# ANALYSIS OF ZERO-HEAVY DATA USING A MIXTURE MODEL APPROACH

By

Shin Cheng Wang

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

**Doctor of Philosophy**

**in**

**Statistics**

APPROVED

Eric P. Smith, Chairman
Jesse C. Arnold
Clint W. Coakley
Klaus H. Hinkelmann
Keying Ye

March 18, 1998
Blacksburg, Virginia

Key Words: *Ceriodaphnia Dubia*, Chronic toxicity testing, Generalized Estimating Equations, Inhibition Concentration, Longitudinal Data, Principal Component Analysis, Zero-inflated Poisson

# Analysis of Zero-Heavy Data Using A Mixture Model Approach

by

Shin Cheng Wang

Eric P. Smith, Chair

Department of Statistics
Virginia Polytechnic Institute and State University

(ABSTRACT)

The problem of high proportion of zeroes has long been an interest in data analysis and modeling, however, there are no unique solutions to this problem. The solution to the individual problem really depends on its particular situation and the design of the experiment. For example, different biological, chemical, or physical processes may follow different distributions and behave differently. Different mechanisms may generate the zeroes and require different modeling approaches. So it would be quite impossible and inflexible to come up with a unique or a general solution.

In this dissertation, I focus on cases where zeroes are produced by mechanisms that create distinct sub-populations of zeroes. The dissertation is motivated from problems of chronic toxicity testing which has a data set that contains a high proportion of zeroes. The analysis of chronic test data is complicated because there are two different sources of zeroes: mortality and non-reproduction in the data. So researchers have to separate zeroes from mortality and fecundity. The use of mixture model approach which combines the two mechanisms to model the data here is appropriate because it can incorporate the mortality kind of extra zeroes.

A zero inflated Poisson (ZIP) model is used for modeling the fecundity in *Ceriodaphnia dubia* toxicity test. A generalized estimating equation (GEE) based ZIP model is developed to handle longitudinal data with zeroes due to mortality. A joint estimate of inhibition concentration ($IC_x$) is also developed as potency estimation based on the mixture model approach. It is found that the ZIP model would perform better than the regular Poisson model if the mortality is high. This kind of toxicity testing also involves longitudinal data where the same subject is measured for a period of seven days. The GEE model allows the flexiblity to incorporate the extra zeroes and a correlation structure among the repeated measures.

The problem of zero-heavy data also exists in environmental studies in which the growth or reproduction rates of multi-species are measured. This gives rise to multivariate data. Since the inter-relationships between different species are imbedded in the correlation structure, the study of the information in the correlation of the variables, which is often accessed through principal component analysis, is one of the major interests in multi-variate data. In the case where mortality influences the variables of interests, but mortality is not the subject of interests, the use of the mixture approach can be applied to recover the information of the correlation structure. In order to investigate the effect of zeroes on multi-variate data, simulation studies on principal component analysis are performed. A method that recovers the information of the correlation structure is also presented.

*Dedicated to my wife, Julia*

# ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Eric Smith for his patience and guidance. Despite his busy schedule and many responsibilities, he always made time for my research and to offer encouragement. His invaluable knowledge and support were indispensable for this work. I would also like to express my gratitude to Dr. Jesse Arnold, Dr. Clint Coakley, Dr. Klaus Hinkelmann, and Dr. Keying Ye for serving on my committee.

I am also indebted to my parents, Lai Foo Wang and Sue Chin Chung, my brother, Shinho, and my sister, Wen Hsien for their love, encouragement, and support throughout the years. I would also like to thank my best friend, Dr. Ian Yip for his support and faith in me.

Finally, to my lovely wife, Julia, I would like to thank her for her patience, encouragement and sacrifice. Without her support, this dissertation would not have come to existence.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES