Statistica Neerlandica

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

Rob Eisinga, Manfred Te Grotenhuis and Ben Pelzer

Department of Social Science Research Methods and Department of Sociology,
Radboud University Nijmegen, PO Box 9104, 6500 HE Nijmegen, The Netherlands

January 18, 2012

Abstract

We discuss saddlepoint approximations to the distribution of the sum of independent non-identically distributed binomial random variables. The saddlepoint solution is the root of a polynomial equation. The paper provides an expression for the coefficients of a polynomial of any degree, the root of which can be found using a simple root-finding algorithm. We examine the accuracy of the saddlepoint methods for a sum of ten 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

Running head: Saddlepoint approximations for sum of non-identical binomials

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).

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 however if the number of outcomes with non-zero probability is large. While exact calculation is feasible with computer algebra systems such as MATHEMATICA, approximation methods continue to be widely used and explored in the literature (e.g., BENNEYAN and TAŞELI, 2010; HONG, 2011). There are several approximations for a single binomial distribution, comprehensively discussed by JOHNSON, KEMP and KOTZ (2005). Some of these approximations provide highly accurate estimates, but additional research is required, such as reported in BUTLER and STEPHENS (1993), to determine whether this accuracy generalizes to a distribution of a sum of binomial random variables, each with different success probability. Also, the distribution of the convolution can be evaluated to any degree of accuracy using Monte Carlo simulation. However, this alternative is likewise inefficient due to the large number of samples required to obtain meaningful estimates. There are computationally efficient simulation-based approaches such as importance sampling and the cross entropy method (RUBINSTEIN and KROESE, 2004), but these methods require an additional layer of computational effort.

This paper 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. Although their derivation is fairly complicated, the resulting equations are straightforward to use. An accessible and detailed introduction to saddlepoint approximations with many applications is provided by BUTLER (2007). PAOLLELA (2007) offers a computational approach.

The remainder of the paper is organized as follows. Section 2 considers the probability distribution of convoluted binomial variables and discusses saddlepoint approximations. The saddlepoint approach implies finding the root of a polynomial equation and the section provides an expression for the coefficients of a polynomial of any degree. 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 r mutually independent binomially distributed discrete random variables taking integer values 0,1,2,..., with X_i having index n_i and probability p_i, i.e., $X_i \sim \text{Bin}(n_i, p_i)$. The probability mass function (pmf) of the sum $S = \sum_{i=1}^{r} X_i$ of the r binomials is then given by (BENNEYAN and TAŞELI, 2010)

$$P(S=s) = P\left(\sum_{i=1}^{r} X_i = s\right) = \sum_{x_1=\max(0,s-\sum_{i=2}^{r} n_i)}^{\min(s,n_1)} \Bigg\{ P(X_1 = x_1) \Bigg(\sum_{x_2=\max(0,s-x_1-\sum_{i=3}^{r} n_i)}^{\min(s-x_1,n_2)} P(X_2 = x_2)$$

$$\cdots \Bigg[\sum_{x_k=\max(0,s-\sum_{i=1}^{k-1} x_i - \sum_{i=k+1}^{r} n_i)}^{\min(s-\sum_{i=1}^{k-1} x_i, n_k)} P(X_k = x_k) \cdots \Bigg(\sum_{x_{r-1}=\max(0,s-\sum_{i=1}^{r-2} x_i - n_r)}^{\min(s-\sum_{i=1}^{r-2} x_i, n_{r-1})} P(X_{r-1} = x_{r-1}) P\left(X_r = s - \sum_{i=1}^{r-1} x_i\right) \Bigg) \Bigg] \Bigg) \Bigg\}.$$

The computational impediment is in the $r-1$ nested summations required for complete enumeration over all possible observations consistent with the sum. As indicated, such calculation is infeasible 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 (pgf) for the random variable X_i and ∂ denotes jth-order partial differentiation. As the pgf 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

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

While the use of this recurrence formula requires far less computation than 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 $\left[p_i (1 - p_i)^{-1} \right]^j$ 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 probability mass function of sum S. The cumulant generating function of the convolution is

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

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

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

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,

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

where $a = (n_1 + n_2 - s)p_1 p_2$, $b = -(n_1 + n_2 - 2s)p_1 p_2 + (n_1 - s)p_1 + (n_2 - s)p_2$, and $c = -sp_1 p_2 + s(p_1 + p_2) - s$. However, for larger values of r the saddlepoint $\hat{u}$ must be determined by numerically solving the saddlepoint equation $K^{'}(\hat{u}) - s = 0$ for u. As shown in Appendix A.1, the saddlepoint is the root of the polynomial equation

$$G_r(\hat{v}) = a_r \hat{v}^r + a_{r-1}\hat{v}^{r-1} + a_{r-2}\hat{v}^{r-2} + \ \dots \ + a_1\hat{v} + a_0 = s,$$

with coefficients given by

$$a_{r-k} = (-1)^k \sum_{\substack{j=0 \\ t\in T_{r-j}:\ t_i = n_i, p_i}}^{r-1} \binom{r-j}{r-k} (-1)^j \ \tfrac{r-k}{r-j} \ \sum n_i(t) - s \quad \prod p_i(t),\ k = 0,\dots,r,$$

where T_{r-j} denotes the set of all subsets of $(r - j)$ integers that can be selected from r. For instance, if $r = 4$, then for $j = 0,1,2,3$ we have T_4={{1,2,3,4}}, T_3={{1,2,3},{1,2,4},{1,3,4},{2,3,4}}, T_2={{1,2},{1,3},{1,4},{2,3},{2,4},{3,4}}, and T_1={{1},{2},{3},{4}}, respectively. As an example, Appendix A.2 provides the coefficients of the polynomial of degree 4. There always exists a unique real root for the polynomial 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 rather easy. The root can be found with a simple root-finding algorithm such as Newton's method (RIDGWAY SCOTT, 2011), which is monotonically convergent from a suitable starting value.

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

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

and

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

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

$$\begin{aligned} P(S=0) &= \prod\nolimits_{i=1}^{r} P(X_i = 0) = \prod\nolimits_{i=1}^{r} (1-p_i)^{n_i}, \\ P(S=\max(s)) &= \prod\nolimits_{i=1}^{r} P(X_i = n_i) = \prod\nolimits_{i=1}^{r} p_i^{n_i}. \end{aligned}$$

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

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

The approximate tail probabilities of S can be determined by numerically integrating $\overline{P}_2(s)$. 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

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

provided that $s \neq E(S)$. The symbols Φ and ϕ denote, respectively, the distribution and density function of a standard normal random variable, $\hat{w} = \mathrm{sign}(\hat{u})\{2\hat{u}K'(\hat{u}) - 2K(\hat{u})\}^{1/2}$, where $\mathrm{sign}(\hat{u})$ captures the sign ± for $\hat{u}$, $\hat{u}_1 = \{1-\exp(-\hat{u})\}\{K''(\hat{u})\}^{1/2}$, 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

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

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 $\hat{u}_2 = \hat{u}\{K''(\hat{u})\}^{1/2}$, $\hat{\kappa}_3 = K'''(\hat{u})\{K''(\hat{u})\}^{-3/2}$, and $\hat{\kappa}_4 = K''''(\hat{u})\{K''(\hat{u})\}^{-2}$. 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

We examined the accuracy of the saddlepoint approximations 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 obtained by Newton's method, 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 Σn_i and probability $r^{-1}\Sigma p_i$, the normal approximation, matching the first two moments, and the Poisson distribution, matching the mean of S. The computation of the cumulant-based GC approximation $\hat{P}_6(s)$ is presented in Appendix A.3. The fitted normal density approximation with mean $\mu = K'(0)$ and variance $\sigma^2 = K''(0)$ is of the form

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

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

Table 1 about here

As can be seen in the top panel of Table 1, the normalized second-order saddlepoint approximation $\overline{P}_2(s)$ 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, since 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 $\overline{P}_2(s)$, 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 accuracy displayed in Table 1. For the extreme right tail it provides results that agree to the tenth 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 $\overline{P}_2(S \geq s)$, the DANIELS (1987) second-order continuity-corrected saddlepoint approximation to the right-tail probability $\hat{P}_4(S \geq s)$, the Gram-Charlier 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.

Table 2 about here

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 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 solution is the root of a polynomial equation and we introduced an expression for the coefficients of a polynomial of any degree, the root of which can easily be determined with a root-finding algorithm such as Newton's method. The saddlepoint methods were shown to provide very accurate estimates for the probability mass function and the right-tail probabilities for the cumulative distribution function of the sum.

The saddlepoint approximation requires an iterative procedure to obtain the root of the saddlepoint equation and once available the method is straightforward to apply. However, if the number of binomial variables gets large, the expression for the polynomial equation becomes lengthy, which in turn makes determining the approximate probabilities difficult, if not impossible. In that case a truncated, as opposed to full, saddlepoint approximation introduced by RENSHAW (1998) may be used, which yields a compact approximate expression for the probability mass function. As shown by MATIS and GUARDIOLA (2005) this truncated saddlepoint method is also straightforward to implement.

Appendix

A.1 Coefficients a_{r-k} in $G_r(\hat{v})$

The saddlepoint $\hat{u}$ must be determined by solving the saddlepoint equation

$$K^{'}(\hat{u}) - s = \sum\nolimits_{i=1}^{r} \frac{n_i p_i \hat{v}}{1 - p_i + p_i \hat{v}} - s = \prod\nolimits_{i=1}^{r} (1 - p_i + p_i \hat{v})^{-1} \left\{ \sum\nolimits_{k=0}^{r} a_{r-k} \hat{v}^{r-k} - s \right\} = 0$$

for u, where $\hat{v} = \exp(\hat{u})$ and a_{r-k} are the polynomial coefficients to be determined. This equation can only be zero if the term in curly brackets is zero, i.e.,

$$G_r(\hat{v}) = \sum\nolimits_{k=0}^{r} a_{r-k} \hat{v}^{r-k} = s.$$

If we apply the saddlepoint equation to r non-identical binomials, collect the coefficients of $\hat{v}^{r-k}$, replace the numbers in the coefficients of the polynomial by binomial coefficients, take powers of -1 to capture the sign ±, and use the notation $\{t \in T_{r-j} : t_i = n_i, p_i\}$ to denote the elements of a subset of a set as explained in the main text, then we have

$$G_r(\hat{v}) = \{[\sum\nolimits_{i=1}^{r} n_i - s] \prod\nolimits_{i=1}^{r} p_i\} \hat{v}^r +$$

$$\cdots$$

$$\{(-1)^m \sum_{\substack{j=0 \\ t \in T_{r-j}:\ t_i = n_i, p_i}}^{r-1} \begin{pmatrix} r-j \\ r-m \end{pmatrix} (-1)^j \ \tfrac{r-m}{r-j} \sum n_i(t) - s \quad \prod p_i(t)\} \hat{v}^{r-m} +$$

$$\cdots$$

$$\{(-1)^r \sum_{\substack{j=0 \\ t \in T_{r-j}:\ t_i = n_i, p_i}}^{r-1} (-1)^j - s \prod p_i(t)\} = s,$$

where $m = 1, ..., r-1$. These three expressions may subsequently be collapsed into the expression for a_{r-k} presented in the main text. To illustrate, for $r = 1, 2, 3$ binomials we have

$r = 1:$
$G_1(\hat{v}) = a_1 \hat{v} + a_0 = \{(n_1 - s) p_1\} \hat{v} + \{s p_1\} = s$

$r = 2:$
$$G_2(\hat{v}) = a_2\hat{v}^2 + a_1\hat{v} + a_0 = \{(n_1 + n_2 - s)p_1p_2\}\hat{v}^2 + \{-(n_1 + n_2 - 2s)p_1p_2 + (n_1 - s)p_1 + (n_2 - s)p_2\}\hat{v}$$
$$+ \{-sp_1p_2 + s(p_1 + p_2)\} = s,$$

$r = 3:$
$$G_3(\hat{v}) = a_3\hat{v}^3 + a_2\hat{v}^2 + a_1\hat{v} + a_0 = \{(n_1 + n_2 + n_3 - s)p_1p_2p_3\}\hat{v}^3 +$$
$$\{-(2(n_1 + n_2 + n_3) - 3s)p_1p_2p_3 + (n_1 + n_2 - s)p_1p_2 + (n_1 + n_3 - s)p_1p_3 + (n_2 + n_3 - s)p_2p_3\}\hat{v}^2 +$$
$$\{(n_1 + n_2 + n_3 - 3s)p_1p_2p_3 - (n_1 + n_2 - 2s)p_1p_2 - (n_1 + n_3 - 2s)p_1p_3 - (n_2 + n_3 - 2s)p_2p_3 +$$
$$(n_1 - s)p_1 + (n_2 - s)p_2 + (n_3 - s)p_3\}\hat{v} +$$
$$\{sp_1p_2p_3 - s(p_1p_2 + p_1p_3 + p_2p_3) + s(p_1 + p_2 + p_3)\} = s,$$

The coefficients of the polynomial of degree 4 are given below.

A.2 Polynomial of degree 4

$$G_4(\hat{v}) = a_4\hat{v}^4 + a_3\hat{v}^3 + a_2\hat{v}^2 + a_1\hat{v} + a_0 = s,$$

where $\hat{v} = \exp(\hat{u})$ and

$$\begin{aligned}
a_4 &= (n_1 + n_2 + n_3 + n_4 - s)p_1p_2p_3p_4,\\
a_3 &= -(3(n_1 + n_2 + n_3 + n_4) - 4s)p_1p_2p_3p_4 + (n_1 + n_2 + n_3 - s)p_1p_2p_3 + (n_1 + n_2 + n_4 - s)p_1p_2p_4\\
&\quad + (n_1 + n_3 + n_4 - s)p_1p_3p_4 + (n_2 + n_3 + n_4 - s)p_2p_3p_4,\\
a_2 &= (3(n_1 + n_2 + n_3 + n_4) - 6s)p_1p_2p_3p_4 - (2(n_1 + n_2 + n_3) - 3s)p_1p_2p_3 - (2(n_1 + n_2 + n_4) - 3s)p_1p_2p_4\\
&\quad - (2(n_1 + n_3 + n_4) - 3s)p_1p_3p_4 - (2(n_2 + n_3 + n_4) - 3s)p_2p_3p_4 + (n_1 + n_2 - s)p_1p_2 + (n_1 + n_3 - s)p_1p_3\\
&\quad + (n_1 + n_4 - s)p_1p_4 + (n_2 + n_3 - s)p_2p_3 + (n_2 + n_4 - s)p_2p_4 + (n_3 + n_4 - s)p_3p_4,\\
a_1 &= -(n_1 + n_2 + n_3 + n_4 - 4s)p_1p_2p_3p_4 + (n_1 + n_2 + n_3 - 3s)p_1p_2p_3 + (n_1 + n_2 + n_4 - 3s)p_1p_2p_4\\
&\quad + (n_1 + n_3 + n_4 - 3s)p_1p_3p_4 + (n_2 + n_3 + n_4 - 3s)p_2p_3p_4 - (n_1 + n_2 - 2s)p_1p_2 - (n_1 + n_3 - 2s)p_1p_3\\
&\quad - (n_1 + n_4 - 2s)p_1p_4 - (n_2 + n_3 - 2s)p_2p_3 - (n_2 + n_4 - 2s)p_2p_4 - (n_3 + n_4 - 2s)p_3p_4 + (n_1 - s)p_1\\
&\quad + (n_2 - s)p_2 + (n_3 - s)p_3 + (n_4 - s)p_4,\\
a_0 &= -sp_1p_2p_3p_4 + s(p_1p_2p_3 + p_1p_2p_4 + p_1p_3p_4 + p_2p_3p_4) - s(p_1p_2 + p_1p_3 + p_1p_4 + p_2p_3 + p_2p_3 + p_3p_4)\\
&\quad + s(p_1 + p_2 + p_3 + p_4).
\end{aligned}$$

A.3 Gram-Charlier type A series of order 6

$$\hat{P}_6(S=s)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(s-\mu)^2}{2\sigma^2}\right\}\left\{1+\frac{\kappa_3}{6\sigma^3}H_3(z)+\frac{\kappa_4}{24\sigma^4}H_4(z)+\frac{\kappa_5}{120\sigma^5}H_5(z)+\frac{1}{720}\left[\frac{\kappa_6}{\sigma^6}+10\left(\frac{\kappa_3}{\sigma^3}\right)^2\right]H_6(z)\right\},$$

where

$$\kappa_3=\sum_{i=1}^{r}n_i(p_i-3p_i^2+2p_i^3),$$

$$\kappa_4=\sum_{i=1}^{r}n_i(p_i-7p_i^2+12p_i^3-6p_i^4),$$

$$\kappa_5=\sum_{i=1}^{r}n_i(p_i-15p_i^2+50p_i^3-60p_i^4+24p_i^5),$$

$$\kappa_6=\sum_{i=1}^{r}n_i(p_i-31p_i^2+180p_i^3-390p_i^4+360p_i^5-120p_i^6),$$

$$H_3(z)=z^3-3z,$$

$$H_4(z)=z^4-6z^2+3,$$

$$H_5(z)=z^5-10z^3+15z,$$

$$H_6(z)=z^6-15z^4+45z^2-15,$$

$z=(s-\mu)/\sigma$, with $\mu=\sum_{i=1}^{r}n_iP_i$ and $\sigma^2=\sum_{i=1}^{r}n_ip_i(1-p_i)$.

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, 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.

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

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.

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.

Matis, T. I. and I. G. Guardiola (2005), *Estimating rare event probabilities using truncated saddlepoint approximations*, Department of Industrial Engineering, Texas Tech University, Lubbock, TX.

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.

Renshaw, E. (1998), Saddlepoint approximations for stochastic processes with truncated cumulant generating functions. *IMA Journal of Mathematics Applied in Medicine and Biology* **15**, 41–52.

Ridgway Scott, L. (2011), *Numerical analysis*, Princeton University Press, Princeton, NJ.

Rubenstein, R. V. and D. P. Kroese (2004), *The cross-entropy method,* Springer, New York.

Shah, B. K. (1973), On the distribution of the sum of independent integer valued random variables, *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.

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

s	$\hat{u}(s)$	$P(s)$	$\overline{P}_2(s)$	$\hat{P}_6(s)$	Bin $(s; \sum n_i, \overline{p}_i)$	$N(s; \mu, \sigma^2)$	Pois $(s; \mu)$
$n_i =$ 12, 14, 4, 2, 20, 17, 11, 1, 8, 11 p_i =.074, .039, .095, .039, .053, .043, .067, .018, .099, .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 =.00074, .00039, .00095, .00039, .00053, .00043, .00067, .00018, .00099, .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 =.74, .39, .95, .39, .53, .43, .67, .18, .99, .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
590	0.084	0.0133	0.0133	0.0133	0.0132	0.0133	0.0126
600	0.132	$0.0^{2}4501$	$0.0^{2}4501$	$0.0^{2}4501$	$0.0^{2}5141$	$0.0^{2}4501$	$0.0^{2}8500$
610	0.181	$0.0^{3}9419$	$0.0^{3}9419$	$0.0^{3}9419$	$0.0^{2}1321$	$0.0^{3}9460$	$0.0^{2}4854$
620	0.230	$0.0^{3}1213$	$0.0^{3}1213$	$0.0^{3}1213$	$0.0^{3}2230$	$0.0^{3}1230$	$0.0^{3}2353$
630	0.279	$0.0^{5}9581$	$0.0^{5}9581$	$0.0^{5}9581$	$0.0^{4}2463$	$0.0^{5}9902$	$0.0^{3}9708$
640	0.328	$0.0^{6}4630$	$0.0^{6}4630$	$0.0^{6}4628$	$0.0^{5}1773$	$0.0^{6}4931$	$0.0^{3}3418$

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

s	$P(S \geq s)$	$\bar{P}_2(S \geq s)$	$\hat{P}_4(S \geq s)$	$\hat{P}_6(S \geq s)$	$\mathrm{Bin}(s; \sum n_i, \bar{p}_i)$	$N(s+\frac{1}{2}; \mu, \sigma^2)$	$\mathrm{Pois}(s; \mu)$
$n_i =$ 12, 14, 4, 2, 20, 17, 11, 1, 8, 11 $p_i =$.074, .039, .095, .039, .053, .043, .067, .018, .099, .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 $p_i =$.00074, .00039, .00095, .00039, .00053, .00043, .00067, .00018, .00099, .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 $p_i =$.74, .39, .95, .39, .53, .43, .67, .18, .99, .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
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	$0.0^{5}1599$	$0.0^{5}1635$	$0.0^{5}1599$	$0.0^{5}1598$	$0.0^{5}5091$	$0.0^{5}1728$	$0.0^{2}2594$