SIGNIFICANCE TESTING IN AUTOMATIC INTERACTION DETECTION (A.I.D.)

by

KEITH JOHN WORSLEY

A Thesis Submitted
for the Degree of
Doctor of Philosophy
at the
University of Auckland

Department of Mathematics

June 1978

ACKNOWLEDGEMENTS

I would like to thank my supervisors Professors A. J. Scott and G. A. F. Seber, but my special thanks must go to Professor Scott, whose constant help and encouragement made this thesis possible.

ABSTRACT

Automatic Interaction Detection (A.I.D.) is the name of a computer program, first used in the social sciences, to find the interaction between a set of predictor variables and a single dependent variable. The program proceeds in stages, and at each stage the categories of a predictor variable induce a split of the dependent variable into two groups, so that the between groups sum of squares (BSS) is a maximum. In this way, the optimum split defines the interaction between predictor and dependent variable, and the criterion BSS is taken as a measure of the explanatory power of the split.

One of the strengths of A.I.D. is that this interaction is established without any reference to a specific model, and for this reason it is widely used in practice. However this strength is also its weakness; with no model there is no measure of its significance. Barnard (1974) has said:

> "... nowadays with more and more apparently sophisticated computer programs for social science, failure to take account of possible sampling fluctuations is leading to a glut of unsound analyses ... I have in mind procedures such as A.I.D., the automatic interaction detector, which guarantees to get significance out of any data whatsoever. Methods of this kind require validation ..."

The aim of this thesis is to supply part of that validation by investigating the null distribution of the optimum BSS for a single predictor at a single stage of A.I.D., so that the significance of any particular split can be judged. The problem of the overall significance of a complete A.I.D. analysis, combining many stages, still remains to be solved.

In Chapter 1 the A.I.D. method is described in more detail and an example is presented to illustrate its use. A null hypothesis that the dependent variable observations have independent and identical normal distributions is proposed as a model for no interaction. In Chapters 2 and 3 the null distributions of the optimum BSS for a single predictor are derived and tables of percentage points are given. In Chapter 4 the normal assumption is dropped and non-parametric A.I.D. criteria, based on ranks, are proposed. Tables of percentage points, found by direct enumeration and by Monte Carlo methods, are given. In Chapter 5 the example presented in Chapter 1 is used to illustrate the application of the theory and tables in Chapters 2, 3 and 4 and some final conclusions are drawn.

CONTENTS