CHAPTER 3

A DISTRIBUTION FOR MODELLING DISPERSION

3.0 Introduction

The Poisson distribution is a popular model for count data and is characterized by the equality of its mean and variance. However, observed data tend to have unequal mean and variance. Therefore, the Poisson distribution has been extended and modified in different ways. For example, through the process of mixtures (Gupta and Ong, 2005) leading to distributions like the negative binomial (NB), Neyman Type A, generalized Poisson (Consul, 1989), generalizations of the NB (Gupta and Ong, 2004) and many more. The mixed Poisson distributions have variances greater than the means, that is, they exhibit over-dispersion.

One of the few distributions which are able to represent under-, equi- and over-dispersion is the generalized Poisson distribution (GPD). The GPD has probability mass function (pmf)

$$f(X=x)=\begin{cases}\dfrac{\theta(\theta+x\lambda)^{x-1}e^{-\theta-x\lambda}}{x!} & \text{for } x=0,\,1,\,2,\,\ldots \\ 0 & \text{for } x>m, \text{ when } \lambda<0\end{cases}$$

where $\theta>0$, max $(-1,-\theta/m)<\lambda\le 1$ and m (≥ 4) is the largest positive integer satisfying $\theta+m\lambda>0$ when λ is negative. As mentioned, the parameters are restricted to ensure that the GPD is a proper distribution for under-dispersion.

Another distribution of interest is the COM-Poisson introduced by Conway and Maxwell (1962) and further work was done by Shmueli *et al.* (2005). The pmf of the

COM-Poisson distribution is given by

$$f(X=x)=\frac{\lambda^x}{(x!)^v}\frac{1}{z(\lambda,v)}$$

where $z(\lambda,v)=\sum_{j=0}^{\infty}\frac{\lambda^j}{(j!)^v}$, for $\lambda>0$ and $v\geq 0$. The COM-Poisson distribution does not have closed form expression for its moments. More properties and discussion of the COM-Poisson distribution are given in Chapter 5. In addition, the double Poisson distribution is proposed to model the issue of under-, equi- and over-dispersion in the data. However, the distribution is not a proper distribution as explained previously.

In this chapter, we consider a particular case of the GIT denoted by $GIT_{3,1}$ which also has the flexibility to model under-, equi- and over-dispersion. An interesting feature of the distribution is that for the case of equi-dispersion the distribution does not reduce to the Poisson distribution. Under equi-dispersion, the GPD and COM-Poisson distribution reduce to the Poisson distribution. It is known that the equality of the mean and variance (equi-dispersion) characterizes the Poisson distribution among the power series distributions (Patil, 1962).

The chapter is organized as follows. Firstly the properties of the $GIT_{3,1}$ distribution and GPD distributions are given. Parameter estimation, testing of hypothesis for equi- and over-dispersion by Rao's score test and LR test, and simulation study of their power are provided. Then, applications to analyze real life data sets are presented where the $GIT_{3,1}$ distribution is compared to the GPD and COM-Poisson distributions.

3.1 Subclass of the Generalized Shifted Inverse Trinomial Distribution, $GIT_{3,1}$

Aoyama *et al.* (2008) have derived the GIT distribution by considering a first-passage time distribution of a modified random walk on the half plane with five transition probabilities p_1, p_2, p_3, p_4, p_5 ($p_i \geq 0$ for $i = 1, 2, ..., 5$; $\sum_{i=1}^{5} p_i = 1$) with barrier at $y = n$ (positive integer). Let *X* be a random variable which represents the number of steps *x* in the random walk. The GIT distribution arises from the movement of a particle from the origin with steps according to the transition probabilities until it first reaches the barrier *n* at the *x*-th step. The GIT family contains twenty-two possible distributions and a particular class designated as $GIT_{3,1}$ distribution has been examined in section 4 of Aoyama *et al.* (2008). The $GIT_{3,1}$ distribution includes the binomial, negative binomial and shifted negative binomial distributions as special cases.

The pmf of the $GIT_{3,1}$ distribution (Aoyama et al, 2008) is given by

$$f_n(x) = \sum_{i=0}^{\min(n,x)} \frac{n}{n+x-i} \binom{n+x-i}{n-i, i, x-i} p_1^{n-i} p_2^{i} p_3^{x-i} \tag{3.1}$$

for $x = 0, 1, 2, ...$, $n \in \mathbb{Z}^+$ and $p_i \geq 0$ for $i = 1, 2, 3$; $\sum_{i=1}^{3} p_i = 1$.

Equation (3.1) may also be expressed as

$$f_n(x) = \binom{n+x-1}{x} p_1^{n} p_3^{x} \, {}_2F_1\left(-n, -x; -n-x+1; -\frac{p_2}{p_1 p_3}\right) \tag{3.2}$$

in terms of the Gauss hypergeometric function ${}_2F_1$. Clearly, the expression of the pmf in (3.2) is a negative binomial pmf weighted by ${}_2F_1$. Another alternative expression showed that it as a binomial pmf weighted by ${}_2F_1$:

$$
\begin{aligned}
f_n(x) &= \sum_{j=\max(0,x-n)}^{x} \frac{n}{n+j}\binom{n+j}{n-x+j,x-j,j} p_1^{n-x+j} p_2^{x-j} p_3^{j} \\
&= \binom{n}{x} p_1^{n-x} p_2^{x}\, {}_2F_1\left(n,-x;n-x+1;-\frac{p_1p_3}{p_2}\right) \quad \text{if } n \geq x \\
&= \binom{x-1}{n-1} p_2^{n} p_3^{x-n}\, {}_2F_1\left(x,-n;x-n+1;-\frac{p_1p_3}{p_2}\right) \quad \text{if } x > n
\end{aligned}
$$

The probability generating function (pgf) is

$$
\varphi_n(t) = \left(\frac{p_1+p_2t}{1-p_3t}\right)^n = \left(\frac{p_1+p_2t}{p_1+p_2}\right)^n \left(\frac{1-p_3}{1-p_3t}\right)^n \quad , \; 1-p_3 = p_1+p_2 \tag{3.3}
$$

It reduces to a binomial pgf if $p_3 = 0$, negative binomial pgf if $p_2 = 0$ and shifted negative binomial (shifted *n* steps to the right) pgf if $p_1 = 0$.

The pmf $f_n(x)$ satisfies the following recurrence relation in *x*

$$
f_n(x) = \left(a+\frac{b}{x}\right) f_n(x-1) + c\left(1-\frac{2}{x}\right) f_n(x-2), \; x \geq 2 \tag{3.4}
$$

with $f_n(0) = p_1^n$, $f_n(1) = np_1^{n-1}(p_1p_3+p_2)$, where

$$
a = \frac{p_1p_3-p_2}{p_1}, \; b = \frac{\left(n(p_1p_3+p_2)-(p_1p_3-p_2)\right)}{p_1}, \; c = \frac{p_2p_3}{p_1}, \; p_1 > 0.
$$

The computation of probabilities by (3.4) is found to be stable with good accuracy. For a discussion on the stability of computation by three-term recurrence formula the interested reader is referred to Ong (1995) and Ong and Muthaloo (1995).

The *r*th descending factorial moment of *X* is

$$
\begin{aligned}
\mu'_{[r]} &= E(X(X-1)...(X-r+1)) \\
&= \frac{n}{(p_1+p_2)^r} \sum_{i=0}^{r} \binom{r}{i} \frac{(n+r-i-1)!}{(n-i)!} p_2^i p_3^{r-i}
\end{aligned}
$$

for $r \geq 1$ and it satisfies the recursion formula

$$(p_1+p_2)^2\,\mu'_{[r+1]}+\{r(p_1+p_2)(p_2-p_3)-n(p_1+p_2)(p_2+p_3)\}\mu'_{[r]}-r(r-1)p_2p_3\mu'_{[r-1]}=0$$

with initial conditions $\mu'_{[0]}=1$, $\mu'_{[1]}=n\dfrac{p_2+p_3}{p_1+p_2}$.

The mean and variance are found to be

$$E(X)=n\frac{p_2+p_3}{p_1+p_2},\qquad Var(X)=n\frac{p_1p_2+p_3}{(p_1+p_2)^2},$$

and these lead to the index of dispersion (ID)

$$\text{ID}=\frac{Var(X)}{E(X)}=\frac{p_1p_2+p_3}{(p_1+p_2)(p_2+p_3)}\quad\begin{cases}>1, & p_3>p_2\\ <1, & p_3<p_2\end{cases}\tag{3.5}$$

Note that if $p_2=p_3$, then ID=1 but $\text{GIT}_{3,1}(n;1-2p,p,p)$ with $0<p<1/2$ is not a Poisson distribution.

In the limit the $\text{GIT}_{3,1}$ distribution goes to a Poisson distribution with parameter $\lambda_2+\lambda_3$ as *n* tends to infinity, provided $np_2=\lambda_2$ and $np_3=\lambda_3$.

3.2 Behaviour of the Subclass of the Generalized Shifted Inverse Trinomial Distribution, $\text{GIT}_{3,1}$

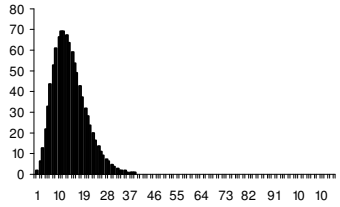
The parameters p_2 and p_3 play an important role in determining the dispersion of the model. To study the behavior of the $\text{GIT}_{3,1}$ distribution, the probabilities are computed and plotted for various values of the parameters as displayed in Table 3.1.

Table 3.1: Dispersion under the $\text{GIT}_{3,1}$ distribution

	Over-dispersion $p_3>p_2$		Under-dispersion $p_3<p_2$		Equal-dispersion $p_3=p_2$
p_1	0.2	0.6	0.2	0.6	0.8
p_2	0.1	0.1	0.7	0.3	0.1
p_3	0.7	0.3	0.1	0.1	0.1

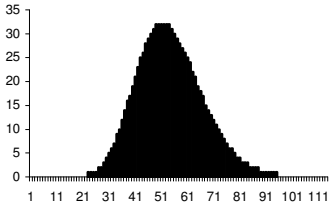
The positive-integer parameter *n* is set as 5 and 20. Thus two graphs are plotted for each set of parameters corresponding to over-, under- and equi-dispersion data.

Over-dispersion



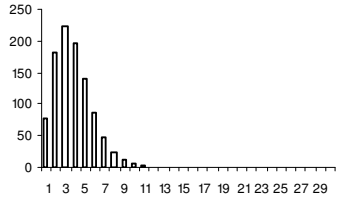
$\mu = 13,\ \sigma^2 = 40,\ \mathrm{ID} = 3$

Figure 3.1: p_1=0.2, p_2=0.1, p_3=0.7,*n*=5



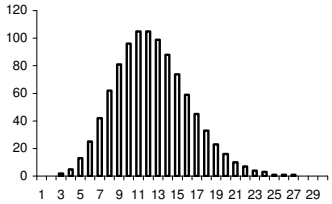
$\mu = 53.33,\ \sigma^2 = 160,\ \mathrm{ID} = 3$

Figure 3.2: p_1=0.2, p_2=0.1, p_3=0.7, *n*=20



$\mu = 2.857,\ \sigma^2 = 3.673,\ \mathrm{ID} = 1.286$

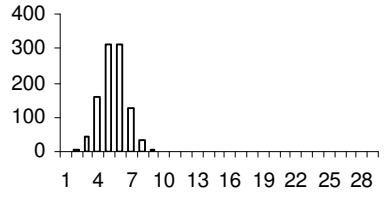
Figure 3.3: p_1=0.6, p_2=0.1, p_3=0.3,*n*=5



$\mu = 11.429,\ \sigma^2 = 14.69,\ \mathrm{ID} = 1.286$

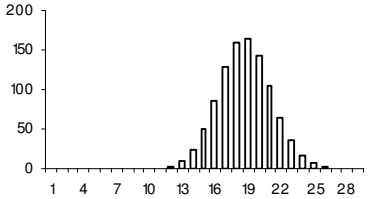
Figure 3.4: p_1=0.6, p_2=0.1, p_3=0.3,*n*=20

Under-dispersion



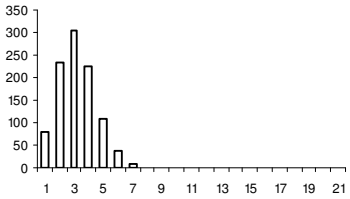
$\mu = 4.444,\ \sigma^2 = 1.481,\ \mathrm{ID} = 0.333$

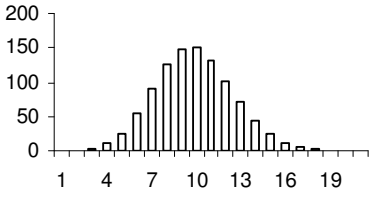
Figure 3.5: p_1=0.2, p_2=0.7, p_3=0.1,*n*=5



$\mu = 17.778,\ \sigma^2 = 5.926,\ \mathrm{ID} = 0.333$

Figure 3.6: p_1=0.2, p_2=0.7, p_3=0.1,*n*=20





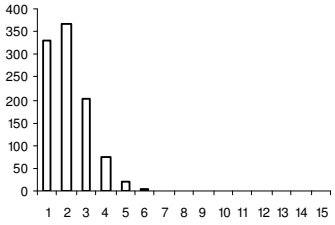
$\mu = 2.222,\ \sigma^2 = 1.728,\ \mathrm{ID} = 0.778$

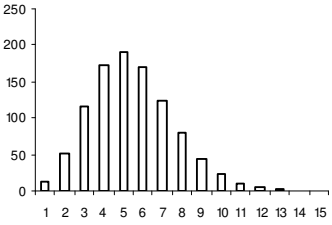
Figure 3.7: p_1=0.6, p_2=0.3, p_3=0.1,n=5

$\mu = 8.888,\ \sigma^2 = 6.913,\ \mathrm{ID} = 0.778$

Figure 3.8: p_1=0.6, p_2=0.3, p_3=0.1,n=20

Equi-dispersion data





$\mu = 1.11,\ \sigma^2 = 1.11,\ \mathrm{ID} = 1$

Figure 3.9: p_1=0.8, p_2=0.1, p_3=0.1,n=5

$\mu = 4.444,\ \sigma^2 = 4.444,\ \mathrm{ID} = 1$

Figure 3.10: p_1=0.8, p_2=0.1, p_3=0.1,n=20

Figures 3.1 to 3.10 show that for small n (=5), the $\mathrm{GIT}_{3,1}$ distributions are skewed to the right. For large n (=20), they are symmetrical and approach a bell-shaped distribution.

An empirical study shows that, for any given values of p_1 , p_2 and p_3 , increasing the value of n shifts the mode to the right where a larger span on the x axis is acquired by moving the distribution to the right side. The graphs gradually lose asymmetry and achieve a bell-shaped form.

3.3 Extension of the $GIT_{3,1}$ to Non-Integer *n*

The generating function (Kemp, 1979)

$$\left(\frac{1-Q_1 t}{1-Q_1}\right)^{U_1}\left(\frac{1-Q_2 t}{1-Q_2}\right)^{U_2} \tag{3.6}$$

is a valid pgf for a non-degenerate distribution on the non-negative integers provided that certain sets of conditions are satisfied. A particular set of conditions of interest is $U_2 < 0 < U_1,\ 0 < -Q_1 < 1,\ 0 < Q_2 < 1$.

From (3.3)

$$\varphi_n(t) = \left(\frac{p_1 + p_2 t}{p_1 + p_2}\right)^n \left(\frac{1-p_3}{1-p_3 t}\right)^n$$

$$= \left(\frac{1-(-p_2 t / p_1)}{1-(-p_2 / p_1)}\right)^n \left(\frac{1-p_3 t}{1-p_3}\right)^{-n} \quad ,\ 1-p_3 = p_1 + p_2 \quad . \tag{3.7}$$

A direct comparison between equation (3.6) and (3.7) gives

$Q_1 = -\frac{p_2}{p_1}$ where $0 < -Q_1 = \frac{p_2}{p_1} < 1$ and $0 < p_2 < p_1$

$Q_2 = p_3$ where $0 < p_3 < 1,\ -n = U_2 < 0 < U_1 = n$.

Observed that the positive-integer parameter *n* can only be extended to a real number when $0 < p_2 < p_1$. Clearly, for *n* to be real, the restriction is required.

3.4 Log-concavity, Unimodality and Reliability Properties

A distribution is said to be log-concave if its pmf $\{f(k)\}$, $f(k) > 0$, for all *k* satisfies $f^2(k) \geq f(k+1) f(k-1)$ for all *k*. The failure rate is defined by $r(k) = f(k) / \sum_{i \geq k} f(i)$.

The pgf (3.3) may be written as

$$\varphi_n(t) = \left(\frac{p_1}{p_1 + p_2} + \frac{p_2 t}{p_1 + p_2} \right)^n \left(\frac{1 - p_3}{1 - p_3 t} \right)^n$$

This shows that $\mathrm{GIT}_{3,1}$ distribution can be regarded as a convolution of binomial $\mathrm{B}\left(n, p_1/(p_1 + p_2)\right)$ and negative binomial $\mathrm{NB}(n, p_3)$ distributions. From the results of Keilson and Geber (1971, page 388), the $\mathrm{GIT}_{3,1}$ distribution is log-concave as both of the binomial and negative binomial distributions are log-concave.

It follows from log-concavity that the $\mathrm{GIT}_{3,1}$ distribution has an increasing failure rate (*IFR*) (Gupta *et al.*, 2008, page 527). Furthermore the ensuing implications hold

$$IFR \Rightarrow IFRA \Rightarrow NBU \Rightarrow NBUE \Rightarrow HNBUE$$

where *IFRA* (increasing failure rate average), *NBU* (new better than used), *NBUE* (new better than used in expectation) and *HNBUE* (harmonic new better than used in expectation). Hence the $\mathrm{GIT}_{3,1}$ distribution is *IFR*, *IFRA*, *NBU*, *NBUE* and *HNBUE*.

From Theorem 3 of Keilson and Geber (1971, page 386) which states that a necessary and sufficient condition for pmf $\{f(k)\}$ be strongly unimodal is that $f(k)$ be log-concave for all *k*, it follows that the $\mathrm{GIT}_{3,1}$ distribution is strongly unimodal.

3.5 Parameter Estimation

3.5.1 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is a very popular statistical estimation method which provides consistent and efficient estimators. We shall consider the case

where the parameter n is fixed. Let the random sample $x_1, x_2, ..., x_N$ arises from the $\mathrm{GIT}_{3,1}$ distribution with pmf $f(x)$. The log-likelihood function is

$$\ln L = \sum_{s=0}^{N} \pi_s \ln f(s) \ , \qquad \pi_s = \text{observed frequency} \tag{3.8}$$

The likelihood score equation of p_2 is

$$\frac{\partial \ln L}{\partial p_2} = \frac{\pi_0}{f(0)} \frac{\partial f(0)}{\partial p_2} + \sum_{k=1}^{N} \frac{\pi_k}{f(k)} \frac{\partial f(k)}{\partial p_2} \tag{3.9}$$

where N = sample size

$$\frac{\partial f(0)}{\partial p_2} = -\phi f(1)$$

$$\frac{\partial f(k)}{\partial p_2} = \phi\big((1+p_3)kf(k) - p_3(k-1)f(k-1) - (k+1)f(k+1)\big), \qquad k \geq 1$$

and $\phi = \dfrac{1}{(1-p_3)(p_2+p_3)}$

The likelihood score equation of p_3 is found to be

$$\frac{\partial \ln L}{\partial p_3} = \frac{\pi_0}{f(0)} \frac{\partial f(0)}{\partial p_3} + \sum_{k=1}^{n} \frac{\pi_k}{f(k)} \frac{\partial f(k)}{\partial p_3} \tag{3.10}$$

where $\dfrac{\partial f(0)}{\partial p_3} = -\phi f(1)$ and

$$\frac{\partial f(k)}{\partial p_3} = \phi\big((1-p_2)kf(k) + p_2(k-1)f(k-1) - (k+1)f(k+1)\big), \qquad k \geq 1$$

The methodology to solve unknown parameters under MLE is simple but the implementation is mathematically involved due to the complicated form of the $\mathrm{GIT}_{3,1}$ pmf. To mitigate this, numerical optimization is employed to find the parameters that will maximize the log-likelihood function. The simulated annealing (SA) algorithm (Metropolis *et al*., 1953) is used in the optimization as it is powerful in searching for a global extremum of a non-smooth function with has many local extrema. Moreover, SA does not require calculation of derivatives.

3.5.2 A Probability Generating Function Based Estimation

Since the pgf of the $\mathrm{GIT}_{3,1}$ distribution is of a simple form, an attractive method of estimation is the pgf-based estimation examined by Sim and Ong (2010). In Sim and Ong (2010), a minimum Hellinger-type distance and squared distance statistics based upon pgf have been considered to provide quick and consistent estimation of discrete distributions.

The pgf based estimators are obtained by minimizing the following quantities

$$T_1 = \int_0^1 \left(\sqrt{f_N(t)} - \sqrt{\varphi(t)} \right)^2 dt \tag{3.11}$$

$$T_2 = \int_0^1 \left(f_N(t) - \varphi(t) \right)^2 dt \tag{3.12}$$

where $f_N(t) = \frac{1}{N} \sum_{i=1}^{N} t^{x_i}$ is the empirical pgf and $\varphi(t)$ is the pgf of the distribution.

3.6 Test for Equi-Dispersion

The index of dispersion shows that when $p_2 = p_3$, the $\mathrm{GIT}_{3,1}$ distribution is equi dispersed. Thus, to test for equi-dispersion we consider the following set of hypotheses.

$$H_0 : p_2 = p_3 \qquad \text{or} \qquad H_0 : p_2 - p_3 = 0$$
$$H_1 : p_2 \neq p_3 \qquad\qquad H_1 : p_2 - p_3 \neq 0$$

The test of hypothesis may be based upon Rao's score test, the likelihood ratio test or the Wald test where their test statistics are asymptotically χ^2 distributed. In this work we consider Rao's score test and likelihood ratio test. A comparative study of the power of these two tests will be given in the next subsection. Meanwhile, the partial derivatives for the pmf can be obtained from the partial derivatives of pgf with respect to the parameters of interest, p_2 and p_3, as given in the following section.

3.6.1 Partial Differentiation

The first order partial derivatives of the pgf with respect to its parameters are

$$\frac{\partial G_n(t)}{\partial p_2} = \phi G'_n(t)(t-1-p_3t^2+p_3t) \tag{3.13}$$

$$\frac{\partial G_n(t)}{\partial p_3} = \phi G'_n(t)\left(t(1-p_2)+p_2t^2-1\right) \tag{3.14}$$

where $G'_n(t) = \left(\frac{nG_n(t)}{1-p_2-p_3+p_2t}\right)\frac{(p_2+p_3)(1-p_3)}{(1-p_3t)}$

From $\frac{\partial G_n(t)}{\partial p_i} = \sum_{k=0}^{\infty}\frac{\partial f(k)}{\partial p_i}t^k$, $i=1,2$, the partial derivative for the pmf of the $GIT_{3,1}$ distribution can be obtained as follows.

From (3.13)

$$\frac{\partial f(0)}{\partial p_2} = -\phi f(1) \tag{3.15}$$

$$\frac{\partial f(k)}{\partial p_2} = \phi\left((1+p_3)kf(k)-p_3(k-1)f(k-1)-(k+1)f(k+1)\right), \qquad k \geq 1 \tag{3.16}$$

From (3.14)

$$\frac{\partial f(0)}{\partial p_3} = -\phi f(1) \tag{3.17}$$

$$\frac{\partial f(k)}{\partial p_3} = \phi\left((1-p_2)kf(k)+p_2(k-1)f(k-1)-(k+1)f(k+1)\right), \qquad k \geq 1 \tag{3.18}$$

Then, from equation (3.15) to (3.18), the second order partial derivatives of the pmf are

$$\frac{\partial^2 f(0)}{\partial p_2^2} = -\phi^2\left[2p_3f(1)-2f(2)\right] = \frac{\partial^2 f(0)}{\partial p_3^2}$$

$$\frac{\partial^2 f(k)}{\partial p_2^2} = \frac{\partial \phi}{\partial p_2}\left((1+p_3)kf(k)-p_3(k-1)f(k-1)-(k+1)f(k+1)\right)+$$
$$\phi\left((1+p_3)k\frac{\partial f(k)}{\partial p_2}-p_3(k-1)\frac{\partial f(k-1)}{\partial p_2}-(k+1)\frac{\partial f(k+1)}{\partial p_2}\right)$$

$$\frac{\partial^2 f(k)}{\partial p_3^2} = \frac{\partial \phi}{\partial p_3}\left((1-p_2)kf(k) + p_2(k-1)f(k-1) - (k+1)f(k+1)\right) +$$
$$\phi\left((1-p_2)k\frac{\partial f(k)}{\partial p_3} + p_2(k-1)\frac{\partial f(k-1)}{\partial p_3} - (k+1)\frac{\partial f(k+1)}{\partial p_3}\right)$$
$$\frac{\partial^2 f(0)}{\partial p_2 \partial p_3} = -\phi^2\left[2p_3 f(1) - 2f(2)\right] = \frac{\partial^2 f(0)}{\partial p_3 \partial p_2}$$

$$\frac{\partial^2 f(k)}{\partial p_2 \partial p_3} = \frac{\partial \phi}{\partial p_2}\left((1-p_2)kf(k) + p_2(k-1)f(k-1) - (k+1)f(k+1)\right) +$$
$$\phi\left(-kf(k) + (1-p_2)k\frac{\partial f(k)}{\partial p_2} + (k-1)f(k-1)\; p_2(k-1)\frac{\partial f(k-1)}{\partial p_2} - (k+1)\frac{\partial f(k+1)}{\partial p_2}\right)$$

$$\frac{\partial^2 f(k)}{\partial p_3 \partial p_2} = \frac{\partial^2 f(k)}{\partial p_2 \partial p_3}$$

where $\frac{\partial \phi}{\partial p_2} = -(1-p_3)\phi^2$ and $\frac{\partial \phi}{\partial p_3} = (p_2 + 2p_3 - 1)\phi^2$

3.6.2 Rao's Score Test

Rao's score test statistic is given by

$$T = VI^{-1}V^T$$

where the score vector V and information matrix I are evaluated at the restricted maximum likelihood estimates. The score vector is given as

$$V = \left(\frac{\partial \ln L}{\partial p_2}, \frac{\partial \ln L}{\partial p_3}\right)$$

where $\frac{\partial \ln L}{\partial p_2}$ and $\frac{\partial \ln L}{\partial p_2}$ are given by equation (3.9) and (3.10) respectively.

The information matrix I, is defined as

$$I = -\begin{bmatrix} E\left[\frac{\partial^2 \ln L}{\partial p_2{}^2}\right] & E\left[\frac{\partial^2 \ln L}{\partial p_2 \partial p_3}\right] \\ E\left[\frac{\partial^2 \ln L}{\partial p_3 \partial p_2}\right] & E\left[\frac{\partial^2 \ln L}{\partial p_2{}^2}\right] \end{bmatrix}$$

where $E\left[-\frac{\partial^2 \ln L}{\partial p_3 \partial p_2}\right] = N\sum_{k=0}^{\infty}\frac{1}{f(k)}\frac{\partial f(k)}{\partial p_3}\frac{\partial f(k)}{\partial p_2}$ and N is sample size.

The partial derivatives and elements of information matrix for $GIT_{3,1}$ distribution are available in Appendix A.

3.6.3 Likelihood Ratio Test (LRT)

To perform a LRT, estimation of the hypothesized model under the null and alternative hypotheses is needed unlike Rao's score test which requires only the restricted ML estimator in the null hypothesis. The LR test begins with a comparison of the log likelihood scores of the two models and gives evidence on whether the difference is statistically significant. The LR test statistic is

$$\text{LR} = -2\ln\left(\frac{L(\hat{\boldsymbol{\theta}}^*; x)}{L(\hat{\boldsymbol{\theta}}; x)}\right)$$

where $\hat{\boldsymbol{\theta}}^*$ is the restricted ML estimator and $\hat{\boldsymbol{\theta}}$ is the unrestricted ML estimator.

3.7 Power Analysis of the Rao's Score Test and Likelihood Ratio Test

The power of a statistical test of hypothesis is the probability of rejecting the null hypothesis when the null hypothesis is untrue. In this section, a simulation study is conducted with 10,000 Monte Carlo repetitions. Under the ML estimates, the asymptotic properties are true when the sample size *N* is sufficiently large. *N* is set at 100, 500 and 1000, to represent small, medium and large sample sizes. The significance level α is chosen to be 5% and 10%. The effect size which is the index of departure from null hypothesis ($|p_2 - p_3|$) is incremented from 0.1 to 0.7 in steps of 0.1.

The results of the power study are displayed in Tables 3.2 and 3.3 for two cases: over- and under-dispersion. In addition, the simulation study is also designed to provide

the estimated empirical level. The results are presented in Table 3.4. The estimated power is given by the number of rejections divided by the number of repetitions.

The Wald test is not considered here because it requires estimation from the unrestricted model which is more complicated. Moreover, the performance of the Rao's score test and LRT for small sample sizes is seen to be satisfactory.

Table 3.2: Simulated power of score test and LRT for $GIT_{3,1}$: over-dispersion

p_1			0.7	0.6	0.5	0.4	0.3	0.2	0.1
p_2			0.1	0.1	0.1	0.1	0.1	0.1	0.1
p_3			0.2	0.3	0.4	0.5	0.6	0.7	0.8
ID			1.1250	1.2857	1.5000	1.8000	2.2500	3.0000	4.5000
N	α	Method	Power						
100	0.05	score	0.1460	0.4539	0.8260	0.9831	0.9994	1.0000	1.0000
		LR	0.1175	0.4021	0.7936	0.9771	0.9993	1.0000	1.0000
	0.10	score	0.2203	0.5628	0.8818	0.9915	0.9996	1.0000	1.0000
		LR	0.1955	0.5208	0.8622	0.9885	0.9996	1.0000	1.0000
500	0.05	score	0.4749	0.9753	1.0000	1.0000	1.0000	1.0000	1.0000
		LR	0.4447	0.9709	1.0000	1.0000	1.0000	1.0000	1.0000
	0.10	score	0.5843	0.9873	1.0000	1.0000	1.0000	1.0000	1.0000
		LR	0.5642	0.9864	1.0000	1.0000	1.0000	1.0000	1.0000
1000	0.05	score	0.7436	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		LR	0.7273	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
	0.10	score	0.8255	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		LR	0.8171	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 3.3: Simulated power of score test and LRT for $GIT_{3,1}$: under-dispersion

p_1			0.7	0.6	0.5	0.4	0.3	0.2	0.1
p_2			0.2	0.3	0.4	0.5	0.6	0.7	0.8
p_3			0.1	0.1	0.1	0.1	0.1	0.1	0.1
ID			0.8889	0.7778	0.6667	0.5556	0.4444	0.3333	0.2222
N	α	Method	Power						
100	0.05	score	0.1056	0.3490	0.7707	0.9814	0.9999	1.0000	1.0000
		LR	0.1505	0.4322	0.8314	0.9883	0.9999	1.0000	1.0000
	0.10	score	0.1977	0.5141	0.8777	0.9914	0.9999	1.0000	1.0000
		LR	0.2345	0.5741	0.9018	0.9955	1.0000	1.0000	1.0000
500	0.05	score	0.4336	0.9769	1.0000	1.0000	1.0000	1.0000	1.0000
		LR	0.4682	0.9808	1.0000	1.0000	1.0000	1.0000	1.0000
	0.10	score	0.5781	0.9913	1.0000	1.0000	1.0000	1.0000	1.0000
		LR	0.6008	0.9924	1.0000	1.0000	1.0000	1.0000	1.0000
1000	0.05	score	0.7418	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		LR	0.7620	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0.10	score	0.8431	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		LR	0.8533	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 3.4: Estimated empirical level of score test and LRT for $GIT_{3,1}$

$p_1 = 0.8, p_2 = 0.1, p_3 = 0.1$			
ID			1.0
N	α	Method	Size
100	0.05	score	0.0487
		LR	0.0508
	0.10	score	0.0961
		LR	0.1046
500	0.05	score	0.0467
		LR	0.0496
	0.10	score	0.0987
		LR	0.0993
1000	0.05	score	0.0463
		LR	0.0475
	0.10	score	0.0954
		LR	0.0973

Table 3.2, for the case of over-dispersion, shows that the powers of the score test is marginally higher than the LR test for any tested sample sizes. For under-dispersion the reverse result is obtained (Table 3.3). For equi-dispersion (Table 3.4), Rao's score and LR tests have estimated empirical levels closer to the specified significance levels of 5% and 10%.

The Monte Carlo results show that as expected, the larger the sample size, the smaller the probability of Type II error and of course the higher the reliability of the power function. Generally, the statistical power can be increased by increasing the sample size. On the other hand, the larger the effect size is posited, the difference in between parameters p_2 and p_3 can be detected even in smaller size. If the effect size is larger or equal to 0.6, 100% detection is achieved and the model will be rejected

strongly even for small sample size of 100. When the sample size increases to 500 (medium sample size), an effect size of at least 0.3 is necessary for the detection. Consequently, for large sample size of N=1000, the difference between the parameters can be detected for an effect size of 0.2.

3.8 Applications

In this section, some real life data sets, representing over-, under- and equi-dispersion are analyzed. The fit of $GIT_{3,1}$ distribution is compared with the GPD and COM-Poisson distribution. Parameters of GPD are estimated by maximum likelihood (Consul, 1989). For comparison the chi-square goodness-of-fit values are shown.

For $GIT_{3,1}$, the positive integer of parameter n is chosen from the fitting which provides the lowest chi-square value. Simultaneously, two pgf based estimators are computed as a comparison with the use of pmf based estimators (MLE and Minimum Hellinger (MHD)) under the $GIT_{3,1}$ distribution. Also, the ID and LR test are provided.

3.8.1 Over-Dispersion Data

Two real life data sets are analyzed:

a) The lesions in rabbit lymphoblasts induced by streptonigrin with exposure 30 ($\mu g / kg$) (Table 3.5)
b) The lesions in rabbit lymphoblasts induced by streptonigrin with exposure 90 ($\mu g / kg$) (Table 3.6)

Table 3.5: Frequency distribution of dicentrics for mammalian cytogenetic dosimetry: lesions in rabbit lymphoblasts induced by streptonigrin with exposure 30 ($\mu g / kg$) (NSC-45383)

Number of dicentrics per cell	Observed frequency	Expected frequency					
	Exposure 30 ($\mu g / kg$)	GPD pmf	COM-Poisson pmf	GIT(3,1) pmf		GIT(3,1) pgf	
		MLE	MLE	MLE	MHD	T_1	T_2
0	404	403.85	403.89	404.00	403.61	403.99	403.99
1	80	80.30	80.08	80.14	79.42	80.00	80.00
2	13	13.33	13.58	13.24	13.98	13.34	13.34
3	3	2.52	2.45	2.62	2.99	2.67	2.67
Total	500	500	500	500	500	500	500
χ^2		0.10	0.15	0.06	0.07	0.05	0.05
P-value		0.752	0.699	0.806	0.791	0.823	0.823
Estimated parameters		$\hat{\theta}$=0.214	$\hat{\lambda}$=0.198	n=1	n=1	n=1	n=1
		$\hat{\lambda}$=0.072	$\hat{v}$=0.226	$\hat{p}_2$=0.027	$\hat{p}_2$=0.017	$\hat{p}_2$=0.0253	$\hat{p}_2$=0.0253
				$\hat{p}_3$=0.165	$\hat{p}_3$=0.176	$\hat{p}_3$=0.1667	$\hat{p}_3$=0.1667
Mean	0.23						
Variance	0.27						
ID	1.15						
LR test	Reject H_0 with LR=4.76.						

Table 3.6: Frequency distribution of dicentrics for mammalian cytogenetic dosimetry: lesions in rabbit lymphoblasts induced by streptonigrin with exposure 90 ($\mu g / kg$) (NSC-45383)

Number of dicentrics per cell	Observed frequency	Expected frequency					
	Exposure 90 ($\mu g / kg$)	GPD	COM-Poisson	$GIT_{3,1}$		$GIT_{3,1}$ pgf-based	
		MLE	MLE	MLE	MHD	T_1	T_2
0	155	154.64	149.25	155.00	154.72	155.13	155.13
1	83	81.96	85.37	82.13	80.49	81.56	81.54
2	33	36.36	39.67	35.61	35.90	35.64	35.64
3	14	15.52	16.32	15.44	16.01	15.58	15.58
4	11	6.59	6.16	6.69	7.14	6.81	6.81
5	3	2.80	2.17	2.90	3.18	2.98	2.98
6	1	2.12	1.06	2.22	2.56	2.31	2.31
Total	300	300	300	300	300	300	300
χ^2		4.04	5.86	3.78	3.62	3.71	3.70
P-value		0.401	0.210	0.437	0.460	0.447	0.448
Estimated parameters		$\hat{\theta}$=0.663	$\hat{\lambda}$=0.572	n=1	n=1	n=1	n=1
		$\hat{\lambda}$=0.223	$\hat{v}$=0.300	$\hat{p}_2$=0.050	$\hat{p}_2$=0.038	$\hat{p}_2$=0.0459	$\hat{p}_2$=0.0458
				$\hat{p}_3$=0.434	$\hat{p}_3$=0.446	$\hat{p}_3$=0.4370	$\hat{p}_3$=0.4371
Mean	0.85						
Variance	1.37						
ID	1.61						
LR test	Reject H_0 with LR= 34.59						

For the over-dispersed data set in Tables 3.5 and 3.6, the $GIT_{3,1}$ distribution achieves a lower chi-square value compared with the GPD and COM-Poisson distributions. H_0 is rejected by the LR test with a small value of test statistic (Table 3.5) and a very large value of test statistic (Table 3.6). This tallies with a significant difference between parameters p_2 and p_3.

3.8.2 Under-Dispersion Data

In this section, two real life data sets representing under-dispersion are presented.

a) The frequency distributions of dicentrics for dose 1200 (Table 3.7);

b) The frequency distributions of dicentrics for dose 2000 (Table 3.8).

Table 3.7: Frequency distribution of dicentrics for dose 1200

Number of dicentrics per cell	Observed frequency	Expected frequency					
	Dose	GPD	COM-Poisson	$GIT_{3,1}$		$GIT_{3,1}$ pgf based	
	1200	MLE	MLE	MLE	MHD	T_1	T_2
0	0	0.27	0.09	0.18	0.11	0.18	0.22
1	4	2.30	1.69	1.81	1.41	1.93	2.09
2	5	8.75	8.61	8.02	7.29	8.63	8.64
3	23	19.71	21.09	20.05	20.19	21.09	20.32
4	24	29.26	30.65	31.10	32.62	31.14	30.17
5	38	30.15	29.71	31.32	31.88	29.50	29.90
6	21	22.10	20.70	21.25	20.24	19.56	20.77
7	10	11.64	10.90	10.38	9.90	10.31	10.82
8	1	4.39	4.51	4.03	4.10	4.67	4.60
9	4	1.42	2.04	1.87	2.25	2.98	2.47
Total	130	130	130	130	130	130	130
χ^2		14.29	13.40	12.17	13.12	11.52	11.13
P-value		0.046	0.063	0.095	0.069	0.117	0.133
Estimated parameters		$\hat{\theta}$=6.171	$\hat{\lambda}$=17.948	n=6	n=5	n=5	n=6
		$\hat{\lambda}$=-0.317	$\hat{v}$=1.813	$\hat{p}_2$=0.523	$\hat{p}_2$=0.558	$\hat{p}_2$=0.510	$\hat{p}_2$=0.491
				$\hat{p}_3$=0.145	$\hat{p}_3$=0.197	$\hat{p}_3$=0.221	$\hat{p}_3$=0.164
Mean	4.69						
Variance	2.68						
ID	0.57						
LR test	Reject H_0 with LR=16.99.						

Table 3.8: Frequency distribution of dicentrics for dose 2000

Number of dicentrics per cell	Observed frequency	Expected frequency					
	Dose	GPD	COM-Poisson	$GIT_{3,1}$		$GIT_{3,1}$ pgf based	
	2000	MLE	MLE	MLE	MHD	T_1	T_2
0	0	0.00	0.00	0.00	0.00	0.00	0.00
1	0	0.06	0.02	0.02	0.02	0.02	0.02
2	0	0.34	0.22	0.18	0.17	0.16	0.19
3	2	1.32	1.08	0.91	0.89	0.90	0.98
4	3	3.66	3.45	3.17	3.09	3.27	3.30
5	6	7.74	7.85	7.77	7.57	8.14	7.90
6	17	12.98	13.50	14.04	13.69	14.52	14.01
7	15	17.70	18.34	19.28	18.89	19.35	19.04
8	21	20.00	20.31	20.86	20.61	20.28	20.53
9	21	18.96	18.78	18.51	18.51	17.71	18.28
10	14	15.25	14.78	14.02	14.22	13.46	13.96
11	10	10.48	10.05	9.35	9.64	9.19	9.43
12	5	6.19	5.98	5.64	5.92	5.77	5.77
13	1	3.15	3.15	3.14	3.35	3.38	3.26
14	5	2.16	2.47	3.10	3.43	3.84	3.33
Total	120	120	120	120	120	120	120
χ^2		8.74	7.59	6.55	6.36	6.36	6.22
P-value		0.73	0.82	0.89	0.90	0.90	0.90
Estimated parameters		$\hat{\theta}$=10.225	$\hat{\lambda}$=26.731	n=8	n=8	n=7	n=8
		$\hat{\lambda}$=-0.221	$\hat{v}$=1.531	$\hat{p}_2$=0.499	$\hat{p}_2$=0.494	$\hat{p}_2$=0.500	$\hat{p}_2$=0.489
				$\hat{p}_3$=0.267	$\hat{p}_3$=0.273	$\hat{p}_3$=0.318	$\hat{p}_3$=0.273
Mean	8.38						
Variance	5.63						
ID	0.67						
LR test	Reject H_0 with LR=7.52						

The result of the LR test in both Tables 3.7 and 3.8 provides evidence that p_2 is significantly different from p_3. With a smaller χ^2, the fit by the $GIT_{3,1}$ distribution is better than the GPD and COM-Poisson distributions.

3.8.3 Equi-Dispersion Data

Here we present the frequency distributions of dicentrics for dose 800 (Table 3.9) and behaviour of spiders (Table 3.10).

Table 3.9: Frequency distribution of dicentrics for dose 800

Number of dicentrics per cell	Observed frequency	Expected frequency					
	Dose 800	GPD	COM-Poisson	$GIT_{3,1}$		$GIT_{3,1}$ pgf based	
		MLE	MLE	MLE	MHD	T_1	T_2
0	6	8.53	7.25	5.45	5.38	5.85	5.73
1	24	23.32	23.18	24.76	25.42	24.64	24.69
2	34	31.07	32.38	37.64	39.31	35.87	36.43
3	27	26.88	27.89	25.31	25.34	24.93	25.06
4	20	16.99	17.04	13.90	13.27	14.28	14.16
5	5	8.36	7.98	6.95	6.32	7.46	7.30
6	2	3.33	3.01	3.29	2.84	3.70	3.56
7	1	1.11	0.94	1.51	1.24	1.77	1.68
8	0	0.31	0.25	0.67	0.52	0.83	0.77
9	0	0.08	0.06	0.29	0.22	0.38	0.35
10	1	0.02	0.01	0.22	0.15	0.31	0.28
Total	120	120	120	120	120	120	120
χ^2		51.88	100.64	8.17	10.52	7.25	7.41
P-value		0.00	0.00	0.42	0.23	0.51	0.49
Estimated parameters		$\hat{\theta}$=2.644	$\hat{\lambda}$=3.195	n=2	n=2	n=2	n=2
		$\hat{\lambda}$=-0.033	$\hat{\nu}$=1.193	$\hat{p}_2$=0.402	$\hat{p}_2$=0.423	$\hat{p}_2$=0.376	$\hat{p}_2$=0.384
				$\hat{p}_3$=0.385	$\hat{p}_3$=0.366	$\hat{p}_3$=0.403	$\hat{p}_3$=0.397
Mean	2.56						
Variance	2.38						
ID	0.93						
LR test	Do not reject H_0 with LR=0.04.						

Table 3.10: Frequency distribution of 102 spiders under 240 boards

Number of dicentrics per cell	Observed frequency	Expected frequency					
		GPD	COM-Poisson	$GIT_{3,1}$		$GIT_{3,1}$ pgf-based	
		MLE	MLE	MLE	MHD	T_1	T_2
0	159	159.05	159.15	159.00	158.42	158.98	158.98
1	64	65.23	63.11	64.32	63.07	64.08	64.10
2	13	13.58	14.80	13.24	14.31	13.40	13.39
3	4	2.14	2.95	3.43	4.20	3.54	3.54
Total	240	240	240	240	240	240	240
χ^2		1.67	0.60	0.10	0.15	0.07	0.07
P-value		0.20	0.44	0.75	0.70	0.79	0.79
Estimated parameters		$\hat{\theta}$=0.411	$\hat{\lambda}$=0.396	n=1	n=1	n=1	n=1
		$\hat{\lambda}$=0.032	$\hat{v}$=0.758	$\hat{p}_2$=0.132	$\hat{p}_2$=0.113	$\hat{p}_2$=0.1286	$\hat{p}_2$=0.1287
				$\hat{p}_3$=0.206	$\hat{p}_3$=0.227	$\hat{p}_3$=0.2090	$\hat{p}_3$=0.2089
Mean	0.43						
Variance	0.45						
ID	1.07						
LR test	Do not reject H_0 with LR= 0.81						

In Table 3.9, the index of dispersion is very close to 1. This is consistent with the LRT which does not reject H_0. The fit by the $GIT_{3,1}$ distribution is significantly better than the GPD and COM-Poisson distributions based upon the chi-square values.

Table 3.10 displayed another data set with equi-dispersion but has a very short tail. The fits by the $GIT_{3,1}$ distribution and COM-Poisson distributions are comparable and better than the GPD distribution. The LR test concludes that H_0 is not rejected with the calculated ID value quite close to one.

It is to be observed that the fits of the $GIT_{3,1}$ distribution from the pgf-based estimation method are consistently better than those by MLE or MHD.