Chapter 1

Introduction to Zero Inflated Models

1.1 Introduction

Statistical modeling for discrete data is done through well known discrete distributions which include Binomial distribution, Poisson distribution, Geometric distribution etc. In practice experimenters can come across a situation wherein the existing well known models fail to model observed data. Statisticians have to provide suitable model for the data sets where existing standard models are not found suitable. In recent years new models are proposed by modifying existing well known models. Zero inflated models are the models obtained from existing models by appropriately mixing existing model with a singular distribution at zero.

Zero inflation indicates that a data set contains an excessive number of zeros. The word inflation is used to emphasize that the probability mass at the point zero exceeds which is allowed under a standard parametric family of discrete distributions.

In certain applications involving discrete data, we come across data having frequency of an observation 'zero' significantly higher than that predicted by the assumed models. The problem of high proportion of zeros has been of an interest in data analysis and modeling. Examples of such applications are cited in the literature from engineering, manufacturing, economics, public health, epidemiology, psychology, sociology, political science, agriculture, road safety, species abundance, use of recreational facilities, horticulture and criminology. In a highly automated stable production process occurrences of defects is assumed to be Poisson, but we get no defectives in many samples. Count data with high proportion of zeros are abundant in many disciplines. Examples of such data are

1. Read write errors discovered in computer hard disks from a manufacturing process (Xie and Goh, 1993) :

Count	0	1	2	3	4	5	6	≥7
Frequency	180	11	5	2	1	1	2	6

2. The data from the department of motor vehicles master driver license file. (Traffic accidents research – Kuan et al.1991) :

Traffic accidents	0	1	2	3
Number of drivers	4499	766	136	21

3. The data from criminal behaviour (Dickmann- 1981) :

Criminal acts	0	1	2	3	4	5
Number of persons	4037	219	29	9	5	2

4. Spider count data (Kulasekeva K.B. and Tonkyn D. W. 1992) :

Count	0	1	2	3	≥ 4
Frequency	159	64	13	4	0

In the above example there are excess number of zero counts. Poisson distribution has often been used for count related data. However this model does not prove a good fit to actual data when there are a frequent or excessive number of zero counts. In such situations, the Zero Inflated Poisson distribution, obtained by mixing Poisson distribution with a singular distribution at zero could be more appropriate model. Similarly a well known Power Series Distribution can be used instead of Poisson distribution so as to get Zero Inflated Power Series Distribution (ZIPSD). The present study is related to ZIPSD and the following members of ZIPSD.

1. Zero Inflated Poisson Distribution (ZIPD).
2. Zero Inflated Negative Binomial Distribution (ZINBD).
3. Zero Inflated Binomial Distribution (ZIBD).

1.2 Literature Survey

In literature several researchers have worked on Zero Inflated Models. Consul and Jain (1973) have provided a new generalization of Poisson distribution in order to take an account of excess zeros in the usual Poisson

distribution. The generalized Poisson distribution has two parameters and mean and variance of this distribution are different.

Yip (1988) has described situations with an inflated Poisson distribution, dealing with the number of insects per leaf. He discussed the problem of estimating mean of Poisson distribution in the presence of a nuisance parameter. Heilbron (1989) proposed similar Zero altered Poisson and Negative Binomial regression models and applied them to data on high risk behaviour in gay men. He also considered models with an arbitrary probability of zero. Arbitrary zeros are introduced by mixing point mass at zero with a positive Poisson that assigns no mass to zero rather than a standard Poisson.

Recently, Lambert (1992) considered Zero inflated Poisson regression models. Gupta et al. (1995) have studied the structural properties and point estimation of parameters of Zero Inflated Modified Power Series distributions with applications of these models to a simulated data sets. They consider a zero inflated model for more general class of discrete distributions known as Modified Power Series Distributions (MPSD). This class includes among others the generalized Poisson, generalized negative binomial and generalized log series distributions and hence the ordinary Poisson, Binomial and Negative Binomial distributions. However, they have not considered testing and interval estimation of the parameters.

Kale (1998), Kale and Muralidharan (2000) have reported results on optimal estimating equations for discrete data with higher frequencies at a point

and for mixture distributions accommodating instantaneous or early failures respectively.

M. Xie et al. (2001) have studied Zero Inflated Poisson distribution with its application in statistical process control. In particular, various tests of Poisson distribution and zero inflated Poisson alternatives are compared. For example the score test, the likelihood ratio test, the chi-square test, the test based on confidence interval, the Cochran test, and the Rao Chakravarti test. However, they have not considered test based on asymptotic distribution of maximum likelihood estimators.

Gupta et al. (2004) studied a zero adjusted generalized Poisson distribution and developed a score test with and without covariates, to determine whether such an adjustment is necessary. Dutta (2004) studied a wide variety of discrete distributions as the possible models for count data with high proportion of zeros. For maximum likelihood estimation of the model parameters he used a stochastic optimization algorithm named 'simulated annealing'. Further for model selection he used Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC).

Suppose the data is not inflated at 'zero' but it is inflated at any of the support point say 's'. We say such models as Non - Zero inflated models. Pandey (1964-65) has described a situation with an inflated Poisson distribution dealing with the number of flowers of plant of *Primula Veris.* He has shown that the excessive number of plants with eight flowers implies application of Poisson

distribution inflated at the point 8. Murat and Szynal (1996) studied moments of certain inflated probability distribution. Murat and Szynal (1998) considered non - zero inflated modified Power Series distributions and extended the results of Gupta, Gupta and Tripathi (1995) to the discrete distributions inflated at any point 's'.

Literature survey reveals that there is scope to review and to study Inflated Power series models with regards to estimation, testing of hypotheses and interval estimation of the parameters of the models.

1.3. Chapter wise Summary

In Chapter 2, we introduce Power Series Distribution and ZIPSD. We study the estimation of the parameters of ZIPSD in general and the estimation of the parameters of the ZIPD, ZINBD, and ZIBD in particular.

Chapter 3 deals with testing of Inflation parameter π of ZIPSD. There are variety of tests reported in literature (please refer to Xie et al. (2001) for ZIPD). They are namely Score test, the likelihood ratio test, the Chi-square test, test based on confidence interval of π, Cochran test and Rao – Chakravarti test. A new test procedure for testing the parameter π of ZIPSD is given. Some examples are also provided. The performance of the test is studied through simulation study for ZIPD, ZINBD and ZIBD. The power curves are drawn for simulated power based on asymptotic distribution of mle's for the different sample size. Performance of the test is studied through simulation for both the

distributions. Asymptotic Confidence Interval for the parameter π based on the proposed test is obtained and illustrated through examples.

Chapter 4 deals with testing the parameter θ of Power Series Distribution in the Zero Inflated Power Series Model. Though the parameter of PSD is of equal importance as the inflation parameter, literature study reveals that testing and interval estimation about this parameter has not been reported to the best of our knowledge. This Chapter is devoted to these two aspects and entire material in this chapter is our contribution. Section 4.2 deals with the theoretical development of the two tests for testing θ, the first one is based on mle's and the other is based on sample mean. Section 4.3 deals with the theoretical development of the two tests for testing the parameter of Poisson distribution in the Zero Inflated Poisson Model. Performance of these tests is studied through simulation. In section 4.5 Asymptotic Confidence Interval is obtained for the parameter θ and illustrated through examples.

Chapter 5 deals with non – zero inflated models. Inflation may occur at any of the point in the support. Section 5.1 is deals with Introduction to non – zero inflated models. Section 5.2 and 5.3 covers estimation of the parameters of Power Series Distribution Inflated at the point 's'. Section 5.4 deals with the application to Poisson distribution Inflated at the point 's' and application to the Geometric distribution inflated at the point '1' respectively.

In the present study our contribution can be summarized as follows.

(a). Estimation of the parameters, Fisher's information for the Zero Inflated Negative Binomial Distribution and Zero Inflated Binomial Distribution.

(b). A new test based on the mle to test the inflation parameter $\pi = 1$ vs. $\pi \neq 1$ for ZIPSD in general and ZIPD, ZINBD, and ZIBD in particular.

(c). Study of the performance of the test through power function proposed in (b) above using 'C' programs .

(d). Asymptotic confidence interval for inflation parameter π based on mle for ZIPSD in general and for ZIPD in particular.

(e). Two asymptotic tests to test the parameter θ of PSD in ZIPSD. The first one based on the mle and the other based on the sample mean.

(f). Asymptotic confidence interval for θ based on mle and based on sample mean.

(g). Development of the required 'C' programs to simulate and to estimate the parameters π and θ, and also to estimate the power of the tests introduced in this study.

Finally, we discuss areas of future research related to inflated models.