A SIMPLE MODIFICATION OF THE BINOMIAL DISTRIBUTION

by

S. W. DHARMADHIKARI

INTRODUCTION AND SUMMARY

Given any probability distribution, new distributions can be derived from it by assuming its parameters to follow some specific probability distributions. A simple example of this process is provided by the Poisson distribution

\[ P(r \mid \lambda) = e^{-\lambda} \frac{\lambda^r}{r!} \quad (r = 0, 1, 2, \ldots). \]

If the parameter \( \lambda \) is assumed to follow the Pearson's Type III law

\[ f(\lambda) = \frac{(k \mid m)^k}{(k-1)!} \lambda^{k-1} e^{-k\lambda m} \quad (0 < \lambda < \infty), \]

then the probability of \( r \) successes is obtained as

\[
P(r) = \int_{0}^{\infty} P(r \mid \lambda) f(\lambda) d\lambda \]
\[= \frac{(k+r-1)!}{r!(k-1)!} \left( \frac{k}{m+k} \right)^k \left( \frac{m}{m+k} \right)^r.\]

This final distribution is known as the negative binomial distribution.

Such distributions obtained by assuming the parameters to be random variables are known as Compound Distributions.

The present note obtains a simple compound distribution from the usual binomial distribution. Of the two parameters \( n \) and \( p \), \( n \) is kept fixed, while \( p \) is assumed to follow a \( \beta \) distribution. The moments of the resulting distribution are obtained and the estimation of the parameters discussed. The relation between numbers of successes in successive sets of binomial trials is also obtained. Numerical data are also discussed.
MODIFICATION OF THE BINOMIAL DISTRIBUTION

RELATED WORK

The idea of letting \( p \) vary from set to set is not new. Greenwood (1949), while discussing the measles data for Providence, Rhode Island, suggested that the chance \( p \) of infection, of an individual, might vary between households. Working on this suggestion, Bailey (1953) assumed a \( \beta \) distribution for \( p \) and obtained excellent fits to the Providence data.

A paper by Ottestad (1948) discusses, in a different notation, the modified distribution obtained in the present note. In addition, many interesting particular and limiting cases are given. However, the estimation of the parameters by maximum likelihood and the relation between the numbers of successes in successive sets is considered for the first time in the present paper.

THE MODIFIED DISTRIBUTION AND ITS MOMENTS

The binomial distribution is defined by

\[
P(r | p) = _nC_r p^r (1-p)^{n-r} \quad (r = 0, 1, 2, \ldots, n).
\]

Let it be assumed that the probability \( p \) of success remains constant within sets of \( n \) trials, but that between sets it varies and follows some probability distribution \( f(p) \). Since \( p \) ranges from zero to unity, it is natural to take \( f(p) \) to be of the \( \beta \) form; i.e.

\[
f(p) = \frac{p^{l-1} (1-p)^{m-1}}{B(l, m)} \quad (0 < p < 1),
\]

where \( l \) and \( m \) are positive constants independent of \( p \). The probability of \( r \) successes in a random set of \( n \) trials is, therefore, obtained as

\[
P(r) = \int_0^1 P(r | p) f(p) \, dp = _nC_r \frac{B(l+r, m+n-r)}{B(l, m)}, \tag{1}
\]

\[
= \frac{n! \, l_{[r]} \, m_{[n-r]}}{(l+m)_{[n]}}, \tag{2}
\]

where \( x_{[r]} = x(x+1)(x+2)\ldots(x+r-1) \),

with \( x_{[0]} = 1 \).
Since we must have
\[ \sum_{r=0}^{n} P(r) = 1, \]
we get
\[ \sum_{r=0}^{n} nC_r B(l+r, m+n-r) = B(l, m). \] (3)

The \( h \)th factorial moment of the random variable \( r \) is obtained as
\[ \mu'_h = \mathcal{E}^r \left( (r-1) \cdots (r-h+1) \right) \]
\[ = \frac{(n-h+1)_h (l)_h}{(l+m)_h}. \] (4)

(Relation (3) enables us to simplify the summation occurring in the expression for \( \mu'_h \).)

From the factorial moments (4), the raw moments \( \mu'_h \) and the central moments \( \mu_h \) can be calculated using standard formulae (see, for example, Kendall (1948)). The results are:

\[ \mu'_1 = \frac{nl}{l+m}, \] (5)

\[ \mu'_2 = \frac{nlm(l+m+n)}{(l+m)^2 (l+m+1)}, \] (6)

\[ \mu'_3 = \frac{nlm(m-l) (l+m+n) (l+m+2n)}{(l+m)^3 (l+m+1) (l+m+2)}, \] (7)

\[ \mu'_4 = \frac{3nl^2m^2(l+m+n)^3 (l+m-6)}{(l+m)^4 (l+m+1)_{[3]}} \]
\[ + \frac{nlm(l+m+n)}{(l+m)^2 (l+m+1)_{[3]}} \left[ (l+m) (l+m-1) - 6lm + 6n(l+m+n) \right]. \] (8)

These moments immediately give
\[ \beta_1 = \frac{\mu'_3}{\mu'_2} = \frac{(m-l)^2 (l+m+2n)^2 (l+m+1)}{nlm(l+m+2)^2 (l+m+n)}, \] (9)

and
\[ \beta_2 = \frac{\mu'_4}{\mu'_2} = \frac{3(l+m+1) (l+m-6)}{(l+m+2) (l+m+3)} \]
\[ + \frac{(l+m)^2 (l+m+1)}{lm(l+m+2)_{[3]}} \left[ \frac{(l+m) (l+m-1) - 6lm}{n(l+m+n)} + 6 \right]. \] (10)
ESTIMATION OF PARAMETERS

It should be noted that, unlike the usual binomial, negative binomial and hypergeometric distributions, which all have two parameters, the modified distribution obtained in the previous section has three parameters, viz. \( l \), \( m \) and \( n \). Of these, \( n \) is generally known, being the number of binomial trials within a set. The remaining two, \( l \) and \( m \), are generally unknown and have to be estimated.

\((a)\) The method of moments

Let a random sample of \( N \) sets of \( n \) trials each be available, the number of sets giving \( r \) successes being \( F_r \), \((r = 0, \, 1, \, 2, \ldots, \, n)\). The first two sample moments are

\[
\nu_1' = \sum_{r=0}^{n} rF_r/N
\]

and

\[
\nu_2 = \sum_{r=0}^{n} (r - \nu_1')^2 F_r/N.
\]

Equating these moments to the corresponding population moments (5) and (6), we get the following estimates of \( l \) and \( m \).

\[
l = \frac{\nu_1'(n - \nu_1') - \nu_1'\nu_2}{n\nu_2 - \nu_1'(n - \nu_1')}, \quad (11)
\]

\[
m = \frac{\nu_1'(n - \nu_1')^2 - (n - \nu_1')\nu_2}{n\nu_2 - \nu_1'(n - \nu_1')}. \quad (12)
\]

\((b)\) The method of maximum likelihood

The maximum likelihood estimates have to be obtained by an iterative process. Using the multinomial distribution, the likelihood of the sample can be written as

\[
L = N! \prod_{r=0}^{n} \frac{[P(r)]^{F_r}}{F_r!}.
\]

Substituting the expression (2) for \( P(r) \) and extracting the logarithm, one gets

\[
\log L = \sum_{r=0}^{n} F_r \left[ \sum_{i=0}^{r-1} \log (l+i) + \sum_{j=0}^{n-r-1} \log (m+j) \right] - N \sum_{k=0}^{n-1} \log (l+m+k) + Q, \quad (13)
\]
where $Q$ comprises terms independent of $l$ and $m$. Differentiation of (13) with respect to $l$ and $m$ yields the following maximum likelihood equations (for the remainder of this section, the summation limits are the same as in (13)):

\[ S_i = \frac{\partial}{\partial l} \log L = \sum_r F_r \sum_i (l+i)^{-1} - N \sum_k (l+m+k)^{-1}, \quad (14) \]

\[ S_m = \frac{\partial}{\partial m} \log L = \sum_r F_r \sum_j (m+j)^{-1} - N \sum_k (l+m+k)^{-1}. \quad (15) \]

Equations (14) and (15), when $S_i$ and $S_m$ are equal to zero, are not explicitly solvable. An iterative process has therefore to be employed.

We define the quantities

\[ I_{ll} = -\frac{\partial^2}{\partial l^2} \log L = \sum_r F_r \sum_i (l+i)^{-2} - N \sum_k (l+m+k)^{-2}, \]

\[ I_{mm} = -\frac{\partial^2}{\partial m^2} \log L = \sum_r F_r \sum_j (m+j)^{-2} - N \sum_k (l+m+k)^{-2} \]

and $I_{lm} = -\frac{\partial^2}{\partial l \partial m} \log L = - N \sum_k (l+m+k)^{-2}.$

Starting with trial values $l_0, m_0$ of $l$ and $m$, the corrections $\delta l_0$ and $\delta m_0$ to be applied to them are obtained as

\[ \delta l_0 = (I_{mm} S_i - I_{lm} S_m) / \Delta, \quad (16) \]

\[ \delta m_0 = (-I_{lm} S_i + I_{ll} S_m) / \Delta, \quad (17) \]

where $\Delta = (I_{ll} I_{mm} - I_{lm}^2)$ and $l$, $m$ are replaced by $l_0$, $m_0$ in all the quantities occurring in (16) and (17).

This procedure is continued until convergence is obtained. The estimates (11) and (12) of $l$ and $m$ by the method of moments may form a convenient starting-point of the calculations. The computations involved are, however, laborious. These can be considerably simplified when $l$ and $m$ are large. For, we can then write

\[ [l(r)] = l(l+1) \ldots (l+r-1) \div [l+(r-1)/2]^r, \]
with corresponding approximations for $m_{[n-r]}$ and $(l+m)_{[n]}$. The logarithm of the likelihood (13) changes to

$$\log L = \sum_{r=0}^{n} F_r \{ r \log [l+(r-1)/2] + (n-r) \log [m+(n-r-1)/2] \} - nN \log [l+m+(n-1)/2] + \mathcal{Q}. \quad (18)$$

The likelihood equations (14) and (15) are modified as

$$S_l = \frac{\partial}{\partial l} \log L = \sum_{r} \frac{rF_r}{l+(r-1)/2} - \frac{nN}{l+m+(n-1)/2} = 0, \quad (19)$$
and

$$S_m = \frac{\partial}{\partial m} \log L = \sum_{r} \frac{(n-r)F_r}{m+(n-r-1)/2} - \frac{nN}{l+m+(n-1)/2} = 0. \quad (20)$$

These, although not explicitly solvable, are much simpler than equations (14) and (15). The iterative procedure [equations (16) and (17)] can still be used, with the following simpler expressions for $I_{ll}$, $I_{mm}$ and $I_{lm}$.

$$I_{ll} = \frac{\partial^2}{\partial l^2} \log L = \sum_{r} \frac{rF_r}{[l+(r-1)/2]^2} - \frac{nN}{[l+m+(n-1)/2]^2},$$

$$I_{mm} = \frac{\partial^2}{\partial m^2} \log L = \sum_{r} \frac{(n-r)F_r}{[m+(n-r-1)/2]^2} - \frac{nN}{[l+m+(n-1)/2]^2},$$

$$I_{lm} = \frac{\partial^2}{\partial l \partial m} \log L = - \frac{nN}{[l+m+(n-1)/2]^2}. $$

In any given situation, the estimates of $l$ and $m$ by the method of moments will suggest whether or not to use these simpler likelihood equations and the iterative procedure.

**RELATION BETWEEN NUMBERS OF SUCCESSES IN SUCCESSIVE SETS**

Let the number of successes in a series of $n$ trials on an individual be observed to be $s$. The posterior probability distribution of $p$ is then

$$f(p \mid s) = \frac{p^s (1-p)^{n-s} f(p)}{\int_0^1 p^s (1-p)^{n-s} f(p) \, dp} = \frac{p^{l+s-1} (1-p)^{m+n-s-1}}{B(l+s, m+n-s)},$$
using the beta form for \( f(p) \). Thus \( f(p | s) \) is of the same form as \( f(p) \) with \( l \) and \( m \) replaced by \((l+s)\) and \((m+n-s)\) respectively. Hence if \( r \) is the number of successes in a further series of \( n' \) trials,

\[
E(r | s) = \frac{n'(l+s)}{l+m+n} \quad \text{(using (5))},
\]

which is a linear function of \( s \). An exactly similar result holds for the negative binomial distribution obtained earlier as a modification of the Poisson distribution (Johnson & Garwood 1957). For this last case, the variance of \( r \) given \( s \) is also a linear function of \( s \), whereas in the present case, we have

\[
\text{var}(r | s) = \frac{n'(l+s)(m+n-s)(l+m+n+n')}{(l+m+n)^2(l+m+n+1)} \quad \text{(from (6))},
\]

which is a quadratic function of \( s \).

### NUMERICAL EXAMPLE

The data are taken from Moore (1958). In a biochemical experiment twenty insects were put in each of 100 jars. After being subject to a fumigant for 3 hr. the number alive in each jar was counted. The resulting frequency distribution is set out in the first two columns of Table 1. The first two sample moments are:

\[
\nu_1' = 4.39 \quad \text{and} \quad \nu_2 = 4.9379.
\]

If the chance \( p \) that an insect survives the fumigant remains constant both within and between jars, the observed frequency distribution should be well approximated by the usual binomial distribution. Since \( n \) equals 20, the estimate of \( p \) from the sample is

\[
\nu_1'/n = 4.39/20 = 0.2195.
\]

This value of \( p \) was used to obtain the expected frequencies shown in column no. 3. The value of \( \chi^2 \) with 5 degrees of freedom is 12.011. Since the probability of obtaining a higher value of \( \chi^2 \) is only about 0.03, the simple binomial model has to be rejected.

The hypothesis that \( p \) may vary from jar to jar is examined next. Substituting the values of \( \nu_1' \) and \( \nu_2 \) in the equations (11) and (12), the following estimates of \( l \) and \( m \) are obtained by the method of moments:

\[
l = 9.2, \quad m = 32.8.
\]
MODIFICATION OF THE BINOMIAL DISTRIBUTION

Table 1

<table>
<thead>
<tr>
<th>No. alive</th>
<th>No. of jars</th>
<th>Ordinary binomial</th>
<th>Modified binomial</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>0.704*</td>
<td>1.835*</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>3.931*</td>
<td>6.518*</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>10.576*</td>
<td>12.434</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>17.846</td>
<td>16.778</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>21.330</td>
<td>17.826</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>19.196</td>
<td>15.753</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>13.496</td>
<td>11.949</td>
</tr>
<tr>
<td>7</td>
<td>11</td>
<td>7.591</td>
<td>7.931</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>5.330</td>
<td>8.976</td>
</tr>
<tr>
<td>10 and over</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100.000</td>
<td>100.000</td>
</tr>
</tbody>
</table>

[Note. Frequencies marked with asterisks were pooled for the purpose of calculating $\chi^2$.]

Using these values in the expression (1) for $P(r)$, the expected frequencies shown in column (4) were obtained. It is evident that these are in good agreement with observed frequencies. The value of $\chi^2$ with 5 degrees of freedom is 2.879, and the probability of exceeding such a value is about 0.70.

Fitting by the method of maximum likelihood was also carried out. Starting with the trial values

$$l_0 = 9.2, \quad m_0 = 32.8,$$

the successive corrections were obtained as

$$(-1.34, -4.83), (0.37, 1.31), (0.02, 0.09), (0.01, 0.02), (0.00, -0.01).$$

The maximum likelihood estimates are therefore

$$l = 8.26, \quad m = 29.38.$$

The expected frequencies based on these estimates of $l$ and $m$ are shown in column (5). The value of $\chi^2$ with 5 degrees of freedom is 2.506 and the probability of observing a higher value is about 0.78.
The modified binomial distribution thus provides a good fit to the observed data.

In the present example, there is very little difference between the results of fitting by moments and fitting by maximum likelihood. Such a situation may not always obtain and for theoretical reasons it may sometimes be more desirable to follow the maximum likelihood method rather than the moments approach.

ACKNOWLEDGEMENT

The author is grateful to S. R. Adke for some useful suggestions and to D. V. Gokhale for help in computations.

REFERENCES