Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

A generalized binomial distribution

Zvi Drezner [a] & Nicholas Farnum [a]
[a] Department of Management Science , California State University-Fullerton , Fullerton, CA, 92634

Published online: 27 Jun 2007.

COMMUN. STATIST.—THEORY METH., 22(11), 3051-3063 (1993)

A GENERALIZED BINOMIAL DISTRIBUTION

Zvi Drezner and Nicholas Farnum
Department of Management Science
California State University-Fullerton
Fullerton, CA 92634.

Key Words: Correlated Bernoulli Trials, Bimodal Distribution.

ABSTRACT

A new generalization of the binomial distribution is introduced that allows dependence between trials, nonconstant probabilities of success from trial to trial, and which contains the usual binomial distribution as a special case. Along with the number of trials and an initial probability of 'success', an additional parameter that controls the degree of correlation between trials is introduced. The resulting class of distributions includes the binomial, unimodal distributions, and bimodal distributions. Formulas for the moments, mean, and variance of this distribution are given along with a method for fitting the distribution to sample data.

1. INTRODUCTION

In this paper we give a new generalization of the binomial distribution, one that allows dependence between trials, nonconstant probabilities of success from trial to trial, and which contains the usual binomial distribution as a special case. Along with the number of trials, n, and an initial probability of 'success', p, the development proceeds by introducing an additional parameter, θ, that controls the degree of correlation between trials. The pdf is found by defining the new distribution to be the solution to a recursive equation similar to that studied by Woodbury (1949) and other authors (Samuels 1965, Rutherford 1954).

Formulas for the moments, mean, and variance of this distribution are given along with a method for fitting the distribution to sample data. Like the binomial, the mean of the new distribution is always np. A closed-form expression for the variance of the new distribution is given and shown to lie between np(1-p) and $n^2p(1-p)$. The resulting class of distributions includes the binomial (when $\theta = 0$), unimodal distributions (for small θ), and bimodal distributions (for large θ).

2. BACKGROUND

Generalizations of the binomial distribution typically involve modifying either the assumption of a constant 'success' probability and/or the assumption of independence between trials in the underlying Bernoulli process. The easiest modification is to allow the probability of 'success' to vary from trial to trial, while retaining the assumption of independence between trials. In this setting, with probability p_i of success at trial i $\{i = 1,2,3,...,n\}$, the underlying trials are called Poisson trials (Feller 1968, pp 230-231). With Poisson trials, the mean and variance of the number of successes X in n trials are given by (Feller 1968, p 231; Nedelman and Wallenius 1986, p 286) $E(X) = n\bar{p}$ *and* $Var(X) = n\bar{p}(1-\bar{p}) - ns^2$

where $\bar{p} = \frac{\sum_{i=1}^{n} p_i}{n}$ *and* $s^2 = \frac{\sum_{i=1}^{n} (p_i-\bar{p})^2}{n}$. Note that with Poisson trials Var(X)

is always less than that of a binomial distribution with $p = \bar{p}$ unless all the p_i's are equal. It can further be shown that the distribution of X under Poisson trials is always unimodal (Samuels 1965, p 1272-1273). This fact also follows from examining the generating function of X which is given by $\prod_{i=1}^{n} (1-p_i+p_it)$.

Since the roots of the generating function are all real and negative, the coefficients of X must form a log-concave sequence (Wilf 1990, p 127), which implies unimodality.

The other approach to generalizing the binomial involves relaxing the assumption of independence between trials. One method of accomplishing this is to consider the recursive relationship between the probabilities at trials n-1 and n,

$$P(x,n) = P(S_n|x-1,n-1)P(x-1,n-1) + P(F_n|x,n-1)P(x,n-1) \tag{2.1}$$

where P(x,n) denotes the probability of x successes in n trials, S_n denotes the event 'success on the n^{th} trial', and F_n denotes the event 'failure on the n^{th} trial'. Woodbury (1949) considered (2.1) in the case where both $P(S_n \mid x\text{-}1,n\text{-}1)$ and $P(F_n \mid x,n\text{-}1)$ were assumed to be functions of x alone. While closed-form expressions of the pdf of this distribution are not available (Johnson and Kots 1969, p 233), Woodbury (1949, p 313) did find its generating function. Since, with the Woodbury distribution, the probability of success at trial i depends on the prior history of successes, Rutherford (1954) referred to such distributions as 'contagious' and proceeded to analyze the special case where the probability of success grows linearly with the number of previous successes.

Recently, Ng (1989) presented a multiparameter generalization called the modified binomial distribution (MBD) whose pdf is the solution of

$$MB_n(x)=\sum_{i=1}^{m} [q\cdot B_{n-i}(x)+p\cdot B_{n-i}(x-i)]\theta_{i-1}(1-\theta_i)\dots(1-\theta_{n-1}) \tag{2.2}$$

where $B_n(x)$ denotes the usual binomial probability of x successes in n trials with success probability p, $\theta_0 = 1$, and the θ_i's are constrained by the fact that $MB_n(x)$ must be nonnegative for each x = 0,1,2,...,n. Even in the interesting case where $\theta_i = \theta$ for every i, there is no closed-form for the pdf in (2.2). However, for illustration, Ng provides the exact pdf in the simple cases n = 2 and n = 4. An interesting feature of the MBD's distribution is that it can either be unimodal or bimodal depending upon the particular values of the θ_i's. In the case where the indicator variables at each trial can be considered exchangeable, it can further be shown that the MBD distribution has mean and variance given by

$$E(X) = np \quad and \quad Var(X) = np(1-p)+n(n-1)p(1-p)\tau_n \tag{2.3}$$

where τ_n is the correlation between any two indicator variables.

There are many practical situations that could be modeled as a sequence of dependent Bernoulli trials. Woodbury (1949) for example, suggests that the distribution of the number of plants in a given area might follow such a model. Ng (1989) showed that the MBD distribution fits binary data from certain toxicological experiments. As another example, the distribution of grades on a multiple choice test might also follow this pattern if a student's ability to correctly answer one question is thought to be related to his or her ability to answer correctly any other question. Another example is attendance at meetings. A congressman who does not attend one meeting is more likely not to attend another one, and there might be some degree of correlation between these Bernoulli trials (attending or not attending meetings). In sporting competitions and games, the a priori probability of a win is 50%. However, as more games (trials) are played, a team's probability of winning successive games could be a function of the winning record to date.

3. THE GENERALIZED BINOMIAL DISTRIBUTION (GBD)

The generalized binomial distribution (hereafter abbreviated GBD) proposed in this paper results from a different approach to solving the recursive equations in (2.1). In place of Woodbury's assumption that only the past number of successes x influences $P(S_n \mid x\text{-}1,n\text{-}1)$ and $P(F_n \mid x,n\text{-}1)$, we substitute expressions that also take account of the previous numbers of trials. Specifically, we set

$$\begin{aligned} P(S_n|x-1,n-1) &= (1-\theta_n)p + \theta_n\frac{x-1}{n-1} \\ P(F_n|x,n-1) &= (1-\theta_n)(1-p) + \theta_n(1 - \frac{x}{n-1}) \end{aligned} \tag{3.1}$$

yielding

$$P(x,n)=[(1-\theta_n)p+\theta_n\frac{x-1}{n-1}]P(x-1,n-1)+[(1-\theta_n)(1-p)+\theta_n(1-\frac{x}{n-1})]P(x,n-1) \tag{3.2}$$

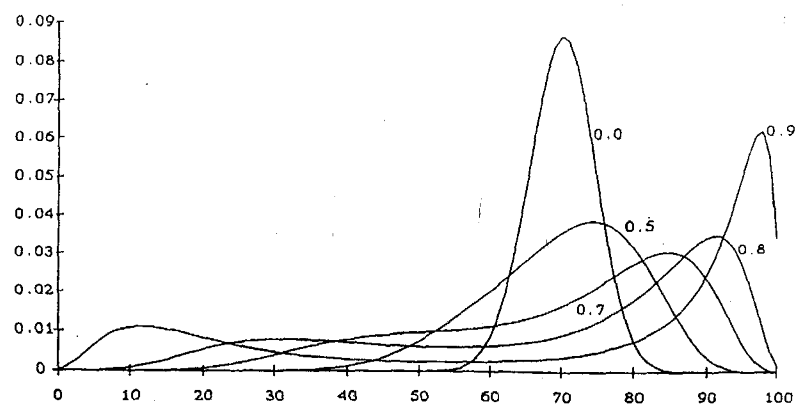

Figure 1: The GBD for p=0.7 and Various Values of θ

along with the starting values $P(0,1)=1-p$, $P(1,1)=p$ and boundary conditions $P(-1,n)=P(n+1,n)=0$. The value of the parameters θ_n are bounded above by 1 but, as is also the case for Ng's MBD distribution, they may also assume negative values, constrained only by the nonnegativity implied in (3.1).

The case where each $\theta_i = \theta$ for $i = 1,2,3,\ldots,n$ is of special interest. As mentioned in the introduction, when $\theta = 0$ the GBD becomes the binomial distribution with parameters n and p. When $\theta = 1$, the GBD is concentrated at the endpoints $X = 0$ (with probability p) and $X = n$ (with probability 1-p). For other values of θ it is possible for the distribution to be either unimodal or bimodal. For example, the GBD with $p = .7$ and $n = 100$ and various values of θ is depicted in Figure 1.

4. ANALYSIS OF THE GENERALIZED BINOMIAL DISTRIBUTION

In this section we analyze the probability distribution defined by equation (3.2). Define for $k\geq 0$ the k^{th} moment of the distribution:

$$E_n(X^k) = \sum_{x=0}^{n} x^k P(x,n) \tag{4.1}$$

Lemma 1: $E_n(X^k)=E_{n-1}(X^k)+(1-\theta_n)p\sum_{i=0}^{k-1}\binom{k}{i}E_{n-1}(X^i)+\frac{\theta_n}{n-1}\sum_{i=0}^{k-1}\binom{k}{i}E_{n-1}(X^{i+1})$

Proof: By equation (3.2): $E_n(X^k) = \sum_{x=0}^{n}x^kP(x,n)=$

$$= (1-\theta_n)p\sum_{x=0}^{n}x^kP(x-1,n-1) + \theta_n\sum_{x=0}^{n}\frac{x-1}{n-1}x^kP(x-1,n-1) +$$

$$+ (1-\theta_n)(1-p)\sum_{x=0}^{n}x^kP(x,n-1) + \theta_n\sum_{x=0}^{n}x^k(1-\frac{x}{n-1})P(x,n-1).$$

Using the boundary conditions P(-1,n-1)=P(n,n-1)=0 and replacing x-1 by x in the first two sums we get:

$$E_n(X^k)=(1-\theta_n)p\sum_{x=0}^{n-1}(x+1)^kP(x-1,n-1)+\theta_n\sum_{x=0}^{n-1}\frac{x}{n-1}(x+1)^kP(x-1,n-1)+$$

$$+ (1-\theta_n)(1-p)\sum_{x=0}^{n-1}x^kP(x,n-1) + \theta_n\sum_{x=0}^{n-1}x^k(1-\frac{x}{n-1})P(x,n-1)$$

Rearranging the sums and combining terms yields:

$$E_n(X^k)=\sum_{x=0}^{n-1}x^kP(x,n-1)+\sum_{x=0}^{n-1}[(1-\theta_n)p+\theta_n\frac{x}{n-1}][(x+1)^k-x^k]P(x,n-1)=$$

$$=E_{n-1}(X^k)+(1-\theta_n)p\sum_{i=1}^{k-1}\binom{k}{i}E_{n-1}(X^i)+\frac{\theta_n}{n-1}\sum_{i=0}^{k-1}\binom{k}{i}E_{n-1}(X^{i+1}). \ \square$$

Theorem 1: $\sum_{x=0}^{n}P(x,n) = 1.$

Proof: From the starting values P(0,1)=1-p and P(1,1)=p, the Theorem is true for n=1. Next assuming the Theorem is true for n-1, it must also hold for n, since (4.1) implies that the sum in Theorem 1 is $E_n(X^0)$ and Lemma 1 yields $E_n(X^0)=E_{n-1}(X^0)=1$. By induction the theorem holds for all n. □

Theorem 1 confirms that P(x,n) defined by equation (3.2) is indeed a probability distribution.

Theorem 2: For the probability distribution (3.2): E(X)=np.

Proof: Proceeding again by induction, the Theorem is obviously true for n=1. Next, assuming the theorem is true for n-1, (4.1) implies $E(X)=E_n(X^1)$ and Lemma 1 yields $E_n(X) = E_{n-1}(X) + (1-\theta_n)pE_{n-1}(X^0) + \frac{\theta_n}{n-1}E_{n-1}(X) =$

$= (n-1)p + (1-\theta_n)p + \theta_n p = np.$

Thus the theorem is true for n and, by induction, for any integer. □

We now obtain a formula for the variance of the distribution. The proof is more complicated and therefore we break it down into a sequence of lemmas first. The first two lemmas are true for any θ_n. The final Theorem is calculated for a fixed θ only. By Lemma 1 we get:

Lemma 2: $E_n(X^2) = (1+\frac{2\theta_n}{n-1})E_{n-1}(X^2) + 2(1-\theta_n)(n-1)p^2 + p.$

Lemma 3: $E_n(X^2) = p\sum_{j=1}^{n} \{ [2(j-1)(1-\theta_j)p + 1]\prod_{i=j+1}^{n} (1+\frac{2\theta_i}{i-1}) \}$

Proof: We prove the Lemma by mathematical induction. The Lemma is true for n=1: The sum is just for j=1 and the product is empty and therefore equal to 1, and $E_1(X^2)=p$. Assuming that the Lemma is true for n-1, we prove it for n:

$$E_n(X^2)=(1+\frac{2\theta_n}{n-1})p\sum_{j=1}^{n-1} \{ [2(j-1)(1-\theta_j)p + 1]\prod_{i=j+1}^{n-1} (1+\frac{2\theta_i}{i-1}) \}+$$

$$+2(1-\theta_n)(n-1)p^2+p =$$

$$=p\sum_{j=1}^{n-1} \{ [2(j-1)(1-\theta_j)p+1]\prod_{i=j+1}^{n} (1+\frac{2\theta_i}{i-1})\} +2(1-\theta_n)(n-1)p^2+p=$$

$$=p\sum_{j=1}^{n} \{ [2(j-1)(1-\theta_j)p + 1]\prod_{i=j+1}^{n} (1+\frac{2\theta_i}{i-1}) \}. \square$$

In the following we assume that $\theta_n=\theta$ and use the Beta function:

$$B(n,m) = \frac{\Gamma(n)\Gamma(m)}{\Gamma(n+m)} = \frac{\Gamma(n)}{\prod_{k=0}^{n-1}(k+m)} \tag{4.2}$$

along with the well known recursive relationship

$$B(n+1,m) = \frac{n}{n+m} B(n,m) \tag{4.3}$$

Lemma 4: $E_n(X^2) = \dfrac{p[n(1-p+np-2\theta np) - \dfrac{1-p}{B(n,2\theta)}]}{1-2\theta}$

Proof: The proof again proceeds by mathematical induction. The Lemma is true for n=1: by (4.2) B(1,2θ)=1/2θ and $E_1(X^2)$=p. Assuming that the Lemma is true for n-1, then by Lemma 2: $E_n(X^2) = [1+\frac{2\theta}{n-1}]E_{n-1}(X^2) + 2(1-\theta)(n-1)p^2 + p$. From the assumption that the lemma is true for n-1:

$$E_n(X^2)=(1+\frac{2\theta}{n-1})\frac{(n-1)(1-2p+np-2\theta(n-1)p) - \dfrac{1-p}{B(n-1,2\theta)}}{1-2\theta}p+2(1-\theta)(n-1)p^2+p$$

and from the recursion formula (4.3): $(1+\frac{2\theta}{n-1})\frac{1-p}{B(n-1,2\theta)} = \frac{1-p}{B(n,2\theta)}$. By extensive but straight forward manipulations it can be verified that:

$$(1+\frac{2\theta}{n-1})\frac{(n-1)[1-2p+np-2\theta(n-1)p]}{1-2\theta}p+2(1-\theta)(n-1)p^2+p =$$

$$= \frac{n(1-p+np-2\theta np)p}{1-2\theta}.$$

so the lemma is true for n, and by induction the Lemma follows. □

Theorem 3: $Var(X) = p(1-p)\dfrac{n - \dfrac{1}{B(n,2\theta)}}{1-2\theta}$

Proof: Var(X) = $E(X^2)-[E(X)]^2 = E_n(X^2)-n^2p^2$, and the Theorem follows by Lemma 4. □

By formula (4.2) the variance Var(X) can be written as:

$$Var(X) = p(1-p)\frac{n - \frac{\prod_{k=0}^{n-1}(k+2\theta)}{(n-1)!}}{1-2\theta} \tag{4.4}$$

An interesting case is θ=½. For this case both Theorem 3 and (4.4) yield a 0/0 expression, which by L'Hopital's rule can be easily be shown to be

$$For\ \theta = \frac{1}{2}:\quad Var(X) = np(1-p)\sum_{k=1}^{n}\frac{1}{k} \tag{4.5}$$

<u>Lemma 5</u>: Var(X) is a monotonically increasing function of θ.

<u>Proof</u>: Define: $h(n,\theta) = \dfrac{n - \frac{1}{B(n,2\theta)}}{1-2\theta}$. We need to show that h(n,θ) is monotonically increasing in θ. Since $h(n,\theta) = \lim\limits_{p\to 0}\dfrac{Var(X)}{p} =$

$= \lim\limits_{p\to 0}\dfrac{E_n(X^2)-n^2p^2}{p} = \lim\limits_{p\to 0}\dfrac{E_n(X^2)}{p}$, then by Lemma 2 the following recursive relationship holds: $h(n,\theta) = [1 + \frac{2\theta}{n-1}]h(n-1,\theta) + 1$. Since h(1,θ)=1, the Lemma is true for n=1. Using the recursion above, the Lemma follows by mathematical induction. ◻

<u>Theorem 4</u>: np(1-p) ≤ Var(X) ≤ n^2p(1-p).

<u>Proof</u>: For θ=0 the distribution is binomial and either Theorem 3 or (4.4) yields Var(X)=np(1-p). For θ=1 the distribution is P(0,n)=1-p; P(n,n)=p and the other probabilities vanish. This case yields Var(X)=n^2p(1-p). The Theorem follows from the results for θ=0 and θ=1, and Lemma 5. ◻

5. USING NEGATIVE θ'S

The results of the previous section (except Theorem 4) apply also for negative θ's. Since the solution to (3.2) will be a probability distribution only if $0 \le P(x,n) \le 1$, it is sufficient to ensure that $P(x,n) \ge 0$, because this condition implies $P(x,n) \le 1$ by Theorem 1. Since the terms multiplying $P(x-1,n-1)$ and $P(x,n-1)$ in (3.2) are interpreted as conditional probabilities, we require that these terms be non-negative, which ensures that $P(x,n) \ge 0$. The required conditions are then: $(1-\theta_n)p + \theta_n \frac{x-1}{n-1} \ge 0$; $(1-\theta_n)(1-p) + \theta_n(1-\frac{x}{n-1}) \ge 0$. These inequalities imply that $\theta_n \le 1$. For negative θ's we get (since the inequalities must be valid for every x): $(1-\theta_n)p+\theta_n \ge 0$; $(1-\theta_n)(1-p)+\theta_n \ge 0$ which yields:

$$\theta_n \ge 1 - \frac{1}{\max\{p,1-p\}} \tag{5.1}$$

Inequality (5.1) allows for negative θ's for $0<p<1$. For example, for $p=½$ we could have $\theta_n \ge -1$.

6. ON THE SELECTION OF THE BEST θ

For a given starting value, p, one can fit the GBD to sample data by selecting the θ that results in the same variance as the sample's variance. This implies solving implicitly (4.4) for θ. The process can be simplified by the following.

Define $r(\theta,n)$ by:

$$r(\theta,n) = \frac{\frac{Var(X)}{np(1-p)} - 1}{n-1} \tag{6.1}$$

From Theorem 3 and (4.4):

$$r(\theta,n) = \frac{2\theta - \frac{1}{nB(n,2\theta)}}{(1-2\theta)(n-1)} = \frac{2\theta[n-\prod_{k=1}^{n-1}(1+\frac{2\theta}{k})]}{(1-2\theta)n(n-1)} \tag{6.2}$$

Table 1: Values of K(n,r)

n	r=0.01	r=0.02	r=0.03	r=0.05	r=0.07	r=0.1	r=0.2	r=0.3	r=0.5	r=0.7	r=0.9
10	.236	.277	.309	.362	.408	.431	.466	.485	.506	.520	.529
20	.309	.364	.406	.449	.459	.469	.484	.491	.498	.502	.505
30	.359	.423	.462	.475	.481	.485	.491	.494	.497	.499	.500
40	.397	.468	.482	.489	.492	.494	.496	.496	.497	.498	.498
50	.429	.488	.494	.497	.498	.499	.498	.498	.498	.498	.498
60	.455	.499	.502	.503	.503	.502	.500	.499	.498	.498	.498
80	.497	.512	.512	.510	.508	.506	.502	.500	.499	.498	.498
100	.519	.519	.517	.513	.511	.508	.503	.501	.499	.498	.498
200	.540	.530	.525	.518	.514	.511	.505	.503	.500	.499	.498
500	.542	.530	.524	.517	.514	.511	.506	.503	.501	.500	.499
1000	.536	.525	.520	.515	.512	.509	.505	.503	.501	.500	.500
5000	.522	.515	.512	.509	.508	.506	.504	.502	.501	.500	.500
10000	.517	.512	.510	.508	.506	.505	.503	.502	.501	.500	.500

Theorem 5: $0 \le r(\theta,n) \le 1$.

Proof: Follows directly from Theorem 4 and (6.1). ◻

$r(\theta,n)$ can be estimated by substituting the sample variance for Var(X) in (6.1), and the best θ can be implicitly obtained by solving (6.2). A table giving the best θ for values of $r(\theta,n)$ and n can be constructed. By defining the variable $r(\theta,n)$, we have circumvented the need to set up a table for each p separately.

It can be shown that for a given θ: $\lim_{n\to\infty}\left\{\dfrac{\log n}{\log\dfrac{1}{r(\theta,n)}}\right\} = \dfrac{0.5}{1-\theta}$. Therefore we chose the following expression for estimating θ:

$$\theta = 1 - \frac{K(n,r)\log\frac{1}{r}}{\log(\alpha n+\beta)} \tag{6.3}$$

By regression analysis we empirically found the best α and β when K(n,r)=0.5 is used in (6.3). The parameters $\alpha=0.241$; $\beta=1.84$ fit this model the best. Table 1 gives values for K(n,r) for various values of n and r and these values of α and β.

7. AN EXAMPLE

Data of the final standing of teams in major league baseball for the years 1961-1991 (except 1981 when a strike shortened the season) were used for

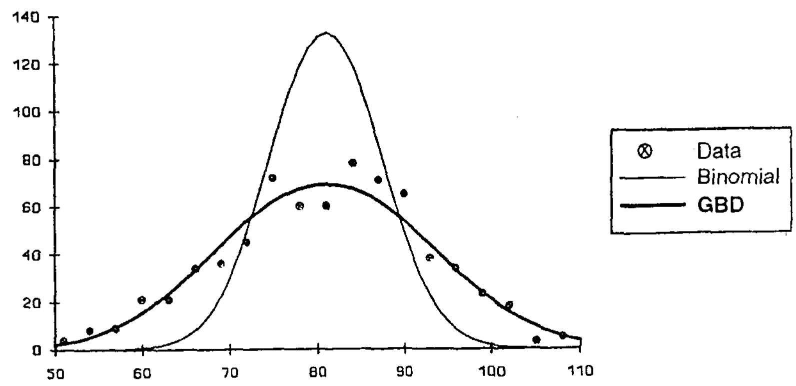

Figure 2: Modeling of Baseball results

modeling by the GBD. Since 1961, n=162 games have been played each season. There are a total of 706 records (number of wins at the end of the season). The mean is very close to 81 wins (in some cases a team did not complete all n=162 games when it did not matter for the pennant race). The variance was found to be 132.595. If the Bernoulli trials are assumed to be independent and p is assumed to be ½, the variance would be np(1-p)=40.5. By comparison (6.1) gives $r(\theta,n)=0.0146$, which from Table 1 yields $K(n,r)=0.529$. Then, by (6.3), $\theta=0.397$. Figure 2 depicts the expected distribution if the Bernoulli trials were independent (i.e. a binomial distribution with n=162 and p=0.5), along with the GBD (with n=162, p=0.5 and $\theta=0.397$), and the actual data (grouped in intervals of 3). The figure clearly shows that the GBD provides a good fit with the actual data.

REFERENCES

Feller, W. (1968) An Introduction to Probability Theory and Its Applications, Vol. I, New York, NY: John Wiley.

Johnson, N.L. and Kots, S. (1969), Discrete Distributions, New York, NY: Houghton Mifflin.

Nedelman, J. and Wallenius, T. (1986) "Bernoulli Trials, Poisson Trials, Surprising Variances, and Jensen's Inequality", American Statistician, 40(4), 286-289.

Ng, T. (1989), "A New Class of Modified Binomial Distributions with Applications to Certain Toxicological Experiments", Communications in Statistics, Part A - Theory and Methods, 18(9), 3477-3492.

Rutherford, R.S.G. (1954), "On a Contagious Distribution", Annals of Mathematical Statistics, 25, 703-713.

Samuels, S.M. (1965), "On the Number of Successes in Independent Trials", Annals of Mathematical Statistics, 36, 1272-1278.

Wilf, H.S. (1990), Generatingfunctionology, San Diego, CA: Academic Press.

Woodbury, M.A. (1949), "On a Probability Distribution", Annals of Mathematical Statistics, 20, 311-313.

Received March 1992; Revised March 1993.