



<http://researchspace.auckland.ac.nz>

### *ResearchSpace@Auckland*

#### **Copyright Statement**

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage.

<http://researchspace.auckland.ac.nz/feedback>

#### **General copyright and disclaimer**

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form.

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

78-171  
cop. 1

## 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.

THESIS

78 6015873X01

78-171

cop. 1

## 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

### CHAPTER 1 THE A.I.D. ALGORITHM

§1.1	The aims of A.I.D.	1
§1.2	The A.I.D. criterion	3
§1.3	Example	7
§1.4	Significance testing: The null hypothesis	10
	Table and Figure	12

### CHAPTER 2 THE MONOTONIC PREDICTOR

§2.1	Likelihood ratio statistics	16
§2.2	One observation per category: Location shift detection	20
§2.3	The null distribution of $K_M$	22
§2.4	Bonferroni approximations to percentage points of $K_M$ and $F_M$	28
§2.5	The null distribution of $F_M$	32
§2.6	The calculation of percentage points of $F_M$	41
§2.7	Unequal numbers of observations per category	53
§2.8	Concluding remarks	56
	Tables and Figure	58

### CHAPTER 3 THE FREE PREDICTOR

§3.1	Likelihood ratio statistics	66
§3.2	The null distribution of $K_F$	70
§3.3	Bonferroni approximations to percentage points of $K_F$ and $F_F$	73
§3.4	The null distribution of $F_F$	75

§3.5	The calculation of percentage points of $F_F$	75
§3.6	Asymptotic null distributions	84
§3.7	Concluding remarks: Optimum stratification and cluster analysis	89
	Tables and Figure	97
CHAPTER 4 DISTRIBUTION-FREE A.I.D.		
§4.1	Introduction	105
§4.2	Distribution-free statistics for the monotonic case	106
§4.3	Distribution-free statistics for the free case	108
§4.4	Asymptotic null distributions	110
§4.5	A note on Bonferroni approximates	113
§4.6	Concluding remarks: robustness	114
	Tables	117
CHAPTER 5 SIGNIFICANCE TESTING AT A SINGLE STAGE		
§5.1	Example	121
§5.2	One-sided monotonic A.I.D.	126
§5.3	Concluding remarks: Some criticisms of A.I.D.	129
	Figure	131
BIBLIOGRAPHY		132