Nonparametric analysis of signal detection confidence ratings

THOMAS C. B. DAVISON
and
RICHARD J. JAGACINSKI
Ohio State University, Columbus, Ohio 43210

The objective of the program is to provide statistics appropriate to nonparametric analysis of receiver operating characteristic (ROC) curves obtained from confidence ratings. The area under the ROC curve, A_g, is computed by the trapezoidal rule (Bamber, 1975; Pollack, Norman, & Galanter, 1964) to provide a nonparametric measure of sensitivity. This measure can be regarded either as an estimate of the true area under the ROC curve, or independent of signal detection theory as simply a measure of the relative overlap of two distributions of categorical responses. The standard error of A_g is estimated in two different ways: (1) by an unbiased nonparametric statistic, s_a, derived by Bamber (1975, p. 402) from the work of Noether (1967); (2) by a binomial approximation, $\sqrt{A_g(1 - A_g)/(n - 1)}$, in which n is the number of noise or signal trials, whichever is smaller.

The concept of sensitivity in signal detection theory refers to the relative overlap of two distributions, namely, for confidence rating data, $A_g = p(S > N) + \frac{1}{2}p(S = N)$ (after Bamber, 1975, p. 401), where $p(S > N)$ is the proportion of signal-trial confidence ratings that exceed noise-trial confidence ratings, and $p(S = N)$ is defined similarly. The standard error, s_a, is thus closely related to the variance of the Mann-Whitney U statistic.

Regarding the binomial approximation, Pollack and Hsieh (1969) used Monte Carlo simulations to show that for normal, uniform, and negative exponential distributions with equal numbers of signal and noise trials, $\sqrt{A(1 - A)/n}$ slightly overestimates the size of the standard error, where A is the expected value of A_g. Birnbaum and Klose (1957) and Van Dantzig (1951, Footnote 4) proved that this expression provides an upper bound for the standard error of A_g if the underlying distributions are continuous. Bamber (1975) proved that this upper bound is also valid for finitely discrete underlying distributions, and recommended using the conservative n − 1 in the denominator. The difference between s_a and the binomial approximation is generally slight, but may be considerable, such as

when most of the confidence ratings fall in a single category, or when A_g is very close to 1.0 (Pollack & Hsieh, 1969). For a discussion of using these estimates of variability to form confidence intervals, see Bamber (1975).

Measures of response bias are often not used in confidence rating experiments. The program does, however, provide a rough measure of response bias by calculating the interpolated category rating at which the ROC curve crosses the negative diagonal (McNicol, 1972). The negative diagonal is the locus of points where the observer is equally disposed toward signal or noise responses, and corresponds to the point where β, the likelihood ratio, equals 1.0 for parametric signal detection theory with normal, equal variance signal and noise distributions. This point also corresponds to the median interpolated category rating if there are equal numbers of signal and noise trials.

The program also tabulates other information. For each rating category, the proportions of signal and noise trials are computed, as are the cumulative proportions used to plot the ROC curve. As a rough indication of the form of the ROC curve, the number of points lying on or above the positive diagonal are counted. When the rating scale has an even number of response categories, the program outputs a 2 by 2 stimulus by response matrix as if yes-no responses had been made.

Input. For each analysis the inputs are the number of response categories and either the frequencies or proportions of responses in each category for both the signal and noise trials. If proportions are input, the variance estimators cannot be calculated. Two data cards are input for each analysis, a title card and a data card.

Limitations. For complete analysis and tabulation of data, the program accepts between two and eight response categories. The program calculates A_g and variance estimates for data with as many as 30 categories, but does not provide complete tabular output of stimulus by response matrices. Instructions for modifying the program to handle more than 30 categories are included.

Language and Computers. The program is largely written in ANSI standard FORTRAN IV, with two exceptions: Mixed-mode arithmetic is used, and character strings for output are enclosed in quotes rather than being specified by Hollerith specifications. The program has been tested on IBM 370/168 and CDC 6400 computers.

Availability. A card deck and listing of the program including instructions for use are available at no cost from Richard J. Jagacinski, Human Performance Center, 404-B West 17th Avenue, Columbus, Ohio 43210.

REFERENCES

BAMBER. D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*. 1975. **12**. 387-415.

BIRNBAUM. Z. W.. & KLOSE. O. M. Bounds for the variance of the Mann-Whitney statistic. *Annals of Mathematical Statistics*. 1957. **28**. 933-945.

McNICOL. D. *A primer of signal detection theory*. London: George Allen & Unwin. 1972.

NOETHER. G. E. *Elements of nonparametric statistics*. New York: Wiley. 1967.

POLLACK. I.. & HSIEH. R. Sampling variability of the area under the ROC curve and of d'_e. *Psychological Bulletin*. 1969. **71**. 161-173.

POLLACK. I.. NORMAN. D. A.. & GALANTER. E. An efficient nonparametric analysis of recognition memory. *Psychonomic Science*. 1964. **1**. 327-328.

VAN DANTZIG. D. On the consistency and power of Wilcoxon's two sample test. *Koninklyke Nederlandse Akademie van Wetenschappen. Proceedings. Series A*. 1951. **54**. (Also: Indagationes Mathematical. **13**). 1-8. As cited in Bamber. D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*. 1975. **12**. 387-415.