J. Appl. Prob. **46,** 1073–1085 (2009)
Printed in England

A THREE-PARAMETER BINOMIAL APPROXIMATION

EROL A. PEKÖZ,* ** *Boston University*
ADRIAN RÖLLIN,*** *National University of Singapore*
VYDAS ČEKANAVIČIUS,**** *Vilnius University*
MICHAEL SHWARTZ,* *Veterans' Health Administration, Boston, and Boston University*

Abstract

We approximate the distribution of the sum of independent but not necessarily identically distributed Bernoulli random variables using a shifted binomial distribution, where the three parameters (the number of trials, the probability of success, and the shift amount) are chosen to match the first three moments of the two distributions. We give a bound on the approximation error in terms of the total variation metric using Stein's method. A numerical study is discussed that shows shifted binomial approximations are typically more accurate than Poisson or standard binomial approximations. The application of the approximation to solving a problem arising in Bayesian hierarchical modeling is also discussed.

Keywords: Stein's method; Poisson binomial distribution; binomial approximation

2000 Mathematics Subject Classification: Primary 60E15
Secondary 60G50

1. Introduction

A common method for improving the accuracy of an approximation is the construction of an asymptotic expansion. In practice, however, this can be more time consuming and much less convenient than calculating the values of a known distribution. An alternative approach is thus to modify a common approximating distribution by introducing some new parameters which then can be used to achieve a better fit. The use of common distributions can make it easy to avoid the need for specialized programing when using standard statistical packages to model data.

One of the simplest modifications is shifting, and this approach works well in the Poisson case. For a large number of independent rare events, the distribution of the number of them that occur is often well approximated by a Poisson distribution. If some of the events are not in fact so rare, this approximation is likely to be poor: the expected number of events occurring may not be close to the variance, but these are equal for the Poisson distribution. One easy way to address this problem is to introduce a shift by adding or subtracting a constant from the Poisson random variable. This then gives essentially two parameters that can be fitted to match the first two moments (subject to the constraint that the shift is an integer). Shifted (also referred to

Received 4 March 2008; revision received 22 July 2009.
* Postal address: Boston University School of Management, 595 Commonwealth Avenue, Boston, MA 02215, USA.
** Email address: pekoz@bu.edu
*** Postal address: National University of Singapore, 2 Science Drive 2, 117543, Singapore.
**** Postal address: Vilnius University, Faculty of Mathematics and Informatics, Naugarduko 24, Vilnius LT-03223, Lithuania.

as translated or centered) Poisson approximations have been studied in many papers; see, for example, Čekanavičius and Vaĭtkus (2001), Barbour and Čekanavičius (2002), Röllin (2005), Barbour and Lindvall (2006), and the references therein.

One of the goals of this paper is to investigate the effect of shifting applied to a two-parameter distribution. It is clear that shifting changes a distribution's mean but not its variance and higher centered moments. Can we expect by shifting to conveniently obtain a three-parameter distribution and match three corresponding moments? In the case of the normal approximation, the obvious answer is no. The normal distribution already has a parameter that can be treated as shifting. Since both parameters of two-parameter distributions are usually closely related to their first two moments, it seems important to show that there are natural cases where shifting can be successfully applied. Below we use a shifted (centered, translated) binomial approximation for the sum of Bernoulli variables. Our primary interest for the statistical application we consider is in the case when the variables are independent.

In the literature, the distribution of the sum of independent Bernoulli random variables with not necessarily identical probabilities is called a Poisson binomial distribution. This distribution is widely applicable and widely studied, and bounds on approximation errors for various approximations have been developed. See Chen and Liu (1997) for an overview of the Poisson binomial distribution and Pitman (1997) for applications, as well as Le Cam (1960) and Barbour *et al.* (1992b, pp. 33–40) for some Poisson approximation results. A number of researchers have studied the binomial distribution as an approximation for the Poisson binomial distribution. For example, Choi and Xia (2002) argued that binomial approximations are better than Poisson approximations.

Before discussing some previously obtained results, we need to introduce some necessary notation. Let $X_1, \dots, X_m$ be independent Bernoulli random variables with $\mathrm{P}(X_i = 1) = p_i$, $W = \sum_{i=1}^m X_i$. Let

$$\lambda_j = \sum_{i=1}^m p_i^j, \quad j = 1, 2, \dots, \qquad \sigma^2 = \operatorname{var} W = \lambda_1 - \lambda_2.$$

The total variation metric distance between two random variables X and Y is defined as

$$d_{\mathrm{TV}}(\mathcal{L}(X), \mathcal{L}(Y)) = \sup_A |\mathrm{P}(X \in A) - \mathrm{P}(Y \in A)|,$$

where the supremum is taken over all Borel sets. Note that if X and Y are integer valued then $d_{\mathrm{TV}}(X, Y) = \frac{1}{2} \sum_{i \in \mathbb{Z}} |\mathrm{P}(X = i) - \mathrm{P}(Y = i)|$. We also define a local metric

$$d_{\mathrm{loc}}(\mathcal{L}(X), \mathcal{L}(Y)) = \sup_{j \in \mathbb{Z}} |\mathrm{P}(X = j) - \mathrm{P}(Y = j)|.$$

The notation $\lfloor \cdot \rfloor$ and $\{\cdot\}$ are used for integral and fractional parts, respectively.

Ehm (1991) gave results for the binomial approximation where the number of trials equals the number of Bernoulli variables and the success probability is chosen to match the first moment. More precisely,

$$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Bi}(m, p)) \le \frac{1 - p^{m+1} - (1-p)^{m+1}}{(m+1)p(1-p)} \sum_{i=1}^m (p_i - p)^2, \tag{1.1}$$

where $p = \lambda_1/m$. Thus, the binomial approximation here is a one-parameter approximation. Ehm's approach was later extended to a Krawtchouk asymptotic expansion in Roos (2000).

Barbour *et al.* (1992b, p. 190) treated the binomial distribution as a two-parameter approximation. Their result was improved in Čekanavičius and Vaĭtkus (2001, Section 4), who showed that

$$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Bi}(n, p)) \leq \frac{4}{1-p} \min\left(1, \frac{\sqrt{\mathrm{e}}}{\sigma}\right)\left(\frac{\lambda_3}{\lambda_1} - \frac{\lambda_2^2}{\lambda_1^2}\right) + \frac{\lambda_2\{\lambda_1^2/\lambda_2\}}{\lambda_1(1-p)n} + \mathrm{P}(W > n). \quad (1.2)$$

Here $n = \lfloor \lambda_1^2/\lambda_2 \rfloor$ and $p = \lambda_1/n$. Note that Čekanavičius and Vaĭtkus (2001) (as well as Barbour *et al.* (1992b) and Soon (1996)), in formulations of their results, overlooked the term $\mathrm{P}(W > n)$, which is necessary because the support of W is typically larger than the support of the approximating binomial distribution.

It is easy to see that both estimates (1.1) and (1.2) are small if all the p_i are close to each other. On the other hand, the second estimate can be sharper than the first one. Indeed, let $p_i = \frac{1}{2}$ for $i \leq m/2$ and $p_i = \frac{1}{3}$ otherwise. Then the right-hand side of (1.1) equals some absolute constant C_1, meanwhile the right-hand side of (1.2), after application of Chebyshev's inequality, becomes $C_2 m^{-1/2}$.

Note that two-parameter binomial approximations are also applied in settings with dependence; see Soon (1996) and Čekanavičius and Roos (2007). Röllin (2008) used a shifted $\mathrm{Bi}(n, \frac{1}{2})$ to approximate sums of locally dependent random variables.

In this paper we study the shifted binomial approximation where the shift, the number of trials, and the success probability are selected to match the first three moments of the shifted binomial and the Poisson binomial distributions. We then give an upper bound on the approximation error by adapting Stein's method to the shifted binomial distribution. This is, to the best of the authors' knowledge, the first time Stein's method has been used to approximate a distribution that fits the first three moments. We also discuss the results of a numerical study showing that a shifted binomial approximation is typically more accurate than the Poisson or other standard binomial approximations discussed in Soon (1996) and Ehm (1991).

At the end of the paper we describe the motivating statistical application in healthcare provider profiling that led to the need for a more accurate approximation. See Peköz *et al.* (2009) for more detail on the application. An introduction to the use of Bayesian hierarchical models for healthcare provider profiling can be found in Ash *et al.* (2003).

Stein's method was introduced in the context of a normal approximation in Stein (1972) and developed for the Poisson distribution in Chen (1974) and Chen (1975). The method is particularly interesting since results in the complex setting of dependent random variables are often not much more difficult to obtain than results for independent variables. Barbour *et al.* (1992b) detailed how the method can be applied to Poisson approximations, Ehm (1991) and Loh (1992) respectively applied the method to binomial and multinomial approximations, Barbour *et al.* (1992a) and Barbour and Chryssaphinou (2001) to the compound Poisson approximation, Barbour and Brown (1992) to the Poisson process approximation, and Peköz (1996) to the geometric approximation. A discussion of the many other distributions and settings that the technique can be applied to can be found in, for example, Barbour and Chen (2005) and Reinert (2005). An elementary introduction to Stein's method can be found in Chapter 2 of Ross and Peköz (2007).

This paper is organized as follows. In Section 2 we give the main approximation theorems by adapting Stein's method to the shifted binomial distribution and in Section 3 we prove these results. In Section 4 we discuss numerical results illustrating the accuracy of several approximations, and in Section 5 we discuss the statistical application in Bayesian hierarchical modeling that motivated our initial interest in this approximation.

2. Main results

Let Y be a shifted binomial random variable with parameters n, p, and integer shift s, that is,

$$Y \sim \mathrm{Bi}(n, p) * \delta_s, \tag{2.1}$$

where '$*$' denotes convolution of measures and δ_s is the measure with mass 1 at s. In this paper we study the approximation of W using Y with parameters n, p, and s chosen so that the first three moments of W and Y are approximately equal. Due to the integer natures of n and s, it will not always be possible to exactly match the three moments—so we match them as closely as possible. We first estimate these parameters, and then give a theorem bounding the approximation error. It is easy to check that

$$\begin{aligned} \mathrm{E}\, Y &= np + s, & \operatorname{var} Y &= np(1-p), & \mathrm{E}(Y - \mathrm{E}\, Y)^3 &= (1-2p)\operatorname{var} Y, \\ \mathrm{E}\, W &= \lambda_1, & \operatorname{var} W &= \lambda_1 - \lambda_2, & \mathrm{E}(W - \mathrm{E}\, W)^3 &= \lambda_1 - 3\lambda_2 + 2\lambda_3. \end{aligned}$$

In order to find the values n, p, and s that match the moments best under the constraint that n and s are integer valued, let us first solve the system of equations $\mathrm{E}\, W = \mathrm{E}\, Y$, $\operatorname{var} W = \operatorname{var} Y$, and $\mathrm{E}(W - \mathrm{E}\, W)^3 = \mathrm{E}(Y - \mathrm{E}\, Y)^3$ for real valued n^*, p^*, and s^*. The system of equations

$$\begin{aligned} s^* + n^* p^* &= \lambda_1, \\ n^* p^*(1-p^*) &= \lambda_1 - \lambda_2, \\ n^* p^*(1-p^*)(1-2p^*) &= \lambda_1 - 3\lambda_2 + 2\lambda_3, \end{aligned}$$

yields the solution

$$p^* = \frac{\lambda_2 - \lambda_3}{\lambda_1 - \lambda_2}, \qquad n^* = \frac{\lambda_1 - \lambda_2}{p^*(1-p^*)}, \qquad s^* = \lambda_1 - n^* p^*. \tag{2.2}$$

We now choose

$$n = \lfloor n^* \rfloor, \qquad s = \lfloor s^* \rfloor, \qquad p = \frac{n^* p^* + \{s^*\}}{n} = p^* + \frac{\{n^*\}p^* + \{s^*\}}{n}$$

(in the last expression we indeed divide by n and not by n^*), and then let Y be as in (2.1). Although p is real valued and therefore does not need any rounding correction with respect to p^*, a small perturbation is still necessary in order to fit the mean exactly, which is crucial to obtain better rates of convergence. For convenience, whenever we use a variable p (or p_i, p^*, etc.) to denote a probability, the variable q (or q_i, q^*, etc.) will denote the counter probability $1 - p$. Let $v = \sum_{i=1}^m (p_i \wedge q_i)$. Then our main result is the following.

Theorem 2.1. *Suppose that $X_1, \dots, X_m$ are independent Bernoulli random variables with $\mathrm{P}(X_i = 1) = p_i$. With the definitions above we have*

$$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Bi}(n, p) * \delta_s) \le K(4A_1 + 2A_2) + \eta, \tag{2.3}$$

where

$$\begin{aligned} K &= \frac{1 - p^{n+1} - q^{n+1}}{\sigma^2}, \qquad (2.4) \\ A_1 &= \frac{\sigma^2(\lambda_3 - \lambda_4) - (\lambda_2 - \lambda_3)^2}{\sigma^2\big(1 \vee (v/2 - 1)\big)}, \qquad A_2 = \frac{\lambda_1[\{n^*\} + \{s^*\}] + n\{s^*\}}{n}, \\ \eta &= \Big(s \max_{i \le s} p_i\Big) \wedge \mathrm{e}^{-\sigma^2/4} + ((m - n - s) \max_{i > n+s} p_i) \wedge \mathrm{e}^{-\sigma^2/4 + 1}. \end{aligned}$$

Furthermore,

$$d_{\mathrm{loc}}(\mathcal{L}(W), \mathrm{Bi}(n, p) * \delta_s) \le K(8A_3 + 4A_4) + \eta, \tag{2.5}$$

where

$$A_3 = \frac{\sigma^2(\lambda_3 - \lambda_4) - (\lambda_2 - \lambda_3)^2}{\sigma^2(1 \vee (v/3 - 2))^{3/2}}, \qquad A_4 = \frac{\lambda_1[\{n^*\} + \{s^*\}] + n\{s^*\}}{n(1 \vee (v-1))^{1/2}}.$$

If $p_1, p_2, \ldots, p_m$ are such that, for some fixed p, we have, for all i, either $p_i = p$ or $p_i = 1$, then W and Y have the same shifted binomial distribution and $d_{\mathrm{TV}}(\mathcal{L}(W), \mathcal{L}(Y)) = 0$. In this case, after omitting η, the right-hand sides of (2.3) and (2.5) also both equal 0. Dropping negative terms, using $(\lambda_3 - \lambda_4) \le \sigma^2 \le v$ and $1 \vee (av - b) \ge av/(1 + b)$, and replacing all fractional parts by unity, we obtain the following simplified bounds.

Corollary 2.1. *Under the conditions of Theorem 2.1, we have*

$$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Bi}(n, p) * \delta_s) \le \frac{17 + 2\lambda_1 n^{-1}}{\sigma^2} + 2\mathrm{e}^{-\sigma^2/4+1}$$

and

$$d_{\mathrm{loc}}(\mathcal{L}(W), \mathrm{Bi}(n, p) * \delta_s) \le \frac{222 + 12\lambda_1 n^{-1}}{\sigma^2 v^{1/2}} + 2\mathrm{e}^{-\sigma^2/4+1}.$$

It is clear from this corollary that, when $c < p_i < d$ for all i and some absolute constants c and d, the order of the upper bound on $d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Bi}(n, p) * \delta_s)$ is $O(n^{-1})$ while, for $d_{\mathrm{loc}}(\mathcal{L}(W), \mathrm{Bi}(n, p) * \delta_s)$, it is $O(n^{-3/2})$. Thus, we obtain a significant improvement over $O(n^{-1/2})$, which can be obtained by a two-parameter binomial approximation (1.2) or by a shifted $\mathrm{Bi}(n, \frac{1}{2})$ distribution as in Röllin (2008).

3. Proofs of the main results

If Stein's method for a normal approximation $N(0, \sigma^2)$ is applied to a random variable X, we typically need to bound the quantity

$$\mathrm{E}[\sigma^2 f'(X) - Xf(X)] \tag{3.1}$$

for some specific functions f, where X is assumed to be centered and $\operatorname{var} X = \sigma^2$. This corresponds to fitting the first two moments. If three moments have to be matched, we need a different approximating distribution and a canonical candidate would be a centered $\Gamma(r, \lambda)$ distribution. This would lead to bounding the quantity

$$\mathrm{E}[(r\lambda^{-2} + \lambda^{-1}X)f'(X) - Xf(X)] \tag{3.2}$$

(cf. Luk (1994, Equation (17))), where the parameters r and λ are chosen to fit the second and third moments of W, that is, $\operatorname{var} W = r\lambda^{-2}$ and $\mathrm{E}\, W^3 / \operatorname{var} W = 2\lambda^{-1}$ (this obviously is only possible if W is skewed to the right, which we can always achieve by considering either W or $-W$). We can see that (3.2) is in some sense a more general form of (3.1), having an additional parameter for skewness. On the integers, we can take a shifted binomial distribution as in this paper. Not surprisingly, the Stein operator for a binomial distribution, shifted to have expectation λ_1 (ignoring rounding problems), can be written in a way similar to (3.2); see (3.3), below. In the following lemma we give the basic arguments to show how to handle expressions of type (3.2) in the discrete case for sums of independent indicators where all the involved

parameters are allowed to be continuous. We will deal with the rounding problems in the main proof of Theorem 2.1.

We need some notation first. For any function g, define the operators $\Delta^k g(w) := \Delta^{k-1} g(w+1) - \Delta^{k-1} g(w)$ with $\Delta^0 g := g$ and $\Theta g(w) := (g(w+1) + g(w))/2$. Note that $\Theta\Delta = \Delta\Theta$. We introduce the operator Θ in order to present the Stein operator of the shifted binomial in a symmetrized form, so that the connection with (3.2) should become more apparent. For the choice $p^* = \frac{1}{2}$, the linear part in the Δg part will vanish, so that the operator indeed becomes symmetric, and, hence, corresponds to the symmetric distribution $\mathrm{Bi}(n^*, \frac{1}{2})$ shifted by $-n^*/2$.

Lemma 3.1. *Let W be defined as before, and let*

$$\hat{\mathcal{B}}^* g(w) := \big(n^* p^* q^* + \big(\tfrac{1}{2} - p^*\big)(w - \lambda_1)\big)\Delta g(w) - (w - \lambda_1)\Theta g(w). \tag{3.3}$$

Then, for n^* and p^* defined as in (2.2), we have, for any bounded function $g\colon \mathbb{Z} \to \mathbb{R}$,

$$\begin{aligned} \mathrm{E}\,\hat{\mathcal{B}}^* g(W) &= \sum_{i=1}^{m} (p^* - p_i) p_i^2 q_i \,\mathrm{E}\,\Delta^3 g(W_i) \\ &= \frac{1}{2\sigma^2} \sum_{i,j=1}^{m} p_i p_j q_i q_j (p_i - p_j)^2 \,\mathrm{E}\,\Delta^3 g(W_{ij}), \end{aligned}$$

where $W_i := W - X_i$ and $W_{ij} := W - X_i - X_j$.

Proof. It is easy to prove that, for any bounded function $h\colon \mathbb{Z} \to \mathbb{R}$, the following identities hold:

$$\mathrm{E}[(X_i - p_i)h(W)] = p_i q_i \,\mathrm{E}[\Delta h(W_i)], \tag{3.4}$$

$$\mathrm{E}[h(W) - \Theta h(W_i)] = -\big(\tfrac{1}{2} - p_i\big)\,\mathrm{E}\,\Delta h(W_i), \tag{3.5}$$

$$\mathrm{E}[h(W) - h(W_i)] = p_i \,\mathrm{E}\,\Delta h(W_i). \tag{3.6}$$

In what follows summation is always assumed to range over $i = 1, \ldots, m$. Using first (3.4) and then (3.5), we obtain

$$\begin{aligned} \mathrm{E}[(W - \lambda_1)\Theta g(W)] &= \sum (X_i - p_i)\Theta g(W) \\ &= \sum p_i q_i \,\mathrm{E}\,\Delta\Theta g(W_i) \\ &= \sum p_i q_i \,\mathrm{E}\,\Delta g(W) + \sum p_i q_i \big(\tfrac{1}{2} - p_i\big)\,\mathrm{E}\,\Delta^2 g(W_i). \end{aligned}$$

From (3.4) we also deduce that

$$\mathrm{E}[(W - \lambda_1)\Delta g(W)] = \sum p_i q_i \,\mathrm{E}\,\Delta^2 g(W_i).$$

Combining these two identities and recalling that $n^* p^* q^* = \lambda_1 - \lambda_2$,

$$\mathrm{E}\,\hat{\mathcal{B}}^* g(W) = \sum p_i q_i (p_i - p^*)\,\mathrm{E}\,\Delta^2 g(W_i).$$

Applying (3.6) and noting that $\sum p_i q_i (p_i - p^*) = 0$ proves the first equality. For the second

equality, we proceed with

$$\begin{aligned}
&\sum_{i=1}^{m}(p_i - p^*)p_i^2 q_i \,\mathrm{E}\,\Delta^3 g(W_i) \\
&\quad = \frac{1}{\sigma^2}\sum_{i,j=1}^{m} p_i^2 p_j q_i q_j (p_i - p_j)\,\mathrm{E}\,\Delta^3 g(W_i) \\
&\quad = \frac{1}{2\sigma^2}\sum_{i,j=1}^{m} p_i p_j q_i q_j (p_i - p_j)(p_i\,\mathrm{E}\,\Delta^3 g(W_i) - p_j\,\mathrm{E}\,\Delta^3 g(W_j)) \\
&\quad = \frac{1}{2\sigma^2}\sum_{i,j=1}^{m} p_i p_j q_i q_j (p_i - p_j)(p_i\,\mathrm{E}\,\Delta^3 g(W_{ij}) + p_i p_j\,\mathrm{E}\,\Delta^4 g(W_{ij}) \\
&\qquad\qquad - p_j\,\mathrm{E}\,\Delta^3 g(W_{ij}) - p_i p_j\,\mathrm{E}\,\Delta^4 g(W_{ij})) \\
&\quad = \frac{1}{2\sigma^2}\sum_{i,j=1}^{m} p_i p_j q_i q_j (p_i - p_j)^2\,\mathrm{E}\,\Delta^3 g(W_{ij}).
\end{aligned}$$

The following fact was implicitly used in Röllin (2008). We give a quick proof here. It is a simple extension of the result in Ehm (1991), and is necessary, as W may have a larger support than Y. Below we use the notation I_A to denote the indicator function for the event A.

Lemma 3.2. *Let $A \subset \mathbb{Z}$, and define the operator $\mathcal{B}f(k) := p(n-k)f(k+1) - qkf(k)$. Let $f\colon \mathbb{Z} \to \mathbb{R}$ be the solution to*

$$\mathcal{B}f(k) = I_{k\in A} - \mathrm{Bi}(n, p)\{A\} \quad \text{if } 0 \le k \le n, \tag{3.7}$$

and let $f(k) = 0$ for $k \notin \{0, 1, \ldots, n\}$. Then, with K as defined in (2.4),

$$\|\Delta f\| \le K. \tag{3.8}$$

Furthermore, if $A = \{k\}$ for some $k \in \mathbb{Z}$, we also have

$$\|f\| \le K. \tag{3.9}$$

Proof. Note that, for $1 \le k \le n$, $f(k)$ coincides with the definition in Ehm (1991), who showed that

$$\sup_{k\in\{1,\ldots,n-1\}} |\Delta f(k)| \le \frac{1 - p^{n+1} - q^{n+1}}{(n+1)pq} < K.$$

It remains to bound $\Delta f(0) = f(1)$ and $\Delta f(n) = -f(n)$ as, obviously, $\Delta f(k) = 0$ if $k < 0$ or $k > n+1$.

Let $\mu := \mathrm{Bi}(n, p)$ be the binomial probability measure. Then, from Barbour *et al.* (1992b, p. 189) we have, for $1 \le k \le n$ and where $U_k := \{0, 1, \ldots, k\}$,

$$\begin{aligned}
f(k) &= \frac{\mu\{A \cap U_{k-1}\} - \mu\{A\}\mu\{U_{k-1}\}}{kq\mu\{k\}} \\
&= \frac{\mu\{A \cap U_{k-1}\}\mu\{U_{k-1}^{\mathrm{c}}\} - \mu\{A \cap U_{k-1}^{\mathrm{c}}\}\mu\{U_{k-1}\}}{kq\mu\{k\}}.
\end{aligned}$$

From this we have

$$|f(k)| \le \frac{\mu\{U_{k-1}^c\}\mu\{U_{k-1}\}}{kq\mu\{k\}};$$

in particular, for $k = 1$,

$$|f(1)| \le \frac{(1-q^n)q}{npq} \le K.$$

For the corresponding bound at the upper boundary, we have, again from Barbour *et al.* (1992b, p. 189),

$$\begin{aligned} f(k) &= -\frac{\mu\{A \cap U_{k-1}^c\} - \mu\{A\}\mu\{U_{k-1}^c\}}{(n-k+1)p\mu\{k-1\}} \\ &= -\frac{\mu\{A \cap U_{k-1}\}\mu\{U_{k-1}^c\} - \mu\{A \cap U_{k-1}^c\}\mu\{U_{k-1}\}}{(n-k+1)p\mu\{k-1\}}, \end{aligned}$$

which, applying it for $k = n$, leads to the same bound on $\Delta f(n)$, so that (3.8) follows. The bound in (3.9) is immediate from the proof of Ehm (1991, Lemma 1).

Proof of Theorem 2.1. We need to bound $|\mathrm{P}(W - s \in A) - \mathrm{Bi}(n, p)\{A\}|$ for any set $A \subset \mathbb{Z}$. Let $f : \mathbb{Z} \to \mathbb{R}$ be such that (3.7) holds. Then we can write

$$\mathrm{P}(W - s \in A) - \mathrm{Bi}(n, p)\{A\} = \mathrm{P}(W - s \in A \setminus \{0, 1, \ldots, n\}) + \mathrm{E}\,\mathcal{B}f(W - s), \quad (3.10)$$

and note that this equation holds because $f = 0$ outside of $\{0, 1, \ldots, n\}$.

Let the operator $\mathcal{B}^*$ be defined as $\mathcal{B}$ in Lemma 3.2, but replacing n by n^* and p by p^*. Let $g(w) := f(w - s)$, and recall that $w - s = w - \lambda_1 + n^*p^* + \{s^*\}$. Then,

$$\begin{aligned} \mathcal{B}f(w-s) &= \mathcal{B}^*f(w-s) + \{s^*\}g(w+1) + (p^*-p)(w-s)\Delta g(w) \\ &=: \mathcal{B}^*f(w-s) + R_1(w). \end{aligned}$$

Note further that

$$\begin{aligned} \mathcal{B}^*f(w-s) &= (n^*p^*q^* - p^*(w-s-n^*p^*))\Delta f(w-s) - (w-s-n^*p^*)f(w-s) \\ &= \hat{\mathcal{B}}^*g(w) - p^*\{s^*\}\Delta g(w) - \{s^*\}g(w) \\ &=: \hat{\mathcal{B}}^*g(w) + R_2(w), \end{aligned}$$

where $\hat{\mathcal{B}}^*$ is as in Lemma 3.1. Hence,

$$\mathcal{B}f(w-s) = \hat{\mathcal{B}}^*g(w) + R_1(w) + R_2(w). \quad (3.11)$$

Let us first deal with the error terms R_1 and R_2 (which arise only due to the necessity that n and s have to be integers). Now,

$$R_1(w) + R_2(w) = (p^*-p)w\Delta g(w) + (\{s^*\}(1-p^*) - s(p^*-p))\Delta g(w).$$

Noting that $\mathrm{E}[W\Delta g(W)] = \sum_i p_i\,\mathrm{E}\,\Delta g(W_i + 1)$ and recalling (3.8), we have

$$\begin{aligned} |\mathrm{E}[R_1(W) + R_2(W)]| &\le 2K(\lambda_1|p^* - p| + \{s^*\}) \\ &\le 2K\left(\frac{\lambda_1(\{n^*\} + \{s^*\})}{n} + \{s^*\}\right), \end{aligned} \quad (3.12)$$

where we have used

$$\begin{aligned}|\{s^*\}(1-p^*)-s(p^*-p)| &= |s^*(1-p^*)-s(1-p)|\\ &\le |s-s^*|+|sp-s^*p^*|\\ &\le \{s^*\}+s^*|p-p^*|+|s-s^*|p\\ &\le 2\{s^*\}+s^*|p-p^*|\\ &\le 2\{s^*\}+\lambda_1|p-p^*|.\end{aligned}$$

To estimate $\mathrm{E}\,\hat{\mathcal{B}}^*(W)$, we use Lemma 3.1. Estimation of $\mathrm{E}\,\Delta^3 g(W_{i,j})$ goes along the lines given in Barbour and Čekanavičius (2002, pp. 521 and 541). For a random variable X, define first

$$D^k(X) = \|\mathcal{L}(X) * (\delta_0-\delta_1)^{*k}\|,$$

where $\|\cdot\|$ denotes the total variation norm when applied to measures. Note that $D^1(X) = 2d_{\mathrm{TV}}(\mathcal{L}(X), \mathcal{L}(X+1))$. We can decompose $W_{i,j} = S_{i,j,1} + S_{i,j,2}$ in such a way that both sums of the $(p_i \wedge q_i)$ corresponding to $S_{i,j,1}$ and $S_{i,j,2}$ are greater than or equal to $v/2 - v^*$, where $v^* = \max_{1\le i\le m}(p_i \wedge q_i)$. We have

$$\begin{aligned}|\mathrm{E}\Delta^3 g(W_{i,j})| &\le \|\Delta g\| D^2(W_{i,j})\\ &\le \|\Delta g\| D^1(S_{i,j,1}) D^1(S_{i,j,2})\\ &\le \frac{4K}{1\vee(v/2-1)}. \end{aligned} \tag{3.13}$$

In the last line we used Barbour and Xia (1999, Proposition 4.6) (later improved in Mattner and Roos (2007, Corollary 1.6) by roughly a constant factor of 0.8 in our case) and Barbour and Čekanavičius (2002, p. 521, Estimate (4.9)).

So, starting from (3.10), then using identity (3.11) along with Lemma 3.1 and estimate (3.13) and also estimate (3.12), we obtain

$$\begin{aligned}&|\mathrm{P}(W-s\in A) - \mathrm{Bi}(n,p)\{A\}|\\ &\quad\le \frac{4K}{2\sigma^2(1\vee(v/2-1))}\sum_{i,j} p_i p_j q_i q_j (p_i-p_j)^2\\ &\qquad + 2K\left(\frac{\lambda_1(\{n^*\}+\{s^*\})}{n} + \{s^*\}\right) + \mathrm{P}(W<s) + \mathrm{P}(W>n+s).\end{aligned}$$

Note now that

$$\frac{1}{2}\sum_{i,j} p_i p_j q_i q_j (p_i-p_j)^2 = (\lambda_1-\lambda_2)(\lambda_3-\lambda_4) - (\lambda_2-\lambda_3)^2.$$

Consequently, to complete the proof for the total variation distance, we need to estimate tails of W. Note that $X_i - p_i$ satisfies Bernstein's inequality with parameter $\tau = 1$. Therefore,

$$\mathrm{P}(W<s) = \mathrm{P}(W-\lambda_1 < s-\lambda_1) \le \exp\left\{-\frac{\sigma^4}{4\sum p_j(1-p_j)^2}\right\} \le \exp\left\{-\frac{\sigma^2}{4}\right\}.$$

Similarly, by applying estimate

$$\mathrm{P}(W-\lambda_1 > x) \le \exp\left\{\frac{\sigma^2}{4} - \frac{x}{2}\right\}$$

(see Equation (4.3) of Arak and Zaĭtsev (1988)), we obtain

$$\mathrm{P}(W > n + s) \le \exp\left\{-\frac{\sigma^2}{4} + 1\right\}.$$

Estimate $\mathrm{P}(W < s) \le s \max i < s p_i$ is straightforward.

To obtain the result for the d_{loc} metric, the proof is similar, except that we now have $A = k$ for some $k \in \mathbb{Z}$ and bound (3.9). We need some refinements of the estimates of $\mathrm{E}[R_1(W) + R_2(W)]$ and $\mathrm{E}\,\hat{\mathcal{B}}^*(W)$. Similar to (3.13),

$$|\mathrm{E}\Delta g(W_i)| \le \|g\| D^1(W_i) \le \frac{2K}{(1 \vee (v-1))^{1/2}},$$

and, choosing $S_{i,j,k}$, $k = 1, 2, 3$, so that the corresponding $(p_i \wedge q_i)$ sum up to at least $(v/3 - 2v^*)$,

$$|\mathrm{E}\Delta^3 g(W_{i,j})| \le \|g\| D^3(W_{i,j}) \le \|g\| \prod_{k=1}^{3} D^1(S_{i,j,k}) \le \frac{8K}{(1 \vee (v/3-2))^{3/2}}.$$

Substituting these estimates into the corresponding inequalities, the final estimate (2.5) is easily obtained.

4. Numerical results

In this section we study the sum of Bernoulli random variables $X_1, \ldots, X_{100}$ with uniformly spread probabilities from 0 to some parameter M, so that $p_i = iM/101$, $i = 1, 2, \ldots, 100$. We analytically compute the exact distribution of $W = \sum_{i=1}^{100} X_i$ and then the exact total variation distance between W and several different approximations for different values of M. Figure 1 shows a graph of the exact total variation approximation error for several different approximations versus M, referred to in the graph on the x-axis as the 'maximum probability'. In the graph 'Poisson' refers to the standard Poisson approximation where the parameter is chosen to match the first moment. 'Binomial' refers to a binomial approximation where a number of trials, n, is fixed to equal 100 but the probability of success, p, is chosen to match the first moment (this is the approximation studied in Ehm (1991)). 'Shifted Poisson' refers to the approximation where a constant is added to a Poisson random variable and the two parameters—the constant and the Poisson rate—are chosen to match the first two moments (this is the approximation studied in Čekanavičius and Vaĭtkus (2001)). The 'normal' approximation is the standard normal approximation to the binomial distribution using the continuity correction. '2 parameter binomial' refers to the approximation where the two binomial parameters n and p are chosen to match the first two moments (this is the approximation studied in Soon (1996)). Finally, 'shifted binomial' refers to the approximation we propose in this paper—where the shift, the number of trials, and the probability of success are chosen to match the first three moments.

We see in Figure 1 that the normal approximation performs well when probabilities are widely spread out but performs very poorly when probabilities are very small. We see that the Poisson, shifted Poisson, and binomial approximations are best for small probabilities but not otherwise. The two-parameter binomial approximation is quite good, but the shifted binomial approximation performs the best over the widest range of values of M. Since the value of M can be viewed as varying widely in our statistical application, this would be the preferred approximation.

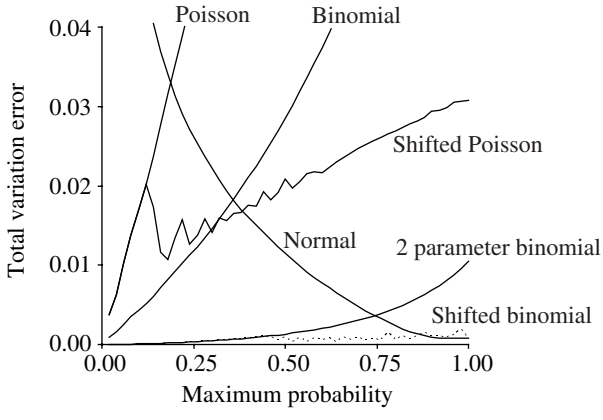

FIGURE 1: Exact total variation distance error between W using $p_i = iM/101$, $i = 1, \ldots, 100$, and six approximations as a function of the maximum probability M.

In summary, we see that over a range of different Poisson binomial random variables that the shifted binomial approximation performs very well—usually better than the other two standard binomial approximations studied previously in the literature. The advantage of the shifted binomial approximation seems to increase as the spread among the Bernoulli probabilities increases.

5. Application to Bayesian hierarchical modeling

The study of shifted binomial approximations is motivated by a statistical problem (see Peköz *et al.* (2009)) of ranking a large number of hospitals with respect to quality as measured by the risk of adverse events at the hospitals. Let X_{ij} be a binary data variable that equals 1 if an adverse event of a particular type happens to patient i in hospital facility j, and equals 0 otherwise. We are interested in the following model where X_{ij} are the data values, p_{ij} are known constants, and θ_j and σ^2 are unknown parameters that we would like to estimate:

$$X_{ij} \mid p_{ij}, \theta_j \sim \text{Be}(\text{logit}^{-1}(\text{logit}(p_{ij}) + \theta_j)),$$

where

$$\theta_j \mid \sigma^2 \sim N(0, \sigma^2).$$

In this model p_{ij} is a risk-adjusted probability that has been previously calculated by taking into account various patient specific indicators and it represents the chance that patient i has an adverse event at a typical hospital. The parameter θ_j is a hospital specific factor that increases or decreases the probability of an adverse event for its patients. Hospitals with a high value of θ_j are poorly performing hospitals. Our goal is to rank hospitals by the values of θ_j. The standard Bayesian hierarchical modeling approach is to put prior distributions on the unspecified parameters and estimate the posterior means of all the parameters conditional on the data.

The difficulty in this situation is that the values of X_{ij} and p_{ij} are both confidential and are too numerous to conveniently transmit from each of the hospitals to the main research facility that would be performing the analysis. We need a method for summarizing each of these so that each facility need only report a few summary statistics. In our application we have

thousands of hospitals, thousands of people in each hospital, and a number of different types of adverse events. A rough approximation of 5000 hospitals with 1000 people each yields a total of $5000 \times 1000 = 5\,000\,000$ random variables—too many to be conveniently computable by standard software.

To circumvent this difficulty, we propose that each hospital aggregate its patients and compute $Y_j = \sum_i X_{ij}$, the number of people in hospital j who have an adverse event. We then use the shifted binomial approximation above for Y_j. This will then yield a total of 5000 random variables—much more easily manageable computationally.

To implement the approximation, in the preparation stage, hospital j also stores and submits the values of $\lambda_{jm} \equiv \sum_i p_{ij}^m$ for $m = 1, 2, 3$ and all j. Then we can easily compute the shifted binomial approximation to Y_j from these as a function of θ_j. This results in the following model:

$$\theta_j \mid \sigma^2 \sim N(0, \sigma^2), \qquad Y_j - s_j \mid \theta_j, n_j, p_j \sim \mathrm{Bi}(n_j, \mathrm{logit}^{-1}(\mathrm{logit}(p_j) + \theta_j))$$

with

$$p_j = \frac{\lambda_{j2} - \lambda_{j3}}{\lambda_{j1} - \lambda_{j2}}, \qquad n_j = \frac{\lambda_{j1} - \lambda_{j2}}{p_j(1 - p_j)}, \qquad s_j = \lambda_{j1} - n_j p_j$$

being the parameters for the shifted binomial approximation designed to match three moments.

Remark 5.1. Though the binomial distribution is not defined for fractional values of the parameter n, we can use a fractional parameter in the likelihood function for the data to obtain in some sense an interpolation of the likelihood functions under the two closest binomial models having integer parameters. For many statistical parameter estimation software packages using likelihood-based approaches, such as maximum likelihood or the Metropolis algorithm, such fractional values of the binomial parameter n can be used this way to yield better approximations.

For example, in the simple model for the data $X \mid n, p \sim \mathrm{Bi}(n, p)$, the likelihood function for the data as a function of the unknown parameter p is $L(p) \propto p^X(1-p)^{n-X}$. Under likelihood-based approaches, this function is all that is used from the model to estimate the parameters, and so with the use of noninteger n the function $L(p)$ can be viewed as yielding an interpolation of the likelihood functions $L_1(p) \propto p^X(1-p)^{\lceil n \rceil - X}$ and $L_2(p) \propto p^X(1-p)^{\lfloor n \rfloor - X}$.

Acknowledgements

VC, EP, and AR would like to express gratitude for the gracious hospitality of Andrew Barbour and Louis Chen during a visit to the National University of Singapore in January 2009 (where a portion of this work was completed), as well as gratitude for generous support from the Institute for Mathematical Sciences of the National University of Singapore. EP and MS would like to thank the Center for Organization, Leadership and Management Research at the Veterans' Health Administration also for generous support: this research was supported in part by a grant from the Department of Veterans Affairs Health Services Research and Development Service, IIR 06-260. Thanks are also due to the anonymous referee for many valuable comments that have led to significant improvements in the paper.

References

Arak, T. V. and Zaĭtsev, A. Yu. (1988). Uniform limit theorems for sums of independent random variables. *Proc. Steklov Inst. Math.* **174,** viii+222 pp.

Ash, A., Shwartz, M. and Peköz, E. (2003). Comparing outcomes across providers. In *Risk Adjustment for Measuring Health Care Outcomes*, 3rd edn. Health Administration Press, Chicago, IL, pp. 297–333.

Barbour, A. D. and Brown, T. C. (1992). Stein's method and point process approximation. *Stoch. Process. Appl.* **43,** 9–31.

Barbour, A. D. and Čekanavičius, V. (2002). Total variation asymptotics for sums of independent integer random variables. *Ann. Prob.* **30,** 509–545.

Barbour, A. D. and Chen, L. H. Y. (eds) (2005). *An Introduction to Stein's Method* (Lecture Notes Ser., Inst. Math. Sci., National Uni. Singapore **4**), Singapore University Press.

Barbour, A. D. and Chryssaphinou, O. (2001). Compound Poisson approximation: a user's guide. *Ann. Appl. Prob.* **11,** 964–1002.

Barbour, A. D. and Lindvall, T. (2006). Translated Poisson approximation for Markov chains. *J. Theoret. Prob.* **19,** 609–630.

Barbour, A. D. and Xia, A. (1999). Poisson perturbations. *ESAIM Prob. Statist.* **3,** 131–150 (electronic).

Barbour, A. D., Chen, L. H. Y. and Loh, W.-L. (1992a). Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Prob.* **20,** 1843–1866.

Barbour, A. D., L. Holst, and Janson, S. (1992b). *Poisson Approximation* (Oxford Stud. Prob. **2**). Clarendon Press, Oxford.

Čekanavičius, V. and Roos, B. (2007). Binomial approximation to the Markov binomial distribution. *Acta Appl. Math.* **96,** 137–146.

Čekanavičius, V. and Vaĭtkus, P. (2001). Centered Poisson approximation by the Stein method. *Lithuanian Math. J.* **41,** 319–329.

Chen, L. H. Y. (1974). On the convergence of Poisson binomial to Poisson distributions. *Ann. Prob.* **2,** 178–180.

Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Prob.* **3,** 534–545.

Chen, S. X. and Liu, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica* **7,** 875–892.

Choi, K. P. and Xia, A. (2002). Approximating the number of successes in independent trials: binomial versus Poisson. *Ann. Appl. Prob.* **12,** 1139–1148.

Ehm, W. (1991). Binomial approximation to the Poisson binomial distribution. *Statist. Prob. Lett.* **11,** 7–16.

Le Cam, L. (1960). An approximation theorem for the Poisson binomial distribution. *Pacific J. Math.* **10,** 1181–1197.

Loh, W.-L. (1992). Stein's method and multinomial approximation. *Ann. Appl. Prob.* **2,** 536–554.

Luk, H. M. (1994). Stein's method for the gamma distribution and related statistical applications. Doctoral Thesis, University of Southern California.

Mattner, L. and Roos, B. (2007). A shorter proof of Kanter's Bessel function concentration bound. *Prob. Theory Relat. Fields* **139,** 191–205.

Peköz, E. A. (1996). Stein's method for geometric approximation. *J. Appl. Prob.* **33,** 707–713.

Peköz, E. A., Shwartz, M., Christiansen, C. and Berlowitz, D. (2009). Approximate Bayesian models for aggregate data when individual-level data is confidential or unavailable. Submitted.

Pitman, J. (1997). Probabilistic bounds on the coefficients of polynomials with only real zeros. *J. Combinatorial Theory A* **77,** 279–303.

Reinert, G. (2005). Three general approaches to Stein's method. In *An Introduction to Stein's Method* (Lecture Notes Ser., Inst. Math. Sci., National Uni. Singapore **4**), Singapore University Press, pp. 183–221.

Röllin, A. (2005). Approximation of sums of conditionally independent variables by the translated Poisson distribution. *Bernoulli* **11,** 1115–1128.

Röllin, A. (2008). Symmetric and centered binomial approximation of sums of locally dependent random variables. *Electron. J. Prob.* **13,** 756–776.

Roos, B. (2000). Binomial approximation to the Poisson binomial distribution: the Krawtchouk expansion. *Theory Prob. Appl.* **45,** 258–272.

Ross, S. and Peköz, E. (2007). *A Second Course in Probability*. ProbabilityBookstore.com, Boston, MA.

Soon, S. Y. T. (1996). Binomial approximation for dependent indicators. *Statistica Sinica* **6,** 703–714.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. 6th Berkeley Symp. Math. Statist. Prob.,* Vol. II, University of California Press, Berkeley, pp. 583–602.