Chapter 1

1 Introduction

The theory of discrete probability distributions is originated from the works of James Bernoulli and Poisson. Now, it has become a very useful and important branch of modern statistics having various important applications in variety of disciplines. In recent years, the mixture of the basic discrete distributions such as Poisson, Negative binomial, geometric etc. have received continued attention for some reasons such as, sometimes the basic distributions which can be created on the basis of simple models, have been found not good enough to describe the situations which occurs in number of phenomenon and practical problems involving mixtures arise in the fields of study. For which now a day, univariate mixture distributions obtained by combining two or more of the basic distributions have become an important branch of statistics.

1.1 Background of the Study

In distribution theory, the mixture of discrete distributions is an important class of distributions which gives more flexibility than the simple basic distributions. The notion of mixing often has a simple and direct interpretation in terms of the physical situation under investigation. Sometimes, the word “mixing” is just a mechanism for constructing new distributions from the given ones for which empirical justification must later be sought.

In some situations, it is found that a simple basic distribution fails to describe a set of data which leads to the belief that the model underlying the distribution has some of the particular characteristics of the mixture model for which further research was made to see if any simpler mixture distribution will describe the data to any degree of satisfaction. In this way a large number of discrete mixture distributions were derived.

A mixture distribution is a superimposition of distributions with different functional forms or different parameters, in specified proportions. Sometimes, however, a mixture of distributions is a mechanism which helps to construct new distributions from the given ones for which empirical justification is sought later on. If $F_j(x_1,\ x_2, \dots, x_n)$, $(j = 0, 1, \dots, m)$ represents different cumulative distribution functions (cdf) and $\omega_j \geq 0$ and $\sum_{j=0}^{m} \omega_j = 1$ then

$$F(x_1, x_2, \dots, x_n) = \sum_{j=0}^{m} \omega_j F_j(x_1, x_2, \dots, x_n)$$

also is a proper cumulative distribution function. This is called a mixture of the distribution $\{F_j\}$. The distribution is finite or infinite according to m is finite or infinite. However, this is only one aspects of mixture distribution, and there are two important categories of mixture distributions, namely

(a) Finite mixture of discrete distributions

(b) Countable or continuous mixture of discrete distributions

(a) Finite mixtures of discrete distributions

Finite mixtures distributions arise in many probabilistic situations. The concept of finite mixture of discrete distributions was introduced by Pearson (1915). Titterington (1990) has commented on their use in speech recognition and in image analysis.

In a k - component finite mixture distribution, there are k different component distributions with cumulative distribution functions (cdf's) $F_1(x), F_2(x), \dots, F_k(x)$ with the weights $\omega_1, \omega_2, \dots, \omega_k$, where $0 < \omega_j < 1$, $\sum_{j=1}^{k} \omega_j = 1$. The cumulative distribution function for the new (mixture) distribution is

$$F(x) = \omega_1 F_1(x) + \omega_2 F_2(x) + \cdots + \omega_k F_k(x)$$

If the component distributions are defined on the nonnegative integers with

$$P_j(x) = F_j(x) - F_j(x-1)$$

Then the mixture distribution is a discrete distribution with probability mass function (pmf)

$$P[X = x] = \sum_{j=1}^{k} \omega_j P_j(x)$$

The support of the outcome for this type of mixture is the union of the supports for the individual components of the mixture.

In case of finite mixture of distributions the problem arises when data are unavailable for each individual component distributions separately but available just for the overall mixture distribution. In these situations interest often gives on estimating the mixing proportions and on estimating the parameters of the component distributions. One of the most important examples of finite mixture of discrete distribution is zero-inflated or zero-modified distribution.

(b) Countable or Continuous mixture of discrete distributions

A mixture distribution also arises when the cumulative distribution of a random variable depends on the parameters $\theta_1, \theta_2, \dots, \theta_m$ and some (or all) of those parameters may vary. Then the new distribution has the cumulative distribution function (cdf)

$$E[F(X|\theta_1, \theta_2, \dots, \theta_m)]$$

where the expectation is with respect to the joint distribution of the k parameters that may vary. This includes situations where the source of a random variable is unknowable. This type of a mixture distribution symbolically represented by

$$\mathcal{F}_A \bigwedge_{\Theta} \mathcal{F}_B$$

where $\mathcal{F}_A$ represents the original distribution whereas $\mathcal{F}_B$ the mixing distribution. When Θ has a discrete distribution with probabilities p_i ($i = 0, 1, \ldots$), then the outcome distribution has a countable mixture of discrete distribution with probability mass function (pmf)

$$P[X = x] = \sum_{i \geq 0} p_i P_i(x)$$

where $P_j(x) = F_j(x) - F_j(x-1)$.

When the points of increase of the mixing distribution are continuous, the outcome distribution is call as a continuous mixture. A continuous mixture of discrete distribution arises when a parameter corresponding to some important part of a model for a discrete distribution can be regarded as a random variable taking continuous values. Greenwood and Yule (1920), Lundberg (1940) first studied the theory of countable and continuous mixture of discrete distribution.

The class of mixtures considered in our study is the finite mixtures of discrete distributions. That is, observations are available from a population which is known to be a mixture of some sub-populations. Each sub-population will be assumed to have the same type of distribution (or different type of distribution) but with different parameter values.

In univariate case a wide class of mixtures of discrete distributions has been constructed by the process of compounding and generalization.

(i) **Compounding:** The term compounding has often used in place of "mixing". Suppose the random variable X_1 have the distribution function $F_1(x_1|\theta)$ for a given value of the parameter θ. Let θ be regarded as a random variable X_2 with the distribution function $F_2(x_2)$. The with the distribution function

$$G(x_1) = \int_T^\infty F_1(x_1|cx_2)dF_2(x_2)$$

(where c is a constant which is arbitrary in some prescribed sense) is said to be a compound with respect to the compounder X_2. This process is referred to as a compound of F_1 with F_2. Symbolically, written as

$$F_1 \bigwedge_\theta F_2$$

where F_1 represents the original distribution, θ be the varying parameter and F_2 be the compounding distribution. The adjective compound was evidently first used by Greenwood and Yule (1920) in describing a population which was thought to be a mixture of Poisson distribution. For discrete distributions "compounding" is commonly used in the place of "mixing" and the resultant distributions are called compound distributions.

(ii) **Generalization:** Generalization is also a process which also generates a verity of distributions. In this process, the new distributions result from the combination of two independent distributions in a particular way. This process is called "generalization" by Feller (1943). The use of the term generalization was reinforced by Gurland (1957), who introduced the symbolic notation that is customarily employed to represent the process.

Let the random variables y_1 and y_2 have the distribution functions $F_1 \; and \; F_2$ and probability generating functions (pgf) $g_1(t)$and $g_2(t)$ respectively. Then the random variable having the pgf

$$G(t) = g_1\{g_2(t)\}$$

is called the generalized y_1 variable with respect to the generalizer y_2. Also, it is called as F_1 distribution generalized by the generalizer F_2. It is written in the symbolic form as

$$F_1 \bigvee F_2$$

It is easy to see that $g_1\{g_2(t)\}$ is a polynomial function of t with non-negative co-efficient. So, there is a probability distribution corresponding to $g_1\{g_2(t)\}$ which is called a "generalized" distribution.

Mixtures of discrete distributions have now received an increasing amount of attention in recent statistical literature. This is due to an increased interest in the mathematics involved in dealing with mixtures and largely due to an increasing number of specific problems encountered in certain applications. In recent years, mixture of discrete distributions is applicable in terms of modeling data (from biology, ecology, geology, etc.) in the different domains of statistical literatures.

In most of the biological application various mixtures of discrete distributions are directly involved. In the biological sciences experimenters frequently encounter mixtures of distributions, particularly in the areas of investigation of natural populations.

Mixtures of discrete distributions are also applicable in the area of life-testing and in acceptance testing. In recent years, many authors have discussed various applications of different mixture of discrete distributions obtained by them in different areas including accident and error data, modeling and waiting survival times, ecology etc.

1.2 Review of Literatures

The origin of the theory of discrete probability distributions began with the work of James Bernoulli and Poisson. The Swiss mathematician, James Bernoulli derived the binomial distribution and published it in the year 1713. Poisson distribution was derived by a French mathematician Simeon D. Poisson as a limiting form of the binomial distribution in 1837. In 1920, Greenwood and Yule obtained negative binomial distribution as a consequence of certain assumptions in accident proneness models.

In the last few years, a large amount of efforts have been made in the area of discrete mixture distributions. Often, the simple basic distributions such as Poisson, binomial, negative binomial etc. have been found to fails to describe some sets of data, which leads to construct new mixture models of the basic distributions. For which further research was made to see if any mixture of basic distributions will describe the data to a better satisfaction. In this way, a large number of discrete mixture distributions, which are classified as generalized, modified and contagious distributions were derived by different authors. These distributions have various important applications in medical sciences, biological sciences, social sciences, engineering and so on.

Detailed information regarding the vast area of discrete mixture distributions, their applications and properties can be found in the books of Everitt and Hand (1981), Consul (1989) and Johnson et al. (2005).

Everitt and Hand (1981) studied finite mixture of distributions in their book and also discuss the problem of parameter estimation by the method of moments and maximum likelihood method. According to Smith (1985) finite mixtures of distributions can be used in economics, in medicine, Fisheries research etc. According to Titterington (1990) finite mixture distribution is used in speech recognition and in image analysis. Ben Nakhi and Kalla (2004) investigated some mixture distributions, which are obtained by mixing discrete distributions with continuous one. These distributions are further extended by them (2005).

Modified distribution is an example of finite mixture of discrete distribution. Other name of modified distribution is inflated distribution. The discrete inflated distribution was first introduced by Singh (1963). He studied inflated Poisson distribution to serve the probabilistic description of an experiment with a slight inflation at a point, say zero.

Pandey (1965) studied the generalized inflated Poisson distribution. Singh (1966) also investigated generalized inflated binomial distribution. A zero-modified geometric distribution was studied by Holgate (1964) as a model for the length of residence of animals in a specified habitat. Mechanisms producing zero-modified distribution have been discussed by Heilbron (1994) in the content of generalized linear model. Lambert (1992) proposed zero-inflated Poisson (ZIP) distribution. For the ZIP model, Bohning (1998) also reviewed the related literature and provided a verity of example from different disciplines. As a generalization of the ZIP model, the zero-inflated negative binomial has been discussed by many authors, such as Ridout et al. (2001). Zeileis et al. (2008) gave a nice overview and comparison of Poisson, negative-binomial and zero-inflated models in the software R. Recently, Younes mouatassim (2012) introduced the zero-modified discrete distributions in the calculation of operational value-at-risk.

The works on countable and continuous mixture of discrete distribution by the accident-proneness theory and actuarial risk theory of Greenwood and Yule (1920) and Lundberg (1940).This leads to the theory of mixtures of Poisson distributions.

Different mixtures of Poisson distributions where the mixing distributions are continuous or countable are discussed in details regarding the vast area and their properties in the book by Johnson et al. (2005).

A Poisson mixture of Poisson distribution known as Neyman Type A distribution has been used to describe plant distributions, especially when reproduction of the species produces clusters. Neyman Type A distribution gave good results for plant distribution which is found by Evans (1953). Cresswell and Froggatt (1963) derived Neyman Type A Distribution in the context of bus driver accidents.

Kemp and Kemp (1965) studied the Hermite distribution which is a Poisson mixture of Bernoulli distribution with applications to the fields of biological sciences, physical science and operation research. Plunkett and Jain (1975) derived the Gegenbauer distribution by mixing the Hermite distribution with the gamma distribution which has a long history in the theory of stochastic processes. Borah (1984) studied the probability and moment properties of the three parameter Gegenbauer distribution and obtained the estimators of the parameters by using different methods. Generalized Gegenbauer distribution studied by Medhi and Borah (1984), Wimmer and Altmann (1995). Medhi and Borah (1984) provided certain adhoc method for estimation of the parameters of Generalized Gegenbauer distribution.

. Lindley (1958) derived a distribution known as Lindley distribution based on Bayes' theorem. A Lindley mixture of Poisson distribution known as Poisson-Lindley distribution was studied by Sankaran (1970), with applications to errors and accidents. He introduced this distribution to model count data and pointed out some difficulties in obtaining the maximum likelihood estimator of the parameter of the distribution. Poisson-Lindley distribution is a special case of Bhattacharya's (1966) more complicated mixed Poisson distribution. Borah and Deka Nath (2001a) studied Inflated Poisson-Lindley (IPL) distribution and applied it successfully to the problem of biological data. Some mixtures of Poisson-Lindley distribution derived by using Gurland's generalization were studied by Borah and Deka Nath (2001b). Borah and Begum (2002) studied some properties of Poisson-Lindley and its derived distributions. Two forms of geometrically infinite divisible two-parameter Poisson-Lindley distribution studied by Borah and Begum (2002). They fitted these distributions to some biological and ecological data for comparison.

Ghitany and Al-Mutairi (2008) obtained the size-biased version of Poisson-Lindley distribution and discussed their various properties and applications. The zero-truncated form of Poisson-Lindley distribution was proposed by Ghitany et al. (2008). Ghitany and Al-Mutairi (2009) also discussed various estimation methods for the Poisson-Lindley distribution. Mahmoudi and Zakerzadeh (2010) obtained an extended version of Poisson-Lindley distribution and discussed its various properties and applications to errors and accidents data. Adhikari and Srivastava (2013) obtained a new form of the size-biased

Poisson-Lindley distribution without considering its size-biased form. Again, the Poisson size –biased Lindley distribution was proposed by Adhikari and Srivastava (2014).

Shanker et al. (2013) introduced a two-parameter Poisson-Lindley distribution, of which Sankaran's (1970) Poisson-Lindley distribution is a special case. Another form of two-parameter Poisson-Lindley distribution was studied by Shanker and Mishra (2014).

A zero-modified form of Poisson –Lindley distribution has been investigated by Dutta and Borah (2014) and fitted the distribution to some well-known data, for empirical comparison. They also reviewed some properties and applications of size-biased Poisson-Lindley distribution. Certain recurrence relations arising in different forms of size-biased Poisson-Lindley distribution has also been discussed by Dutta and Borah (2015).

Shanker and Mishra (2013b) obtained the quasi-Lindley mixture of Poisson distribution known as quasi Poisson-Lindley distribution with application. They also obtained the size-biased form of quasi Poisson-Lindley distribution along with its applications. Shanker and Tekie Asehun Leonida (2014) proposed a new quasi Poisson-Lindley (QPL) distribution. Shanker et al. (2014) investigated another mixture distribution known as Poisson-Janardan distribution (PJD), of which Sankaran's discrete Poisson-Lindley distribution is a particular case. They discussed the method of maximum likelihood and method of moments for estimating the parameter of Poisson-Janardan distribution. They also found that, Poisson-Janardan distribution is more flexible than the Sankaran's one parameter Poisson-Lindley distribution for analyzing different types of count data. Recently, a discrete Poisson-Sushila distribution was studied by Shanker et al. (2014).

The main aim of the thesis is to derive and study, some generalized form of Poisson-Lindley and quasi Poisson-Lindley distributions namely, size-biased, zero-modified and zero-truncated form of the distributions. The distributions have been studied with various applications in different fields of biological and ecological sciences, home injuries and accidents etc.

1.3 Objectives

In this investigation, our objectives were to study mixture of certain discrete probability distributions. To achieve the main goal of the study to be presented in the thesis the following objectives have been undertaken.

- To derive certain mixture distribution for some basic distributions.
- To investigate properties of the generalized derived distributions.
- To study the flexibility of the derived distributions.
- To obtain the recurrence relations for probabilities, moments and cumulants of the derived distributions.
- To estimate the parameters of distributions various method of estimation has been considered. Best estimator has been selected based on Chi-square goodness of fit.
- To fit and see the applications of the distributions some well-known reported data sets have been considered.

1.4 Outline and Organization of the thesis

With relevant to the objectives mentioned above, the present thesis is organized in eight chapters under broad headings.

- Chapter 1 Introduction
- Chapter 2 Some Properties and Application of Size-biased Poisson-Lindley Distribution.
- Chapter 3 On Certain Recurrence Relation Arising in Different Forms of Size-biased Poisson- Lindley Distribution.
- Chapter 4 Zero-modified Poisson-Lindley Distribution.
- Chapter 5 Certain Properties of Generalized Poisson-Lindley Distribution.

- Chapter 6 Two-parameter Quasi Poisson-Lindley Distribution and its Applications.
- Chapter 7 A Comparative Study on Zero-truncated Poisson-Lindley and Quasi Poisson- Lindley Distributions.
- Chapter 8 A Study on Some Properties of Poisson size-biased Quasi Lindley Distribution.

.

A brief summary of each chapter of the thesis mentioned above is highlighted below.

The first chapter is an introductory one which gives an account of the relevant works done earlier in the theory of univariate discrete probability distributions and on different types of finite, continuous and countable mixture of some discrete distributions.

The second chapter concerned a review on size-biased Poisson-Lindley distribution (Ghitany and Al-Mutairi, 2008). The distribution is further investigated by working out some properties like cumulative distribution function, probability recurrence relations, factorial moments recurrence relation, moments, coefficient of variation, reliability function etc. with the estimation procedures of the parameters used by the respective authors. To justify the suitability and applicability, the distribution has been fitted to a number of reported data sets. The resulting fit is found to be good in comparison to fits of other distribution fitted by different authors.

In chapter 3, an attempts have been made for deriving certain recursion formulae involving probabilities, factorial moments and cumulants of size-biased Poisson-Lindley distribution (Adhikari and Srivastava, 2013) and Poisson size-biased Lindley distribution (Adhikari and Srivastava, 2014). The parameters of these distributions are estimated by the method of moments and the ratio of the first two frequencies. A few reported data sets have been considered for empirical fitting of size-biased Poisson-Lindley and Poisson size-biased Lindley distributions with remarkable results.

The Zero-modified Poisson-Lindley distribution has been discussed to serve the probabilistic description of an experiment with slight modification of probability at zero in chapter 4. Some properties of the distribution including index of dispersion, recurrence relations for probabilities, moments and cumulants have been investigated. The parameters of the distribution have been estimated by the method of moments and ratio of the first two relative frequencies. The fitting of the distribution has been also considered by using some published data in different fields of biology and ecology to see the applicability of the distribution.

In chapter 5, an extended version of Poisson-Lindley distribution i.e., Generalized Poisson-Lindley distribution (Mahmoudi and Zakerzadeh, 2010) is revisited to obtain certain distributional properties of the distribution. The problem of estimation of the parameters by using an ad-hoc method has also been discussed. A few sets of error and accident data have been considered for fitting of the distribution and found a better fit in all the cases. The zero-modified form of Generalized Poisson-Lindley distribution is also discussed.

In chapter 6, the two-parameter quasi Poisson-Lindley distribution has been derived. The probability mass function and some statistical properties like recurrence relation for probabilities, factorial moments, moments and related measure, cumulative distribution function are studied. The parameters of two-parameter quasi Poisson-Lindley distribution are estimated by method of moments and maximum likelihood method. For testing validity of the estimate of the parameters of two-parameter quasi Poisson-Lindley distribution, the fitting of this distribution is considered. The zero-modified, zero-truncated and size-biased forms of quasi Poisson-Lindley distribution have been also discussed in this chapter.

In chapter 7, the zero-truncated quasi Poisson-Lindley distribution has been introduced and made a comparative study on zero-truncated Poisson-Lindley and quasi Poisson-Lindley distributions. An ad-hoc method is used for estimating the parameters of zero-truncated quasi Poisson-Lindley distribution. Different applications of zero-truncated quasi Poisson-Lindley distribution are also discussed. The fits are compared with the zero-truncated Poisson-Lindley distribution with good results.

In chapter 8, a study on some properties of the Poisson size-biased quasi-Lindley distribution has been made. The probability mass function of the distribution is investigated. Various statistical properties of Poisson size-biased quasi Lindley distribution with estimation techniques of the parameters have also been studied.

The literatures cited in the thesis are listed at the end.

*****