A Generalized Binomial Model with Possible Applications to Environmental Data

P. Vellaisamy
Department of Mathematics
Indian Institute of Technology, Bombay
Mumbai-400 076, India

Let $\{X_i\}$, $1 \le i \le n$, be a sequence of Bernoulli variables and

$$S_n = \sum_1^n X_i.$$

It is well known that $S_n \sim B(n, p)$, when

(a) X_i's are independent, and

(b) X_i's are identical, i.e., $P(X_i = 1) = p$ for all i.

When X_i's are independent with $P(X_i = 1) = p_i$, $1 \le i \le n$, ((b) is violated) then S_n is said to follow Poisson-binomial. The pmf of Poisson - binomial distribution is

$$P(S_n = k) = \sum_{\sum x_i = k} \prod_{i=1}^n p_i^{x_i} q_i^{1-x_i} \quad 0 \le k \le n.$$

(Samuels (1965) ; Wang (1993)).

1. An extension of the Binomial Model

Note that

$$\begin{aligned} P(S_n = x) &= P(X_n = 1 | S_{n-1} = x-1) P(S_{n-1} = x-1) \\ &\quad + P(X_n = 0 | S_{n-1} = x) P(S_{n-1} = x). \end{aligned} \tag{1.1}$$

Woodbury (1945) considered the case where both $P(X_n = 1|S_{n-1} = x-1)$ and $P(X_n = 0|S_{n-1} = x)$ are functions of x alone.
Rutherford (1954) considered the special case where $P(X_n = 1|S_{n-1} = x) = a + bx$, with certain conditions on a and b, among others.

Recently, Drezner and Farnum (DF) (1993) considered

$$P(X_n = 1|S_{n-1} = x-1) = (1-\theta_n)p + \theta_n \left(\frac{x-1}{n-1}\right),$$

$$P(X_n = 0|S_{n-1} = x) = (1-\theta_n)(1-p) + \theta_n \left(\frac{n-1-x}{n-1}\right),$$

where $P(X_1 = 1) = p$; $\theta_1 = 0$, $\theta_i, 2 \le i \le n$ are such that the above quantities are probabilities.

The probabilities depend both on n and x.
They discussed various practical applications where the above model fits better than the usual binomial model.
Their analysis of the model involves a tedious algebra and $E(S_n)$ and $V(S_n)$ (for equal θ_i's) requires a number of of lemmas.

Questions 1:

(a) Is there any simple approach?

(b) What are $P(X_i = 1)$ and θ_i's ?

A new approach to the the distribution of S_n.

The following results are from Vellaisamy (1996).

Lemma 2.1. Let $X_1, \cdots, X_n$ be any Bernoulli variables, $S_0 = 0$, and $S_k = \sum_{j=1}^{k} X_j$. Then the distribution of S_k is completely known iff $P(X_k = 1|S_{k-1})$ is known, for $1 \leq k \leq n$.

One way of studying the distribution of S_k is through conditional distributions of X_j given $S_{j-1}, 1 \leq j \leq k$.

This approach is much more efficient and also leads to new probabilistic models for analyzing dependent Bernoulli variables.

Question 2 :

Does independence of X_k and S_{k-1}, for $1 \leq k \leq n$, imply the independence of $X_1, \ldots, X_n$?

The Answer is 'No'

A counter example follows:

Example 2.1. Let X_1 and X_2 be iid Bernoulli variables with success probability p. Let X_3 be such that

$P(X_3 = 1|X_1 = 0; X_2 = 0)$

$$= P(X_3 = 1 | X_1 = 1, X_2 = 1) = p;$$

$P(X_3 = 1 | X_1 = 0, X_2 = 1) = a;$ and

$P(X_3 = 1 | X_1 = 1; X_2 = 0) = 2p - a,$

where $\max\{0, (2p-1)\} < a < min\{1, 2p\}$, and $a \neq p$.

Facts:

$a) P(X_3 = 1) = p,$

$b) P(X_3 = 1 | S_2 = j) = p$, for $j = 0, 1, 2;$

$\Longrightarrow X_3$ and S_2 are independent.

But, X_1, X_2 and X_3 are not independent unless $a = p$.

A Characterization of the $B(n, p)$

Theorem 1.1. For $1 \leq k \leq n$,

$S_k \sim B(k, p)$ iff

$$P(X_k = 1 | S_{k-1}) = p, \text{ for all } 1 \leq k \leq n.$$

An important consequence : $B(n, p)$ arises also as the distribution of sum of dependent (but identical) Bernoulli variables.

We will return to related questions later.

Passing Remark: Poisson-binomial distribution also arises from the model

$$P(X_k = 1 | S_{k-1}) = p_k, \text{ for all } 1 \leq k \leq n.$$

Note that DF's (1993) model corresponds to the form

$$P(X_i = 1|S_{i-1}) = (1-\theta_i)p + \frac{\theta_i}{(i-1)}S_{i-1}. \qquad (3.2)$$

Answer to Question 1 (b):

Lemma 1.2. Let X_k's, $1 \le k \le n$, be Bernoulli variables as in (3.2). Then, for $1 \le k \le n$,

(*i*) $E(S_k) = kp$.

(*ii*) $P(X_k = 1) = p$; hence X_k's are identical.

(iii) The parameters θ_k's are given by

$$\theta_k = (k-1)\rho_k C_k,$$

where $\rho_k = Corr(S_{k-1}, X_k)$ and $C_k = \sigma(X_k)/\sigma(S_{k-1})$.
Result (i) above for the case $k = n$ is Theorem 2 in DF(1993). Our proof is much simpler.

Theorem 1.2. For X_i's satisfying (3.2)

$$V(S_n) = \left\{1 + \sum_{j=2}^{n}\prod_{k=0}^{n-j}\left(1 + 2\frac{\theta_{n-k}}{n-k-1}\right)\right\}pq. \qquad (1.2)$$

The proof requires only a few steps.

Extension to Non-identical case is also simple.

Note that Theorem 1.2. implies $S_k \sim B(k,p)$, $1 \le k \le n$, iff

(i) X_i is independent of S_{i-1} $(2 \le i \le n)$, and

(ii) X_i's are identical with $P(X_i = 1) = p$.

Questions 3:

(a) Is it possible to relax the condition (ii) above?

(b) Does the $B(n,p)$ arise as the distribution of the sum of dependent and non-identical Bernoulli variables?

(c) Is there a complete characterization of $B(n,p)$?

questions are addressed.

2. The Nature of the $B(n,p)$

The following results are from Vellaisamy and Punnen (1999).

Lemma 2.1 Let $X_1, \ldots, X_n$ be arbitrary Bernoulli variables, and $S_n = \sum_1^n X_i$.

Let $T_j = \sum_{1 \le i_1 < i_2 < \cdots < i_j \le n} E(X_{i_1} X_{i_2} \ldots X_{i_j})$. Then

$$P(S_n = k) = \sum_{j=k}^{n} (-1)^{j-k} \binom{j}{k} T_j, \tag{2.1}$$

for $k = 0, 1, \ldots, n$.

The above result, called "sieve formula", is well known (cf. Blom, Holst and Sandell (1994, p.30)).

A simple but an interesting characterization of $B(n,p)$ is:

Theorem 2.1 Under the conditions of Lemma 2.1, $S_n = B(n,p)$ iff $T_j = \binom{n}{j} p^j$, $1 \le j \le n$.

Remark 2.1. This result does not identify all the distributions of X_i's, which lead to $B(n,p)$. So $B(2,p)$ and $B(3,p)$ are analyzed in detail.

2.1 The case $B(2, p)$

Let X_1 and X_2 be any Bernoulli variables. Find all the X_1 and X_2 such that $S_2 = X_1 + X_2 \sim B(2, p)$.

Crucial fact : Joint distribution of (X_1, X_2) is completely determined by the vector (call it the pmf)

$(p_1, p_{2.1}, p_{2.0})$

$= (P(X_1{=}1), P(X_2{=}1|X_1{=}1), P(X_2{=}1|X_1{=}0))$.

By Theorem 2.1, enough to find X_1 and X_2 such that $T_1 = 2p$ and $T_2 = p^2$, that is, to find $(p_1, p_{2.1}, p_{2.0})$ satisfying

$$\begin{aligned} P(S_2 = 0) &= 1 - T_1 + T_2 = q^2 && (2.2)\\ \text{and} \qquad P(S_2 = 2) &= T_2 = p^2. && (2.3) \end{aligned}$$

Equation (2.3) implies $p_1 p_{2.1} = p^2$.

Fix now p_1 so that $p_{2.1} = p^2/p_1$. Using (2.2), we get

$$(1 - p_{2.0})(1 - p_1) = q^2$$

which yields $p_{2.0} = 1 - \frac{q^2}{q_1}$. The conditions

$p_{2.0} \geq 0$ and $p_{2.1} \leq 1$ lead to $p^2 \leq p_1 \leq 1 - q^2$ for p_1.

The above analysis can be strengthened as follows:

Lemma 2.2. Let X_1 and X_2 be any Bernoulli variables, and $P(X_1 = 1) = p_1$.Then $S_2 \sim B(2, p)$ iff

$$\begin{aligned} P(X_2 = 1|X_1 = 1) &= \frac{p^2}{p_1}, && (2.4)\\ P(X_2 = 1|X_1 = 0) &= 1 - \frac{q^2}{q_1}, && (2.5) \end{aligned}$$

where $p^2 \le p_1 \le 1-q^2$ and $q_1 = 1-p_1$.

Corollary 2.1 Let $P(X_i = 1) = p_1$, $i = 1, 2$. Then $S_2 \sim B(2,p)$ for some p, iff they are independent.

Remarks 2.1 Let $B_2(p)$ denote the set of all distributions that lead to $B(2,p)$. Then

$$B_2(p) = \left\{ (p_1, \frac{p^2}{p_1}, 1 - \frac{q^2}{q_1}) | p^2 \le p_1 \le 1 - q^2 \right\}. \tag{2.6}$$

The distribution (iid case) $(p,p,p) \in B_2(p)$.

Example 2.1. Observe that

$$B_2(\frac{1}{3}) = \left\{ (p_1, \frac{1}{9p_1}, 1 - \frac{4}{9q_1}) \middle| \frac{1}{9} \le p_1 \le \frac{5}{9} \right\},$$

which is an infinite set. So $B(2, \frac{1}{3})$ arises infinitely many ways. For example, $(\frac{1}{2}, \frac{2}{9}, \frac{1}{9}) \in B_2(\frac{1}{3})$

A simple mnemonic device : To check if $(p_1, p_{2.1}, p_{2.0})$ corresponds to $B(2,p)$:

(a) Find $p = (p_1 p_{2.1})^{1/2}$, and $q = 1 - p$.

(b) If $p_{2.0} = 1 - \frac{q^2}{q_1}$, then $S_2 \sim B(2,p)$.

Note $(\frac{1}{9}, \frac{1}{4}, \frac{1}{4}) \notin B(2,p)$, as (b) is violated.

2.2. The case $B(3,p)$

Same Approach : In addition to p_1, $p_{2.1}$, $p_{2.0}$, define $p_{3.ij} = P(X_3 = 1 | X_1 = i, X_2 = j)$. The distribution of (X_1, X_2, X_3) is determined by the vector

$$(s_1, s_2, s_3, s_4, s_5, s_6, s_7) = (p_1, p_{2.1}, p_{2.0}, p_{3.11}, p_{3.10}, p_{3.01}, p_{3.00}),$$

which we call the distribution of (X_1, X_2, X_3). Then,

$$\begin{aligned} B_3(p) \;=\; & \Big\{(s_1, s_2, s_3, \frac{p^3}{s_1 s_2}, s_5, \\ & \frac{p^2(1+2q) - s_1 s_2 - s_1(1-s_2)s_5}{(1-s_1)s_3}, \\ & 1 - \frac{q^3}{(1-s_1)(1-s_3)}\Big\}, \quad \text{where} \end{aligned}$$

(a) $0 < p^3 < s_1 s_2$;

(b) $0 < q^3 \leq (1-s_1)(1-s_3)$;

(c) $0 < p^2(1+2q) - s_1(s_2 + (1-s_2)s_5) \leq (1-s_1)s_5$.

(a)-(c) ensure that s_4 , s_6 and s_7 are probabilities.

Remarks 2.2. **(i)** A Simple Procedure:

(a) Compute $p = (s_1 s_2 s_4)^{1/3}$, and $q = (1-p)$.

(b) Check if

$$s_6 = \frac{p^2(1+2q) - s_1 s_2 - s_1(1-s_2)s_5}{(1-s_1)s_3} \tag{2.7}$$

(c)Check also if

$$s_7 = 1 - \frac{q^3}{(1-s_1)(1-s_3)}. \tag{2.8}$$

If (2.7) and (2.8) are satisfied, then $S_3 \sim B(3, p)$.

(ii) X_1, X_2 and X_3 are independent, and $S_3 \sim B(3, p)$, implies $B_3(p) = \{(p, \ldots, p)\}$ and hence identical.

Question 4: Do the identicalness and $S_3 \sim B(3,p)$ imply the independence?

Answer is NO unlike in the case $B(2,p)$.

Example 2.2. Let (X_1, X_2, X_3) have the distribution

$$(\frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{3}{4}, \frac{1}{5}, \frac{4}{5}, \frac{1}{4}).$$

$$\text{then} \qquad P(X_2 = 1) = s_1 s_2 + (1 - s_1)s_3 = \frac{1}{2}$$

$$\begin{aligned} P(X_3 = 1)(1 - s_1)\{s_3 s_6 + &= p^3 + p^2(1+2q) - s_1 s_2 + (1 - s_1)(1 - s_3) - q^3 \\ &= \frac{1}{2} \quad \text{(hence identical)} \end{aligned}$$

Also, $S_3 \sim B(3, \frac{1}{2})$, as (2.7) and (2.8) are satisfied. But X_1, X_2 and X_3 are not independent.

In Example 2.2, X_i's are identical and $S_3 \sim B(3, p_1)$. However, identicalness is not necessary to have $B(3,p)$ with $p = p_1$.

Example 2.3. Consider the distribution

$$\left(\frac{1}{3}, \frac{4}{9}, \frac{1}{3}, \frac{1}{4}, \frac{1}{3}, \frac{2}{9}, \frac{1}{3}\right)$$

Then $S_3 \sim B(3,p)$ with $p = p_1$. But X_i's are not identical, as $P(X_2 = 1) = \frac{10}{27}$ and $P(X_3 = 1) = \frac{8}{27}$.

Example 2.4. Let (X_1, X_2, X_3) have distribution

$$(\frac{1}{3}, \frac{1}{6}, \frac{1}{4}, \frac{9}{32}, \frac{3}{10}, \frac{5}{48}, \frac{5}{32}).$$

Then, $S_3 \sim B(3, \frac{1}{4})$ and $P(X_2 = 1) = \frac{2}{9}$ and

$P(X_3 = 1) = \frac{7}{36}$.

This is interesting, as X_i's are neither identical nor independent and also $p \neq p_1$.

3. The General Case

First a result showing the connection between the binomial distributions and the Poisson process.

The 'if' part of the following characterization of Poisson model is not known.

Lemma 3.1. Let $\{X_i\}_{i\geq 1}$ be a sequence of Bernoulli variables, and $\{N(t)\}$, independent of the X_i's, be a Poisson process with rate $\lambda > 0$. Then $S_{N(t)} = \sum_{i=1}^{N(t)} X_i$ follows $P(\lambda pt)$ iff $S_n \sim B(n, p)$ for every $n \geq 1$.

Question 4 : When $S_k \sim B(k, p)$ for every $k \geq 1$?

Answer : A slight modification of Theorem 1.1.

Theorem 3.1. (Vellaisamy, 1996) For $k \geq 1$, $S_k \sim B(k, p)$ iff $P(X_k = 1|S_{k-1}) = p$ for every $k \geq 1$.

Lemma 3.1 and Theorem 3.1 leads to

Corollary 3.1. Under the conditions of Lemma 3.1, $S_N \sim P(\lambda p)$ iff $P(X_i = 1|S_{i-1}) = p$ for every i.

Implication: Poisson distribution could arise as the distribution of a random sum of dependent Bernoulli variables.

Example 3.1 Consider the distribution

$$(\frac{1}{2}, \frac{2}{3}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{1}{5}, \frac{1}{4}).$$

Then, X_i's are identical with $P(X_i = 1) = \frac{1}{2}$, and $S_3 \sim B(3, \frac{1}{2})$. But S_2 does not follow $B(2, p)$, as X_i is not independent of S_{i-1}, $1 \leq i \leq 3$.

So, when we observe a sequence of Bernoulli variables, distribution then $B(n, p)$ could arise at any stage. of S_n. stage.

Difficult to extend the approach used in earlier sections for $n \geq 4$. For example, one has to deal with a vector of 15 coordinates to denote an arbitrary joint distribution of four Bernoulli variables. So, we adopt a slightly different method based on the conditional distribution of X_n given S_{n-1}. Such models occur in the analysis of shock models in reliability theory. DF's(1993) model is another example. Recently, we have used these models for modelling dependent production processes. These models could also be helpful in analyzing environmental data.

Let $d(j) = P(S_{n-1} = j)$ and $D(j) = P(S_{n-1} \leq j)$, $0 \leq j \leq n-1$. Similarly, let $b(j)$ and $B(j)$ respectively denote the pmf and cdf of $B(n, p)$.

Theorem 3.2. Let $X_1, \ldots, X_n$ be any sequence of Bernoulli variables such that $0 < D(k) - B(k) \leq d(k)$ for $1 \leq k \leq n-1$.

Then $S_n \sim B(n,p)$ iff

$$P(X_n = 1|S_{n-1} = k)d(k) = D(k) - B(k), \tag{3.1}$$

for every $k \in \{0, 1, \ldots, n-1\}$.

Remarks 3.1. (i) Let $d(k)$, $0 \leq k \leq n-1$, be any distribution of S_{n-1} such that ($D_l = 0$ for $l < 0$)

$$B_k - D_{k-1} < d_k < B_{k+1} - D_{k-1},\ 0 \leq k \leq n-2, \tag{3.2}$$

and $d_{n-1} = 1 - \sum_{i=0}^{n-2} d_i$.

Let $P(X_n = 1|S_{n-1} = k) = c(k)$, say, satisfy

$$c(k)d(k) = D(k) - B(k) = c(k-1)d(k-1) + d(k) - b(k)$$

for $0 \leq k \leq n-2$, and $c(n-1)d(n-1) = b(n)$. Then by Theorem 3.2, $S_n \sim B(n,p)$.

(ii) As an example, let $d(0)$ be any real with
$B(0) < d(0) < B(1)$, and

$$d(k) = B(k) - D(k-1) + \binom{n-1}{k+1} p^{k+1} q^{n-k-1},$$

for $0 \leq k \leq n-2$. This choice satisfies (3.2).

4. Identical or Independent Summands

Lemma 4.1. Let $X_1, \ldots, X_n$ be identical Bernoulli variables with $P(X_1 = 1) = p_1$. If $S_n \sim B(n,p)$, then $p = p_1$.

The proof is trivial. The result holds for exchangeable rv's also.

Finally, the case of independent Bernoulli variables with $P(X_i = 1) = p_i, \;\; 1 \le i \le n$

For completeness, we state the following results.

Lemma 4.1. Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli variables with $P(X_i = 1) = p_i$, $1 \le i \le n$.
Then $S_n \sim B(n,p)$ iff $p_1 = p_2 = \cdots = p_n = p$.

Theorem 4.1. Let $Y_1, \ldots, Y_k$ be independent binomial random variables, $Y_i \sim B(n_i, p_i)$, and $n = \sum_1^k n_i$. Then $S_k = \sum_1^k Y_i \sim B(n,p)$ iff $p_1 = \ldots = p_k = p$.

5. Concluding Remarks

In the study of
$S_n = \sum_1^n X_i$, it is commonly assumed that X_i's are independent, even though the underlying physical situation may or may not support it. This very assumption of independence, to avoid statistical complexity, has led us to a very narrow or little understanding of the binomial distribution.

As seen earlier, the infinite sets $B_2(p)$ and $B_3(p)$ reduces to the singleton set.

Moreover, when $n = 3$, for example, the set $B_3(p)$ is characterized by seven parameters (probabilities) out of which three have to satisfy certain conditions.

In fact, for a general n, the set $B_n(p)$ is determined by $(2^n -$

1)-dimensional vectors of probabilities and only n coordinates have to satisfy n conditions and the remaining $(2^n - 1 - n)$ coordinates could be arbitrary probabilities.

Hence, for large n, the distribution of S_n is quite likely to follow or to be close to the binomial distribution.

Finally, by Poisson's theorem, it is tempting to conclude that the utility of the Poisson model in a variety of situations dealing with Bernoulli summands (see, Barbour, Holst and Janson (1992)) is partly due to the nature of the binomial distribution.

References

Barbour, A.D, Holst, L.and Janson, S. (1992) *Poisson Approximation.* Oxford University Press, Oxford.

Blom, G., Holst,L. and Sandell, D. (1994). *Problems and Snapshots from the World of Probability*, Springer - Verlag, New York.

Drezner, Z. and Farnum, N. (1993). A generalized binomial distribution. *Commun. Statist.* - *Theory Meth.,* 22, 3051-3063

Edwards, A.W.F. (1960). The meaning of binomial distribution. *Nature,* 186, 1074.

Feller, W. (1968). *An Introduction to Probability Theory and Its*

Applications, **Vol. I, Third Edition, John Wiley & Sons, New York.**

Johnson, N.L, Kotz, S. and Kemp, A.W. (1992). *Univariate Discrete Distributions*. **John Wiley & Sons, New York.**

edition, Addision- Wesley,

Nedelman, J. and Wallenius, T. (1986). Bernoulli trials, surprising variances, and Jensen's inequality. *Amer. Statistician*, **40, 286-289.**

Rutherford, R.S.G. (1954). On a contagious distribution. *Ann. Math. Stat.,* **36, 1272-1278.**

Samuels, S.M. (1965). On the number of successes in independent trials. *Ann. Math. Stat.,* **36, 1272-1278.**

Vellaisamy, P. (1996). On the number of successes in dependent trials. *Commun. Statist. - Theory Meth.,* **25, 1745-1756.**

Vellaisamy, P and Punnen, A. P. (1999). On the nature of the binomial distribution. To appear in *Journal of Applied Probability*.

Wang, Y.H. (1993). On the number of successes in independent trials. *Statistica Sinica,* **3, 295-312.**

Woodbury, M.A. (1949). On a probability distribution. *Ann. Math. Stat.,* **20, 311-313.**