Behavior Research Methods & Instrumentation
1979, Vol. 11 (4), 467-468

An algorithm for assessing bimodality vs. unimodality in a univariate distribution

RONALD P. LARKIN
Rockefeller University, New York, New York 10021

Sometimes when dealing with continuous distributions the question arises as to how many populations are represented in a sample. If, in a continuous distribution, two or more modes are present, the experimental or observational techniques used in collecting the data may have inadvertently included a mixture of two or more distinct populations or tendencies. In that case, further analysis of the data must usually proceed along fundamentally different lines than had only one mode been present, or the experiments must be redesigned.

Although a number of techniques address this sort of question for multidimensional data sets (e.g., Tryon & Bailey, 1970), few techniques seem to be available for analyzing one-dimensional distributions in terms of degree of multimodality. A method for determining the "distinctness" of clusters by estimating the degree of overlap is given by Sneath (1977), but the method is not directly applicable to the bimodality problem. A more suitable approach was introduced by Engleman and Hartigan (1969). Following a suggestion by Hartigan (1975, 1977), the algorithm described here generates an F ratio that is small if a population is unimodal and larger if it is bimodal. The algorithm can be modified to handle more than two modes. It has been applied to a data set consisting of about 200 distributions of speeds of animal locomotion. In several cases the program located bimodal distributions that were not obvious when examining histograms by eye, in addition to providing a quantitative index of the degree of bimodality for each distribution.

Assumptions. Both modes within the overall distribution are assumed to fall somewhere near the maxima of normal or quasinormal subdistributions. However, the assumption of normality is more a convenient way to estimate clumping than a fundamental principle of the method. Neither mode in the distribution should be located near an extreme; the distribution must be shaped like an inverted "W" rather than like an "M" or a "V." This assumption is often met if the measurement techniques have been appropriate. The algorithm takes data tallied into an integer array in the form of a histogram.

Algorithm. The function BIMODF first computes the variance for the entire distribution considered as a unimodal one. It then locates two end points which are spaced an equal distance in from the two extremes of the distribution. The distribution is then divided into two parts repeatedly, once at each bin between the end points. The mean of the variances of the two parts is computed each time. The lowest such mean variance is used to represent the variance of the distribution considered as bimodal. The F ratio between the unimodal and bimodal variances is returned as an index of bimodality. The modes themselves are not located, but this can easily be done by the user if desired.

Performance. A unimodal, a bimodal, and an intermediate distribution are shown in Figure 1. The distributions having F ratios significantly bimodal at the 1% level were usually the same ones that seemed to be bimodal by eye, with some important exceptions. Highly skewed unimodal distributions sometimes gave large F ratios; this problem could easily be eliminated by adding refinements to the algorithm. Strongly leptokortic distributions that happened to have one or two

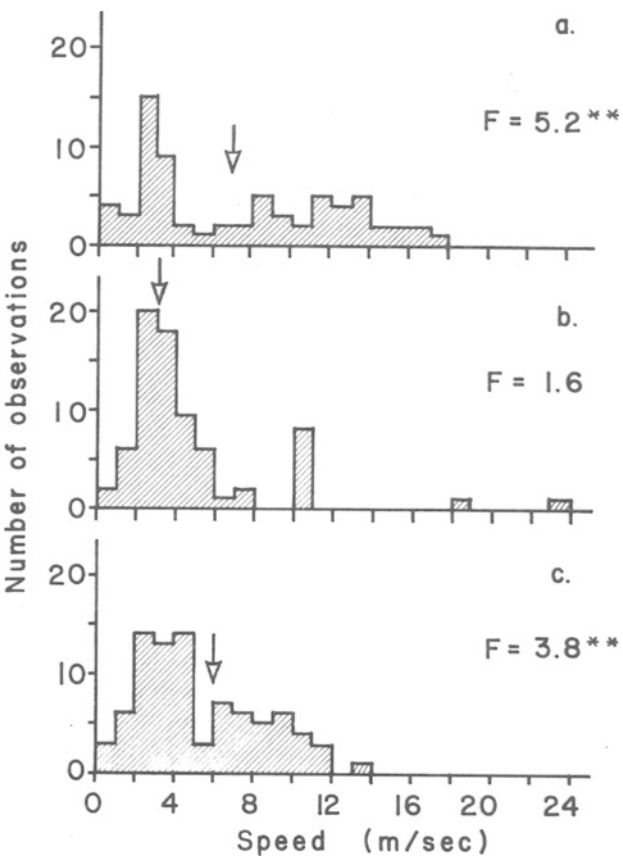

Figure 1. Histograms illustrating the performance of the bimodality algorithm. Arrows indicate the locations along the histograms where lowest F ratios were found. Data are the speeds of flight of migrating birds relative to the air. (a) 5/7/77, a bimodal case, N = 69; (b) 10/8/76, a unimodal case, N = 75; (c) 4/29/78, a distribution that is either unimodal and skewed or bimodal, N = 85. Distributions a and c are significantly bimodal at $p < .01$.

J. Cohen and R. Schor gave helpful comments on the manuscript.

0005-7878/79/040467-02$00.45/0

low values near the mode (making a deep notch in a single high peak) gave high F ratios. This could result in a scientifically misleading outcome even though the algorithm functioned correctly.

The test is a great asset in finding distributions having a second mode with only a few representatives, well separated from the primary mode. This type of bimodality is difficult to spot by eye.

The algorithm gives F ratios near 4.0 for rectangular distributions, almost independent of N. Curiously, 4.0 is roughly the value at which bimodality appears by eye and at the 1% level in actual samples of size 30-100.

Coding. Functions BIMODF and XMEAN are written in standard FORTRAN IV, using about 150 lines including comments. Inputs are the distribution and its limits and the fraction of the total N that should be excluded at each extreme. Outputs are the F ratio and the location of the bin at which the distribution is best separated into two distributions. In addition, some means and variances are returned to the calling program for use in debugging and location of actual modes. No input/output statements are employed; errors are signaled by F ratios and subdistribution means of zero. The subroutines took negligible computing time on a minicomputer without floating-point hardware using sample sizes less than 100.

Program Availability. A listing of the program and sample input (printed histograms) and output may be obtained at no cost from Ronald P. Larkin, Rockefeller University, New York, New York 10021.

REFERENCES

ENGLEMAN, L., & HARTIGAN, J. A. Percentage points of a test for clusters. *Journal of the American Statistical Association*, 1969, **64**, 1647-1648.

HARTIGAN, J. A. *Clustering algorithms*. New York: Wiley, 1975.

HARTIGAN, J. A. Distribution problems in clustering. In J. Van Ryzin (Ed.), *Classification of clustering*, New York: Academic Press, 1977.

SNEATH, P. H. A. A method for testing the distinctness of clusters: A test of the disjunction of two clusters in Euclidean space as measured by their overlap. *Mathematical Geology*, 1977, **9**, 123-143.

TRYON, R. C., & BAILEY, D. E. *Cluster analysis*. New York: McGraw-Hill, 1970.

(Accepted for publication June 22, 1979.)