Properties and Applications of a New Discrete Probability Distribution for Survival Data

David Hitchcock

A paper submitted to the
Department of Mathematical Sciences
of Clemson University

in partial fulfillment of the
requirements for the degree of

Master of Science

Mathematical Sciences

Approved: _____
Major Advisor

May, 1999

Acknowledgements

I would like to thank my advisor, Dr. K.B. Kulasekera, for his great help and encouragement as I worked on this project. I also thank my committee members, Dr. Peter Kiessler and Dr. Herman Senter, for their support and advice. I have truly enjoyed working with all three of these professors in classes and as part of this project.

I would also like to thank my family for supporting and inspiring me both in school and in life. I also thank my friends and fellow math graduate students at Clemson, who have made getting a master's degree here a super experience.

Table of Contents

PROPERTIES AND APPLICATIONS OF A NEW DISCRETE PROBABILITY DISTRIBUTION FOR SURVIVAL DATA

1. Introduction / Background Information

Modeling survival data with probability distributions is a common statistical problem, and many distributions are suitable for this kind of data. The major feature of survival data is that the survival time (i.e., lifetime) of an individual or item is a random variable that can take on only nonnegative values. Actually, survival times are often only measured discretely (e.g., to the nearest year, month, etc.), because of limitations in the process of collecting data. This is typical in studies conducted outside a laboratory setting. For example, in studies of wildlife survival times, it is usually impossible to report lifetimes except on a discrete scale.

For instance, Caughley (1966) gathered data on the lifelengths of the Himalayan thar, a goatlike mammal that lives in Asia and New Zealand. The lifelengths of 623 of these creatures were measured in this experiment. As is typical in many wildlife studies, the lifetimes were reported only to the nearest year, possibly because of the difficulties in measuring the lifetimes. In the case of the thar, its locale is so remote that scientists probably cannot study the animal's exact survival pattern; they may have been forced to assess lifelengths by looking at evidence such as skull size.

Even when the survival data involves less exotic wildlife, the data may be measured and reported on a more precise discrete scale. For example, survival tables of the snowshoe hare were reported in which the lifelengths were measured to the nearest half-year by Krebs (1989). If the species involved is a plant or very small animal, the lifetimes may be given in terms of the number of months, days, or even hours. The lifelengths need not be integer values; they may (as with the snowshoe hare data) be given as $k = 0, 0.5, 1, 1.5, \ldots$, or as any other discrete set. If the lifetime values k are evenly spaced, they may be transformed into a new variable k' which is defined over 1, 2, ... Also, the lifetime may merely be reported to have fallen within some interval, say, 2-3 years. In any case, one needs a distribution which is defined discretely to properly model these situations.

In survival models, the lifetimes are random variables; that is, they take on possible values according to some probability structure. (In general, lifetimes can be continuous or discrete random variables; this report focuses on the case in which the lifetimes are discrete.) Since the survival times are random, we must model them with a probability distribution. Such a model can specify the probability that the lifetime random variable will fall within a specified interval of values. Distributions commonly used to model continuously measured lifetimes include the exponential, Weibull, Gamma, or log-normal (see Lawless, 1982).

Let T represent the lifetime variable and let t represent any specific value it may take on. Of particular interest to statisticians studying lifetime data is the survivor function S(t). The survivor function gives the probability of the lifetime being greater than t (that is, the probability of an individual surviving past time t). Therefore

$$S(t) = P(T > t) = 1 - F(t), \tag{1}$$

where F(t) is the cumulative distribution function of the probability distribution. Whether T is continuous or discrete, S(t) is nonincreasing, with $S(0) = 1$ and $\lim_{t\to\infty} S(t) \equiv S(\infty) = 0$ (Lawless, 1982). That is, the individual is certain not to die before time 0 and cannot live forever.

Another important function in survival analysis is the hazard function h(t). The hazard function measures the instantaneous rate of failure at time t given that the individual has survived until t (Lawless, 1982). For continuous T with density function f(t),

$$h(t) = f(t)/S(t). \tag{2}$$

For discrete T with mass function p(t), h(t) is defined as

$$h(t) = p(t)/S(t). \tag{3}$$

The shape of the hazard function (increasing, decreasing, constant, or a combination or these) is often an important distinguishing characteristic of populations. Items from populations with a constant hazard function have the same instantaneous risk of failure no matter how long they have lived. An increasing hazard function corresponds to items whose rate of instantaneous failure increases with age, for example, automobiles, which

wear out as they age. A decreasing hazard function would correspond to an item that improves with age. Another important hazard function is the "bathtub-shaped" or U-shaped hazard. It may be used to chart human populations, whose members have a high instantaneous rate of death as infants, a low death rate throughout middle age, and a high death rate when they reach old age (Lawless, 1982).

The hazard rate or failure rate is simply the value of the hazard function at a specific time t. While the hazard rate gives information about the instantaneous behavior of a system, sometimes the "long-run" behavior or "typical" behavior of a system may be of interest. Indices reflecting long-run behavior are usually defined by meaningful averages or quantiles.

An important measure of this type is the mean residual lifetime. Given an individual who has survived until time t, the mean residual lifetime is how much longer we would expect that individual to live. Formally, the mean residual life function m(t) is defined as

$$m(t) = E(T - t \mid T \geq t). \tag{4}$$

Another is the median survival time. This is the time for which the probability of an individual surviving this long is exactly .50. In the context of a group of identically distributed observations, the median survival time is the time at which we would expect exactly 50 percent of the group to be still living.

If we have a model that properly describes the behavior of a population of survival times, it is often straightforward to obtain expressions or values for any of these measures. Therefore a good probability model is crucial to studying lifetime data.

Several probability distributions have traditionally been used to model discretely measured survival data. Depending on certain characteristics of the data, the binomial, the Poisson, and the negative binomial distributions have been used more than others (geometric, discrete uniform). Generally, the binomial distribution is used to model data whose mean exceeds the variance; the Poisson is used when the mean equals the variance; and the negative binomial is used when the mean is less than the variance. This report examines the performance in modeling survival data using a new discrete probability distribution:

$$p(k) = \frac{k^{\alpha} q^{k}}{b} \qquad k = 1, 2, 3, \ldots \qquad (5)$$

where $-\infty < \alpha < \infty$, $0 < q < 1$, and $b = \sum_k k^{\alpha} q^{k} > 0$.

This distribution was introduced by Kulasekera and Tonkyn (1992). The variable k in this model represents the (discretely measured) lifelength, while p(k) represents the probability that an observation from a population following this distribution has lifelength k. The distribution has two free parameters, α, a shape parameter, and q, a scale parameter. The third parameter, b, is the normalizing constant. For fixed α, increasing q rescales the distribution, shifting its peak to the right, although the left tail of the mass function always stays anchored at k = 1 (see Figures 1a-1e). For fixed q, increasing α makes the distribution appear more peaked (see Figures 1f-1i). When $\alpha < 1$, the mass function tends to decrease monotonically in k. When $\alpha > 1$, it tends to increase to a peak (more gradually for large α), then decrease in k. (This is not a hard and fast threshold point, however; if α is slightly less than 1, the mass function may rise almost immediately to a peak and then decrease.) For instance, a histogram of the thar survival data shows the frequencies rising quickly to a maximum at k = 3, and then decreasing fairly consistently with k. Hence we would expect the model fitted to this data set to have α near 1. An estimate of α for this example indicates this is reasonable.

Several relationships to other distributions are given by Kulasekera and Tonkyn (1992). For fixed α, it is in the family of power series distributions. When we rewrite the probability distribution as

$$\frac{p(k+1)}{p(k)} = \left(\frac{k+1}{k}\right)^{\alpha} q, \qquad k = 1, 2, \ldots, \; -\infty < \alpha < \infty, \; 0 < q < 1 \qquad (6)$$

we see that when $\alpha = 1$, the distribution is the shifted negative binomial (q, m = 2); when $\alpha = 0$, it becomes the geometric (m = 1) (see Johnson and Kotz, 1969). When $\alpha = -1$, it

Figure 1a: alpha = -0.5, q = 0.6

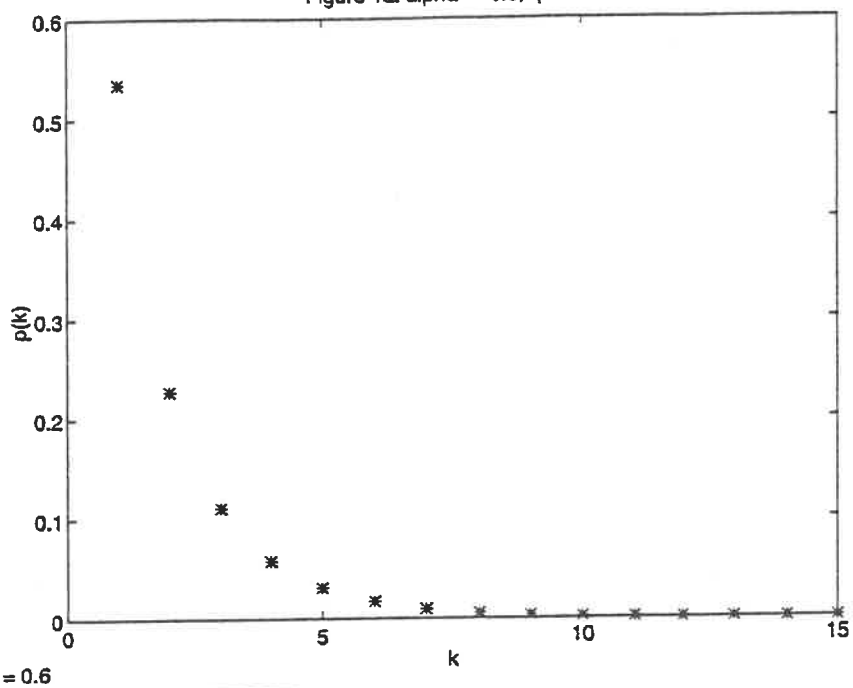


Figure 1b: alpha = 0.5, q = 0.6

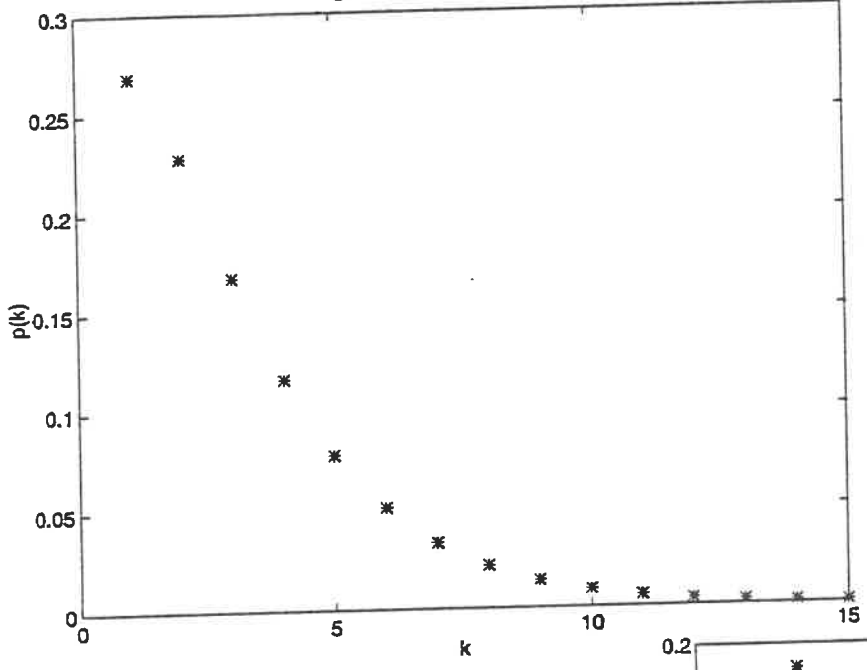
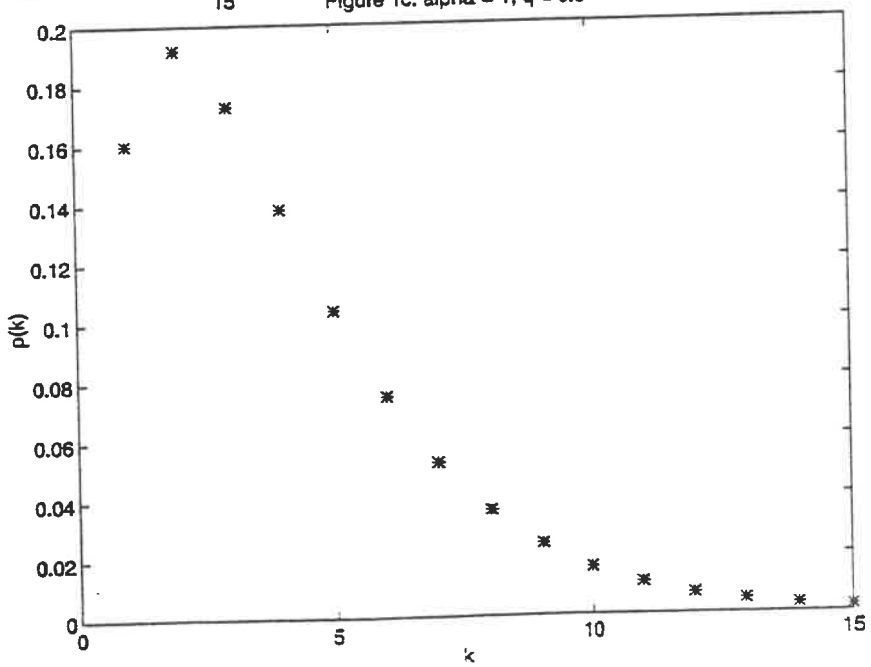


Figure 1c: alpha = 1, q = 0.6



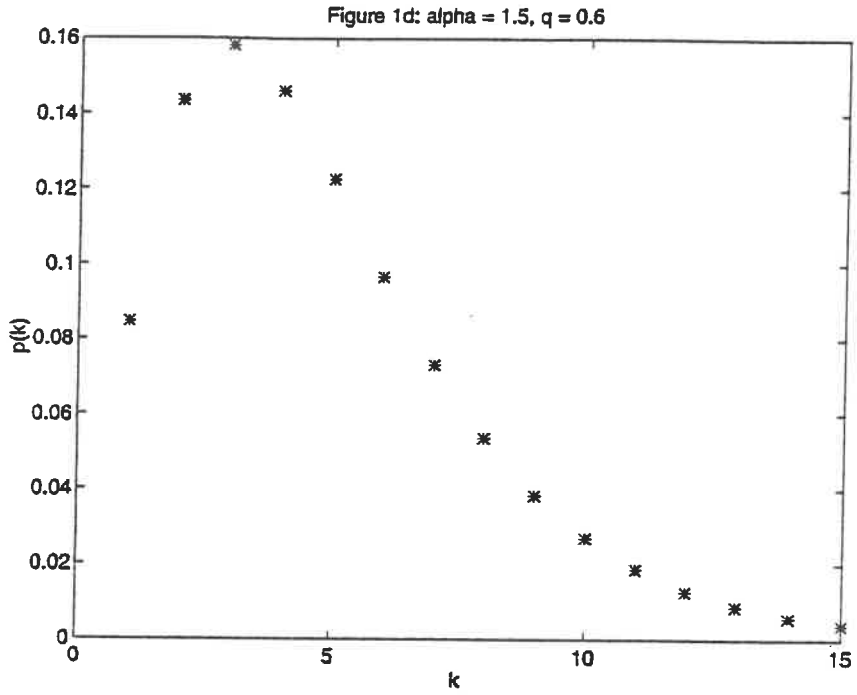


Figure 1d: alpha = 1.5, q = 0.6
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
p(k)
0
5
10
15
k

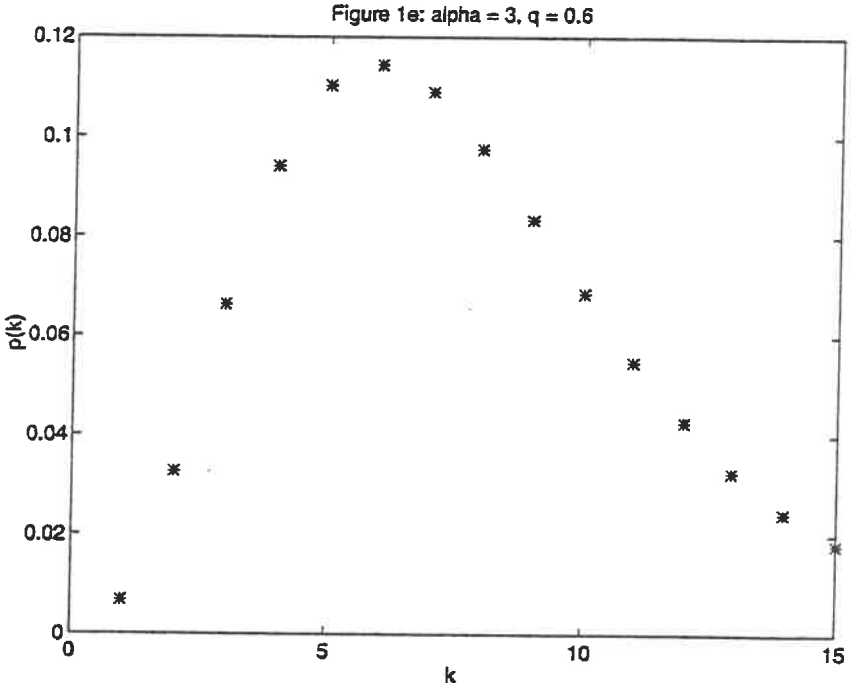


Figure 1e: alpha = 3, q = 0.6
0.12
0.1
0.08
0.06
0.04
0.02
0
p(k)
0
5
10
15
k

Figure 1f: alpha = 2, q = 0.2

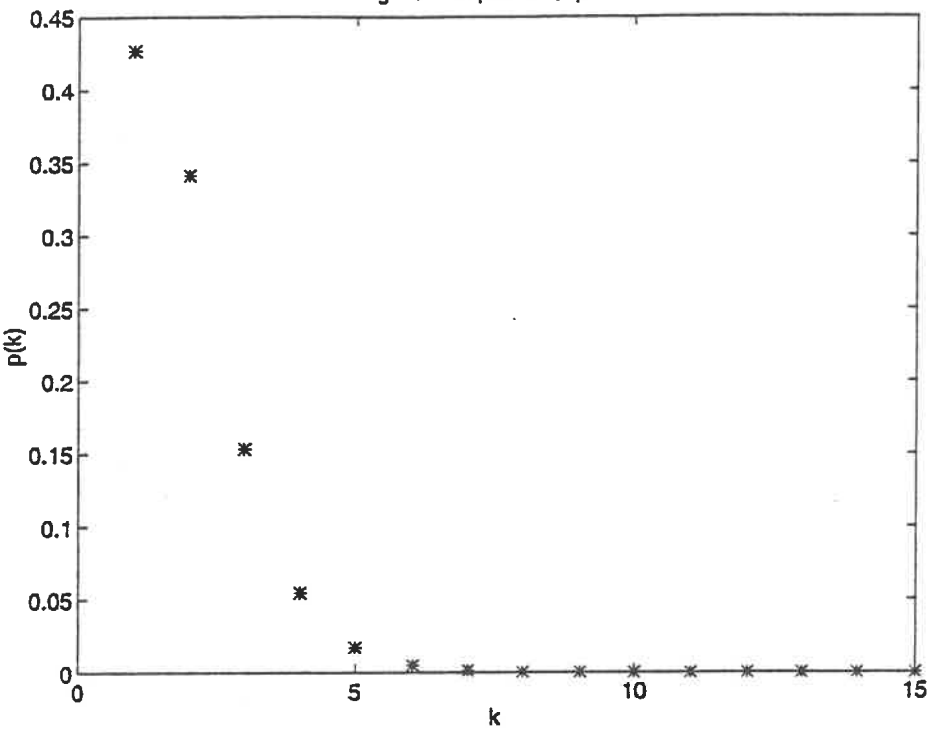


Figure 1g: alpha = 2, q = 0.4

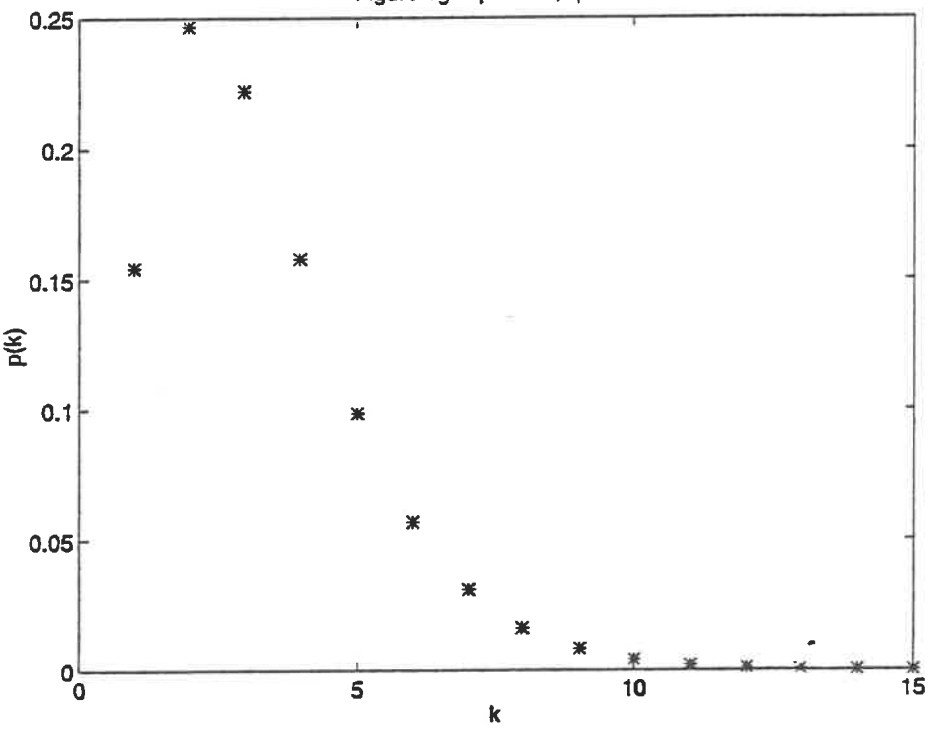


Figure 1h: alpha = 2, q = 0.6

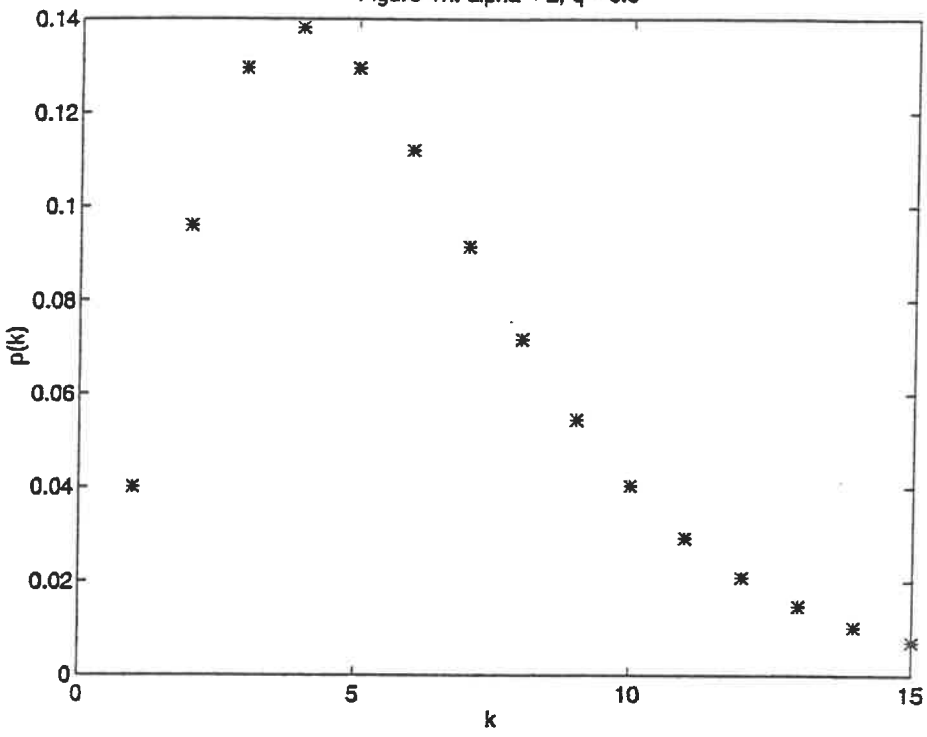
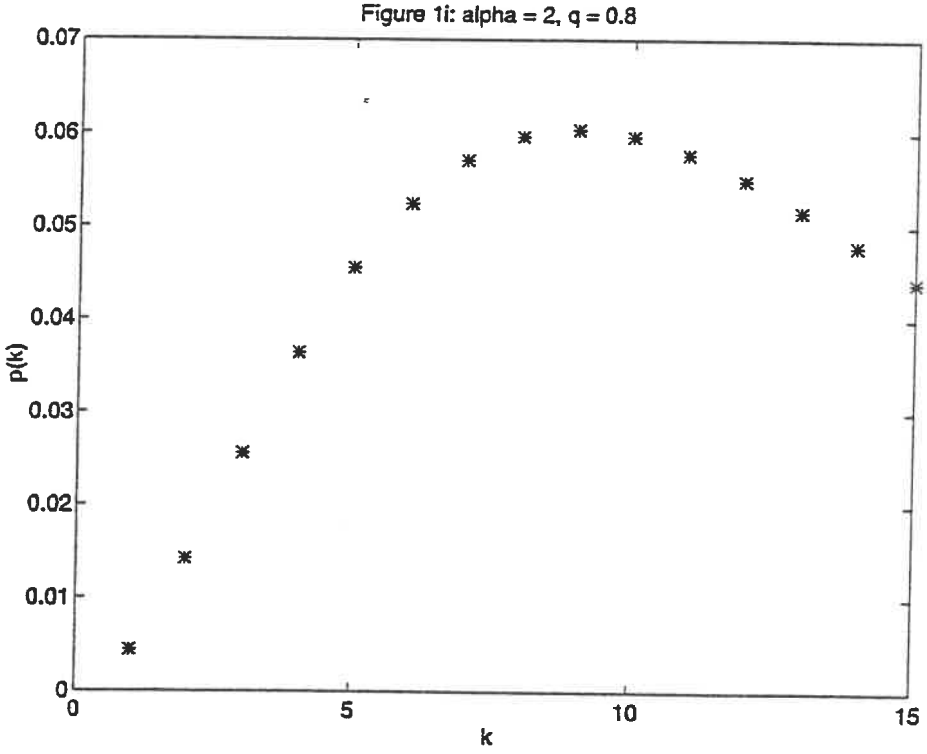


Figure 1i: alpha = 2, q = 0.8



forms the logarithmic series distribution, and when $q = 1$ and $\alpha < -1$, it becomes the discrete Pareto (zeta) distribution.

Also, this distribution has been shown to be a special case of the Lerch distribution (see Zornig and Altman (1995), Doray and Luong, 1997) which has the mass function

$$p(k) = \frac{q^k}{[\Phi(q, b, c)](b + k)^c}, \qquad k = 1, 2, \ldots, b \neq -1, -2, \ldots$$

with $b = 0$ and $c = -\alpha$. Here, $\Phi(q, b, c)$ is the Lerch zeta function (see Gradshetyn and Ryzhik (1980), p. 1075) defined by $\Phi(q, b, c) = \Sigma_i \; q^i/(b + i)^c$, $b \neq -1, -2, \ldots$

The distribution can be thought of as a generalization of the geometric model, which is used to describe territorial dispersal distances in which the dispersal situation has a constant hazard function (see Miller and Carroll, 1989; Tonkyn and Plissner, 1991). In some sense this is analogous to the generalization of the exponential distribution to the gamma distribution in the continuous case. In fact, the motivation for introducing this distribution has been this exponential-gamma type of connection. It been been shown by Kulasekera and Tonkyn (1992) that this model allows for situations in which the hazard function is monotonically increasing ($\alpha > 1$), constant ($\alpha = 1$), or monotonically decreasing ($\alpha < 1$).

Of course, the distribution is not limited to use in dispersal models. It can generally be applied to model survival data or failure time data which is measured discretely. In terms of dispersal models, the crucial property of this distribution is that it can model data whose mean is greater than, less than, or equal to the variance. Ecologists characterize a population as overdispersed or underdispersed depending on whether the mean is less than the variance or greater than the variance.

Typically in ecological studies, discrete survival data sets whose mean exceeds the variance are modeled with the binomial; those whose mean is equal to the variance are modeled with the Poisson; and those whose mean is less than the variance are modeled with the negative binomial. Since this new distribution is capable of modeling all three types of data, it is more flexible than any of the other three. Hence, presumably one could use this single distribution to model related data sets in which the mean may exceed, equal, or be less than the variance. This property of versatility is similar to that

of the generalized Poisson (Consul, 1989). In other words, the distribution can describe lifetime data or other count data which is overdispersed or underdispersed (Kulasekera and Tonkyn, 1992).

There are no general closed form expressions for the mean and variance of the distribution. Kulasekera and Tonkyn (1992) gave convenient analytical expressions in terms of the polylogarithm function for when α is an integer. However, the formula for the r-th moment about the origin (Kulasekera and Tonkyn, 1992)

$$E(X^r) = \mu_r' = \frac{b(\alpha + r, q)}{b(\alpha, q)} \tag{8}$$

for positive or negative r yields expressions for the mean and variance in terms of α and q.

$$\mu = \frac{b(\alpha + 1, q)}{b(\alpha, q)} \tag{9}$$

$$\sigma^2 = \frac{b(\alpha + 2, q)}{b(\alpha, q)} - \left(\frac{b(\alpha + 1, q)}{b(\alpha, q)}\right)^2 \tag{10}$$

Given any parameter pair (α, q), one can numerically calculate the mean and variance of the distribution fairly easily using these infinite sums. In fact, since the sums typically converge quickly, one can use partial sums in place of the infinite sums and find the mean and variance with any reasonable degree of accuracy.

The knowledge of the properties of the population parameters is critical in accurate modeling. In almost all applications, one must estimate the parameter values from observed data. There are at least two important reasons why the estimation techniques must produce good estimates. The first quality one hopes for is that the estimated model parameters can be used to predict probabilities (and other measures) well. Having found a fitted model, we would hope that we could use the model to estimate, say, the median survival time of another (very similar) population, or perhaps to compare characteristics

with a similar population. Since the true parameter values are unknown, the best we can do is to hope that our sample is representative of the population it came from and then choose a parameter estimation technique that will yield accurate estimates.

Another use of the estimated model is for testing hypotheses about the broad nature of the population. We hope that the parameter estimates can tell us something about the type of population the data come from. For example, if we fit this discrete distribution to a set of data and obtained a point estimate of α which (along with the variability of the estimate) indicated that the true value of α was significantly different from zero, we could reject the hypothesis that the data in fact come from a geometric distribution. This type of inference can be very important in applications. Thus accurate and precise estimates which allow us to make correct inferences are highly desirable. The search for a method that yields good parameter estimates for this model is one of the issues considered in this report. In the subsequent sections, we examine:

- method of parameter estimation
- properties of the parameters of the model
- performance of the distribution compared with the traditional distributions
- advantages of using this distribution over the others

We begin with the estimation problem.

2. Method of Parameter Estimation

Estimating the parameters of this model presents an obstacle in working with this distribution. One must estimate the two parameters α and q. (The parameter b is merely a normalizing constant which is a function of α and q; thus an estimate of b can be obtained using the estimates of α and q.)

The maximum likelihood estimators of α and q are sometimes difficult to obtain; the MLE's must be found numerically, and the computations are difficult. Also, we will show that there may be parameter combinations that are indistinguishable with respect to

the maximum log-likelihood. That is, the likelihood surface (for a given sample) may not have a unique or obvious maximum; it may reach a plateau instead, hindering the chance that the maximum likelihood method will numerically converge to the true parameter values. This property of indistinguishable distributions (which Wallenius and Korkotsides (1990) studied for the three-parameter Weibull distribution) will be discussed further later in this paper.

The sample arithmetic mean $\overline{X}$ and the sample geometric mean $\tilde{X}$ are jointly sufficient statistics for α and q, as shown by Doray and Luong (1997). Further, they are jointly complete sufficient statistics, since the new distribution is a member of the exponential family. They also showed that the MLE's are the solution (α, q) to the following system of two equations:

$$\overline{X} = \frac{\Phi(q, 0, -(\alpha+1))}{\Phi(q, 0, -\alpha)} \tag{11}$$

$$\ln(\tilde{X}) = \frac{\Sigma\, k^{\alpha} q^{k} \ln(k)}{\Phi(q, 0, -\alpha)} \tag{12}$$

where $\Phi(q, b, c)$ is the Lerch zeta function (see Gradshetyn and Ryzhik (1980), p. 1075) defined by $\Phi(q, b, c) = \Sigma\, q^{i}/(b+i)^{c}$, $b \neq -1, -2, \ldots$

This system must be solved numerically, and even with a good computer package, well chosen initial values are essential to obtain the correct solution (even if the likelihood function has a unique maximum).

Other methods of estimation have been proposed, some simple and some fairly involved. Kulasekera and Tonkyn proposed several ad hoc methods. One of these is a least squares method based on the transformed variables

$$X_k = \ln\left(\frac{k+1}{k}\right) \quad \text{and} \quad Y_k \approx \ln\left[\frac{f(k+1)}{f(k)}\right] \tag{13}$$

where f(k) is the observed relative frequency for a lifelength. After a logarithmic transformation of (6), one gets the relation:

$$Y_k = \alpha X_k + \ln(q) \tag{14}$$

Therefore a least squares regression of Y on X yields estimates for α and ln(q). Another method also uses the above relationship between Y and X; one writes the equation for two successive sample values, i and j, of k, and solves the resulting pair of equations for α and ln(q). This method has the disadvantage of using very little of the sample to estimate the parameters. A third method involves proportions and uses (6) to obtain the relation:

$$\frac{p^2(k+1)}{p(k)\,p(k+2)} = \left(\frac{(k+1)^2}{k\,(k+2)}\right)^{\alpha} \tag{15}$$

Solving this equation for α gives

$$\alpha_k = (Y_{k+1} - Y_k)/(X_{k+1} - X_k) \tag{16}$$

One may find (m-2) separate estimates of α by letting k = 1, 2, ..., m-2, for some m, and these estimates can be averaged to find an overall estimate of α. From this estimate one can solve for q.

Doray and Luong (1997) gave a method of quadratic distance estimation (QDE), which is an iteratively reweighted least squares estimation. This method obtained an estimate of the parameter vector $\theta = (\ln(q), \alpha)'$ by minimizing the quadratic form

$$[Y - X\theta]'\Sigma^{-1}[Y - X\theta]$$

where $Y = (\ln(f_2/f_1), \ldots, \ln(f_{k+1}/f_k))'$, $X = \begin{bmatrix} 1 & \ldots & 1 \\ \ln 2 & \ldots & \ln((k+1)/k) \end{bmatrix}'$ (17)

f_i = the frequency for class i, i = 1, ..., k, and Σ = the covariance matrix for an error vector ε. The QDE is then

$$\tilde{\theta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y. \tag{18}$$

where $\Sigma =$

$$\frac{1}{n}\begin{pmatrix} \frac{p_1+p_2}{p_1p_2} & -\frac{1}{p_2} & 0 & 0 & \cdots & 0 \\ -\frac{1}{p_2} & \frac{p_2+p_3}{p_2p_3} & -\frac{1}{p_3} & 0 & \cdots & 0 \\ 0 & -\frac{1}{p_3} & \frac{p_3+p_4}{p_3p_4} & -\frac{1}{p_4} & \ddots & \vdots \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & -\frac{1}{p_{k-1}} & \frac{p_{k-1}+p_k}{p_{k-1}p_k} & -\frac{1}{p_k} \\ 0 & 0 & \cdots & 0 & -\frac{1}{p_k} & \frac{p_k+p_{k+1}}{p_kp_{k+1}} \end{pmatrix} \tag{18a}$$

A problem with these methods which use ratios of relative frequencies is that in practice, unless the data are unusually tidy, the ratios may not display any trend or relationship that can be exploited with regression. The method proposed and used in this paper is a least squares method based on the actual frequency counts, rather than the ratios of relative frequencies. The rationale behind the method is as follows:

The model to be examined is:

$$p(k) = \frac{k^{\alpha}q^k}{b} \tag{19}$$

Since $b = \Sigma\, k^{\alpha}q^k$ is a normalizing constant which makes the probabilities sum to one, we can rewrite the equation in terms of the frequencies:

$$freq(k) \approx c\, k^{\alpha}q^k \tag{20}$$

where c is a constant of proportionality and freq(k) are the observed frequencies of each lifetime k.

Therefore

$$\ln [freq(k)] \approx \ln(c\, k^{\alpha}q^k) \tag{21}$$

and

$$\ln [freq(k)] \approx \ln(c) + \alpha\ln(k) + k\ln(q) \tag{22}$$

So from a least squares regression of ln[freq(k)] on ln(k) and k, we obtain (from the regression coefficients) estimates of α and ln(q), and thus q. As an abbreviation, the estimates obtained by this process will be called the log-frequency least squares (LFLS) estimates and the LFLS estimates will be denoted α^* and q^*. Both Kulasekera and Tonkyn's least squares method and the QDE method use simple regression, while the LFLS uses a multiple regression.

Since it is ln[freq(k)] that is the dependent variable in the regression, the direct antilog of [ln(q)]* may yield a slightly biased estimate of q. To improve the accuracy of the estimate, one may let $q^* = \exp([\ln(q)]^* + MSE/2)$, where MSE is the mean squared error of the regression. Kmenta (1986) reported this adjustment in the context of a transformed ordinary linear regression model, where the transformed model has the typical error assumptions as in the classical linear regression model. The adjustment may be useful here as a remedial measure, although we must exercise caution in using it. When the MSE is large or when [ln(q)]* is near 0, it is not advisable to use this adjustment, since it may yield a $q^* > 1$, as the parameter space for q is (0,1).

Once we have estimates for α and q, we can use these to get a value for b*, since $b = \Sigma k^{\alpha} q^k$. In practice, we merely use a partial sum of this series to get b* since the series converges quickly. Summing the first 500 terms of the series is virtually always sufficient to give an extremely accurate answer; a computer package like Maple can calculate this partial sum in a few seconds.

Since the distribution is defined for k = 1, 2, 3, ..., it may be necessary to align the lifetimes in the data set to put them in this form. For example, if the given lifetimes are 0 years, 1 year, 2 years,..., then one should use k + 1 in place of k.

This method has the advantage of using the frequencies directly rather than ratios of successive relative frequencies, which may not perform well when any successive relative frequencies are fairly far apart or when the number of lifetimes is small. The least squares method used here typically yields an estimated model that fits the data well, which is understandable given the nature of the least squares method. Another advantage of this method is its computational simplicity compared to finding the MLE's.

The log transformation encounters a problem when some lifetimes in the data have frequency zero. This problem can be partially overcome by an iterative method that adjusts the problematic frequency and refits the model. The model would first be fit with missing values in place of the zero-frequency classes and a temporary estimated model obtained. This model would be used to get an expected frequency for the lifetime, and the model would be fit again, with the expected frequency in place of the observed frequency 0. It is important to note that this procedure may not improve the accuracy if the observed frequency of zero occurs in the extreme tails of the distribution, where the expected frequencies will be very close to zero anyway. If the frequency of zero does not occur in the tails, this iterative method enables one to use the information in that point to estimate the model rather than deleting the point completely.

This method can be illustrated with the following data, in which the "lifetime variable" k is actually the number of sowbugs found under boards, along with the observed number of boards containing k sowbugs (original data from Janardan, et al., 1979). (This data set has observed frequencies of zero for k = 12 and k = 16 sowbugs.) These values are presented in a table with the expected frequencies from the model fitted by various types of parameter estimation. The three estimation methods used are:

(1) the LFLS method with the "zero frequency classes" k = 12 and k = 16 replaced in the manner described above

(2) maximum likelihood estimation (MLE) (expected frequencies taken from Doray and Luong, 1997)

(3) quadratic distance estimation (QDE) (Doray and Luong) in which the "zero classes" are deleted

Table 1. Different estimation techniques

Estimates for:	α	q
LFLS	-0.4798	0.803916
MLE	-0.355221	0.819788
QDE	-0.743481	0.9117893

Table 2. Fit of this distribution to sowbug data (with different estimates)

k	obs(k)	LFLS e(k)	MLE e(k)	QDE e(k)
1	28	31.189	26.01	26.87
2	14	17.980	16.67	14.63
3	11	11.899	11.83	9.87
4	8	8.332	8.76	7.27
5	11	6.018	6.63	5.61
6	2	4.433	5.10	4.47
7	3	3.310	3.95	3.63
8	3	2.496	3.09	3.00
9	3	1.896	2.43	2.51
10	3	1.449	1.92	2.11
11	2	1.113	1.52	1.79
12	0	0.858	1.21	1.53
13	1	0.664	0.96	1.32
14	2	0.515	0.77	1.14
15	1	0.401	0.62	0.98
16	0	0.312	0.49	0.86
17	2	0.244	0.40	0.75
18+	0	0.891	1.64	5.65

Pearson's Chi-square goodness-of-fit tests were performed for each fitted model, with classes chosen so that each class had an expected frequency greater than 2.

Table 3. Goodness-of-fit results for sowbug data (different estimates)

	LFLS	MLE	QDE
Classes (cells)	k = 1, 2,..., 8, 9, 11, 12+	k = 1, 2,..., 9, 10-12, 13+	k = 1, 2,..., 10, 14-17, 18+
χ^2 statistic	10.84	6.12	14.06
Critical point	$\chi^2(7,.05) = 14.07$	$\chi^2(8,.05) = 15.51$	$\chi^2(9,.05) = 16.92$
p-value	.14575	.63379	.12021

If we assume the data come from this distribution, the results of this test (based on the p-values) tell us that the MLE's yield the smallest normalized discrepancy between the observed frequencies and the corresponding expected frequencies. Hence we may, in a

sense, conclude that the MLE's perform best in this case. This rationale is somewhat similar to the minimum distance estimation method in which we would select the estimator by minimizing a distance measure with respect to a parameter of a specified distribution. If we take the p-values as an index of estimator performance, the LFLS estimates, obtained by adjusting the zero-frequency classes, perform better than the QDE estimates, obtained after simply deleting the two classes which have observed frequencies of zero. All three methods give a χ^2 statistic better than the critical point at the 5% significance level. It should be noted that using the LFLS estimates having deleted the zero-frequency classes would yield a poor fit in this example; the fitted mass function would contain too much probability in the right-hand tail (k = 18+). So the method of accounting for the classes with observed frequency zero is clearly beneficial in this case.

An unusual property of the parameters

To test the performance of the LFLS estimation technique, a few random samples were generated from the discrete distribution with a "true" parameter pair (α_0, q_0). The goal was that if many such samples were generated, and the distribution was fitted to each model, in the long run, the average of the estimates of the two parameters would be very close to the true parameters from which the data were generated. As it turned out, this did not happen, and the results revealed an interesting property of the distribution and its parameters.

Fitting the model to the simulated sample data actually yielded parameter estimates that were quite different than the original pair, yet the probabilities calculated with them were very close to the probabilities calculated with the original pair. Experimentation revealed that there exist a variety of very different parameter pairs (α, q) which yield nearly indistinguishable probability distributions. This complicates the problem of parameter estimation, since in fitting the distribution to sample data, there may be many different parameter pairs that give similarly good fits. Particularly, this would hinder the process of numerically obtaining the MLE's, since the maximum likelihood surface, given a sample, may not have a unique maximum. The likelihood surface may instead reach a kind of plateau, so that an iterative maximization method would produce a

different maximum depending on the initial values. These different maxima would correspond to the different parameter pairs that yield indistinguishable distributions.

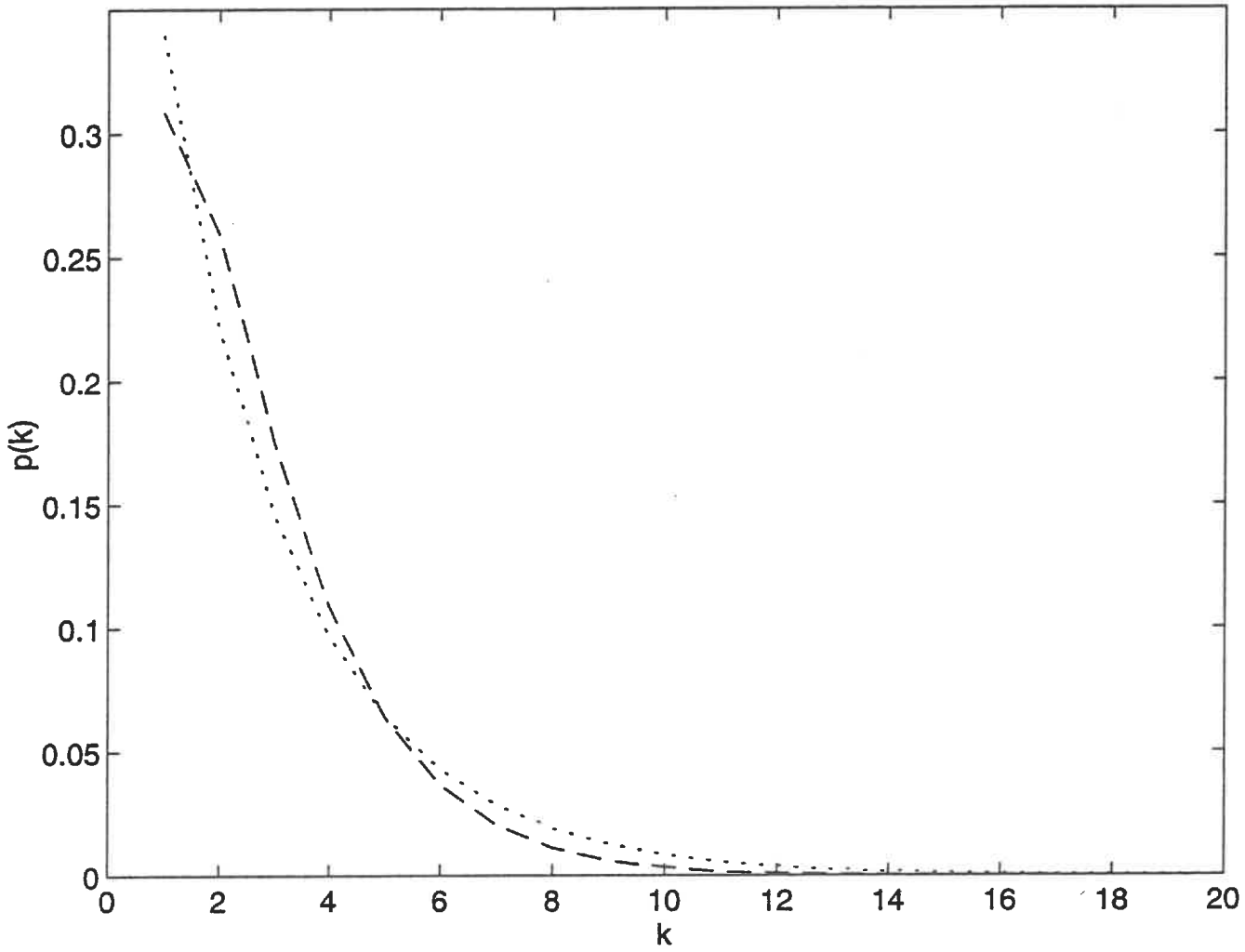
Wallenius and Korkotsides (1990) studied the indistinguishability property in the three-parameter Weibull distribution and noted these problems. In that model, as with this distribution, widely varying parameter values correspond to very similar probability distributions. Wallenius and Korkotsides (1990) noted the connection between the indistinguishability property and multicollinearity in linear regression. They also discussed the fact that this problem makes finding the MLE's more difficult, since an iterative maximization algorithm may stop short of finding the true maximum of the likelihood function. Zanakis (1979) noted that the likelihood surface may have a "long, flat ceiling," which inhibits finding the maximum computationally.

For example, one simulation started with a seed parameter pair $(\alpha_0, q_0) = (1.9, 0.5)$ and generated 150 random samples of data (each sample having 100 observations) from this "population". The LFLS estimation technique was then used on each sample and the 150 resulting estimates averaged to get an average (α^*, q^*). After several repetitions of simulating 150 samples and obtaining averaged estimates, the resulting averaged estimates were all fairly close to $(\alpha^*, q^*) = (1.2, 0.61)$: clearly different from the true pair. When the mass functions corresponding to $(\alpha = 1.9, q = 0.5)$ and corresponding to the estimated parameter pairs were plotted, it was apparent that the mass functions were very close to each other, almost identical. For the purposes of fitting the model to sample data, the distributions defined by these quite different parameter pairs are indistinguishable. Each would fit the sample data equally well.

Another instance of this began with the parameter pair $(\alpha_0 = 0.75, q_0 = 0.5)$. Generating sample data from this distribution and refitting the model yielded estimated parameter pairs near $(\alpha^*, q^*) = (-0.05, 0.67)$. Again, a plot of the mass functions corresponding to the seed parameter pair and the estimated pair (see Figure 2) shows that the mass functions are indistinguishable for the purposes of fitting the model to data.

These and many other similar cases led to the question of whether there were in fact sets of parameter pairs that yield "indistinguishable" distributions. If so, was it possible to describe a relationship between these pairs (α, q) that yielded distributions that looked almost the same? For example, could a linear equation $q = A\alpha + B$ (where A and B are

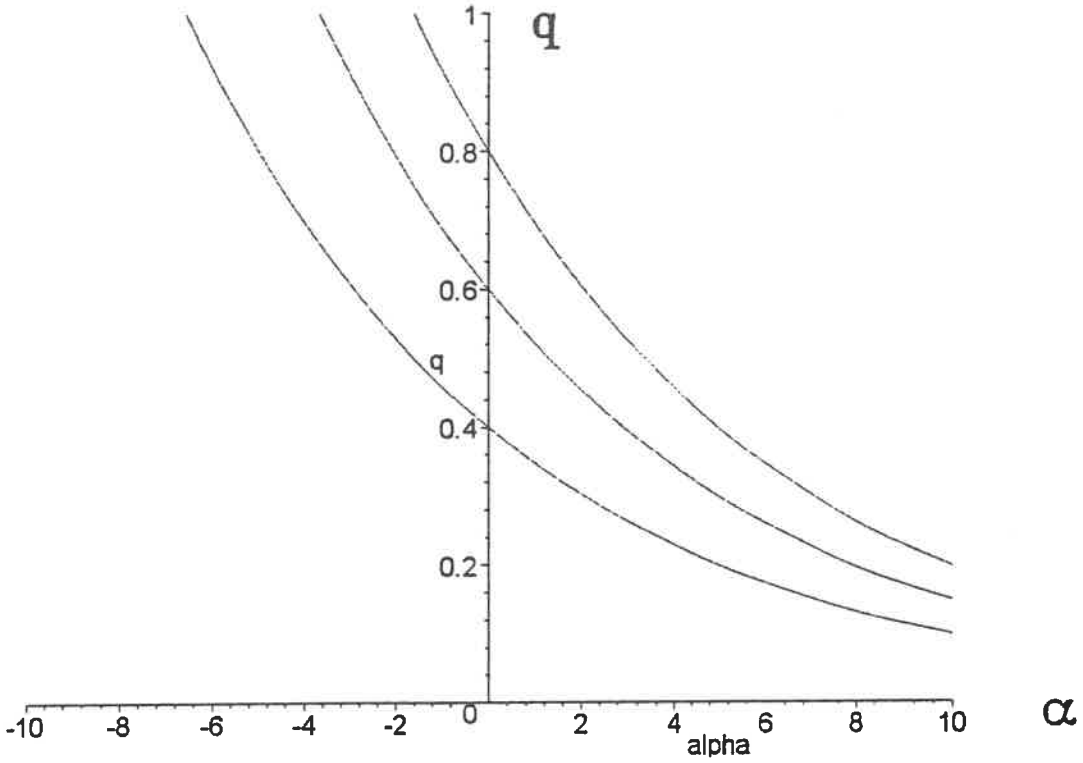
Figure 2: Two parameter pairs which yield nearly indistinguishable distributions



— — — $\alpha=0.75,\ q=0.5$

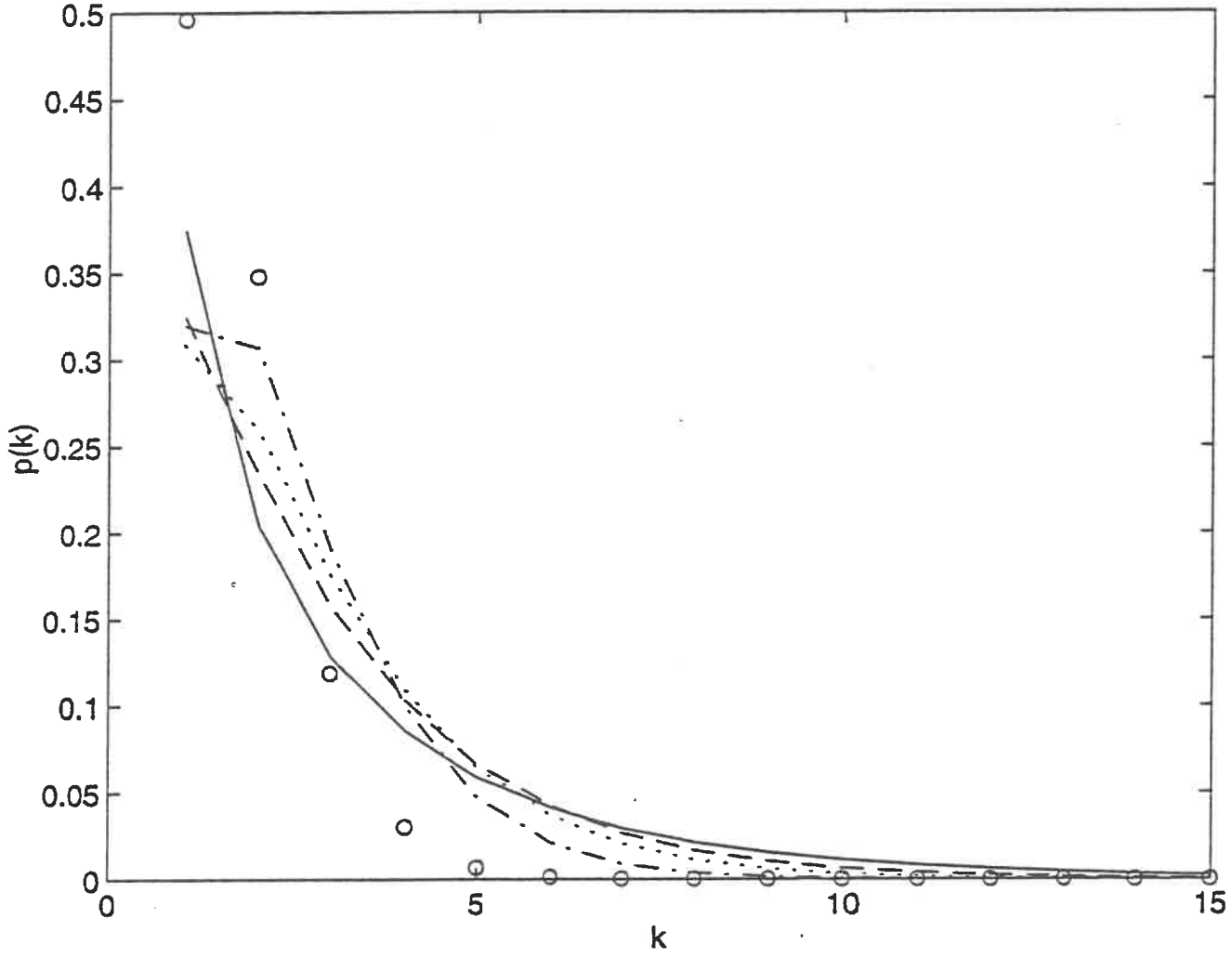
- - - - - - - $\alpha=-0.0473,\ q=0.6736$

Figure 3: possible indistinguishability regions



It is possible that curves something like these define "indistinguishability" regions, such that any choices of α and q that lie on the same curve will yield almost identical probability distributions.

Figure 4: Testing for a linear relationship between alpha and q



α = -0.5, q = 0.7695

α = 0.25, q = 0.6081

α = 0.75, q = 0.5

α = 1.5, q = 0.3391

α = 2.5, q = 0.1239

constants) describe a deterministic relationship such that any α and q satisfying this equation would yield nearly identical probability distributions? Or was there a nonlinear relationship between α and q that had the same effect? If so, would it be possible to create some sort of contour curves (see Figure 3), each of which defines an "indistinguishability region", such that any choices of α and q that lie on the curve will yield almost identical probability distributions?

For the most part, these questions remain unanswered, but after some experimentation on the topic, we are able to make some conjectures about this indistinguishability property of the distribution.

Hypothesizing that there was a linear relationship $q = A\alpha + B$ between the parameters, we can perform a linear regression, using as the data pairs several parameter pairs that gave indistinguishable distributions. Let the data pairs be the averaged estimates (α^*, q^*) obtained by repeating the simulation described above. For the case described above in which the seed parameter pair was $(\alpha_0 = 0.75, q_0 = 0.5)$, the regression equation was $q = -0.2152\alpha + .6619$. If the linear relationship was valid, then any reasonable pairs (α, q) that satisfied this equation should yield nearly identical distributions. Picking a variety of pairs that satisfied the equation and plotting the resulting distributions, we see that the distributions were not nearly identical (see Figure 4). Still, there was a definite pattern, a nonrandom progression, to their difference, indicating that there is in fact a relationship between the "indistinguishable pairs", but not a linear one. This is the conjecture of this report (possibly the true relationship is like the curves drawn in Figure 3), but any definite answer will require further study.

3. Modeling of Survival Data and Comparisons with other Distributions

This section discusses the fitting of the model to some actual data sets. Both real ecological survival data and some simulated data sets are modeled with this distribution. The primary purpose of the modeling is to determine how well the distribution models actual survival data. A secondary purpose is to compare this distribution with the three traditional distributions, the binomial, Poisson, and negative binomial.

Because the application of this distribution considered in this report is modeling ecological survival data, a great emphasis is put on the performance of the distribution in modeling real-life data as opposed to simulated data. However, simulated data supplemented the ecological data for a couple of reasons. Data sets whose mean was greater than, or equal to, the variance were relatively less numerous in the available ecological literature than data whose mean was less than the variance. Thus simulated data with these characteristics was generated so that the distribution's performance could be compared with the binomial and the Poisson. Also, using the simulated data afforded a measure of control over some aspects of the data sets which were useful in determining exactly how, and with exactly what type of data, the new distribution's performance compared with the traditional distributions' performance.

Comparing the new distribution with the binomial

The sample mean and sample variance of a moderate or large data set are typically good indicators of the population mean and population variance. Therefore in determining which model to use, ecologists have used sample values to pick a model.

To test the performance of the fitted models, Pearson's Chi-square goodness-of-fit tests were performed to examine the closeness of the expected frequencies to the observed frequencies. It should be noted that the Chi-square tests were performed on estimated models whose parameters may have been estimated differently. For example, the Poisson parameter was estimated by the sample mean (the MLE) while the new distribution's parameters were estimated with the LFLS method. When we compare the results of a test of the goodness-of-fit of the Poisson with a test of the goodness-of-fit of the new distribution, we should understand there is a difference in the estimation methods.

Also, when the parameters of the null distribution are estimated, Pearson's χ^2 statistic

$$\Sigma_i [(O_i - e_i)^2/e_i]$$

where O_i = observed frequency for class i and e_i = expected frequency for class i under the hypothesized distribution, has an asymptotically χ^2-distribution only if the estimators are asymptotically normal. We assume this property for the LFLS estimators. We treat the test statistic's distribution as if the χ^2 approximation is close enough to be adequate

for these tests, as is commonly done in practice (Gibbons and Chakraborti, 1992). Generally, the classes, or cells, of k-values were chosen so that the expected frequency for each class would be greater than two.

A data set involving the number of sperm on a sea urchin egg, which can be modeled with the binomial distribution, or with the new distribution, was given in Janardan, et al. (1979). The lifetime variable k represents the number of sperm on sea urchin eggs 40 seconds after initial contact with the egg, and the frequency variable represents the number of eggs containing zero sperm, one sperm, etc. It may be modeled with the binomial since the sample mean = 0.7625 and the sample variance = 0.4311. For the following data, and for any subsequent data in which the "lifetimes" are given as 0, 1, 2,..., we make a transformation to get the following form of the new distribution:

$$p(k) = \frac{(k+1)^{\alpha} q^{k+1}}{b}, \qquad k = 0, 1, 2, \ldots \tag{23}$$

The binomial distribution is defined for k = 0, 1, 2, ... , n by

$$p(k) = [n!/(k!(n-k)!)]\, p^k (1-p)^{n-k}$$

Table 4. Fit of the models to sea urchin egg data

k (sperm)	0	1	2	3+
obs(k)	28	44	7	1
LFLS e(k)				
($\alpha^* = 6.14913$, $q^* = 0.0185$)	30.2934	39.7660	8.9005	1.0402
binomial e(k)	33.184	33.936	11.568	1.312

Table 5. Goodness-of-fit results for sea urchin egg data

	New distribution	Binomial
χ^2 statistic	1.0318	5.6724
Critical point	$\chi^2(1,.05) = 3.84$	$\chi^2(2,.05) = 5.99$
p-value	.30974	.058648

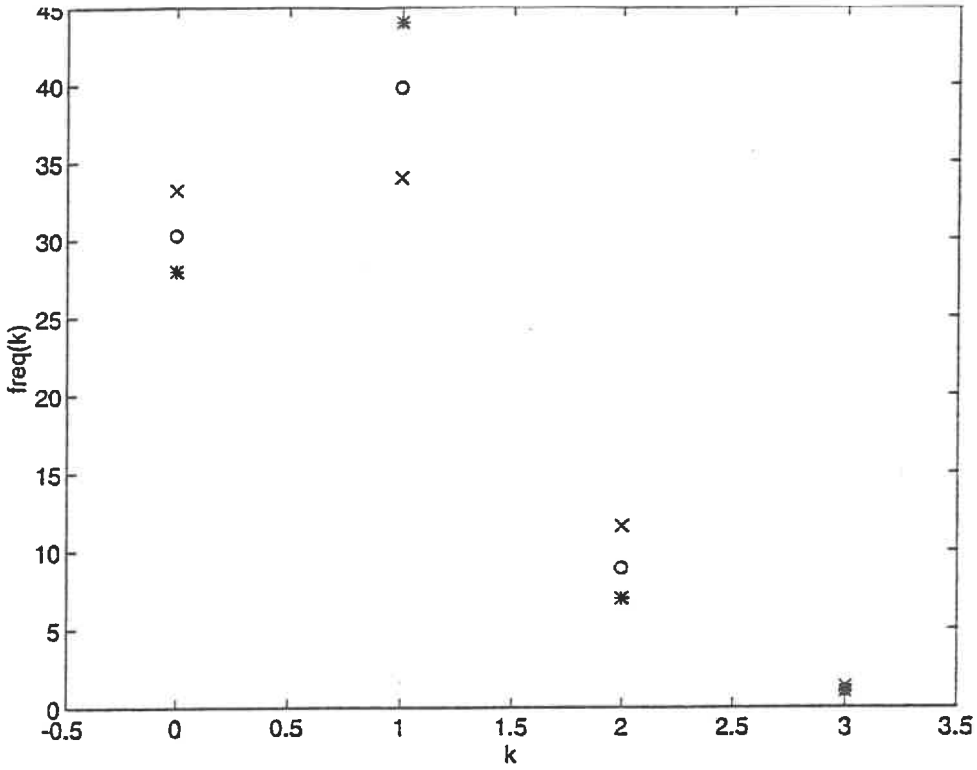
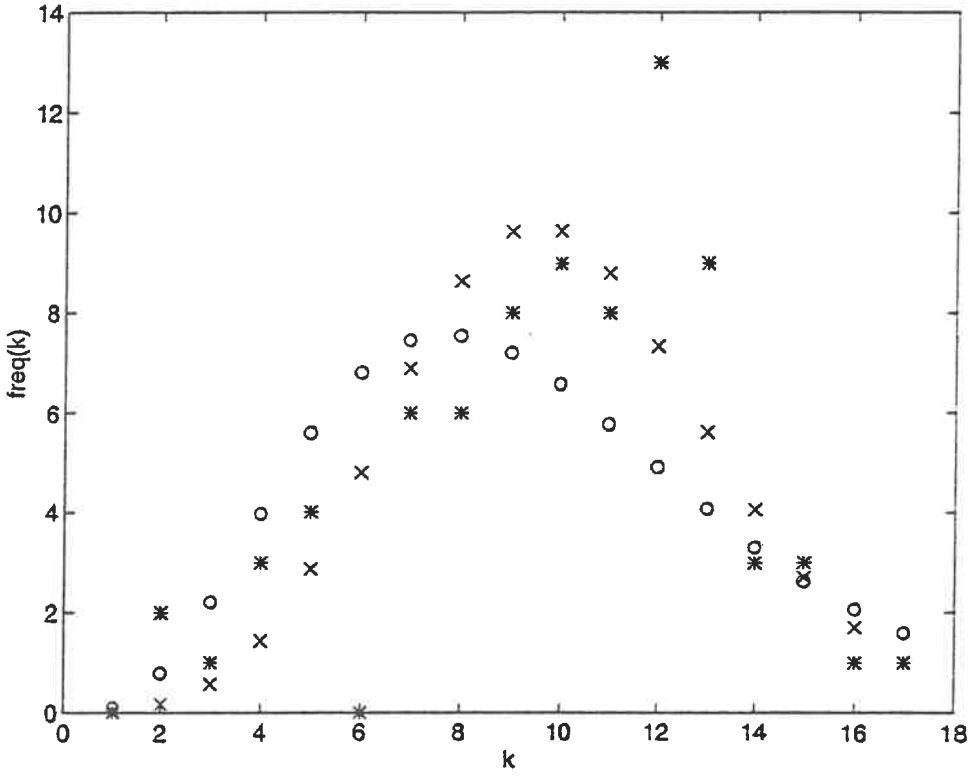


Figure 6: Modeling bighorn ram survival data



The goodness-of-fit test shows that both distributions give a good fit at the 5% significance level. In this example, however, the goodness-of-fit hypothesis is much closer to being rejected in the test about the binomial distribution, indicating that, in a sense, the new distribution gives a better fit to these data than the binomial (see Figure 5). In other words, based on the higher p-value of the test involving the new distribution, the goodness-of-fit of the binomial model would be rejected much more readily than the goodness-of-fit of the new distribution.

Janardan, et al. (1979) also collected data on the number of weevil eggs laid on beans, a data set which can be modeled as discrete lifetime data. The lifetime variable k represents the number of eggs (0, 1, 2, ...) while the frequency variable represents the number of beans containing zero eggs, one egg, etc. The sample mean of the data is 1.788, while the sample variance is 0.566, so we may model the data with the binomial distribution, as well as with the new discrete distribution.

Table 6. Fit of the models to weevil egg data

k (eggs)	0	1	2	3	4
obs(k)	5	68	88	32	0
LFLS e(k)	4.756	69.12	78.86	31.74	8.522
($\alpha^* = 8.8453$, $q^* = 0.031599$)					
binomial e(k)	18.92	59.58	70.33	36.90	7.261

Table 7. Goodness-of-fit results for weevil egg data

	New distribution	Binomial
χ^2 statistic	9.614	23.7863
Critical point	$\chi^2(2,.05) = 5.99$	$\chi^2(3,.05) = 7.81$
p-value	0.008172	0.000028

A Chi-square goodness-of-fit test shows that neither the new distribution nor the binomial distribution gives a good fit at the 5% significance level in this example. Closer inspection shows that the new distribution gives a relatively good fit to the data over most of the values of k (a better fit, in fact, than the binomial), but both models give a poor fit in the right-hand tail. The significant discrepancy between the observed frequencies and

the expected frequencies in the $k \geq 4$ class causes the overall goodness-of-fit hypothesis to be rejected in both models. The new distribution comes closer to giving a good fit than the binomial distribution. It should be noted that the number of classes in this goodness-of-fit test is somewhat smaller than is ideal, which inflates the importance of a single severe discrepancy between observed and expected frequencies of a class.

Comparing the new distribution with the Poisson

Janardan (1979) also reported lifetime data in which the lifetime variable k is the number of spiders under boards, with the observed numbers of boards containing k spiders. Since the sample mean and sample variance are approximately equal (mean = 0.425, variance = 0.453), we model the data with the Poisson distribution along with the new distribution.

Table 8. Fit of the models to spider data

k (spiders)	0	1	2	3+
obs(k)	159	64	13	4
LFLS e(k)	163.323	56.544	15.287	4.8449
($\alpha^* = 0.85954$, $q^* = .190806$)				
Poisson e(k)	156.905	66.684	14.1704	2.2403

Table 9. Goodness-of-fit results for spider data

	New distribution	Poisson
χ^2 statistic	1.5872	1.6149
Critical point	$\chi^2(1,.05) = 3.84$	$\chi^2(2,.05) = 5.99$
p-value	.20773	.44599

In this example, the Chi-square test reveals that both the new distribution and the Poisson distribution give good fits to the data at the 5% significance level. The Poisson model appears to fit slightly better.

An instructive ecological example is the survival data for the bighorn ram reported by Krebs (1989). The sample mean for this data set is 10.026 and the sample variance is 10.649, close enough that we may model the ram data with the Poisson, along with the new distribution. Lifetimes of the rams are given in years, with observed frequencies of rams having lifetimes k, for k = 1, 2, ...

Table 10. Fit of the models to bighorn ram data

k (years)	obs(k)	LFLS e(k) ($\alpha^* = 3.735$, $q^* = 0.61503$)	Poisson e(k)
1	0	0.0958	0.0342
2	2	0.7848	0.1712
3	1	2.1949	0.5721
4	3	3.9534	1.4340
5	4	5.5955	2.8755
6	0	6.7998	4.8050
7	6	7.4379	6.8821
8	6	7.5328	8.6250
9	8	7.1931	9.6082
10	9	6.5572	9.6331
11	8	5.7574	8.7802
12	13	4.9008	7.3358
13	9	4.0645	5.6125
14	3	3.2970	4.0516
15	3	2.6238	2.7081
16	1	2.0536	1.6970
17	1	1.5840	1.0008

Table 11. Goodness-of-fit results for bighorn ram data

	New distribution	Poisson
χ^2 statistic	29.92	19.90
Critical point	$\chi^2(11,.05) = 19.675$	$\chi^2(2,.05) = 19.675$
p-value	.001631	.046731

In this example, neither model gives a good fit at the 5% level (the Poisson comes close), but it is interesting to examine why this may be. For the new distribution, the chief culprit appears to be the unusually high observed frequency of 13 at the lifetime k = 12

(see Figure 6). Because of the fact that the LFLS method uses a log transformation of the frequencies, at such a large observation, there could be a large discrepancy between the observed and expected frequencies (as was the case here), aiding the rejection of the goodness-of-fit hypothesis. A likely reason that the expected frequency is not close to the observed value here is as follows. In a typical case of fitting a regression line to data by least squares, such an outlying value would pull the regression line toward the value unduly, because of the inherent nature of minimizing the sum of squared deviations. However, in the LFLS, the dependent variable is the logarithm of the frequencies, rather than the frequencies themselves. The log transformation tends to compress the values of the frequencies, and so the log-frequency at k = 12 is not so influential, not so relatively large in magnitude. Hence the fitted LFLS model is not pulled toward this observation as much.

We can assume that this fact holds in general: when data sets have a single unusually large frequency, the LFLS model will not be pulled close to this observation. There are negative and positive consequences of this fact. As seen is the above example, the large discrepancy between the observed and expected frequencies tends to drive up the χ^2 statistic, leading to probable rejection of the goodness-of-fit hypothesis. On the other hand, this means that the model is not overly sensitive to outlying cases, and that the model is not just fitting the noise in the data. For example, it could be that the large frequency of ram lifetimes at k = 12 is just due to the vagaries of this sample, and is not reflective of the population as a whole.

Goodness-of-fit analysis with simulated data

Since the available literature does not have extensive examples of survival data having mean greater than the variance, or mean equal to the variance, the ecological survival data was supplemented with some simulated data sets having these properties. The simulation procedure was done in Matlab. The procedure was basically to generate some data from the binomial mass function and from the Poisson mass function and to perturb the data with a normal error term to simulate the random noise one might encounter in real-life data. The characteristics of the perturbations were adjusted by adjusting the mean and the standard deviation of the error term. If one wanted data that were "close"

to say, purely Poisson data, the standard deviation of the normal error would be made small. Making the standard deviation of the errors large created data that deviated severely from, say, Poisson data. Giving the normal errors a positive mean created data that were somewhat skewed away from purely Poisson.

Of course, since the data sets analyzed are just perturbed binomial or Poisson data, comparing the performance of the new distribution to these traditional distributions in modeling this simulated data will reflect an advantage for the traditional distributions. We would certainly expect that the Poisson distribution would give a good fit to perturbed Poisson data, as long as the perturbations were not too large. In fact, the experiments revealed that the binomial and Poisson distributions generally fit their respective simulated data better than did the new distribution.

The goodness-of-fit results (obtained by Pearson's Chi-square test) are summarized in the following tables for perturbed data from the binomial distribution. For purposes of brevity, only the p-values of the goodness-of-fit test about the binomial, and the goodness-of-fit test about the new distribution are given. A low p-value leads to rejection of the goodness-of-fit hypothesis, so in a sense a higher p-value corresponds to a better fit. It is common to reject the hypothesis that the data come from the particular distribution if the p-value is less than .05. P-values that would lead us to conclude a good fit at the 5% level are in bold in the table.

In the tables, "skew" represents the mean of the normal perturbations and "sd" is proportional to the standard deviation of the perturbations, so that the greater "sd" is, the more the data deviates from purely binomial data. (The simulated data sets have around 300 observations; because of rounding off the frequencies, the number of observations varies slightly for each data set.)

Table 12. P-values for perturbed binomial data from a population with parameter p = 0.4

	skew = 0	skew = 0.1	skew = 0.25	skew = 0.49
sd = 0.1	**new distr. = .886** **binomial = .999**	**new distr. = .977** **binomial= .9999**	**new dist. = .989** **binomial= .9999**	**new dist. = .682** **binomial= .9997**
sd = 0.1	**new dist. = .76** **binomial = .996**			
sd = 0.25	new dist. = .031 **binomial = .132**			
sd = 0.5	new dist. ≈ 0 binomial = .0002			

For a fixed sd = 0.1 and a fixed skew = 0.25, the following table gives p-values of the goodness-of-fit tests on perturbed binomial data in which the parameter p of the population from which the data was generated is p = 0.1 and p = 0.25.

Table 13. More p-values for perturbed binomial data

p = 0.1	p = 0.25
new distr. P-value = .07689 **binomial P-value = .77348**	**new distr. P-value = .65834** **binomial P-value = .99904**

These goodness-of-fit results show that the binomial distribution typically fits the data better (in terms of the test about the binomial having a higher p-value). Yet the new distribution usually fits the data quite well, nearly always giving a good fit at the 5% significance level. We see that increasing the spread of the perturbations (by increasing sd) hinders the fit of either model. At least in these examples, the new distribution fits nearly as well as the binomial when the perturbations are slightly skewed (skew = 0.1 and skew = 0.25 in the first table).

For data sets whose mean approximately equaled the variance (the perturbed Poisson data), the goodness-of-fit results are summarized in the following tables for perturbed data from the Poisson distribution. Again, for purposes of brevity, only the p-values of the goodness-of-fit test about the binomial, and the goodness-of-fit test about the new

distribution are given. P-values that would lead us to conclude a good fit at the 5% level are in bold in the table.

In the tables, "skew" represents the mean of the normal perturbations and "sd" is proportional to the standard deviation of the perturbations, so that the greater "sd" is, the more the data deviates from purely Poisson data. (The simulated data sets have around 300 observations; because of rounding off the frequencies, the number of observations varies slightly for each data set.)

Table 14. P-values for perturbed Poisson data from a population with parameter $\lambda = 2$

	skew = 0	skew = 0.1	skew = 0.25	skew = 0.49
sd = 0.1	**new distr. = .397** **Poisson = .971**	**new distr. = .214** **Poisson = .894**	new dist. = .026 **Poisson = .531**	**new dist. = .378** **Poisson = .415**
sd = 0.1	**new dist. = .76** **Poisson = .996**			
sd = 0.25	new dist.=.00002 Poisson=.000006			
sd = 0.5	new dist. = .0332 **Poisson = .242**			

For a fixed sd = 0.1 and and fixed skew = 0, the following table gives p-values of the goodness-of-fit tests on perturbed Poisson data in which the parameter λ of the population from which the data was generated is $\lambda = 0.5$, 3.5, 5, and 8, respectively.

Table 15. More P-values for perturbed Poisson data.

$\lambda = 0.5$	$\lambda = 3.5$	$\lambda = 5$	$\lambda = 8$
new distr. = .651 **Poisson = .9438**	**new distr. = .669** **Poisson = .999**	**new distr. = .3134** **Poisson = .9952**	**new distr. = .679** **Poisson = .9982**

The results show that the Poisson model typically fits the data better than the new distribution does. Of course, we would expect the Poisson distribution to fit the

perturbed Poisson data quite well. The new distribution usually gives a good fit, particularly when the perturbations are not heavily spread out.

Comparing the new distribution with the negative binomial

Survival data whose mean is less than the variance are relatively numerous in the ecological literature. This type of data can be modeled with the negative binomial distribution, as well as with the new distribution. Several real ecological data sets are presented here, along with the fits of the negative binomial and the new distribution, which are compared using the Chi-square goodness-of-fit test.

Lack (1943a) collected survival data on the song thrush, a British bird. The life tables were reproduced in Deevey (1947). The life tables in Deevey list the survival frequencies per 1000 individuals, and for this report these numbers were converted to the raw frequencies by using the stated actual sample size of the study. The sample mean of the thrush data was 0.9305 years, and the sample variance was 1.877. The data were modeled with the new distribution and the negative binomial, and the results are summarized.

Table 16. Fit of models to the song thrush data

k (years)	obs(k)	LFLS e(k) ($\alpha^* = -0.33805$, $q^* = 0.551166$)	negative binom. e(k)
0	208	200.801	197.038
1	69	87.556	90.625
2	51	42.076	43.763
3	27	21.042	21.460
4	8	10.75	10.606
5	5	5.573	5.267
6	4	2.916	2.623
7	1	1.536	1.309
8	1	0.814	0.655

Table 17. Goodness-of-fit results for song thrush data.

	New distribution	Negative binomial
χ^2 statistic	8.99	9.486
Critical point	$\chi^2(5,.05) = 11.07$	$\chi^2(4,.05) = 9.49$
p-value	.10946	.050036

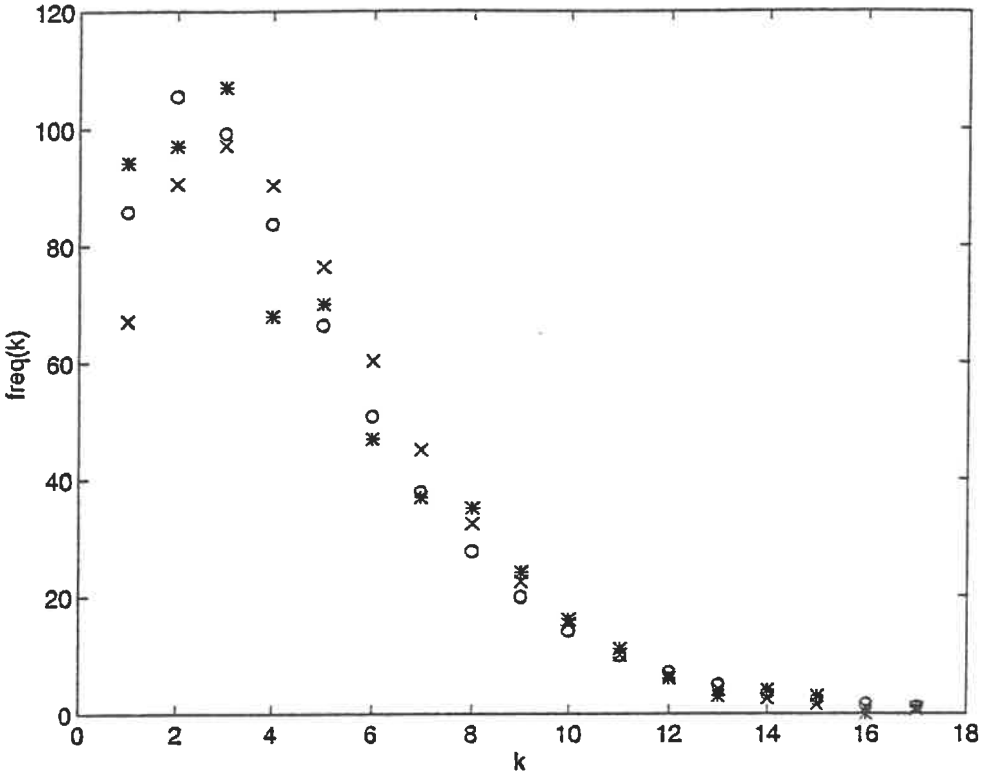
In this example, the new distribution easily gives a good fit at the 5% level, but the negative binomial barely gives a good fit to the data. The χ^2 statistic for the test about the negative binomial is barely lower than the critical point, meaning the goodness-of-fit hypothesis was nearly rejected for the negative binomial.

Another data set whose mean is less than the variance is Caughley's thar survival data (1966), whose sample mean = 4.46 years, with sample variance = 9.04. The data can be modeled with the new distribution and with the negative binomial.

Table 18. Fit of models to the thar data

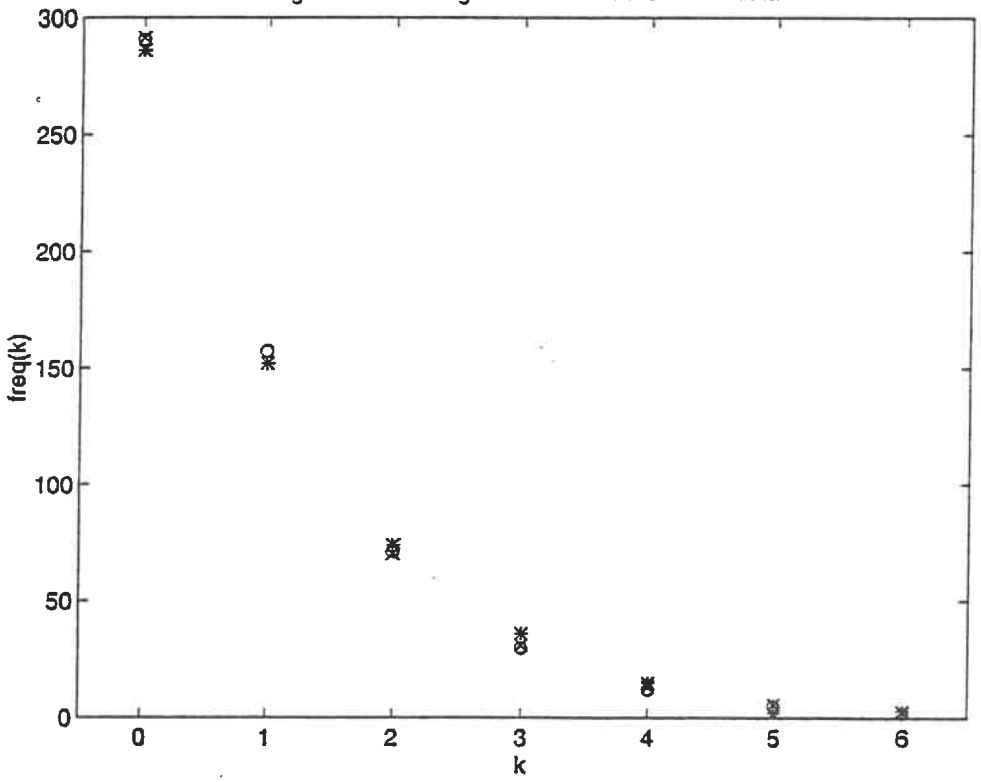
k (years)	obs(k)	LFLS e(k) ($\alpha^* = 0.9289$, $q^* = 0.6456$)	negative binom. e(k)
1	94	85.712	67.024
2	97	105.349	90.641
3	107	99.119	97.044
4	68	83.593	90.228
5	70	66.397	76.265
6	47	50.775	60.166
7	37	37.826	45.043
8	35	27.645	32.362
9	24	19.911	22.491
10	16	14.178	15.209
11	11	9.999	10.051
12	6	6.998	6.513
13	3	4.867	4.150
14	4	3.366	2.606
15	3	2.317	1.615
16	0	1.588	0.990
17	1	1.085	0.600

Figure 7: Modeling thar survival data



Observed frequencies = *
Expected frequencies (new distr.) = o
Expected frequencies (negative binomial) = ×

Figure 8: Modeling American robin survival data



Observed frequencies = *
Expected frequencies (new distr.) = o
Expected frequencies (negative binomial) = ×

Table 19. Goodness-of-fit results for thar data.

	New distribution	Negative binomial
χ^2 statistic	10.85	24.382
Critical point	$\chi^2(13,.05) = 22.36$	$\chi^2(12,.05) = 21.03$
p-value	.62338	.01804

The new distribution, in this example, gives a very good fit at the 5% level (see Figure 7). The negative binomial gives a poor fit at the 5% level. The new distribution again appears to give a better fit to the data than the negative binomial.

Lack (1943b) gathered survival data for the blackbird which is reproduced in Deevey (1947). The sample mean for the blackbird data = 1.074 and the sample variance = 2.6305, so the data is modeled with the new distribution and the negative binomial.

Table 20. Fit of models to the blackbird data

k (years)	obs(k)	LFLS e(k) ($\alpha^* = -1.0956$, $q^* = 0.7854$)	negative binom. e(k)
0	192	189.413	181.498
1	60	69.615	79.255
2	50	35.065	40.825
3	20	20.096	22.099
4	12	12.360	12.247
5	7	7.950	6.883
6	6	5.274	3.904
7	3	3.579	2.229
8	0	2.470	1.279
9	2	1.729	0.737

Table 21. Goodness-of-fit results for blackbird data.

	New distribution	Negative binomial
χ^2 statistic	9.194	8.946
Critical point	$\chi^2(6,.05) = 12.59$	$\chi^2(6,.05) = 12.59$
p-value	.16296	.17664

Both models give quite good fits to the blackbird data. The p-value for the negative binomial fit is slightly better than for the new distribution, but both distributions fit the data almost equally well.

The sowbug data (Janardan, et al., 1979) which was presented in Table 2 can also be modeled with the new distribution and the negative binomial.

Table 22. Fit of models to sowbug data

k (years)	obs(k)	LFLS e(k) ($\alpha^* = -0.4798$, $q^* = 0.8039$)	negative binom. e(k)
1	28	31.189	23.822
2	14	17.980	16.651
3	11	11.899	12.368
4	8	8.332	9.368
5	11	6.018	7.164
6	2	4.433	5.510
7	3	3.310	4.254
8	3	2.496	3.294
9	3	1.896	2.555
10	3	1.449	1.985
11	2	1.113	1.544
12	0	0.858	1.202
13	1	0.664	0.937
14	2	0.515	0.731
15	1	0.401	0.570
16	0	0.312	0.445
17	2	0.244	0.347
18+	0	0.891	1.251

Table 23. Goodness-of-fit results for sowbug data.

	New distribution	Negative binomial
χ^2 statistic	10.84	6.945
Critical point	$\chi^2(7,.05) = 14.07$	$\chi^2(9,.05) = 16.92$
p-value	.14575	.64285

Both models provide a good fit at the 5% level to the sowbug data. Based on the p-values, the negative binomial gives an exceptional fit, better than the new distribution in this example. It should be noted that the p-value of the test about the negative binomial in this example is comparable to the p-value of the goodness-of-fit test for the new distribution fitted with the MLE's (p-value = .63379, from Table 3).

Deevey (1947) reproduced the survival data for the lapwing, another British bird, collected by Lack (1943c). The sample mean for this data = 1.863, while the sample variance = 5.4802. The data can be modeled with the new distribution and with the negative binomial.

Table 24. Fit of models to the lapwing data

k (years)	obs(k)	LFLS e(k) ($\alpha^* = -0.69084$, $q^* = 0.79496$)	negative binom. e(k)
0	380	406.850	355.355
1	213	200.361	224.821
2	128	120.366	145.355
3	78	78.440	94.649
4	67	53.448	61.850
5	56	37.460	40.503
6	24	26.771	26.561
7	20	19.406	17.4359
8	7	14.222	11.454
9	9	10.512	7.529
10	7	7.824	4.951
11	11	5.857	3.257

Table 25. Goodness-of-fit results for lapwing data

	New distribution	Negative binomial
χ^2 statistic	24.331	35.58
Critical point	$\chi^2(9,.05) = 16.92$	$\chi^2(9,.05) = 16.92$
p-value	.0038	.000047

Both models give poor fits to the lapwing data at the 5% level.

Krebs (1989) reported lifetimes for the snowshoe hare in six-month intervals, such as 0, 0.5, 1, 1.5, ... years. To fit the new distribution and the negative binomial to the data, the lifetimes were transformed into half-year units, so that k = 0, 1, 2, ... was the transformed variable used in fitting the negative binomial. The sample mean of the hare data was 2.1746 half-years, with a sample variance of 4.4298.

Table 26. Fit of models to the snowshoe hare data

k (years)	obs(k)	LFLS e(k) ($\alpha^* = -0.29424$, $q^* = 0.83971$)	negative binom. e(k)
0	21	14.912	14.261
0.5	6	10.211	15.110
1	12	7.610	11.877
1.5	9	5.872	8.254
2	4	4.617	5.361
2.5	4	3.675	3.335
3	5	2.949	2.014
3.5	2	2.381	1.190

Table 27. Goodness-of-fit results for snowshoe hare data

	New distribution	Negative binomial
χ^2 statistic	10.02	13.719
Critical point	$\chi^2(5,.05) = 11.07$	$\chi^2(4,.05) = 9.49$
p-value	.07467	.008248

For the snowshoe hare data, the new distribution gives a good fit at the 5% level and the negative binomial does not.

Farner (1945) collected survival data for the American robin, which were reproduced by Deevey (1947). The sample mean for the robin data was 0.87324 years, with a sample variance of 1.2973. The data can be modeled with the new distribution and with the negative binomial.

Table 28. Fit of models to the American robin data

k (years)	obs(k)	LFLS e(k) ($\alpha^* = 0.61813$, $q^* = 0.35229$)	negative binom. e(k)
0	286	290.262	291.316
1	152	156.954	152.008
2	74	71.043	70.254
3	36	29.899	31.068
4	15	12.091	13.428
5	2	4.768	5.723
6	3	1.848	2.417

Table 29. Goodness-of-fit results for American robin data

	New distribution	Negative binomial
χ^2 statistic	2.681	3.827
Critical point	$\chi^2(3,.05) = 7.81$	$\chi^2(4,.05) = 9.49$
p-value	.44347	.42992

Both models fit the American robin data very well (see Figure 8). Based on the p-values, the new distribution gives a slightly better fit than the negative binomial.

4. Conclusions about the performance of the new distribution vs. the traditional distributions in modeling survival data

In the real ecological examples discussed in this report, the new distribution fares well relative to the traditional distribution in fitting the survival data. In some cases, the new distribution gave a better fit than its traditional counterpart, and in other cases a worse fit, but typically the results were fairly close; in modeling the real survival data, neither the new nor the traditional distributions had an obvious advantage.

The binomial and Poisson distributions did better in modeling the simulated data, but this is no real surprise, since the simulated data were perturbed binomial and Poisson data.

The real advantage of the new distribution is its flexibility. While the binomial, Poisson, and negative binomial are each restricted to modeling a certain type of data, the

new distribution can model all three types with good results. This property of versatility has definite advantages. Suppose a statistician wished to model the survival data of several groups that were related; however, the mean exceeded the variance in some groups, the mean equaled the variance in other groups, and the mean was less than the variance in other groups. Rather than having to use three different distributions to model the groups, the statistician could use the new distribution to model all of them. The groups could then be compared by, say, comparing the estimated parameter values α^* and q^* among all the groups. Tests for differences among the groups would be much easier if all the groups were modeled with the same distribution. The results in this report indicate that using this flexible distribution would yield reasonable models about as often as using the traditional distributions.

References

Caughley, G. (1966). Mortality patterns in mammals. *Ecology* 47, 906-918.

Consul, P. (1989). *Generalized Poisson distributions: Properties and Applications.* Marcel Dekker, Inc., New York.

Deevey, E. S. (1947). Life tables for natural populations of animals. *Quarterly Review of Biology* 22, 283-314.

Doray, L. G. and Luong, A. (1995). Efficient estimators for the Good family. *Communications in Statistics: Simulation and Computation* 26, 1075-1088.

Farner, D. S. (1945). Age groups and longevity in the American robin. *Wilson Bull.* 57, 56-74.

Gibbons, J. D. and Chakraborti, S. (1992). *Nonparametric Statistical Inference.* 3rd edition. Marcel Dekker, New York.

Gradshetyn, I. S. and Ryszhik, I. M. (1980). *Tables of Integrals, Series, and Products.* Academic Press, New York.

Janardan, K. G., Kerster, H. W. and Schaeffer, D. J. (1979). Biological applications of the Lagrangian Poisson distribution. *Bioscience* 29, 599-602.

Johnson, N. L. and Kotz, S. (1969). *Discrete Distributions.* John Wiley and Sons, New York.

Kmenta, J. (1986). *Elements of Econometrics.* 2nd Edition. Macmillan Publishing Company, New York.

Krebs, C. J. (1989). *Ecological Methodology.* Harper and Row, New York.

Kulasekera, K. B. and Tonkyn, D. W. (1992). A new discrete distribution, with applications to survival, dispersal and dispersion. *Communications in Statistics: Simulation and Computation* 21, 499-518.

Lack, D. (1943a). *The life of the robin.* H. F. & G. Witherby, London.

Lack, D. (1943b). The age of the blackbird. *Brit. Birds* 36, 166-175.

Lack, D. (1943c). The age of some more British birds. *Brit. Birds* 36, 193-197, 214-221.

Miller, G. L. and Carroll, B. W. (1989). Modelling vertebrate dispersal distances: alternatives to the geometric distribution. *Ecology* 70, 977-986.

Tonkyn, D. W. and Plissner, J. H. (1991). Models of multiple dispersers from the nest: predictions and inference. *Ecology* 72, 1721-1730.

Wallenius, K. T. and Korkotsides, A. S. (1990). Exploratory model analysis using cdf knotting with applications to distinguishability, limiting forms, and moment ratios to the three-parameter Weibull family. *Journal of Statistical Computation and Simulation* 35, 121-133.

Weisstein, Eric W. (1996-1999). *CRC Concise Encyclopedia of Mathematics.* CRC Press. (web page: http://www.astro.virginia.edu/~eww6n/math/).

Zanakis, S. H. (1979). Extended pattern search with transformations for the three parameter Weibull family. *Management Science* 25, 1149-1161.

Zornig, P. and Altman, G. (1995). Unified representation of Zipf distributions. *Computational Statistics and Data Analysis* 19, 461-473.