THE POWER OF A PROCEDURE FOR DETECTING MIXTURE DISTRIBUTIONS IN LATERALITY DATA.

I. C. McManus

(Department of Psychology, Bedford College, University of London, and Department of Psychiatry, St. Mary's Hospital Medical School, London W2)

In psychology and biology one frequently suspects that a distribution is not a single normal, log.normal, exponential, Poisson or other distribution, but rather that it is a mixture of two or more simpler distributions, with the parameters of the component distributions unknown in advance. On theoretical grounds we may expect such a situation to occur in psychology if, say, subjects are using different strategies to carry out a task or if, in the specific case of lateralisation of function, we expect a subset of the population to have a reversed pattern of lateralisation (McManus, 1983a). In biology we may well expect mixture distributions if a single major gene locus is affecting a character, in the presence of a host of relatively minor polygenic or environmental factors. In medicine we might expect a mixture distribution if a disease is heterogenous in aetiology or presentation (see e.g. Everitt, 1981a; Kendell and Brockington, 1980), or if a variable such as blood pressure is indicative of a sub-group of the population with different aetiology, or disease course (McManus, 1982a; 1983b).

In searching for mixture distributions one requires not just a procedure that determines whether or not a distribution is fitted by a single distribution (e.g. as has erroneously been used by Asberg et al., 1976) but rather it is required that a mixture distribution should be shown to be a significantly better fit than a single distribution. Everitt and Hand (1981) have recently reviewed many aspects of mixture distributions, and they suggest (p. 117) that the power of tests in relatively low. Since *ad hoc* use of a program of my own has suggested that power is fairly low unless substantial sample sizes are used, the present study has systematically investigated the power of a maximum-likelihood method. The importance of such analyses is that they will prevent type II errors by researchers; i.e. the inadvertent conclusion that a distribution is not a mixture when in fact the study has little power to detect a mixture. In view of the general analytic intractability of mixture distributions, a Monte Carlo method has been used.

Cortex (1984) *20*, 421-426

MATERIALS AND METHOD

Consider a data set, x_i, $i=1, n$. We may fit a normal distribution to these data, and estimate the mean and standard deviation, m and s i.e. $x_i=N(m, s)$. Alternatively we may fit a two-component mixture to x, where $x_i=p.N(m_1, s_1)+(1-p).N(m_2,s_2)$, with five free parameters, m_1 and m_2 being the means and s_1 and s_2 the standard deviations of the component distributions, and p is the proportion of individuals in the first component. In general one is willing to assume that $s_1=s_2$, and hence there are four parameters to be estimated from the data. Since in the present case I was specifically interested in the problems of laterality research (McManus, 1983a) I have considered the situation $x_1=(1-p).N(k/2,s)+p.N(-k/2),s)$, in which the two groups are symmetrically arranged with respect to zero (and hence $m=k(1-2p)/2$), and thus there are only three free parameters, p; k, and s; if $p=0$ then $m=k/2$. k can be considered as the *separation* between the component means. Maximum likelihood estimates of the parameters were found by a quasi-Newton-Raphson iterative method, implemented using the subroutine EO4JBF of the NAG program library (Numerical Algorithms Group, 1981). At the maximum the sub-routine gives estimates of the parameters, and also an estimate of the Hessian matrix; square roots of the diagonal elements of the inverse of the Hessian matrix were used as estimates of the standard error of each parameter.

The significance of the improvement of fit of the mixture distribution as opposed to the single normal distribution was found by calculating $-2.(\ln(L_1)-\ln(L_2))$, where L_1 and L_2 are the likelihoods of the single and the two-component distribution respectively, and treating this value as a chi-square distribution, the number of degrees of freedom being the difference between the number of free parameters of the two distributions (i.e. 1 degree of freedom in the present case).

Although on theoretical grounds the likelihood ratio statistic is not well defined in the present case since the parameter estimates are on the edge of the parameter space (see Aitkin et al., 1981), in the present situation I have used the conventional likelihood ratio test, since Everitt (1981b) has suggested that Wolfe (1971)'s modification of it is an adequate approximation, and even for the univariate case with sample size equal to 50 the chi-squared values obtained with the conventional formula and with Wolfe's modified formula are within 6% of one another, with the difference shrinking to 0.4% for $N=800$.

In order to avoid negative standard deviations and proportions the iterative procedure actually used the logarithm of the standard deviations and the logit of the proportion, so that all possible estimates after back transformation were valid; thus standard error estimates were actually on the transformed scales and hence confidence intervals were not necessarily symmetric about the estimate on the back-transformed scales.

For a particular sample size, n, and probability of an item coming from the minor distribution, p, the program was run for separation values (k) of 0, .5, 1, 1.5, 2, 3, 4 and 5, 100 replications being used for n values of 50, 100 or 200, and 50 replications being used for values of 400 and 800. For each value of k the proportion of significant results at an alpha level of .05 was calculated, and this proportion as a function of k was fitted by a cumulative normal distribution, from which estimates of the necessary separation for a power (beta) of .5 and .9 were calculated.

RESULTS

Figure 1 summarises the results of the study for n values of 50, 100, 200, 400, and 800, and for p values of .05, .1, .2, and .5. It can be seen that the necessary separation for detection of a difference at low n values is relatively large (3 — 5 standard deviations), even though n values of 50 or 100 are conventionally regarded as moderately high in empirica research. Furthermore the necessary separation for detection falls only relatively slowly as n increases. It can also be seen that p values of .2 are better detected than those of .5 or .1, and that p values of .05 are particularly difficult to detect. Since in the particular case of laterality experments on speech dominance we might well expect p values in the range 5 to 10%, this aspect has been studied in more detail. Figure 2 shows for a value of n of 100, and a separation of 3 standard deviations between the means, the power of detection of mixing for an alpha of .05, .01, and .00., and the median chi-squared values obtained for p values of .01(.01).1, .1(.05).5. It can be seen that power is optimal at a probability of between 2 and .25, and that the power falls away fairly rapidly below probabilities of about 0.1.

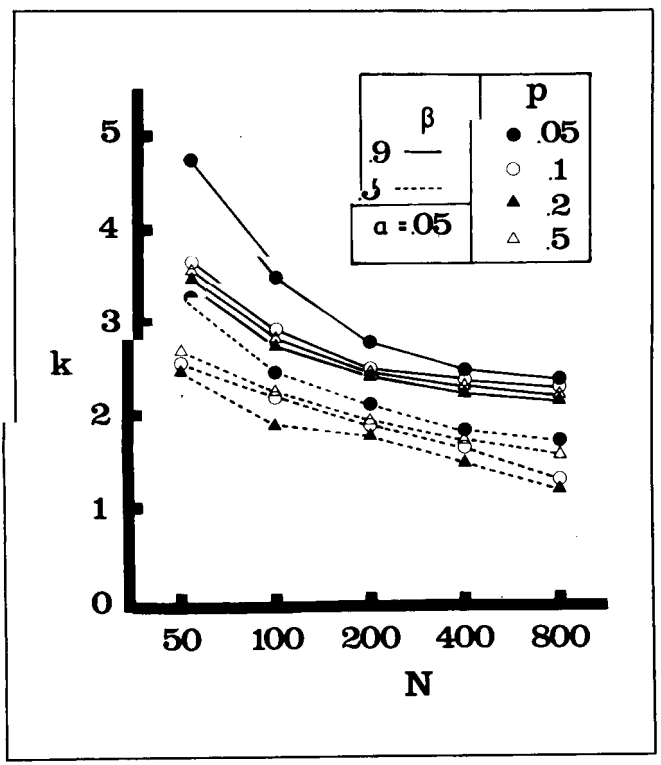

Fig. 1 — *Shows the standardised separation between the means, k, for achieving an alpha of 0.05 with a power, Beta, of 0.5 (——) or 0.9 (-----) when p=0.05, .1, .2 or .5.*

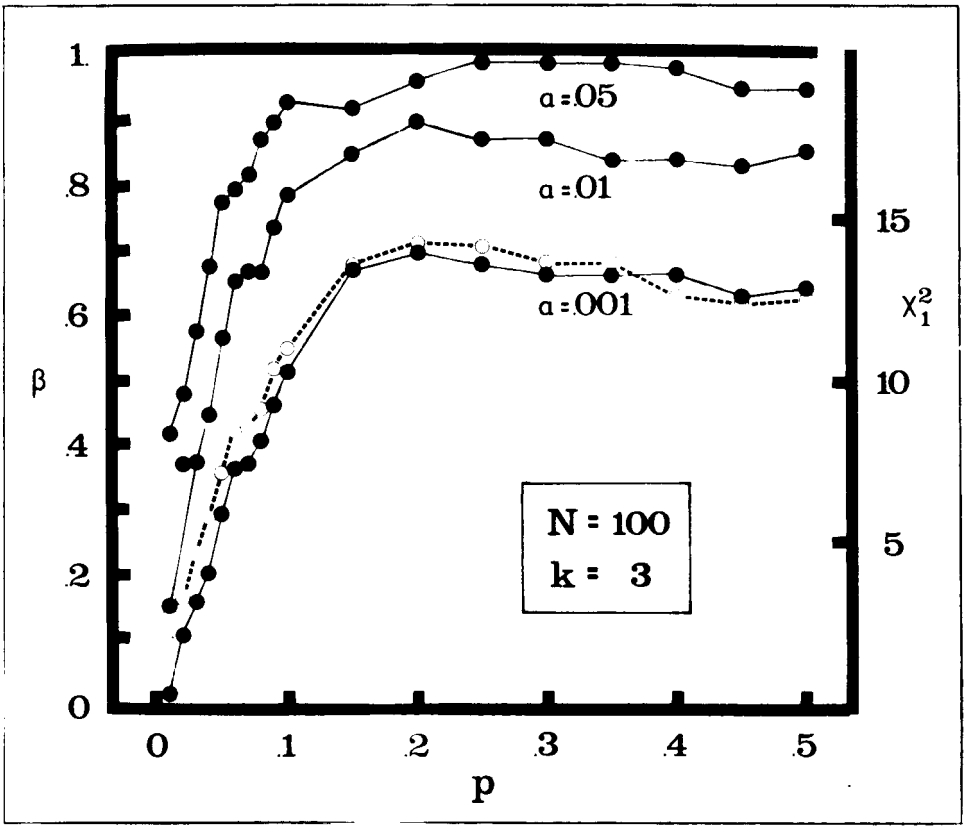

Fig. 2 — *Solid points show the proportion of significant results (Beta, left-hand ordinate) for an alpha level of .05, .01, and .001 for N = 100 and k = 3, and different levels of p (abscissa). The open points show the median chi-squared value (right-hand ordinate) for α = .001 in the same analysis.*

Estimation of Parameters

For each simulation the 95% confidence intervals of the estimated parameters were calculated and compared with the actual parameters of the distribution. For those simulations in which significant differences were found at the 0.05 level in at least 50% of cases, and excluding individual simulations in which a minor proportion of less than 0.001 was estimated, 5.92%, 5.88% and 0.13% of confidence intervals for the mean, standard deviation and proportion did not include the actual value. It is clear therefore that the method of estimation is adequate for the mean and standard deviation, but that the confidence interval for the proportion is erring on the conservative side, and this may reflect the need, for computational reasons, to use a logistic transformation of the proportion in

the actual calculations, combined with a back transformation at the end of the calculations. The estimates themselves showed no consistent bias, e.g. for $n=200$, $k=3$, $s=1$ and $p=.1$ after 100 replications the means of the estimates of k, s and p were 3.010, 1.001 and 0.099 respectively).

DISCUSSION

The present study suggests that fairly substantial sample sizes should be used in order to achieve an adequate power for detecting mixtures. Two questions arise; what sample size should be recommended for experiments on laterality; and how might the efficiency of the procedure be improved upon?

A recommendation on sample size is difficult to make. An estimate of the separation of the expected means is required. Five estimates may be considered. Data on skull asymmetry (McManus, 1982b) suggest that that the means are about 1.86 SD's apart, while studies of dichotic listening in bilinguals (Gordon, 1980; re-analysed in McManus, 1983a) suggest separations in the two languages of 4.27 and 4.13 SD's. A re-analysis of the data of Annett (1983) on asymmetry in a peg-moving task (Mc-Manus, 1984) suggests separations of 1.82 and 2.40 SD.s in males and females respectively. Taking the median of these five studies (2.40) as an exceptionally crude measure of expected effect size, then in order to achieve a power of 0.5 or 0.9 with an expected proportion int the sub-group of 0.05, and an alpha level of 0.05, then sample sizes of about 150 and 300 respectively would be required.

Might the large samples required be a reflection of an inefficient estimation procedure? This seems unlikely with respect to the statistical method itself, since it is a maximum likelihood procedure, and hence uses all of the available information in order to distinguish the two hypotheses. It is however possible that the numerical 'hill-climbing' algorithm is occasionally sub-optimal and failing to converge at an optimal solution. This will on occasion be inevitable if there is a small separation between the means, and if by chance all of the data points just happen to cluster close to the mean of the major distribution. An estimate of the frequency with which a complete failure of convergence occurs (i.e. estimated proportion in the minor distribution is less than 0.001) is 45%, 40%, 17% and 3% of occasions when $n=200$, $p=.1$, and separations are .5, 1, 1.5 and 2 respectively. For most reasonable effect sizes therefore this should not be a problem. It should also be borne in mind that such convergence failures have been excluded from the calculations of figures 1 and 2, and thus at low values of p and k it is necesary to reduce 'empirical' power-levels by a factor of from 3 to 50% to account for convergence failure.

ABSTRACT

The results of a Monte Carlo simulation of the power of an iterative version of a Maximum Likelihood procedure for the detection of mixture distributions are described. In general the power of such techniques is low, and for mixtures typical of those found in laterality research a sample size of about 150 or 300 is required to have a 50% or 90% probability of detecting the presence of a mixture.

REFERENCES

AITKIN, M., ANDERSON, D., and HINDE, J. Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, A, 144*: 419-461, 1981.

ANNETT, M. Right and left kand skill. II: Estimating the parameters of the distribution of L-R differences in males and females. *British Journal of Psychology, 74*: 269-283, 1983.

ASBERG, M., THOREN, P., TRASKMAN, L., BERTILSON, L., and RINGBERGER, V. "Serotonin depression" - a biochemical sub-group within the affective disorders. *Science, 191*: 478-480, 1976.

EVERITT, B.S. Bimodality and the nature of depression. *British Journal of Psychiatry, 138*: 336-339, 1981a.

EVERITT, B.S. A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioural Research, 16*: 171-180, 1981b.

EVERITT, B.S., and HAND, D.J. *Finite Mixture Distributions*. London: Chapman and Hall, 1981.

GORDON, H.W. Cerebral organisation in bilinguals. I: lateralisation. *Brain and Language, 9*: 255-268, 1980.

KENDELL, R.E., and BROCKINGTON, I.F. The identification of disease entities and the relationship between schizophrenia and affective psychoses. *British Journal of Psychiatry, 137*: 324-331, 1980.

MCMANUS, I.C. The distribution of blood pressure. *Clinical Science, 62:* 30P-31P, 1982a.

MCMANUS, I.C. The distribution of skull asymmetry in man. *Annals of Human Biology, 9*: 167-170, 1982b.

MCMANUS, I.C. The interpretation of laterality. *Cortex, 19*: 187-214, 1983a.

MCMANUS, I.C. Bimodality of blood pressure levels. *Statistics in Medicine, 2*: 253-258, 1983b.

MCMANUS, I.C. Right- and Left-hand Skill: Failure of the right shift model. *British Journal of Psychology*, (1984 (in press).

NUMERICAL ALGORITHMS GROUP *NAG Fortran Library Manual, Mark 8*, volume 3. Numerical Algorithms Group: Oxford, 1981.

WOLFE, J.H. A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. *Naval Personnel and Training Research Laboratory, Technical Bulletin, STB72-2*, San Diego, California (cited in Everitt, 1981b), 1981.

I.C. McManus, Department of Psychiatry, St. Mary's Hospital, London W2, U.K.