ESAIM: PS 16 (2012) 425–435
DOI: 10.1051/ps/2010027

ESAIM: Probability and Statistics
www.esaim-ps.org

TAIL APPROXIMATIONS FOR SAMPLES FROM A FINITE POPULATION WITH APPLICATIONS TO PERMUTATION TESTS

Zhishui Hu[1], John Robinson[2] and Qiying Wang[2]

Abstract. This paper derives an explicit approximation for the tail probability of a sum of sample values taken without replacement from an unrestricted finite population. The approximation is shown to hold under no conditions in a wide range with relative error given in terms of the standardized absolute third moment of the population, β_{3N}. This approximation is used to obtain a result comparable to the well-known Cramér large deviation result in the independent case, but with no restrictions on the sampled population and an error term depending only on β_{3N}. Application to permutation tests is investigated giving a new limit result for the tail conditional probability of the statistic given order statistics under mild conditions. Some numerical results are given to illustrate the accuracy of the approximation by comparing our results to saddlepoint approximations requiring strong conditions.

Mathematics Subject Classification. 62E20, 60F05.

Received September 11, 2009. Revised June 17, 2010.

1. Introduction

Let $X_1, X_2, \ldots, X_n$ be a simple random sample drawn without replacement from a finite population $\{a\}_N = \{a_1, \ldots, a_N\}$ with $\sum a_k = 0$ and $\sum a_k^2 = N$, where $n < N$ and throughout the paper $\sum$ without limits denotes summation over k from 1 to N. Let

$$S_n = \sum_{j=1}^{n} X_j, \qquad p = n/N, \qquad q = 1 - p, \qquad \omega_N^2 = Npq, \qquad b_N = \max_k |a_k|.$$

Under appropriate conditions, the so-called finite central limit theorem (see [7,8]) states that, as $n \to \infty$,

$$\sup_x \left| P(S_n \geq x\omega_N) - \{1 - \Phi(x)\} \right| \to 0,$$

where $\Phi(x)$ is the distribution function of a standard normal variate. The central limit theorem is useful when x is not too large. There are two approaches for estimating the error of the normal approximation. One approach

Keywords and phrases. Cramér large deviation, saddlepoint approximations, moderate deviations, finite population, permutation tests.

[1] Department of Statistics and Finance, University of Science and Technology of China, Hefei 230026, Anhui, P.R. China. huzs@ustc.edu.cn

[2] School of Mathematics and Statistics The University of Sydney, NSW, 2006, Australia. johnr@maths.usyd.edu.au; qiying@maths.usyd.edu.au

Article published by EDP Sciences

© EDP Sciences, SMAI 2012

https://doi.org/10.1051/ps/2010027 Published online by Cambridge University Press

is to investigate the absolute error *via* Berry–Esseen bounds and Edgeworth expansions. This has been done by many researchers. We only refer to Bikelis [3] and Höglund [9] for the rates in the Erdös and Rényi central limit theorem, and Bickel and van Zwet [2], Robinson [12], Babu and Bai [1] as well as Bloznelis [4,5] for the Edgeworth expansions.

Another approach is to investigate the relative error of $P\big(S_n \geq x\omega_N\big)$ to $1-\Phi(x)$. In this direction, Hu *et al.* (HRW) [10] derived the following result: there is an absolute constant $A > 0$ such that

$$\exp\big\{-A(1+x)^3\beta_{3N}/\omega_N\big\} \leq \frac{P\big(S_n \geq x\omega_N\big)}{1-\Phi(x)} \leq \exp\big\{A(1+x)^3\beta_{3N}/\omega_N\big\}, \tag{1.1}$$

for $0 \leq x \leq (1/A)\omega_N/b_N$, where $\beta_{3N} = E|X_1|^3 = \sum |a_k|^3/N$. As a direct consequence of (1.1), HRW [10] also established the following Cramér-type large deviation result: there exists an absolute constant $A > 0$ such that

$$\frac{P\big(S_n \geq x\omega_N\big)}{1-\Phi(x)} = 1 + O(1)(1+x)^3\beta_{3N}/\omega_N, \tag{1.2}$$

for $0 \leq x \leq (1/A)\min\big\{\omega_N/b_N, (\omega_N/\beta_{3N})^{1/3}\big\}$.

Results (1.1) and (1.2) are useful because they provide not only the relative error but also a Berry–Esseen rate of convergence. It is also interesting to note that the results depend only on β_{3N} with an absolute constant and we have no restrictions on the $\{a\}_N$ and p. As mentioned in Remark 1.4 of HRW [10], however, the results (1.1) and (1.2) do not provide sufficiently precise results on the accuracy of the normal approximation in the large deviation region.

In the present paper, an analogue of Cramér's large deviation result in the independent case (see Chap. 8 of Petrov (1975), for example) is obtained in Theorem 2.2, which essentially improves the results (1.1) and (1.2) in the situation that an explicit large deviation range for x is provided and the error term is improved by reducing the power of x. Rather than studying the accuracy of the normal approximation, we derive in Theorem 2.1 an approximation for the tail probability $P\big(S_n \geq x\omega_N\big)$ with no restrictions on the population or on p, with a relative error depending only on β_{3N}, and giving an explicit large deviation range for x. This result is comparable to those of Robinson [13], Robinson *et al.* [14] and Wang [15] which require a number of very strong restrictive conditions. Furthermore, numerical examples show that the tail approximation of Theorem 2.1 is effectively as accurate as the saddle-point approximations of Booth and Butler [6] and Wang [15]. This indicates that the tail approximation of Theorem 2.1 can be used to replace the saddle-point approximations in many applications since these require a number of very strong restrictive conditions on the sampled data.

In Section 2 we present, in Theorem 2.1, a tail probability approximation and from this we derive, in Theorem 2.2, a result analogous to that of Cramér for independent and identically distributed random variables. The result of Theorem 2.1 may be considered a first order saddlepoint approximation, which can be applied with no conditions. It gives a precise region for existence of a unique solution of the saddlepoint equations and a relative error for an explicit large deviation region. Theorem 2.2 improves essentially Theorem 1 of Robinson [11] in two aspects. First we provide an explicit range for the x and, second, the error term in Robinson [11] depends on p, which limits essential applications of the result, whereas our results hold true with no restrictions on the $\{a\}_N$ and p. Section 3 applies our main results to two-sample permutation tests, obtaining in Theorem 3.1 an approximation for the conditional probability given the order statistics with small relative error almost surely. Some numerical comparisons are presented in Section 4. Finally, in Section 5, we present the proofs of the main results. Throughout $A, A_1, A_2, \ldots$ are absolute constants.

2. Main results

The complex moment generating function of S_n is shown in (1) of Robinson [11] to be

$$Q_n(u+iv) = (2\pi B_n)^{-1}\int_{-\pi}^{\pi} \exp[\sum K((u+iv)a_k(pq)^{-1/2} + \alpha + i\theta)]d\theta,$$

https://doi.org/10.1051/ps/2010027 Published online by Cambridge University Press

for any α, where $B_n = \binom{N}{n} p^n q^{N-n}$, $K(z) = \log(p\ \mathrm{e}^{\ qz} + q\ \mathrm{e}^{\ -pz})$, $p, q > 0$ and $p + q = 1$. Noting that

$$K'(t) = pq(\ \mathrm{e}^{\ t} - 1)/(p\ \mathrm{e}^{\ t} + q) \quad \text{and} \quad K''(t) = pq\ \mathrm{e}^{\ t}/(p\ \mathrm{e}^{\ t} + q)^2, \tag{2.1}$$

we have that $K''(t) > 0$ for any $t \in R$ and $K'(t) > 0$ or < 0 if $t > 0$ or $t < 0$. These facts imply that, for each $u \in R$, $\sum K'(ua_k + \alpha)$ is strictly increasing in α and the equation

$$\sum K'(u\, a_k + \alpha) = 0 \tag{2.2}$$

has a unique solution $\alpha_N(u)$. Furthermore, $\alpha_N(0) = 0$ and $|\alpha_N(u)| < C$ if $0 < u < C/b_N$.

In the following, let $a_{(1)} \le a_{(2)} \le \ldots \le a_{(N)}$ be the ordered values of $a_1, \ldots, a_N$, and $K_k(u), K_k'(u)$ and $K_k''(u)$ be the values of $K(t), K'(t)$ and $K''(t)$ evaluated at $t = u\, a_k + \alpha_N(u)$. Throughout we take $x > 0$ and $u > 0$. Our main result is given as follows.

Theorem 2.1.

(i) *Assume that* $\min\{a_{(k+1)} - a_{(k)} : k = 1, \ldots, N-1\} > \delta_N > 0$, *we have* $\alpha_N(u)/u \to -(a_{(N-n+1)} + a_{(N-n)})/2$ *as* $u \to \infty$.

(ii) *The function* $m_N(u) = \sum a_k K_k'(u)$ *is strictly increasing,* $m_N(u) \to a_{(N)} + \ldots + a_{(N-n+1)}$ *as* $u \to \infty$, *and the equation*

$$m_N(u) = x\omega_N, \tag{2.3}$$

has a unique solution u_x *whenever* $0 < x < \left[a_{(N)} + \ldots + a_{(N-n+1)}\right]/\omega_N$.

(iii) *For any given constant* $C > 0$, *if* $0 < x < m_N(C/b_N)/\omega_N$, *then*

$$P(S_n > x\omega_N) \quad = \quad (1 - \Phi(u_x\sigma_N))\, \Lambda_n(x) \left(1 + O_1(1 + u_x\sigma_N)\beta_{3N}/\omega_N\right), \tag{2.4}$$

where

$$\Lambda_n(x) = \frac{\omega_N}{\sqrt{\sum K_k''(u_x)}} \exp\left\{\sum K_k(u_x) - u_x\, x\, \omega_N + u_x^2\sigma_N^2/2\right\},$$

u_x *is the solution of equation* (2.3),

$$\sigma_N^2 \equiv \sigma_N^2(u_x) = \sum a_k^2 K_k''(u_x) - \left(\sum a_k K_k''(u_x)\right)^2 / \sum K_k''(u_x),$$

$\beta_{3N} = E|X_1|^3$ *and* O_1 *is bounded by a constant depending only on* C.

(iv) *The result* (2.4) *holds true if* $0 < x < \omega_N/(2eb_N)$.

Theorem 2.1 provides an approximation for the tail probability $P(S_n > x\omega_N)$ in an explicit wide large deviation range for the x. As mentioned in Section 1, the tail approximation of Theorem 2.1 is comparable in accuracy to the well-known saddle-point approximations developed in [6,13,15]. The numerical examples to illustrate this statement are given in Section 4. However, unlike these results which require a number of very strong restrictive conditions on the sampled data, our result (2.4) is established under no conditions since we always have $E|X_1|^3 = \sum |a_k|^3/N < \infty$. The error can be improved under a smoothness condition and further terms to improve accuracy can then be obtained.

As a consequence of this result, we also have the following Theorem 2.2, which is comparable to the classical Cramér large deviation result as in the independent case. This result also provides an approximation for the tail probability $P(S_n > x\omega_N)$ under no conditions and has theoretical interest. This approximation gives less accuracy compared to the saddle-point approximation (2.4) as illustrated in Section 4, and its purpose is to give a rate of convergence for the Central Limit Theorem rather than to provide an approximation.

https://doi.org/10.1051/ps/2010027 Published online by Cambridge University Press

Theorem 2.2. *For $0 < x < \omega_N/(2eb_N)$, we have*

$$\frac{P(S_n > x\omega_N)}{1-\Phi(x)} = \exp\left\{x^3\lambda_N(x)\right\}\left(1 + O_2(1+x)\beta_{3N}/\omega_N\right), \tag{2.5}$$

where $x^3\lambda_N(x) = \sum K_k(u_x) - u_x\, x\, \omega_N + x^2/2$, u_x is the solution of equation (2.3) and O_2 is bounded by an absolute positive constant.

Remark 2.3. As in [11], $\lambda_N(x)$ may be represented as a power series of x with coefficients not depending on p and convergent in some circle. For more details on these lines, we refer to (45) and (46) in [11].

Remark 2.4. For an asymptotic version of these results, consider a sequence of finite populations $\{a_{N1}, \ldots, a_{NN}\}$, $N = 1, 2, 3, \ldots$ The results of Theorems 2.1 and 2.2 hold for each member of this sequence. The relative error term in (2.4), $O_1(1 + u_x\sigma_N)\beta_{3N}/\omega_N$ can then be replaced by an order term $O((1 + u_x\sigma_N)\beta_{3N}/\omega_N)$ and for the results to give appropriate limit theorems we need to assume that $(1 + u_x\sigma_N)\beta_{3N}/\omega_N \to 0$ for $0 < x < \omega_N/2eb_N$ as $N \to \infty$. In the particular example of the Wilcoxon two sample statistic considered in Section 4, $b_N = \sqrt{3(N-1)/(N+1)}$, $0 < u_x < C/b_N$, σ_N can be bounded above and below by $D\sqrt{pq}$ and $d\sqrt{pq}$, where d and D depend only on C, and $|\beta_{3N}| < 1$, so the relative error is $O(1/\omega_N)$. The range of x in this case is $0 < x < c\omega_N$. Here then, the relative error term is $O((1+x)/\sqrt{n})$ so it is possible to have $p \to 0$ so long as $n = Np \to \infty$. Further asymptotic results are given in Section 3.

3. Applications to permutation tests

Suppose that we observe the random variables $X_1, \ldots, X_n$ and $X_{n+1}, \ldots, X_N$. Consider the model:

> $X_1, \ldots, X_n$ and $X_{n+1}, \ldots, X_N$ are independent and identically distributed (iid) with distribution functions $F(x - \theta_1)$ and $F(x - \theta_2)$, respectively.

We are interested in the hypothesis $H : \theta_1 - \theta_2 = \delta_0$. Write

$$\widetilde{a}_k = (Y_k - \bar{Y}) \Bigg/ \left[\frac{1}{N}\sum (Y_i - \bar{Y})^2\right]^{1/2}, \qquad k = 1, \ldots, N, \tag{3.1}$$

where $\bar{Y} = \frac{1}{N}\sum Y_k$ and

$$Y_i = \begin{cases} X_i - \delta_0, & i = 1, \ldots, n, \\ X_i, & i = n+1, \ldots, N. \end{cases}$$

The typical permutation test statistic for the hypothesis H is given by

$$T_N = \frac{1}{(Npq)^{1/2}} \sum_{i=1}^{n} \widetilde{a}_{R_i},$$

where $(R_1, \ldots, R_N)$ is a random vector, independent of all preceding random variables, taking each permutation of $(1, \ldots, N)$ with equal probability. Given the order statistics $Y_{(1)}, \ldots, Y_{(N)}$, the conditional distribution of T_N has been investigated widely. For instance, Bickel and van Zwet [2] and Robinson [12] investigated Edgeworth expansions. Robinson [13] discussed the saddle point approximation under smoothness conditions. By using Theorems 2.1 and 2.2, this section derives explicit approximations of $P^*(T_N \geq x)$ in the large deviation range assuming only the existence of a finite moment of the sampled population, where P^* indicates the conditional distribution given $Y_{(1)}, \ldots, Y_{(N)}$. Simulations in Section 4 show that the approximation of Theorem 2.1 provides accuracy which is comparable to the saddlepoint approximation given in Booth and Butler [6] and Wang [15].

https://doi.org/10.1051/ps/2010027 Published online by Cambridge University Press

In order to use Theorems 2.1 and 2.2, we still adopt the notation as in Section 2 except that $a_k = \widetilde{a}_k$, where $\widetilde{a}_k$ is given by (3.1). The main result in this section is as follows.

Theorem 3.1. *Under the assumption that the hypothesis H is true, if $E|X_1|^m < \infty$ for some $m \geq 3$, then for $j = 1, 2$,*

$$\frac{P^*(T_N > x)}{\Psi_{jN}(x)} \to 1, \quad \text{a.s.} \quad \textit{as } N \to \infty, \tag{3.2}$$

holds true uniformly in $0 < x \leq O\big(\sqrt{Npq}/\tau_N\big)$, where $\tau_N/N^{1/m} \to \infty$, $\Psi_{1N}(x) = (1 - \Phi(u_x\sigma_N))\Lambda_n(x)$ and $\Psi_{2N}(x) = (1 - \Phi(x)) \exp\left\{x^3\lambda_N(x)\right\}$ are defined as in Theorems 2.1 and 2.2 respectively.

Proof. By the result (iv) of Theorem 2.1 and (5.11) below, we have

$$\frac{P^*(T_n > x)}{\Psi_{1N}(x)} = \Big(1 + O'(1+x)\beta^*_{3N}/\omega_N\Big), \tag{3.3}$$

for $0 < x < \omega_N/(2eb^*_N)$, where $b^*_N = \max_k |\widetilde{a}_k| = \max_k |Y_k - \bar{Y}| \Big/ \Big[\frac{1}{N}\sum (Y_k - \bar{Y})^2\Big]^{1/2}$,

$$\beta^*_{3N} = \frac{1}{N}\sum |Y_k - \bar{Y}|^3 \Big/ \left[\frac{1}{N}\sum (Y_k - \bar{Y})^2\right]^{3/2}$$

and O' is bounded by an absolute positive constant. Noting that $W_1 = Y_1 - EY_1, \ldots, W_N = Y_N - EY_N$ are iid random variables with $E|Y_1|^m < \infty$, under the hypothesis H and the assumption $E|X_1|^m < \infty$, for $m \geq 3$, it is readily seen from the law of large numbers that, as $N \to \infty$,

$$\beta^*_{3N} \to E|W_1|^3/(EW_1^2)^{3/2} \quad \text{a.s.} \quad \text{and} \quad b^*_N \geq \frac{1}{2}\max_k |Y_k - \bar{Y}|/(EW_1^2)^{1/2} \quad \text{a.s.} \tag{3.4}$$

From (3.3) and (3.4), Theorem 3.1 for $j = 1$ will follow if we prove, as $N \to \infty$,

$$\tau_N^{-1} \max_{1 \leq k \leq N} |Y_k - \bar{Y}| \to 0, \quad \text{a.s.}$$

This follows from the law of large numbers, since $E|Y_1|^m \leq A\, E|X_1|^m < \infty$ implies that

$$\tau_N^{-1} \max_{1 \leq k \leq N} |Y_k - EY_k| \leq \frac{N^{1/m}}{\tau_N}\left[\frac{1}{N}\sum |Y_k - EY_k|^m\right]^{1/m} \to 0, \quad \text{a.s.},$$

whenever $\tau_N/N^{1/m} \to \infty$. Theorem 3.1 for $j = 2$ is similar except that we replace the result (iv) of Theorem 2.1 by Theorem 2.2 and hence the details are omitted. This also completes the proof of Theorem 3.1. □

Remark 3.2. Note that in the asymptotic case for the bound $0 < x < \sqrt{Npq}/\tau_N$ to be non-trivial we require that p does not decrease as fast as τ_N^2/N.

4. Numerical comparisons

We will present two examples to illustrate the accuracy of the results of Theorems 2.1 and 2.2, comparing these to the full saddlepoint given by Booth and Butler [6] (with no proof of the bounds) and to the approximation of Wang [15], obtained under very strong conditions. We look at the data set used by Wang [15] in his Table 1 where $N = 36$ and $n = 5$, and we also consider the approximation for the Wilcoxon rank sum statistic for

https://doi.org/10.1051/ps/2010027 Published online by Cambridge University Press

TABLE 1. Approximations to $P(\sum_{j=1}^{5} d_{R_j} > y)$ for Wang [15] data set $d_1, d_2, \ldots, d_{36}$.

y	1.2	3.6	6.0	8.8	12.8
MC	0.3350	0.1523	0.0523	0.0110	0.0004
Wang93	0.3371	0.1513	0.0532	0.0109	0.0004
BB	0.3387	0.1523	0.0534	0.0109	0.0004
Theorem 2.1	0.3613	0.1561	0.0528	0.0104	0.0004
Theorem 2.2	0.3602	0.1503	0.0470	0.0079	0.0002
Normal	0.3589	0.1391	0.0353	0.0040	0.0001

TABLE 2. Approximations to $P(\sum_{i=1}^{10} R_i > y)$ for random permutation $R_1, R_2, \ldots, R_{16}$ of $1, 2, \ldots, 16$.

y	115	114	113	112	111	110	109	108
Exact	0.00012	0.00025	0.00050	0.00087	0.00150	0.00237	0.00387	0.00562
BB	0.00011	0.00025	0.00049	0.00088	0.00149	0.00240	0.00370	0.00552
Theorem 2.1	0.00009	0.00021	0.00042	0.00077	0.00132	0.00214	0.00333	0.00499
Theorem 2.2	0.00002	0.00006	0.00017	0.00036	0.00068	0.00122	0.00204	0.00326
Normal	0.00048	0.00070	0.00103	0.00150	0.00214	0.00303	0.00424	0.00586

$P(S_n \geq y)$ where $N = 16$ and $n = 10$. In the first case, in Table 1, we look at Monte Carlo estimates based on 1 000 000 samples (MC), then the results taken from Wang [15] Table 2, an approximation based on the full conditional saddlepoint of Booth and Butler [6] (BB), the approximations from Theorems 2.1 and 2.2 and the normal approximation. For the Wilcoxon, in Table 2, we give the exact probabilities in place of a Monte Carlo approximation, and all other approximations but those of Wang [15], using a continuity correction in this lattice case to improve the approximations.

In each case the results are similar. The results of Wang [15] and of the method of Booth and Butler [6] are remarkably close to the Monte Carlo values throughout the range. The methods of Theorem 2.1 give quite good results throughout the range, again as might be expected since this gives a first order approximation to the full conditional saddlepoint. The second order approximations of Booth and Butler [6] and of Wang [15] have second order relative accuracy only under a smoothness condition, whereas no such restriction applies to the results of Theorem 2.1. Theorem 2.2 does not give particularly good approximations. We might remark that, although the errors in Theorems 2.1 and 2.2 are the same, the result (2.4) of Theorem 2.1 (iii) can be extended, under a smoothness condition, to an indirect Edgeworth approximation with a smaller relative error, as in Robinson [13], but such an improvement is not possible for the approximation from (2.5).

5. PROOFS OF THEOREMS

Proof of Theorem 2.1. (i). From (3)

$$K'(t) + p = \frac{p\mathrm{e}^t}{p\mathrm{e}^t + q} \text{ and } K'(t) - q = \frac{-q\mathrm{e}^{-t}}{p + q\mathrm{e}^{-t}}. \tag{5.1}$$

Let $\Delta = a_{(N-n+1)} - a_{(N-n)}$. If $\alpha/u = -(a_{(N-n+1)} + a_{(N-n)})/2 - \epsilon$, for arbitrarily small $\Delta/2 > \epsilon > 0$, then for $k = 1, \ldots, N - n - 1$,

$$0 < K'(ua_{(k)} + \alpha) + p = \mathrm{e}^{-u\Delta/2 - u\epsilon} O(\mathrm{e}^{-u\delta_N}),$$

for $k = N - n + 2, \ldots, N$,

$$0 < -K'(ua_{(k)} + \alpha) + q = \mathrm{e}^{-u\Delta/2 + u\epsilon} O(\mathrm{e}^{-u\delta_N})$$

https://doi.org/10.1051/ps/2010027 Published online by Cambridge University Press

and

$$K'(ua_{(N-n)}+\alpha)+p+K'(ua_{(N-n+1)}+\alpha)-q=\frac{p^2\mathrm{e}^{-u\Delta/2-u\epsilon}-q^2\mathrm{e}^{-u\Delta/2+u\epsilon}}{(p\mathrm{e}^{ua_{(N-n)}+\alpha}+q)(p+q\mathrm{e}^{-ua_{(N-n+1)}-\alpha})},$$

which is less than 0 for u large. So, noting that $(N-n)p-nq=0$,

$$\sum_{k=1}^{N-n}(K'(ua_{(k)}+\alpha)+p)+\sum_{k=N-n+1}^{N}(K'(ua_{(k)}+\alpha)-q)<0,$$

for u large. In the same way, putting $\alpha/u=-(a_{(N-n+1)}+a_{(N-n)})/2+\epsilon$, we can show that

$$\sum_{k=1}^{N-n}(K'(ua_{(k)}+\alpha)+p)+\sum_{k=N-n+1}^{N}(K'(ua_{(k)}+\alpha)-q)>0.$$

Since ϵ is chosen arbitrarily, this implies $\alpha_N(u)/u\to-(a_{(N-n+1)}+a_{(N-n)})/2$ as $u\to\infty$.

(ii). Since $\alpha'_N(u)=-\sum a_kK''_k(u)/\sum K''_k(u)$ by differentiating (2.2), we have

$$\begin{aligned} m'_N(u) &= \sum a_k\left[a_k+\alpha'_N(u)\right]K''_k(u) \\ &= \sum[a_k+\alpha'_N(u)]^2K''_k(u)=\sigma_N^2(u)>0, \end{aligned} \tag{5.2}$$

for $u\in R$. So $m_N(u)$ is strictly increasing. Then, as $u\to\infty$, using (i) and $\sum a_k=0$,

$$\begin{aligned} m_N(u) &= pq\sum a_k\frac{\mathrm{e}^{q(ua_k+\alpha_N(u))}-\mathrm{e}^{-p(ua_k+\alpha_N(u))}}{p\mathrm{e}^{q(ua_k+\alpha_N(u))}+q\mathrm{e}^{-p(ua_k+\alpha_N(u))}} \\ &\sim pq\Big[\sum a_kI(ua_k+\alpha_N(u)>0)/p-\sum a_kI(ua_k+\alpha_N(u)<0)/q\Big] \\ &= \sum a_kI(ua_k+\alpha_N(u)>0) \\ &\to a_{(N)}+\ldots+a_{(N-n+1)}. \end{aligned}$$

This, together with the fact $m_N(0)=0$ (recalling $\alpha_N(0)=0$), yields that for $0<x\omega_N<a_{(N)}+\ldots+a_{(N-n+1)}$, there is a unique solution of (2.3). This is as might be expected as $a_{(N)}+\ldots+a_{(N-n+1)}$ is the maximum value for S_n.

(iii). From (ii), for any $0<x<m_N\left(C/b_N\right)/\omega_N$, the solution u_x of equation (2.3) is unique and $0<u_x\leq C/b_N$. Let $H_n(t;u_x)=E\mathrm{e}^{u_xS_n}I(S_n\leq t)/E\mathrm{e}^{u_xS_n}$. We have

$$\begin{aligned} P(S_n>x\omega_N) &= Ee^{u_xS_n}\int_{x\omega_N}^{\infty}\mathrm{e}^{-u_xt}\mathrm{d}H_n(t;u_x) \\ &= E\mathrm{e}^{u_xS_n}\int_{x\omega_N}^{\infty}\mathrm{e}^{-u_xt}\mathrm{d}\Phi\left(\frac{t-m_N(u_x)}{\sigma_N}\right) \\ &\quad + Ee^{u_xS_n}\int_{x\omega_N}^{\infty}\mathrm{e}^{-u_xt}\mathrm{d}\left(H_n(t;u_x)-\Phi\left(\frac{t-m_N(u_x)}{\sigma_N}\right)\right) \\ &:= E\mathrm{e}^{u_xS_n}(I_1+I_2). \end{aligned} \tag{5.3}$$

Recalling $0<u_x\leq C/b_N$, it follows from Theorem 3.1 of HRW (2007) that

$$|I_2| \leq 2\,\mathrm{e}^{-x\,u_x\omega_N}\sup_t\left|H_n(t;u_x)-\Phi\left(\frac{t-m_N(u_x)}{\sigma_N}\right)\right|\leq A\,\mathrm{e}^{-x\,u_x\omega_N}\,\beta_{3N}/\omega_N, \tag{5.4}$$

https://doi.org/10.1051/ps/2010027 Published online by Cambridge University Press

and

$$\begin{aligned} E\mathrm{e}^{u_x S_n} &= \frac{1}{G_n(p)\sqrt{\sum K_k''(u_x)}} \exp\Big\{\sum K_k(u_x)\Big\}(1+O_2/\omega_N) \\ &= \frac{\omega_N}{\sqrt{\sum K_k''(u_x)}} \exp\Big\{\sum K_k(u_x)\Big\}(1+O_3/\omega_N), \end{aligned} \tag{5.5}$$

where O_2 and O_3 are constants depending only on C, and we have used the estimate, $G_n(p) = \sqrt{2\pi}\binom{N}{n}p^n q^{N-n} = \omega_N^{-1}(1+O_6/\omega_N^2)$, where $|O_6| \le 1/6$, which follows from Stirling's formula. Also,

$$I_1 = (2\pi)^{-1/2}\mathrm{e}^{-x\,u_x\omega_N}\int_0^\infty \mathrm{e}^{-u_x\sigma_N t - t^2/2}\mathrm{d}t = \mathrm{e}^{-u_x x\omega_N + u_x^2\sigma_N^2/2}(1-\Phi(u_x\sigma_N)). \tag{5.6}$$

Clearly $I_1 \ge \mathrm{e}^{-u_x\,x\omega_N}(1-\Phi(1))$ if $u_x\sigma_N \le 1$. By noting that

$$\frac{1}{x^2}\int_x^\infty \mathrm{e}^{-y^2/2}\mathrm{d}y > \int_x^\infty \frac{1}{y^2}\mathrm{e}^{-y^2/2}\mathrm{d}y = \frac{1}{x}\mathrm{e}^{-x^2/2} - \int_x^\infty \mathrm{e}^{-y^2/2}\mathrm{d}y,$$

we have $1-\Phi(x) > x\Phi'(x)/(1+x^2)$ for $x \ge 1$ and hence $I_1 \ge \frac{1}{2\sqrt{2\pi}}\,\mathrm{e}^{-u_x x\omega_N}(u_x\sigma_N)^{-1}$ if $u_x\sigma_N > 1$. From these facts and (5.3)–(5.6), we obtain

$$\begin{aligned} P(S_n > x\omega_N) &= E\mathrm{e}^{u_x S_n}\mathrm{e}^{-u_x x\omega_N + u_x^2\sigma_N^2/2}(1-\Phi(u_x\sigma_N))\,(1+I_2/I_1) \\ &= (1-\Phi(u_x\sigma_N))\,\Lambda_n(x)\left(1+O_1(1+u_x\sigma_N)\beta_{3N}/\omega_N\right), \end{aligned}$$

O_1 is bounded by a constant depending only on C, which completes the proof of (iii).

(iv). We first show that, for any given $C > 0$, if $0 < u < C/b_N$, then

$$Npq\,\mathrm{e}^{-2C} \le m_N'(u) = \sigma_N^2(u) \le Npq\,\mathrm{e}^{2C}. \tag{5.7}$$

In fact, by recalling $|\alpha_N(u)| \le C$ and $|ua_k + \alpha_N(u)| < 2C$ if $0 < u < C/b_N$, it follows from (2.1) that, for each k and $0 < u < C/b_N$,

$$pq\mathrm{e}^{-2C} \le K_k''(u) \le pq\mathrm{e}^{2C}, \tag{5.8}$$

where we use $pq\mathrm{e}^t(p\mathrm{e}^t+q)^{-2} = pq\mathrm{e}^{-t}(p+q\mathrm{e}^{-t})^{-2}$. This, together with the first equality of (5.2), yields that

$$m_N'(u) = \sigma_N^2(u) \le \sum a_k^2 K_k''(u) \le Npq\,\mathrm{e}^{2C}.$$

Similarly it follows from $\sum a_k = 0$, the second equality of (5.2) and (5.8) that

$$m_N'(u) = \sigma_N^2(u) \ge pq\,\mathrm{e}^{-2C}\sum(a_k + \alpha_N'(u))^2 \ge Npq\,\mathrm{e}^{-2C}.$$

This proves (5.7). From (5.7), if $0 < x < \omega_N/(2eb_N)$, then

$$m_N\,(1/2b_N) = \int_0^{1/2b_N} \sigma_N^2(u)\mathrm{d}u \ge Npq/(2e\,b_N) \ge x\omega_N. \tag{5.9}$$

Thus (2.4) holds from (iii). □

https://doi.org/10.1051/ps/2010027 Published online by Cambridge University Press

Proof of Theorem 2.2. It follows from Theorem 2.1 (iv) that, for $0 < x < \omega_N/(2eb_N)$,

$$\begin{aligned}\frac{P(S_n > x\omega_N)}{1-\Phi(x)} &= \frac{\omega_N}{\sqrt{\sum K_k''(u_x)}}\frac{\psi(u_x\sigma_N)}{\psi(x)}\exp\Big\{\sum K_k(u_x) - u_x\, m_N(u_x) + x^2/2\Big\} \\ &\quad\times\Big(1 + O_1'(1+u_x\,\sigma_N)\beta_{3N}/\omega_N\Big), \end{aligned} \tag{5.10}$$

where $\psi(x) = (1-\Phi(x))/\Phi'(x)$, u_x is the solution of (2.3) and O_1' is bounded by an absolute constant. Now the theorem follows from (5.10), if we show that, for $0 < x < \omega_N/(2eb_N)$,

$$e^{-3/2}x \le u_x\sigma_N(u_x) \le e^{3/2}\,x \tag{5.11}$$

and

$$\frac{\omega_N}{\sqrt{\sum K_k''(u_x)}}\frac{\psi(u_x\sigma_N)}{\psi(x)} = 1 + O_4(1+x)\beta_{3N}/\omega_N, \tag{5.12}$$

where O_4 is bounded by an absolute constant.

We will complete the proof by establishing these two results. First note that $0 < u_x \le 1/(2b_N)$ from (5.9) and Theorem 2.1 (ii). It follows from (5.7) with $C = 1/2$ that

$$u_x\,\omega_N^2\,e^{-1} < x\omega_N = m_N(u_x) = m_N(0) + u_x\,m_N'(u_1) < u_x\,\omega_N^2\,e \tag{5.13}$$

where $0 < u_1 < u_x$. Thus $e^{-1}\omega_N^{-1} \le u_x/x \le e\omega_N^{-1}$. This, together with (5.7) with $C = 1/2$ again, implies (5.11).

We next prove (5.12). Write $\alpha_N \equiv \alpha_N(u)$. Recalling $\sum(a_k + \alpha_N')K_k''(u) = 0$, we have $\sum(a_k+\alpha_N')^2K_k'''(u) + \sum\alpha_N''K_k''(u) = 0$. So $\alpha_N'' = -\sum(a_k+\alpha_N')^2K_k'''(u)/\sum K_k''(u)$. This, together with $\alpha_N'(u) = -\sum a_kK_k''(u)/\sum K_k''(u)$ and (5.2), yields that

$$\begin{aligned} m_N''(u) &= \sum a_k\,\alpha_N''\,K_k''(u) + \sum a_k\,(a_k+\alpha_N')^2\,K_k'''(u) \\ &= \sum(a_k+\alpha_N')^3K_k'''(u). \end{aligned} \tag{5.14}$$

If $0 < u \le C/b_N$, by recalling $|u\,a_k + \alpha_N(u)| \le 2C$, noting that $K_k'''(u) = K_k''(u)(q - p\mathrm{e}^{u\,a_k+\alpha_N(u)})/(p\mathrm{e}^{u\,a_k+\alpha_N(u)} + q)$ and using (5.8), it is readily seen that $|K_k'''(u)| \le pq\mathrm{e}^{2C}$. On the other hand, $|\alpha_N'(u)| \le \mathrm{e}^{4C}\sum|a_k|/N$ by (5.8) whenever $0 < u \le C/b_N$. Taking these facts into (5.14), we obtain that

$$|m_N''(u)| < A_1pq\left[\sum|a_k|^3 + \frac{1}{N^2}\Big(\sum|a_k|\Big)^3\right] \le A_2\omega_N^2\beta_{3N}, \tag{5.15}$$

for $0 < u \le 1/(2b_N)$. Using (5.15), Taylor's expansion and the facts that $m_N(0) = 0$, $m_N'(0) = \omega_N^2$ and $u_x/x \le e/\omega_N$, we have

$$\begin{aligned} |x - u_x\omega_N| &= \frac{1}{\omega_N}\big|m_N(u_x) - m_N(0) - u_xm_N'(0)\big| \\ &\le \frac{u_x^2}{\omega_N}|m_N''(u_2)| \le A_3\,x^2\beta_{3N}/\omega_N, \end{aligned} \tag{5.16}$$

where $0 \le u_2 \le u_x \le 1/(2b_N)$. Similarly,

$$\begin{aligned} u_x\big|\sigma_N(u_x) - \omega_N\big| &\le u_x\,|\sigma_N^2(u_x) - \omega_N^2|/\omega_N \le eu_xx|m_N'(u_x) - m_N'(0)|/\omega_N^2 \\ &\le eu_xx|m_N''(u_3)|/\omega_N^2 \le A_3\,x^2\beta_{3N}/\omega_N, \end{aligned} \tag{5.17}$$

https://doi.org/10.1051/ps/2010027 Published online by Cambridge University Press

where $0 \le u_3 \le u_x \le 1/(2b_N)$. From (5.16) and (5.17), we obtain

$$\big|x - u_x\sigma_N(u_x)\big| \quad \le \quad \big|x - u_x\,\omega_N\big| + u_x\big|\sigma_N(u_x) - \omega_N\big| \le A_4\,x^2\beta_{3N}/\omega_N.$$

This, together with (5.11), yields that

$$\Big|\frac{\psi(u_x\sigma_N)}{\psi(x)} - 1\Big| \quad = \quad \frac{|u_x\sigma_N(u_x) - x||\psi'(\theta)|}{\psi(x)} \le A_5(1+x)\beta_{3N}/\omega_N, \tag{5.18}$$

where $x \le \theta \le u_x\sigma_N(x)$ and we have used the following estimates:

$$\psi(t) \ge \min\{(2t)^{-1},\ 1 - \Phi(1)\} \quad \text{and} \quad |\psi'(t)| = |t\psi(t) - 1| \le t^{-2} \quad \text{for} \quad t > 0.$$

From (5.18), the result (5.12) will follow if we prove

$$\left|\frac{\omega_N}{\sqrt{\sum K_k''(u_x)}} - 1\right| \le A\,x/\omega_N. \tag{5.19}$$

In fact, by noting that for any $0 < u < \infty$

$$\frac{\mathrm{d}\sum K_k''(u)}{\mathrm{d}u} = \sum(a_k + \alpha_N'(u))\,K_k'''(u),$$

it follows from Taylor's expansion that

$$\sum K_k''(u_x) \quad = \quad N\,K''(0) + u_x\,R,$$

where $|R| \le A\,pq\,\sum|a_k|$. Therefore, by recalling $u_x \le ex/\omega_N$ and $K''(0) = pq$, for $0 < x < \omega_N/(2eb_N)$,

$$\Big|\sum K_k''(u_x) - \omega_N^2\Big| \quad \le \quad A\,u_x\,pq\,\sum|a_k| \le A\,x\,pq\sqrt{N\sum a_k^2}/\omega_N \le A\,x\,\omega_N.$$

This, together with (5.8), implies that, for $0 < x < \omega_N/(2eb_N)$,

$$\left|\frac{\omega_N}{\sqrt{\sum K_k''(u_x)}} - 1\right| \le \frac{|\sum K_k''(u_x) - \omega_N^2|}{\sum K_k''(u_x)} \le A\,x/\omega_N,$$

which yields (5.19). Proof of Theorem 2.2 is now complete. □

Acknowledgements. The authors wish to thank the referees for comments which have led to improvements in the paper. Hu is partially supported by NSFC (No. 10801122). Robinson and Wang are partially supported by the Australian Research Council.

References

[1] G.J. Babu and Z.D. Bai, Mixtures of global and local Edgeworth expansions and their applications. *J. Multivariate Anal.* **59** (1996) 282–307.

[2] P.J. Bickel and W.R. van Zwet, Asymptotic expansions for the power of distribution-free tests in the two-sample problem. *Ann. Statist.* **6** (1978) 937–1004.

[3] A. Bikelis, On the estimation of the remainder term in the central limit theorem for samples from finite populations. *Stud. Sci. Math. Hung.* **4** (1969) 345–354 (in Russian).

https://doi.org/10.1051/ps/2010027 Published online by Cambridge University Press

[4] M. Bloznelis, One and two-term Edgeworth expansion for finite population sample mean. Exact results I. *Lith. Math. J.* **40** (2000) 213–227.
[5] M. Bloznelis, One and two-term Edgeworth expansion for finite population sample mean. Exact results II. *Lith. Math. J.* **40** (2000) 329–340.
[6] J.G. Booth and R.W. Butler, Randomization distributions and saddlepoint approximations in generalized linear models. *Biometrika* **77** (1990) 787–796.
[7] P. Erdös and A. Rényi, On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hungarian Acad. Sci.* **4** (1959) 49–61.
[8] J. Hájek, Limiting distributions in simple random sampling for a finite population. *Publ. Math. Inst. Hugar. Acad. Sci.* **5** (1960) 361–374.
[9] T. Höglund, Sampling from a finite population. A remainder term estimate. *Scand. J. Stat.* **5** (1978) 69–71.
[10] Z. Hu, J. Robinson and Q. Wang, Crameŕ-type large deviations for samples from a finite population. *Ann. Statist.* **35** (2007) 673–696.
[11] J. Robinson, Large deviation probabilities for samples from a finite population. *Ann. Probab.* **5** (1977) 913–925.
[12] J. Robinson, An asymptotic expansion for samples from a finite population. *Ann. Statist.* **6** (1978) 1004–1011.
[13] J. Robinson, Saddlepoint approximations for permutation tests and confidence intervals. *J. R. Statist. Soc. B* **20** (1982) 91–101.
[14] J. Robinson, T. Hoglund, L. Holst and M.P. Quine, On approximating probabilities for small and large deviations in R^d. *Ann. Probab.* **18** (1990) 727–753.
[15] S. Wang, Saddlepoint expansions in finite population problems. *Biometrika* **80** (1993) 583–590.

https://doi.org/10.1051/ps/2010027 Published online by Cambridge University Press