CHAPTER ONE
INTRODUCTION

1.1 Zero-heavy data

Data in many controlled experiments and observational studies often exhibit a high proportion of zeroes. This type of data, which can be called zero-heavy data, creates many difficulties in analyses. One of the major concerns of zero-heavy data is about the assumption of the distribution. Since most parametric and nonparametric tests assume the data have a continuous underlying distribution, a large number of zeroes clearly violates the assumption. As a result, the traditional method of data modeling may produce biased estimates and inappropriate results.

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. One should also understand the underlying mechanism that creates the data before starting to model the data. For example, different biological, chemical, or physical processes may follow different distributions and behave differently. It would seem impossible and inflexible to come up with a unique or general solution because different mechanisms may generate the zeroes and require different modeling approaches.

1.2 True Zeroes and Sampling Zeroes

In the case of biological and ecological studies, zeroes typically occur when a species is totally absent in a sampling area or when the species is present but absent in the sample (Smith and Palettas, 1992). The first case corresponds to a true zero or a structural zero (Chester, 1982), while the latter case corresponds to a sampling zero. Lindsey (1995) discusses the structural zeroes in the context of categorical data analysis. Chelius and Hoffmann (1996) develop two-way ANOVA method in the presence of structural zeroes. Sampling zero problems often occur in wildlife and fishery studies, where a population estimate is based on survey data. In these studies, zero observations may often be due to errors or biases in the sampling method. For example, the counting of birds or other animals in a particular area often involves human errors or visual biases. A zero count may be a human mistake, a sampling zero, or a true zero. Unfortunately, in most of the practical situations, it is impossible to separate the sampling zeroes from the true zeroes.

1.3 Zero Creating Mechanisms

In order to solve the problems with heavy zero data, we have to understand how the zeroes arise. As mentioned in the previous section, zeroes may result from both sampling errors and direct measurement. A zero observation can either be a true zero or a sampling zero. From the statistical modeling viewpoint, these zero creating methods can be grouped into two basic mechanisms.

For the first zero creating mechanisms, extra zeroes may be produced by ordinary overdispersion (Heilbron, 1989). This is when the variance of the data exceeds the expected variation. For example, in ecology, the number of organisms or plants in a quadrant is used to model the spatial pattern. These patterns sometimes follow a Poisson distribution. However, very often, the offsprings of plants or organisms tend to

concentrate more at the neighborhood of the parent than at other places. As a result, sampling at the sparsely populated area can yield a large number of zero observations. This kind of clustering behavior creates a higher variability in the spatial pattern of plants than one would expect under the Poisson assumption. This phenomenon is called overdispersion. For a Poisson distribution, overdispersion implies the variance of the Poisson distribution is larger than the mean of the distribution. In this case, zeroes come from only one source, and many techniques are available to model the data based on a single distribution assumption.

For the second zero creating mechanisms, extra zeroes can be produced by a distinct subpopulation of subjects (Heilbron, 1989). Empirically, this may be related to some natural intervention or censoring and truncation of the data. An example can be found in the manufacturing industry where counts of defective items are often recorded as quality control data. When the manufacturing process is reliable, there are very few defects, and so a lot of zeroes in the data. However, the number of non-defect items is sometimes much more than expected. One can think that small, unobservable changes in the environment cause the process to switch randomly back and forth between a perfect state in which defects are extremely rare and an imperfect state in which defects are possible but not inevitable (Lambert, 1992). In this case, the perfect state can be thought of as a subpopulation that creates only zero counts.

Another example is from an economic study concerning household expenditures. Since some households do not spend anything on certain commodities, the data would consist of a mass of zero observations. In this case, the sample population can then be separated into two groups: spenders and non-spenders (Aitchison and Brown, 1957). The non-spenders group would be the subpopulation who creates the zero observations. Here the zeroes come from two different sources, and so it is more appropriate to assume a mixture distribution when modeling the underlying mechanism.

This concept of having two sources of zeroes leads to a mixture model for the data in which zeroes are “mixed” with other data.

This dissertation is motivated by problems occurring in chronic toxicity testing. Since the objective of the chronic toxicity tests is to investigate the effect of toxicant on either growth or reproduction of animals/microorganisms, the analyses of data from these tests are complicated by the presence of many zeroes resulting from mortality. The two sources of zeroes in these chronic tests are: mortality and non-reproduction.

The organization of this dissertation is as follows: Chapter Two presents the background on heavy zero data and a literature review of the mixture models. Chapter Three discusses the use of the mixture models in chronic toxicity testing. Chapter Four provides a discussion of the use of mixture models in repeated measure studies of toxicity testing. Chapter Five presents principal component analysis for zero-heavy data. Chapter Six summarizes results and suggests further research possibilities.