENCY OF HITS, E(G), FOR GROUP PSY IS 17.91401
THE STANDARDIZED NORMAL TEDT STATISTIC, Z, FOR GROUP PSY IS 3.273042

THE CHANCE FREQUENCY OF HITS, E(G), FOR GROUP NOR IS 139.1115
THE STANDARDIZED NORMAL TEST STATISTIC, Z, FOR GROUP NOR IS 1.59646

*** NOTE: TO CONTROL FOR 'PROBABILITY PYRAMIDING' THAT RESULTS
          FROM MULTIPLE TESTS, CONSIDER TESTING THE ABOVE FOR
          SIGNIFICANCE AT THE ALPHA/K LEVEL (ONE-TAILED).

MAXIMUM CHANCE CRITERION

THE STANDARDIZED NORMAL STATISTIC, Z,
    USING THE MAXIMUM CHANCE CRITERION IS −1.076563

THE PERCENT OF CASES CORRECTLY CLASSIFIED
    USING THE MAXIMUM CHANCE CRITERION IS 66.56051%

IMPROVEMENT OVER CHANCE: OVERALL

THE IMPROVEMENT OVER CHANCE, I, USING
    THE PROPORTIONAL CHANCE CRITERION IS 26.02604%

THE IMPROVEMENT OVER CHANCE, I, USING
    THE MAXIMUM CHANCE CRITERION IS −8.571417%

IMPROVEMENT OVER CHANCE: SEPARATE GROUPS

THE IMPROVEMENT OVER CHANCE FOR GROUP OBS IS 63.14554%
THE IMPROVEMENT OVER CHANCE FOR GROUP PSY IS 21.17155%
THE IMPROVEMENT OVER CHANCE FOR GROUP NOR IS 15.57986%

*** NOTE: THE MAXIMUM CHANCE CRITERION IS NOT RELEVANT
          TO SEPARATE GROUP HIT RATES.
Ok