Statistica Neerlandica (2013) Vol. 67, nr. 2, pp. 190–201
doi:10.1111/stan.12002

# Saddlepoint approximations for the sum of independent non-identically distributed binomial random variables

Rob Eisinga*

*Radboud University Nijmegen, Social Science Research Methods, PO Box 9104, Nijmegen, 6500HE, The Netherlands*

Manfred Te Grotenhuis†

*Radboud University Nijmegen, Sociology, PO Box 9104, Nijmegen, 6500HE, The Netherlands*

Ben Pelzer‡

*Radboud University Nijmegen, Social Science Research Methods, PO Box 9104, Nijmegen, 6500HE, The Netherlands*

We discuss saddlepoint approximations to the distribution of the sum of independent non-identically distributed binomial random variables. We examine the accuracy of the saddlepoint methods for a sum of 10 binomials with different sets of parameter values. The numerical results indicate that the saddlepoint approximations provide very accurate estimates for the probability mass function and the right-tail probabilities for the cumulative distribution function of the sum.

*Keywords and Phrases:* sum of non-identical binomial variables, saddlepoint approximation.

## 1 Introduction

We are interested in obtaining the probability distribution of the sum of independent binomial random variables that are not necessarily identically distributed and in estimating the rare event probability that the convolution exceeds some large threshold. Convolutions of non-identical binomial variables occur in a variety of settings as for instance in reliability analysis and quality control, including acceptance sampling (Kotz and Johnson, 1984; Jolayemi, 1992). Other applications include the analysis of DNA matching in the context of a genome search (Smalley, Woodward, and Palmer, 1996) and measures of bundle compliance as indicators of quality in health care organizations (Benneyan and Taşeli, 2010). Several physical and stochastic models that give rise to the convolution of two binomial variables are addressed in Ong (1995).

*r.eisinga@maw.ru.nl
†m.tegrotenhuis@maw.ru.nl
‡b.pelzer@maw.ru.nl

The computation of the exact probability distribution of the sum of non-identical binomials by enumeration involves calculating the probability of all possible elements consistent with the sum. This naive way of computing is intractable in practice if the number of outcomes with non-zero probability is large. Although exact calculation is feasible with computer algebra systems such as *Mathematica* or *Maple* by applying the standard inversion formula to the characteristic function, approximation methods continue to be widely used and explored in the literature (Johnson, Kemp, and Kotz, 2005; Benneyan and Taşeli, 2010; Hong, 2011) for various reasons. An important one is that, from a practical view, the approximations are sometimes as good as exact and straightforward to implement in statistical software programs of any kind.

This note explains how to estimate probabilities of convoluted binomial random variables using saddlepoint mass approximations. Saddlepoint approximations were seminally explored by Daniels (1954) and have received considerable recent attention in the statistical literature. An accessible and detailed introduction into saddlepoint approximations with many applications is provided by Butler (2007). Paolella (2007) offers a computational approach. Although their derivation is involved, the resulting saddlepoint equations are very easy to incorporate in popular statistical software packages such as SAS, SPSS, and R, and they offer a convenient tool, for example, for approximate maximum likelihood model fitting using the saddlepoint method (Davison and Semadeni, 2004).

The remainder of the paper is organized as follows. Section 2 considers the probability distribution of convoluted binomial variables and discusses saddlepoint approximations. Section 3 presents the results of a numerical investigation. Conclusion remarks are in section 4.

## 2 Saddlepoint mass approximations for convoluted binomial variables

Let $X_1, X_2, \ldots, X_r$ be a sequence of mutually independent binomially distributed discrete random variables taking integer values 0,1,2,. . ., with $X_i$ having index $n_i$ and probability $p_i$, that is, $X_i \sim \text{Bin}(n_i,p_i)$. The probability mass function (pmf) of the sum $S = \sum_{i=1}^{r} X_i$ of the $r$ binomials is given by Benneyan and Taşeli (2010). The computational impediment is in the nested summations required for complete enumeration over all possible observations consistent with the sum. Such calculation is unfeasible unless the number of products in the summations is small.

The number of arithmetic operations can efficiently be reduced by calculating the probabilities recursively (Butler and Stephens, 1993; Chen, Dempster, and Liu, 1994; Woodward and Palmer, 1997). Shah (1973) has shown that the probability of the sum of $r$ independent integer valued random variables (not necessarily identically distributed) may be calculated using the recurrence relation

$$P(S=s) = (1/s)\sum_{j=1}^{s} P(S=s-j)(1/j!)\left\{\sum_{i=1}^{r} \frac{\partial^j \ln[A_i(z)]}{\partial z^j}\right\}_{z=0},$$

where $A_i(z)$ is the probability generating function for the random variable $X_i$ and $\partial$ denotes the $j$th-order partial differentiation. As the probability generating function of a binomial random variable $X_i$ is $A_i(z) = (1 - p_i + p_i z)^{n_i}$, the probability of the sum $S$ of $r$ independent non-identical binomials may be obtained as

$$P(S = s) = \begin{cases} \prod_{i=1}^{r}(1-p_i)^{n_i} & s = 0 \\ (1/s)\sum_{j=1}^{s}(-1)^{j-1}\left(P(S = s-j)\sum_{i=1}^{r} n_i\left[\frac{p_i}{1-p_i}\right]^j\right) & s > 0. \end{cases}$$

Although the use of this recurrence formula requires less computation than would the evaluation of each probability directly, the method may be numerically unstable as a result of round-off error in computing $P(S=0)$ if $r$ is large and the explosion of the term $[p_i(1-p_i)^{-1}]^j$ if $s$ is large and $p_i$ is close to 0 or 1 (Hong, 2011).

An alternate procedure that avoids exact computation is to obtain a saddlepoint approximation to the pmf of sum $S$. The cumulant generating function of the convolution is

$$K(u) = \sum_{i=1}^{r} n_i \ln\{1 - p_i + p_i \exp(u)\} \quad u \in (-\infty, +\infty).$$

Let $q_i = p_i \exp(u)/\{1 - p_i + p_i \exp(u)\}$. The first-order saddlepoint approximation to the pmf of $S$ is then given by

$$\hat{P}_1(S = s) = \left\{2\pi K''(\hat{u})\right\}^{-1/2} \exp\{K(\hat{u}) - \hat{u}s\},$$

where the saddlepoint $\hat{u} = \hat{u}(s)$ is the unique value of $u$ satisfying the saddlepoint equation $K'(\hat{u}) = s$, with $K'(u) = \sum_{i=1}^{r} n_i q_i$ being the first-order and $K''(u) = \sum_{i=1}^{r} n_i q_i(1-q_i)$ the second-order derivative of $K(u)$ with respect to $u$. The cumulant generating function $K(u)$ is a strictly convex function when evaluated over $(-\infty, +\infty)$ so $K''(u) > 0$ for all $u$. Also, as the binomial variables are independent, the mean of sum $S$ is $\mu = K'(0) = \sum_{i=1}^{r} n_i p_i$, and the variance is $\sigma^2 = K''(0) = \sum_{i=1}^{r} n_i p_i(1-p_i)$.

The derivative of $K(u)$ set equal to $s$ cannot be solved in closed form, except for small values of $r$, say up to 3 or 4. For example, Eisinga and Pelzer (2011) have shown that for the sum of two binomials, each with different probability,

$$\hat{u} = \ln\left\{\left[-b + \left(b^2 - 4ac\right)^{1/2}\right]2a^{-1}\right\},$$

where $a = (n_1 + n_2 - s)p_1p_2$, $b = -(n_1 + n_2 - 2s)p_1p_2 + (n_1 - s)p_1 + (n_2 - s)p_2$, and $c = -sp_1p_2 + s(p_1 + p_2) - s$. However, for larger values of $r$, the saddlepoint $\hat{u}$ must be determined by solving the saddlepoint equation $K'(\hat{u}) - s = 0$ for $u$. There always exists a unique real root to the equation. The reason for this is that the

convergence strip of the cumulant generating function $K(u)$ is the whole real number line $(-\infty, +\infty)$, and $K(u)$ is strictly convex in $u$ (i.e. $K'(u)$ is strictly increasing) over the whole real line (Butler, 2007). Thus, solving $K'(u) = s$ for any $u$ is not really any problem.

For the first-order saddlepoint approximation, the error is of order $O(n^{-1})$,

$$P(S = s) = \hat{P}_1(S = s)\{1 + O(n^{-1})\},$$

and there are several approaches to further minimize the error of the first-order approximation (Gillespie and Renshaw, 2007). One is to obtain a second-order approximation by including adjustments for the third and fourth cumulants (Daniels, 1987; Akahira, Takahashi, and Takeuchi, 1999; Akahira and Takahashi, 2001). The second-order saddlepoint mass approximation uses the correction term

$$\hat{P}_2(S = s) = \hat{P}_1(S = s)\left\{1 + \frac{1}{8}\frac{K''''(\hat{u})}{\{K''(\hat{u})\}^2} - \frac{5}{24}\frac{\{K'''(\hat{u})\}^2}{\{K''(\hat{u})\}^3} + O(n^{-2})\right\},$$

where

$$K'''(\hat{u}) = \sum_{i=1}^{r} n_i q_i (1 - q_i)(1 - 2q_i),$$

and

$$K''''(\hat{u}) = \sum_{i=1}^{r} n_i q_i (1 - q_i)\{1 - 6q_i(1 - q_i)\}.$$

Further, the saddlepoint equation cannot be solved at the endpoints 0 and $\max(s) = \sum_{i=1}^{r} n_i$ of the support of $S$. This implies that the approximation does not sum to unity, which jeopardizes its accuracy. For a sum of $r$ binomials, the exact boundary probabilities are given by

$$P(S = 0) = \prod_{i=1}^{r} P(X_i = 0) = \prod_{i=1}^{r} (1 - p_i)^{n_i},$$
$$P(S = \max(s)) = \prod_{i=1}^{r} P(X_i = n_i) = \prod_{i=1}^{r} p_i^{n_i}.$$

For small values of $n_i$ or extreme values of $p_i$, a potentially more accurate normalized second-order approximation may be obtained, following Butler (2007), as

$$\bar{P}_2(S = s) = \begin{cases} P(S = 0) & s = 0 \\ [1 - P(S = 0) - P(S = \max(s))]\hat{P}_2(S = s)/\sum_{1 \le j \le \max(s)-1} \hat{P}_2(S = j) & 1 \le s \le \max(s) - 1 \\ P(S = \max(s)) & s = \max(s). \end{cases}$$

The approximate tail probabilities of $S$ can be determined by numerically integrating $\bar{P}_2(s)$. An alternate approach is to use the Lugannani and Rice (1980) formula for the

continuous tail probability approximation. For the discrete setting, Daniels (1987) introduced two continuity-corrected modifications of this tail approximation. One of the first-order approximations to the right-tail probability is

$$\hat{P}_3(S \geq s) = 1 - \Phi(\hat{w}) - \phi(\hat{w})\left\{\frac{1}{\hat{w}} - \frac{1}{\hat{u}_1}\right\},$$

provided that $s \neq E(S)$. The symbols $\Phi$ and $\phi$ denote, respectively, the distribution and density function of a standard normal random variable, $\hat{w} = \text{sign}(\hat{u})\left\{2\hat{u}K'(\hat{u}) - 2K(\hat{u})\right\}^{1/2}$, where $\text{sign}(\hat{u})$ captures the sign $\pm$ for $\hat{u}$, $\hat{u}_1 = \{1 - \exp(-\hat{u})\}\left\{K''(\hat{u})\right\}^{1/2}$, and $\hat{u}$ solves $K'(\hat{u}) = s$. Note that the last term in the expression is undefined if $\hat{w} = \hat{u}_1 = 0$. This occurs if $s = E(S)$ or $\hat{u} = 0$. The approximation at the mean of $S$ or when $\hat{u} = 0$ is

$$\hat{P}_3(S \geq s) = \frac{1}{2} - \{2\pi\}^{-1/2}\left\{\frac{1}{6}K'''(0)\left\{K''(0)\right\}^{-3/2} - \frac{1}{2}\left\{K''(0)\right\}^{-1/2}\right\},$$

where $K''(0) = \sum_{i=1}^{r} n_i p_i (1 - p_i)$ and $K'''(0) = \sum_{i=1}^{r} n_i p_i (1 - p_i)(1 - 2p_i)$. The second-order continuity-corrected saddlepoint approximation to the right-tail probability is given by Daniels (1987) as

$$\hat{P}_4(S \geq s) = \hat{P}_3(S \geq s) - \phi(\hat{w})\left\{\frac{1}{\hat{u}_2}\left(\frac{1}{8}\hat{\kappa}_4 - \frac{5}{24}\hat{\kappa}_3^2\right) - \frac{1}{\hat{u}_2^3} - \frac{\hat{\kappa}_3}{2\hat{u}_2^2} + \frac{1}{\hat{w}^3}\right\},$$

where $\hat{u}_2 = \hat{u}\left\{K''(\hat{u})\right\}^{1/2}$, $\hat{\kappa}_3 = K'''(\hat{u})\left\{K''(\hat{u})\right\}^{-3/2}$, and $\hat{\kappa}_4 = K''''(\hat{u})\left\{K''(\hat{u})\right\}^{-2}$. We finally note that there are other expressions for the right-tail probability approximation in the discrete setting and that these approximations exhibit different accuracies depending on the distribution of $S$ and the selection of $s$. A detailed discussion is given by Butler (2007).

## 3 Numerical example

The accuracy of the saddlepoint approximations was examined numerically for various values of $r$, $n_i$, and $p_i$. We give one example using data from Benneyan and Taşeli (2010). It concerns the sum of $r = 10$ binomial variables with parameters $n_i$ and $p_i$ as listed in the top panel of Table 1. We present the root $\hat{u}(s)$ of the saddlepoint equation, the exact probability $P(s)$, and the normalized second-order saddlepoint approximation $\bar{P}_2(s)$. For comparison, we also obtained the Gram–Charlier (GC) type A series approximation of order 6 employed by Benneyan and Taşeli (2010), the single binomial approximation with index $\sum n_i$ and probability $r^{-1}\sum p_i$, the normal approximation, matching the first two moments, and the Poisson distribution, matching the mean of $S$. The fitted

Table 1. Probability mass function approximations for the sum of $r = 10$ binomial variables

| $s$ | $\hat{u}(s)$ | $P(s)$ | $\bar{P}_2(s)$ | $\hat{P}_6(s)$ | Bin $\left(s; \sum n_i, \bar{p}_i\right)$ | $N(s; \mu, \sigma^2)$ | Pois$(s;\mu)$ |
|---|---|---|---|---|---|---|---|
| $n_i = 12, 14, 4, 2, 20, 17, 11, 1, 8, 11$ | | | | | | | |
| $p_i = 0.074, 0.039, 0.095, 0.039, 0.053, 0.043, 0.067, 0.018, 0.099, 0.045$ | | | | | | | |
| 1 | −1.800 | 0.0165 | 0.0164 | 0.0172 | 0.0168 | 0.0215 | 0.0187 |
| 3 | −0.678 | 0.0994 | 0.0994 | 0.0986 | 0.0999 | 0.0862 | 0.1021 |
| 5 | 0.144 | 0.1716 | 0.1716 | 0.1719 | 0.1712 | 0.1641 | 0.1673 |
| 7 | 0.216 | 0.1346 | 0.1346 | 0.1346 | 0.1340 | 0.1481 | 0.1305 |
| 9 | 0.492 | 0.0587 | 0.0587 | 0.0590 | 0.0586 | 0.0634 | 0.0594 |
| 11 | 0.717 | 0.0160 | 0.0160 | 0.0156 | 0.0161 | 0.0129 | 0.0177 |
| 13 | 0.909 | $0.0^2$2912 | $0.0^2$2913 | $0.0^2$3013 | $0.0^2$2969 | $0.0^2$1237 | $0.0^2$3719 |
| 15 | 1.078 | $0.0^3$3751 | $0.0^3$3752 | $0.0^3$4015 | $0.0^3$3893 | $0.0^4$5646 | $0.0^3$5805 |
| 17 | 1.230 | $0.0^4$3543 | $0.0^4$3544 | $0.0^4$2621 | $0.0^4$3762 | $0.0^5$1222 | $0.0^4$6995 |
| 19 | 1.368 | $0.0^5$2524 | $0.0^5$2525 | $0.0^6$7245 | $0.0^5$2756 | $0.0^7$1253 | $0.0^5$6704 |
| $n_i = 120, 140, 40, 20, 200, 170, 110, 10, 80, 110$ | | | | | | | |
| $p_i = 0.00074, 0.00039, 0.00095, 0.00039, 0.00053, 0.00043, 0.00067, 0.00018, 0.00099, 0.00045$ | | | | | | | |
| 1 | 0.558 | 0.3231 | 0.3227 | 0.3690 | 0.3230 | 0.4496 | 0.3230 |
| 2 | 1.252 | 0.0924 | 0.0928 | 0.0480 | 0.0923 | 0.0889 | 0.0925 |
| 3 | 1.659 | 0.0176 | 0.0177 | 0.0284 | 0.0176 | $0.0^2$3059 | 0.0176 |
| 4 | 1.948 | $0.0^2$2514 | $0.0^2$2525 | $0.0^2$2794 | $0.0^2$2508 | $0.0^4$1834 | $0.0^2$2525 |
| 5 | 2.172 | $0.0^3$2868 | $0.0^3$2881 | $0.0^4$1707 | $0.0^3$2859 | $0.0^7$1915 | $0.0^3$2891 |
| 6 | 2.356 | $0.0^4$2723 | $0.0^4$2735 | $0.0^7$1167 | $0.0^4$2713 | $0.0^{11}$3482 | $0.0^4$2759 |
| 7 | 2.511 | $0.0^5$2213 | $0.0^5$2224 | $0.0^{11}$1077 | $0.0^5$2205 | $0.0^{15}$1103 | $0.0^5$2256 |
| $n_i = 120, 140, 40, 20, 200, 170, 110, 10, 80, 110$ | | | | | | | |
| $p_i = 0.74, 0.39, 0.95, 0.39, 0.53, 0.43, 0.67, 0.18, 0.99, 0.45$ | | | | | | | |
| 510 | −0.300 | $0.0^5$2363 | $0.0^5$2363 | $0.0^5$2363 | $0.0^4$1058 | $0.0^5$2346 | $0.0^3$5109 |
| 520 | −0.252 | $0.0^4$3730 | $0.0^4$3730 | $0.0^4$3730 | $0.0^3$1056 | $0.0^4$3706 | $0.0^2$1458 |
| 530 | −0.204 | $0.0^3$3638 | $0.0^3$3638 | $0.0^3$3638 | $0.0^3$7061 | $0.0^3$3623 | $0.0^2$3436 |
| 540 | −0.156 | $0.0^2$2195 | $0.0^2$2195 | $0.0^2$2195 | $0.0^2$3162 | $0.0^2$2191 | $0.0^2$6702 |
| 550 | −0.108 | $0.0^2$8202 | $0.0^2$8202 | $0.0^2$8202 | $0.0^2$9471 | $0.0^2$8201 | 0.01087 |
| 560 | −0.060 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0190 | 0.0147 |
| 570 | −0.012 | 0.0272 | 0.0272 | 0.0272 | 0.0253 | 0.0272 | 0.0166 |
| 580 | 0.036 | 0.0242 | 0.0242 | 0.0242 | 0.0224 | 0.0241 | 0.0157 |

(*Continues*)

Table 1. (Continued)

| $s$ | $\hat{u}(s)$ | $P(s)$ | $\bar{P}_2(s)$ | $\hat{P}_6(s)$ | $\text{Bin}\left(s; \sum n_i, \bar{p}_i\right)$ | $N\left(s; \mu, \sigma^2\right)$ | $\text{Pois}(s; \mu)$ |
|---|---|---|---|---|---|---|---|
| 590 | 0.084 | 0.0133 | 0.0133 | 0.0133 | 0.0132 | 0.0133 | 0.0126 |
| 600 | 0.132 | $0.0^24501$ | $0.0^24501$ | $0.0^24501$ | $0.0^25141$ | $0.0^24501$ | $0.0^28500$ |
| 610 | 0.181 | $0.0^39419$ | $0.0^39419$ | $0.0^39419$ | $0.0^21321$ | $0.0^39460$ | $0.0^24854$ |
| 620 | 0.230 | $0.0^31213$ | $0.0^31213$ | $0.0^31213$ | $0.0^32230$ | $0.0^31230$ | $0.0^32353$ |
| 630 | 0.279 | $0.0^59581$ | $0.0^59581$ | $0.0^59581$ | $0.0^42463$ | $0.0^59902$ | $0.0^39708$ |
| 640 | 0.328 | $0.0^64630$ | $0.0^64630$ | $0.0^64628$ | $0.0^51773$ | $0.0^64931$ | $0.0^33418$ |

normal density approximation with mean $\mu = K'(0)$ and variance $\sigma^2 = K''(0)$ is of the form

$$N(s;\mu,\sigma^2) = \sigma^{-1}\{2\pi\}^{-1/2}\exp\left\{-(s-\mu)^2/2\sigma^2\right\},$$

and Pois $(s;\mu)$ is the fit of a Poisson variable with mean $\mu$.

As can be seen in the top panel of Table 1, the normalized second-order saddlepoint approximation $\bar{P}_2(s)$ provides a superior fit. It captures both the center of the distribution and the tail behavior of $S$ very well. The GC approximation $\hat{P}_6(s)$ is very accurate near the mean of $S$ but degrades in the tails. The single binomial approximation is slightly over-dispersed but performs rather well overall. The normal and the Poisson approximations perform poorly in comparison. The middle panel of Table 1 presents the approximations of the exact $P(s)$ for $n_i$ multiplied by 10 and $p_i$ divided by 100. On this occasion, we would expect the simple Poisson approximation to work well, because the $p_i$'s are very small and the $n_i$'s are quite large. Both the Poisson and the binomial approximations are seen to adequately capture the distribution as does the saddlepoint approximation $\bar{P}_2(s)$, which performs extremely well especially in the right tail. The normal approximation is again ineffective because of the considerable skewness in the distribution of $S$, whereas the GC approximation fails to assume the correct form in the center and in the extreme right tail. The bottom panel of Table 1 gives the approximate $P(s)$ for both $n_i$ and $p_i$ multiplied by 10. For this distribution, we would expect the normal Gaussian approximation to work well. The normal, the saddlepoint, and the GC approximations all provide accurate estimates near the mean of the distribution, whereas the binomial and the Poisson approximations behave rather poorly. The latter tends to overestimate the tail probabilities at both tails of the distribution. The tail behavior of $S$ is captured well by the normal procedure, but the GC and the saddlepoint approximations are observed to be most accurate. The probability values provided by the latter procedure are the same as the exact values to the four significant digit accuracy displayed. For the extreme right tail, it provides results that agree to the 10th decimal places.

Table 2 presents approximations for the right-tail probabilities of $S$, using the same binomial parameters as in Table 1. It presents the exact probability $P(S \geq s)$, the normalized second-order saddlepoint approximation $\bar{P}_2(S \geq s)$, the Daniels (1987) second-order continuity-corrected saddlepoint approximation to the right-tail probability $\hat{P}_4(S \geq s)$, the GC type A series approximation of order 6 $\hat{P}_6(S \geq s)$, the single binomial, the normal, and the Poisson approximations. For the normalized second-order saddlepoint and the GC approximations, the approximate tail probabilities were obtained by integrating the approximations to the mass function of $S$. The normal approximation uses a continuity correction.

The figures show that the Poisson works fine for very small $p_i$ and quite large $n_i$ (middle panel) and that the normal approximation performs well for larger values

Table 2. Cumulative distribution function approximations for the right tail of the sum of $r = 10$ binomial variables

| $s$ | $P(S \geq s)$ | $\bar{P}_2(S \geq s)$ | $\hat{P}_4(S \geq s)$ | $\hat{P}_6(S \geq s)$ | Bin $\left(s; \sum n_i, \bar{p}_i\right)$ | $N\left(s+\frac{1}{2}; \mu, \sigma^2\right)$ | Pois $(s; \mu)$ |
|---|---|---|---|---|---|---|---|
| $n_i = 12, 14, 4, 2, 20, 17, 11, 1, 8, 11$ | | | | | | | |
| $p_i = 0.074, 0.039, 0.095, 0.039, 0.053, 0.043, 0.067, 0.018, 0.099, 0.045$ | | | | | | | |
| 1 | 0.9973 | 0.9973 | 0.9972 | 0.9968 | 0.9972 | 0.9880 | 0.9967 |
| 3 | 0.9310 | 0.9311 | 0.9308 | 0.9300 | 0.9300 | 0.9182 | 0.9246 |
| 5 | 0.6847 | 0.6847 | 0.6847 | 0.6847 | 0.6831 | 0.7016 | 0.6765 |
| 7 | 0.3481 | 0.3482 | 0.3481 | 0.3479 | 0.3474 | 0.3689 | 0.3496 |
| 9 | 0.1187 | 0.1187 | 0.1187 | 0.1180 | 0.1189 | 0.1154 | 0.1257 |
| 11 | 0.0277 | 0.0277 | 0.0276 | 0.0270 | 0.0280 | 0.0196 | 0.0323 |
| 13 | $0.0^2$4544 | $0.0^2$4545 | $0.0^2$4546 | $0.0^2$4305 | $0.0^2$4654 | $0.0^2$1716 | $0.0^2$6130 |
| 15 | $0.0^3$5432 | $0.0^3$5434 | $0.0^3$5435 | $0.0^3$1122 | $0.0^3$5666 | $0.0^4$7535 | $0.0^3$8897 |
| 17 | $0.0^4$4852 | $0.0^4$4853 | $0.0^4$4855 | $-0.0^3$4031 | $0.0^4$5177 | $0.0^5$1630 | $0.0^3$1015 |
| 19 | $0.0^5$3311 | $0.0^5$3312 | $0.0^5$3313 | $-0.0^3$4341 | $0.0^5$3632 | $0.0^6$1719 | $0.0^5$9326 |
| $n_i = 120, 140, 40, 20, 200, 170, 110, 10, 80, 110$ | | | | | | | |
| $p_i = 0.00074, 0.00039, 0.00095, 0.00039, 0.00053, 0.00043, 0.00067, 0.00018, 0.00099, 0.00045$ | | | | | | | |
| 1 | 0.4360 | 0.4360 | 0.4375 | 0.4339 | 0.4357 | 0.5382 | 0.4359 |
| 2 | 0.1129 | 0.1133 | 0.1133 | 0.0649 | 0.1127 | 0.1101 | 0.1129 |
| 3 | 0.0204 | 0.0205 | 0.0205 | 0.0169 | 0.0204 | $0.0^2$5413 | 0.0205 |
| 4 | $0.0^2$2830 | $0.0^2$2843 | $0.0^2$2840 | −0.01142 | $0.0^2$2823 | $0.0^4$5435 | $0.0^2$2844 |
| 5 | $0.0^3$3164 | $0.0^3$3178 | $0.0^3$3174 | −0.01422 | $0.0^3$3154 | $0.0^6$1038 | $0.0^3$3191 |
| 6 | $0.0^4$2961 | $0.0^4$2975 | $0.0^4$2970 | −0.01423 | $0.0^4$2951 | $0.0^{10}$3648 | $0.0^4$3001 |
| 7 | $0.0^5$2381 | $0.0^5$2392 | $0.0^5$2388 | −0.01423 | $0.0^5$2372 | $0.0^{14}$2331 | $0.0^5$2429 |
| $n_i = 120, 140, 40, 20, 200, 170, 110, 10, 80, 110$ | | | | | | | |
| $p_i = 0.74, 0.39, 0.95, 0.39, 0.53, 0.43, 0.67, 0.18, 0.99, 0.45$ | | | | | | | |
| 510 | $0.9^5$3616 | $0.9^5$3616 | $0.9^5$3616 | $0.9^5$3614 | $0.9^4$5448 | $0.9^5$3631 | $0.9^2$5770 |
| 520 | $0.9^3$8791 | $0.9^3$8791 | $0.9^3$8791 | $0.9^3$8791 | $0.9^3$4795 | $0.9^3$8796 | 0.9861 |
| 530 | $0.9^2$8549 | $0.9^2$8549 | $0.9^2$8549 | $0.9^2$8549 | $0.9^2$5910 | $0.9^2$8554 | 0.9617 |
| 540 | 0.9889 | 0.9889 | 0.9889 | 0.9889 | 0.9778 | 0.9887 | 0.9104 |
| 550 | 0.9444 | 0.9444 | 0.9444 | 0.9444 | 0.9151 | 0.9445 | 0.8208 |
| 560 | 0.8161 | 0.8161 | 0.8161 | 0.8161 | 0.7690 | 0.8161 | 0.6902 |
| 570 | 0.5825 | 0.5825 | 0.5825 | 0.5825 | 0.5388 | 0.5823 | 0.5306 |

(*Continues*)

Table 2. (Continued)

| $s$ | $P(S \geq s)$ | $\bar{P}_2(S \geq s)$ | $\hat{P}_4(S \geq s)$ | $\hat{P}_6(S \geq s)$ | Bin $\left(s; \sum n_i, \bar{p}_i\right)$ | $N\left(s+\frac{1}{2}; \mu, \sigma^2\right)$ | Pois $(s; \mu)$ |
|---|---|---|---|---|---|---|---|
| 580 | 0.3140 | 0.3140 | 0.3140 | 0.3140 | 0.2939 | 0.3139 | 0.3667 |
| 590 | 0.1194 | 0.1194 | 0.1194 | 0.1194 | 0.1184 | 0.1195 | 0.2250 |
| 600 | 0.0306 | 0.0306 | 0.0306 | 0.0306 | 0.0340 | 0.0307 | 0.1214 |
| 610 | $0.0^2$5133 | $0.0^2$5133 | $0.0^2$5133 | $0.0^2$5133 | $0.0^2$6765 | $0.0^2$5184 | 0.0573 |
| 620 | $0.0^3$5522 | $0.0^3$5522 | $0.0^3$5522 | $0.0^3$5522 | $0.0^3$9185 | $0.0^3$5648 | 0.0235 |
| 630 | $0.0^4$3757 | $0.0^4$3761 | $0.0^4$3757 | $0.0^4$3757 | $0.0^4$8356 | $0.0^4$3926 | $0.0^2$8387 |
| 640 | $0.0^5$1599 | $0.0^5$1635 | $0.0^5$1599 | $0.0^5$1598 | $0.0^5$5091 | $0.0^5$1728 | $0.0^2$2594 |

of $p_i$ and $n_i$ (bottom panel). In the latter case, the GC approximation yields extremely accurate results, but for smaller values of $n_i$ (top panel) or $p_i$ (middle panel), it fails to assume the correct form in the long right-hand tail and suffers from negative tail probabilities. Whereas the single binomial approximation provides rather accurate estimates if its over-dispersion relative to the exact distribution is small (top and middle panel), its accuracy deteriorates if the parameters of the individual binomials are less homogeneous (bottom panel). The integrated normalized second-order method performs well, although it fails to capture the extreme right tail if the $p_i$'s and $n_i$'s are quite large (bottom panel). The second-order continuity-corrected saddlepoint approximation yields the most accurate results. In general, this saddlepoint method tends to perform better in the extreme right tail than the integrated normalized second-order approximation. This conclusion not only holds for the current numerical example but for many other realizations of $S$ we investigated, with different values for $r$ and parameters $n_i$ and $p_i$.

## 4 Conclusion

This paper examined saddlepoint approximations to the distribution of the sum of independent binomial random variables with different success probabilities. The saddlepoint methods were shown to provide very accurate estimates for the pmf and the right-tail probabilities for the cumulative distribution function of the sum.

## References

AKAHIRA, M. and K. TAKAHASHI (2001), A higher order large-deviation approximation for the discrete distributions, *Journal of the Japan Statistical Society* **31**, 257–267.

AKAHIRA, M., K. TAKAHASHI and K. TAKEUCHI (1999), The higher order large-deviation approximation for the distribution of the sum of independent discrete random variables, *Communications in Statistics – Theory and Methods* **28**, 705–726.

BENNEYAN, J. C. and A. TAŞELI (2010), Exact and approximate probability distributions of evidence-based bundle composite compliance measures, *Health Care Management Science* **13**, 193–209.

BUTLER, R. W. (2007), *Saddlepoint approximations with applications*, Cambridge University Press, Cambridge, MA.

BUTLER, K. and M. STEPHENS (1993), *The distribution of a sum of binomial random variables*, Technical Report No. 467, Department of Statistics, Stanford University, Stanford, CA.

CHEN, X.–H., A. P. DEMPSTER and J. S. LIU (1994), Weighted finite population sampling to maximize entropy, *Biometrika* **81**, 457–469.

DANIELS, H. E. (1954), Saddlepoint approximations in statistics, *Annals of Mathematical Statistics* **25**, 631–650.

DANIELS, H. E. (1987), Tail probability approximations, *International Statistical Review* **55**, 37–48.

DAVISON, A. C. and C. SEMADENI (2004), Discussion on the paper by Wakefield, *Journal of the Royal Statistical Society. Series A* **167**, 434–435.

EISINGA, R. and B. PELZER (2011), Saddlepoint approximations to the mean and variance of the extended hypergeometric distribution, *Statistica Neerlandica* **65**, 22–31.

GILLESPIE, C. S. and E. RENSHAW (2007), An improved saddlepoint approximation, *Mathematical Biosciences* **208**, 359–374.

HONG, Y. (2011), *On computing the distribution function for the sum of independent and non-identical random indicators*, Technical Report No. 11-2, Department of Statistics, Virginia Tech, Blacksburg, VA.

JOHNSON, N. L., A. W. KEMP and S. KOTZ (2005), *Univariate discrete distributions*, Wiley, Hoboken, NJ.

JOLAYEMI, J. K. (1992), A unified approximation scheme for the convolution of independent binomial variables, *Applied Mathematics and Computation* **49**, 269–297.

KOTZ, S. and N. L. JOHNSON (1984), Effects of false and incomplete identification of defective items on the reliability of acceptance sampling, *Operations Research* **32**, 575–583.

LUGANNANI, R. and S. RICE (1980), Saddlepoint approximation for the distribution of the sum of independent random variables, *Advances in Applied Probability* **12**, 475–490.

ONG, S. H. (1995), Some stochastic models leading to the convolution of two binomial variables, *Statistics and Probability Letters* **22**, 161–166.

PAOLELLA, M. S. (2007), *Intermediate probability. A computational approach*, Wiley, Chicester.

SHAH, B. K. (1973), On the distribution of the sum of independent integer valued random variables, *The American Statistician* **27**, 123–124.

SMALLEY, S. L., J. A. WOODWARD and C. G. S. PALMER (1996), A general statistical model for detecting complex-trait loci by using affected relative pairs in a genome search, *American Journal of Human Genetics* **58**, 844–860.

WOODWARD, J. A. and C. G. S. PALMER (1997), On the exact convolution of discrete random variables, *Applied Mathematics and Computing* **83**, 69–77.

Received: 20 January 2012. Revised: 12 October 2012.