CHAPTER 1

INTRODUCTION

1.0 Discrete distributions in statistical analysis

Discrete models play an extremely important role in probability theory and statistics for modeling count data. The use of discrete models does not only enable the investigators to visualize and discuss the observed data but also to summarize and make predictions or estimate the properties of the samples easily. In addition, the behaviour and patterns observed in many types of phenomena can be described through discrete models. Therefore, it is not surprising to find that the study and application of discrete models has received much attention and became a major research area for decades. The discovery of new models with applications is ongoing. The Poisson, binomial and hypergeometric distributions are a few of the most often used discrete models.

In short, probability models or distributions are very important in statistical data analysis as empirical models or as mathematical models for random variations. Families of univariate and multivariate distributions have been examined by many researchers. In particular for discrete distributions, the difference-equation families, power series distributions and Kemp families (see Johnson *et al*., 2005) are well-known while the bivariate and multivariate distributions have been covered by a number of books (Mardia, 1970; Kocherlakota & Kocherlakota, 1992; Johnson *et al*., 1997; Kotz *et al*., 2000).

The principle of complete randomness is not practical in all situations. The observed patterns tend to have variations and evidences were present in the fields of

biology, ecology, and medicine to indicate contagion in the distributions of the observed data. This has led to a rich class of contagious distributions; for example, the Neyman's Type A, B, C (1939), negative binomial (Fisher, 1941), discrete lognormal (Preston, 1948), Thomas (1949) distributions and so on. Archibald (1948) had shown that the contagious distributions of Neyman often give a good fit to plant distributions. Bliss and Fisher (1953) studied the efficiency of parameter estimation methods for the negative binomial distribution to biological data, and Pipes *et al.* (1977) indicated its suitability for fitting coliform counts (bacterial density in water). Clearly, enumerative studies on contagious distributions in various fields have been well-developed.

1.1 Literature review

This section gives a brief review relevant to the development of this thesis and to provide motivation for research problems considered.

To describe the behavior of different types of phenomena, the Poisson distribution is derived by assuming the events occur under the principle of complete randomness. The model depends upon a single parameter which is the mean as well as the variance. Based on the equality of mean and variance, the Poisson distribution is said to satisfy the equi dispersion property. However, in practice, most of the real data encountered are either over or under dispersed and therefore do not follow a Poisson model. Numerous researchers have tried to explain the variations in the observed data by various modifications and generalization of the Poisson distribution; see Lundberg, (1940), Haight (1967), Patil and Joshi (1968) and Johnson and Kotz (1969). Most of these distributions were developed to explain the unequal mean and variance in the observed data.

Consul and Jain (1970, 1973a, b) introduced a generalized Poisson distribution (GPD) with parameters θ and λ. The variance of this model can be less than, equal to or even greater than the mean depending on the parameter λ, meaning that the GPD can cater for under, equi and over dispersed data. The increase or decrease of parameter θ tends to increase or decrease both the mean and variance. But for under dispersion, the parameters have to be restricted in order for the GPD to be a proper distribution (Nelson, 1975); see also Scollnik (1998) for further discussion.

Efron (1986) proposed the double Poisson distribution to deal with the issue of under, equi and over dispersion in the data. A disadvantage of the double Poisson distribution is that it is not a proper distribution as the probabilities do not sum to one. However, Efron showed that the multiplication constant, which makes the distribution proper, is very close to one virtually over the whole parameter space. Efron also suggested an approximation for this constant which was obtained by Edgeworth expansion.

Another distribution that can model under, equi and over dispersion is the Conway-Maxwell-Poisson distribution (COM-Poisson) (Conway and Maxwell, 1962; Shmueli *et al.*, 2005) which enjoyed a recent revival. The COM-Poisson distribution is a generalization of the Poisson distribution which contains the Bernoulli and geometric distributions as special cases. In addition, it is also a member of the exponential family. The flexibility of this distribution has prompted a fast growth of research in various fields as it possesses only two parameters and allows a wide range of under and over dispersion. The earliest use of the COM-Poisson distribution is found in the field of linguistics. Major development of the COM-Poisson distribution followed after it was re-introduced by Shmueli *et al.* (2005) where the statistical properties of the COM-Poisson were studied in detail. However, the COM-Poisson distribution does not have

closed form expression for its moments and asymptotic approximations are required even for its mean and variance (Shmueli *et al.*, 2005, page 130).

Shimizu and Yanagimoto, (1991) proposed the inverse trinomial (IT) distribution which arises as a random walk on the real line with three transition probabilities. The IT distribution has pmf as

$$\Pr(X = x) = \frac{\lambda p^{\lambda} q^{x}}{x+\lambda} \sum_{t=0}^{\lfloor x/2 \rfloor} \binom{x+\lambda}{t, t+\lambda, x-2t} \left(\frac{pr}{q^2}\right)^t$$

for $x = 0, 1, 2, \ldots$ where $\lambda > 0, p \geq r, p + q + r = 1$ and

$$\binom{x+\lambda}{t, t+\lambda, x-2t} = \frac{(x+\lambda)!}{t!(t+\lambda)!(x-2t)!}$$

Although, this model has a pmf in terms of the Gauss hypergeometric function, the existence of the recurrence formula simplifies its computation. Khang and Ong (2007) have shown the IT distribution to be a Poisson-stopped (generalized Poisson) distribution. Aoyama *et al.* (2008) proposed a generalization of the shifted inverse trinomial (GIT) distribution.

The GIT distribution is formulated as a first-passage time distribution of a modified random walk on the half plane with five transition probabilities. It contains up to twenty-two possible distributions, including the binomial, negative binomial, shifted negative binomial, shifted inverse binomial and shifted inverse trinomial distributions. According to Aoyama *et al.* (2008), one of the subclass of the GIT distribution, denoted by $\mathrm{GIT}_{3,1}$ distribution, also has the flexibility to model under, equi and over dispersion data. This interesting feature has motivated us to further study the $\mathrm{GIT}_{3,1}$ distribution.

The study of discrete distributions has been widened to multivariate distributions. During the 1950's the negative binomial distribution has received considerable attention, mainly in problems concerning accident proneness where it has

been extended to the bivariate case (see Arbous and Kerrich , 1951) and the multivariate case (see Bates and Neyman, 1952). Edward and Gurland (1961) and Subrahmaniam (1966) worked independently on a version of the bivariate negative binomial distribution (BNBD). The joint probability mass function formulated by Subrahmaniam (1966) is given as

$$f(x_1, x_2) = q^{\nu} \sum_{i=0}^{\min(x_1, x_2)} \frac{\Gamma(\nu + x_1, x_2 - i)}{\Gamma(\nu) i!(r-i)!(s-i)!} p_1^{r-i} p_2^{s-i} p_3^{i}$$

where $q = 1 - (p_1 + p_2 + p_3)$ and $x_1,\ x_2 = 0,\ 1,\ 2,\ \ldots$. The disadvantage of this model is the correlation between the two random variables $X_1,\ X_2$ must be positive. A less restrictive model which allows negative correlation is introduced by Mitchell and Paulson (1981) but the shape parameter is limited to integer values.

Based on the trivariate reduction method, Famoye and Consul (1995) developed a bivariate generalized Poisson distribution with $X_1 = Y_1 + Y_3$ and $X_2 = Y_2 + Y_3$ where $Y_1,\ Y_2,\ Y_3$ are independent univariate generalized Poisson random variables. Unfortunately, the model only permits positive correlation. Lee (1999) defined a bivariate negative binomial distribution (BNBD) based on copula function which allows positive or negative correlation. This model suffers from two main drawbacks where the joint probability function is very complicated and it can only be used for over dispersed data.

Lakshaminaraya, Pandit and Rao (1999) formulated a bivariate Poisson distribution as a product of Poisson marginals with a multiplicative factor. The joint probability function is defined as

$$f(x_1, x_2) = \theta_1^{x_1} \theta_2^{x_2} e^{-x_1 - x_2} \frac{1 + \lambda\left(e^{-x_1} - e^{-d\theta_1}\right)\left(e^{-x_2} - e^{-d\theta_2}\right)}{x_1!\, x_2!}$$

where $d = 1 - e^{-1}$, $e^{-d\theta_i}$ is the expectation $E(e^{-X_i})$ (i=1, 2) and x_1, $x_2 = 0, 1, 2, \ldots$. The correlation coefficient is given by $\rho(x_1, x_2) = \lambda\sqrt{\theta_1\theta_2}d^2e^{-d(\theta_1+\theta_2)}$. Clearly, the correlation is flexible as it depends on the values of the multiplicative factor λ which can be negative, zero or positive.

The work of Lakshaminaraya, Pandit and Rao (1999) is extended by Famoye (2010) who introduced a new bivariate generalized Poisson distribution (BGPD) with joint probability function

$$f(x_1, x_2) = \prod_{i=1}^{2}\left\{\frac{\theta_i^{x_i}(1+\alpha_i x_i)^{x_i-1}}{x_i!}\exp\left[-\theta_i(1+\alpha_i x_i)\right]\right\}\left[1+\lambda(e^{-x_1}-c_1)(e^{-x_2}-c_2)\right]$$

where $c_i = E(e^{-X_i}) = \exp\left[\theta_i(s_i - 1)\right]$, $\ln s_i - \alpha_i\theta_i(s_i - 1) + 1 = 0$ (i=1, 2) and x_1, $x_2 = 0, 1$ 2, ... The correlation is also dependent on the value of the multiplicative factor λ. Hence, it can be negative, zero or positive.

Assumed models are almost never exactly true in reality; the goals of a practitioner are to estimate θ, the vector of parameters of a distribution, efficiently when the model is correct and robustly in the case where the true distribution is in the neighbourhood of the model. Various estimation methods such as the method of moments, maximum likelihood (ML) method and so on have been the source of considerable interest as seen in recent statistical literature.

Traditionally, the maximum likelihood (ML) approach is a popular technique in parameter estimation due to its generality and asymptotic efficiency. Various aspects and applications of MLE have been discussed in many papers. However, difficulties arise when the probability function is complicated, not tractable or not bounded over the parameter space. Besides that, MLE does not work well in the presence of outliers. Therefore, minimum Hellinger (MH) estimation which was introduced by Beran (1977)

attracted great attention because of its ability to reconcile the properties of robustness and asymptotic efficiency. Tamura and Boos (1986) have also considered the minimum Hellinger distance estimator for continuous models. Simpson (1989) discussed the minimum Hellinger distance estimation for discrete models and studied the robust hypothesis testing problem for general models using the Hellinger distance. Unfortunately, computations in the MH estimation also depend on the probability function, similar to MLE, where a model with complicated probability function will severely slow down the parameter estimation process.

This problem led researchers to investigate simpler inference procedures one of which is based on the probability generating function (pgf). The pgf based approach has been proposed for testing goodness-of-fit and parameter estimation. Kemp and Kemp (1988) suggested estimators based on the empirical probability generating function (epgf); the methods involve solving estimating equations obtained by equating functions of the epgf and probability generating function (pgf) on a fixed, finite set of values. Dowling and Nakamura (1997) considered the epgf for parameter estimation for families of discrete distributions with support on the nonnegative integers or a subset thereof. The pgf viewpoint leads to appealing graphical procedures and computational advantages.

For the test of goodness-of-fit, much work has been focused on discrete distributions. In general, the Pearson's chi-square is most widely used where it can be applied to count data distribution with a finite number of classes. Kocherlakota and Kocherlakota (1986) proposed goodness-of-fit tests for discrete random variables based on the pgf $G(t;\theta) = E_\theta\left[t^X\right]$ and derived the necessary asymptotic results. In the proposed tests, the epgf has to be evaluated at a number of values of t , and therefore, the performance of the tests depends on the selected t. Marques and Pérez-Abreu (1989)

showed that the process $\xi_N(t,\theta) = \sqrt{N}\{G_N(t) - G(t;\theta)\}$ converges in the space $C[0,1]$ to a continuous Gaussian process centred at 0, where $G_N(t)$ is the empirical pgf (epgf) given by

$$G_N(t) = \frac{1}{N}\sum_{i=1}^{N} t^{x_i}, \ |t| \le 1.$$

By using this result, Rueda *et al*. (1991) proposed a quadratic statistic for testing goodness-of-fit. Rueda *et al*. (1999) derived further results and corrected an error in the expression for the covariance formula.

1.2 Contributions of the Thesis

A new distribution denoted by $GIT_{3,1}$ is proposed to model under, equi and over dispersion. Inference, study of statistical power by Monte Carlo simulation and goodness-of-fit of this $GIT_{3,1}$ model have been considered. A paper based upon this work has been submitted for publication.

Two new bivariate $GIT_{3,1}$ distributions have been derived. The properties and the characteristics of the distributions are presented. The performance of the proposed pgf-based MH type distance estimation is extended and applied to the parameter estimation for the bivariate case. A paper on this part of work is in progress.

The COM-Poisson distribution is also studied. The limitation of computer accuracy for the computation of the infinite sum and the accuracy of the proposed asymptotic approximation are investigated. The study of statistical power by Monte Carlo simulation has been conducted. A paper on the COM-Poisson distribution has been prepared for publication.

A new estimation procedure based upon minimum disparity and probability generating function has been investigated and compared to well-known estimation methods like MLE and minimum Hellinger distance estimation. It is of interest to identify an efficient and simple method of parameter estimation for discrete distributions with complicated probability mass functions. The basic asymptotic theory for the proposed pgf-based MH type distance estimation is provided. An intensive simulation study has been done to investigate the robustness and consistency of the proposed statistic in estimation. This work has been published in Sim and Ong (2010).

1.3 Organization of the Thesis

A general introduction and brief literature survey regarding the thesis are given in Chapter 1. The contributions of the thesis are also stated.

Chapter 2 gives the preliminaries of the thesis by providing the terms, theorems and concepts needed for the following chapters. Our main findings are concentrated in Chapters 3 to 6. A paper is to be written based upon each chapter.

In Chapter 3, we explored and reviewed the properties of the $GIT_{3,1}$ distribution. A proposed pgf-based MH type distance estimation and goodness-of-fit test are considered for the $GIT_{3,1}$ distribution. A test for equi- dispersion by Rao's score test and LR test is provided**.** Then results of a simulation study of the statistical power are given. The fitting of this distribution to real life data sets will be examined where the $GIT_{3,1}$ distribution is compared to the GPD and COM-Poisson distributions.

In Chapter 4 two new bivariate $GIT_{3,1}$ distributions have been derived. The properties and characteristics of the distributions are presented together with parameter estimation and application of the new bivariate distributions to real life data sets.

Further results for the COM-Poisson distribution are given in Chapter 5. Some computational issues, parameter estimation, test of equi- dispersion and statistical power of the score and log-likelihood tests are examined.

In Chapter 6, we proposed a pgf-based MH type distance estimation. A study of various techniques in parameter estimation and goodness-of-fit tests will be carried out. Comparative analysis of the performance of the proposed parameter estimation methods has been done. Simulated data will be used to validate the theoretical results concerning an efficient method for the proposed parameter estimation. Analyses of real life data sets are then provided.

Finally, Chapter 7 concludes the finding of our research and we discussed works for the future.