PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 135, Number 5, May 2007, Pages 1535–1545
S 0002-9939(07)08676-5
Article electronically published on January 5, 2007

A BERRY-ESSEEN THEOREM FOR HYPERGEOMETRIC PROBABILITIES UNDER MINIMAL CONDITIONS

S. N. LAHIRI AND A. CHATTERJEE

(Communicated by Edward C. Waymire)

Abstract. In this paper, we consider simple random sampling without replacement from a dichotomous finite population and derive a necessary and sufficient condition on the finite population parameters for a valid large sample Normal approximation to Hypergeometric probabilities. We then obtain lower and upper bounds on the difference between the Normal and the Hypergeometric distributions solely under this necessary and sufficient condition.

1. Introduction

Consider a dichotomous finite population of size N having M individuals of 'type A' and $N-M$ individuals of 'type B'. Suppose a sample of size n is drawn at random, without replacement from this population. Let X denote the number of 'type A'-individuals in the sample. Then, X is said to have the Hypergeometric distribution with parameters n, M, N, written as $X \sim Hyp(n; M, N)$. The probability mass function (p.m.f.) of X is given by

$$P(X=x) \equiv P(x; n, M, N) = \begin{cases} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} & \text{if} \quad x = 0, 1 \ldots, n, \\ 0 & \text{otherwise,} \end{cases} \tag{1.1}$$

where, for any two integers $r \geq 1$ and s,

$$\binom{r}{s} = \begin{cases} \frac{r!}{s!(r-s)!} & \text{if} \quad 0 \leq s \leq r, \\ 0 & \text{otherwise,} \end{cases} \tag{1.2}$$

with $0! = 1$ and $r! = 1 \cdot 2 \cdots r$. Let $f = \frac{n}{N}$ denote the sampling fraction and let $p = \frac{M}{N}$ denote the proportion of the 'type A' objects in the population. The Hypergeoemetric distribution plays an important role in many areas of statistics, including sample surveys (Burstein (1975) and Wendell and Schmee (1996)), capture-recapture methods (Seber (1970) and Wittes (1972)), analysis of contingency tables (Blyth and Staudte (1997)), statistical quality control (Patel and Samaranayake (1991) and Sohn (1997)), etc. Normal approximations to the Hypergeometric probabilities $P(.; n, M, N)$ of (1.1) are classical in the cases where the sampling fraction f and the proportion p are bounded away from 0 and 1; see, for example, Feller (1971). The nonstandard cases correspond to the extremes where f or p take values near the boundary values 0 and 1. Although the nonstandard cases arise frequently in all

Received by the editors April 12, 2005 and, in revised form, February 16, 2006.
2000 *Mathematics Subject Classification.* Primary 60F05; Secondary 60G10, 62E20, 62D05.
This research was partially supported by NSF grant no. DMS 0306574.

these areas of application, the validity and accuracy of the Normal approximation in such situations are not well studied. This paper is devoted to investigating the behavior of Normal approximation for both standard and nonstandard cases.

The main results of the paper give a necessary and sufficient condition on the parameters f and p for a valid Normal approximation. It is shown that a Normal limit for properly centered and scaled version of X holds if and only if

$$Np(1-p)f(1-f) \to \infty. \tag{1.3}$$

As a consequence, we conclude that for the Normal distribution function to approximate the distribution function of X, all four quantities, namely, (i) the number M $(= Np)$ of 'type A' objects, (ii) the number of 'type B' objects, $N - M$, (iii) the sample size n, as well as (iv) the size of the unselected objects $N - n$ in the population, must tend to infinity.

We next investigate the rate of Normal approximation to the distribution of X. Note that X is the sum of a collection of n *dependent* Bernoulli random variables. In Section 2, we establish a Berry-Esseen Theorem on the rate of Normal approximation to the distribution function of X solely under the necessary and sufficient condition (1.3). It is shown that under (1.3) the rate of approximation is $O([Np(1-p)f(1-f)]^{-1/2})$. It is also shown in Section 2 that this rate is *optimal* in the sense that the (Kolmogorov) distance between the cdfs of the Hypergeometric distribution and the Normal distribution is bounded *below* by a constant multiple of $[Np(1-p)f(1-f)]^{-1/2}$. Thus, the accuracy of Normal approximation necessarily deteriorates as the factor $Np(1-p)f(1-f)$ becomes small. In particular, for a given value of the population size N, the accuracy decreases as either p or f (or both) approach the boundary values 0 and 1. Note that the rate $O([Np(1-p)f(1-f)]^{-1/2})$ is equivalent to the standard rate $O(n^{-1/2})$ (for sums of n *independent* Bernoulli random variables, say) only when p is bounded away from 0 and 1 and f bounded away from 1. However, for p and f close to these boundary points, the rate of approximation can be substantially slower. In such situations, the dependence of the Bernoulli random variables associated with X has a nontrivial effect on the accuracy of the Normal approximation.

The rest of the paper is organized as follows. We conclude Section 1 with a brief literature review. Section 2 introduces the asymptotic framework and contains the results on the validity of the Normal approximation and the Berry-Esseen theorem. Proofs of all the results are given in Section 3. For results on Normal approximations to Hypergeometric probabilities in the standard cases where the sampling fraction f and the proportion p are bounded away from 0 and 1, see Feller (1971). For general p and f, Nicholson (1956) derived some very precise bounds for the point probabilities $P(.; n, M, N)$ (cf. (1.1)) using some special normalizations of the Hypergeometric random variable X. General methods for proving the CLT for sample means under sampling without replacement from finite populations are given by Madow (1948), Erdos and Renyi (1959) and Hajek (1960). In relation to the earlier work, the main contribution of our paper is to establish the theoretical validity of Normal approximation and the Berry-Esseen Theorem under *minimal* conditions.

2. Theoretical results

Let r be a positive integer valued variable and for each $r \in \mathbb{N} = \{1, 2, \ldots\}$, let X_r be a random variable having the Hypergeometric distribution with parameters

(n_r, M_r, N_r). Thus we consider a sequence of dichotomous finite populations indexed by r, with the population of objects of type A and the sampling fraction respectively given by

$$p_r = \frac{M_r}{N_r} \quad \text{and} \quad f_r = \frac{n_r}{N_r} \quad \text{for all} \quad r \in \mathbb{N}. \tag{2.1}$$

To avoid trivialities, all through the paper, we shall assume that for all $r \in \mathbb{N}$,

$$1 \leq M_r < N_r, 1 \leq n_r < N_r, \quad \text{and} \quad N_r^{-1} = o\,(1) \quad r \to \infty. \tag{2.2}$$

Thus, p_r , $f_r \in (0,1)$ for all $r \in \mathbb{N}$. Let

$$\sigma_r^2 \equiv N_r p_r q_r f_r (1 - f_r), \tag{2.3}$$

where $q_r = 1 - p_r$. The first result concerns the validity of the Normal approximation to the distribution of X_r.

Theorem 2.1. *Suppose that* (2.2) *holds and that* $X_r \sim Hyp(n_r, M_r, N_r)$, $r \in \mathbb{N}$. *Then there exists a Normal random variable* $W \sim N(\mu, \sigma^2)$ *for some* $\mu \in \mathbb{R}$ *and* $\sigma \in (0, \infty)$ *such that*

$$\Delta_r \equiv \sup_{x \in \mathbb{R}} \left| P\left(\frac{X_r - n_r p_r}{\sigma_r} \leq x \right) - P\,(W \leq x) \right| \longrightarrow 0 \quad \textit{as} \quad r \to \infty \tag{2.4}$$

if and only if

$$\sigma_r^2 \to \infty \qquad \textit{as} \quad r \to \infty. \tag{2.5}$$

When (2.5) *holds, one must have* $\mu = 0$ *and* $\sigma = 1$.

Theorem 2.1 shows that the Normal approximation to the Hypergeometric distribution holds solely under the condition that the function σ_r^2 of the parameters p_r and f_r goes to infinity with r. In particular, it is not necessary to impose separate conditions on the asymptotic behavior of the three sequences $\{n_r\}_{r \geq 1}$, $\{p_r\}_{r \geq 1}$ and $\{f_r\}_{r \geq 1}$. A necessary condition for (2.5) is that $n_r \to \infty$ and $(N_r - n_r) \to \infty$ as $r \to \infty$. This follows by noting that $\sigma_r^2 = n_r p_r q_r (1 - f_r) = (N_r - n_r) p_r q_r f_r \leq \min\{n_r, N_r - n_r\}$ for all $r \geq 1$. Thus, for the Normal approximation to hold, both the sample size n_r and the residual sample size $(N_r - n_r)$ must become unbounded as $r \to \infty$. By similar arguments, it follows that for the validity of the Normal approximation, we must also have

$$\min\{M_r, (N_r - M_r)\} \longrightarrow \infty \quad \text{as} \quad r \to \infty, \tag{2.6}$$

i.e., the number of objects of type A and type B must go to infinity with r.

Condition (2.5) also allows the proportion p_r of 'type A' objects in the population and the sampling fraction f_r to simultaneously converge to the extreme points 0 and 1 at certain rates. If the sequence $\{f_r\}_{r \geq 1}$ is bounded away from 0 and 1 and (2.2) holds, then the CLT of Theorem 2.1 holds if and only if (iff)

$$\frac{1}{N_r} = o(q_r \wedge p_r) \quad \text{as} \quad r \to \infty, \tag{2.7}$$

i.e., iff (2.6) holds. Similarly, for $\{p_r\}_{\{r \geq 1\}}$ bounded away from 0 and 1, the CLT holds *iff*

$$\frac{1}{N_r} = o(f_r \wedge (1 - f_r)) \quad \text{as} \quad r \to \infty. \tag{2.8}$$

However, when both $\{p_r\}_{\{r \geq 1\}}$ and $\{f_r\}_{\{r \geq 1\}}$ simultaneously converge to some limits in $\{0, 1\}$, neither (2.7) nor (2.8) alone is enough to guarantee the CLT. For

example if $f_r \sim N_r^{-a}$ and $p_r \sim N_r^{-b}$ for some $0 < a, b < 1$ with $a+b > 1$, then (2.7) and (2.8) hold, but the Normal approximation is no longer valid.

Next we obtain a refinement of (2.4) by specifying the rate of convergence of Δ_r to zero.

Theorem 2.2. *Suppose that $X_r \sim Hyp(n_r, M_r, N_r)$, $r \in \mathbb{N}$, and that (2.5) holds. Then there exist constants $C_1, C_2 \in (0, \infty)$ such that for all $r \in \mathbb{N}$ with $\sigma_r > 0$,*

$$\frac{C_1}{\sigma_r} \leq \sup_{x \in \mathbb{R}} \left| P\left(\frac{X_r - n_r p_r}{\sigma_r} \leq x\right) - \Phi(x) \right| \leq \frac{C_2}{\sigma_r}, \tag{2.9}$$

where $\Phi(\cdot)$ denotes the cdf of the standard Normal distribution.

Theorem 2.2 gives a uniform Berry-Esseen theorem that shows that under (2.5), the rate of Normal approximation to the Hypergeometric distribution is uniformly $O\left(\sigma_r^{-1}\right)$ as $r \to \infty$. Further, the lower bound in (2.9) shows that the rate $O\left(\sigma_r^{-1}\right)$ is *optimal* and cannot be improved upon. A second important aspect of Theorem 2.2 is that the bound on Δ_r holds under the same condition (2.5) that is both necessary and sufficient for a Normal limit. Thus, the conditions for the Berry-Esseen theorem is also *minimal*, and this cannot be improved upon either.

When both the sequences $\{p_r\}_{r \geq 1}$ and $\{f_r\}_{r \geq 1}$ are bounded away from 0 and 1, the rate of approximation in Theorem 2.2 matches the standard rate $O(1/\sqrt{n_r})$ of Normal approximation for the sum of n_r *independent and identically distributed* (iid) random variables with a finite third moment. Although the Hypergeometric random variable X_r can be written as a sum of n_r *dependent* Bernoulli (p_r) variables, the lack of independence of the summands does not affect the rate of Normal approximation as long as the sequence $\{p_r\}_{r \geq 1}$ is bounded away from 0 and 1 and $\{f_r\}_{r \geq 1}$ is bounded away from 1. On the other hand, if $\{p_r\}_{r \geq 1}$ converges to one of the extreme values 0 and 1 or if $\{f_r\}_{r \geq 1}$ converges to 1, then $\sigma_r = o(n_r^{1/2})$ as $r \to \infty$. The lower bound in Theorem 2.2 implies that the rate of normal approximation to the Hypergeometric distribution is indeed *worse* than the standard rate $O(n_r^{-1/2})$ in such *nonstandard cases*.

3. Proofs

We now introduce some notation and notational convention to be used in this section. Let $\mathbb{Z}_+ = \{0, 1, \ldots\}$ and $\mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}$. Let $I(\cdot)$ denote the indicator function. For $x, y \in \mathbb{R}$, let $x \wedge y = \min\{x, y\}$, $x \vee y = \max\{x, y\}$, and let $\lfloor x \rfloor$ denote the largest integer not exceeding x. For $a \in (0, \infty)$, write $\phi_a(x) = \frac{1}{a}\phi(\frac{x}{a})$ and $\Phi_a(x) = \Phi(\frac{x}{a})$, $x \in \mathbb{R}$, for the density and distribution functions of a $N(0, a^2)$ variable. Write $\phi_a = \phi$ and $\Phi_a = \Phi$ for $a = 1$. Let

$$\Delta_r^*(x) = P\left(\frac{X_r - n_r p_r}{\sigma_r} \leq x\right) - \Phi(x), \quad x \in \mathbb{R}, \tag{3.1}$$

$$\delta_r = (10\max(a_{1r}, 2))^{-1}, \qquad r \geq 1, \tag{3.2}$$

where $a_{1r} = \frac{\bar{f}_r + 4}{4(1 - \bar{f}_r)}$ and where $\bar{f}_r = f_r$ if $f_r \leq \frac{1}{2}$ and $\bar{f}_r = 1 - f_r$ if $f_r > \frac{1}{2}$. We shall use C to denote a generic positive constant that does not depend on r. Unless otherwise stated, limits in order symbols are taken by letting $r \to \infty$.

The first result gives a basic approximation to Hypergeometric probabilities solely under condition (3.3) stated below.

Lemma 3.1. *Suppose that $X \sim Hyp(n; M, N)$ for a given set of integers $n, M, N \in \mathbb{N}$ such that*

$$0 < f < 1, \quad 0 < p < 1 \quad \text{and} \quad 6(np \wedge nq) \geq 1, \tag{3.3}$$

where $f = \frac{n}{N}$, $p = \frac{M}{N}$ and $q = 1 - p$. Then, for any given $\delta \in (0, \frac{1}{2}]$,

$$\log P(k; n, M, N) = -\frac{x_{k,n}^2}{2(1-f)} - \frac{1}{2}\log\left(2\pi npq(1-f)\right) + R_n^*(k) \tag{3.4}$$

for all $k \in \{0, \ldots, n\}$, where $P(k; n, M, N) = P(X = k)$ (cf. (1.1)), $x_{k,n} = \frac{x-np}{\sqrt{npq}}$ and $a_{k,n} = \frac{x_{k,n}}{(1-f)\sqrt{npq}}$, $0 \leq k \leq n$, and where, for $|a_{k,n}| \leq \delta$, the remainder term $R_n^*(k)$ admits the bound

$$\begin{aligned} |R_n^*(k)| \quad &\leq \quad \frac{1}{6npq(1-\delta)(1-f)} + \left[\frac{1}{2}|a_{k,n}| + a_{k,n}^2\left\{\frac{1}{4} + \frac{2\delta}{(1-\delta)^3}\right\}\right] \\ &\quad + |a_{k,n}|^3 npq\left(\frac{f}{4} + 1\right)\left\{\frac{1}{2} + \frac{2(1+\delta)}{(1-\delta)^3}\right\}. \end{aligned} \tag{3.5}$$

The proof of Lemma 3.1 is based on a long and careful analysis of the Hypergeometric probabilities in (1.1) using Stirling's approximation. For the proofs of Lemma 3.1 and of the next two results, see Lahiri, Chatterjee and Maiti (2004).

Lemma 3.2. *Let $g : \mathbb{R} \longrightarrow [0, \infty)$ be such that g is $\uparrow$ on $(-\infty, a)$ and g is $\downarrow$ on (a, ∞) for some $a \in \mathbb{R}$. Then, for any $k \in \mathbb{N}$, $b \in \mathbb{R}$ and $h \in (0, \infty)$,*

$$\sum_{i=o}^{k} g(b + ih) \leq \int_b^{b+hk} g(x)dx + 2hg(x_0), \tag{3.6}$$

where $g(x_0) = \max\{g(b + ih) : i = 0, 1, \ldots, k\}$.

Lemma 3.3. *Let $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$, $x \in \mathbb{R}$. Then, for any $h \in (0, \infty)$, $b \in [0, \infty)$, $j_0 \in \mathbb{N}$,*

$$\begin{aligned} &\left|h\sum_{i=0}^{j_0}\phi(b + ih) - \int_{b-\frac{h}{2}}^{b+(j_0+\frac{1}{2})h}\phi(x)dx\right| \\ \leq \quad &\frac{h^2}{12}\left[\int_{b-\frac{h}{2}}^{b+j_0h+\frac{h}{2}}|\phi''(x)|dx + (4+h)\sup\left\{|\phi''(x)| : -\frac{h}{2} < x - b < j_0h + \frac{h}{2}\right\}\right]. \end{aligned} \tag{3.7}$$

Proof of Theorem 2.1. Suppose that (2.5) holds. Fix $\epsilon \in (0, 1)$. By Chebyshev's inequality, for all $r \in \mathbb{N}$,

$$P\left(\left|\frac{X_r - n_r p_r}{\sigma_r}\right| > \frac{2}{\epsilon}\right) \quad \leq \quad \frac{\epsilon^2}{4} \cdot \frac{N_r}{N_r - 1}. \tag{3.8}$$

By Lemmas 3.1 and 3.3, for any $r \in \mathbb{N}$ with $f_r \leq \frac{1}{2}$,

$$\begin{aligned}
\Delta_{1r}(\epsilon) &\equiv \sup_{-\frac{2}{\epsilon} \leq a < b \leq \frac{2}{\epsilon}} \left| P\left(a < \frac{X_r - n_r p_r}{\sigma_r} \leq b\right) - [\Phi(b) - \Phi(a)] \right| \\
&\leq \sum_{-\frac{2\sigma_r}{\epsilon} < k - n_r p_r \leq \frac{2\sigma_r}{\epsilon}} \left| P(k; n_r, M_r, N_r) - \frac{1}{\sigma_r}\phi\left(\frac{k - n_r p_r}{\sigma_r}\right) \right| \\
&\quad + \sum_{-\frac{2}{\epsilon} \leq a < b \leq \frac{2}{\epsilon}} \left| \sum_{a\sigma_r < k - n_r p_r \leq b\sigma_r} \frac{1}{\sigma_r}\phi\left(\frac{k - n_r p_r}{\sigma_r}\right) - [\Phi(b) - \Phi(a)] \right| \\
&\leq \frac{C}{\sigma_r^2} \sum_{-\frac{2\sigma_r}{\epsilon} < k - n_r p_r \leq \frac{2\sigma_r}{\epsilon}} \exp\left(\frac{C}{\sigma_r}\right) \exp\left(-\frac{(k - n_r p_r)^2}{\sigma_r^2}\left[\frac{1}{2} - \frac{C}{\sigma_r}\right]\right) \\
&\quad + \frac{C}{\sigma_r^2}\left[\int_{-\infty}^{\infty} |\phi''(x)| dx + 1\right] + \frac{2}{\sqrt{2\pi}\sigma_r} \\
&\leq \frac{C}{\sigma_r}\left[\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{4}\right) dx + 1\right],
\end{aligned}$$

provided $\frac{C}{\sigma_r} < \frac{1}{4}$. Hence, there exists an $r_0 \in \mathbb{N}$ such that for all $r \geq r_0$ with $f_r \leq \frac{1}{2}$, $\Delta_{1r}(\epsilon) < \frac{\epsilon}{4}$. Also by Mill's ratio, $\Phi(-\frac{2}{\epsilon}) + 1 - \Phi(\frac{2}{\epsilon}) < \epsilon\phi(\frac{2}{\epsilon})$. Hence, using (3.8) and the above inequalities, it can be shown that for all $r \geq r_0$ with $f_r \leq \frac{1}{2}$,

$$\Delta_r(\epsilon) < \epsilon. \tag{3.9}$$

Next suppose that $f_r > \frac{1}{2}$. Consider the collection of $N_r - n_r$ objects that are left after the sample of size n_r has been selected from the population of size N_r. Let Y_r be the number of 'type A' objects in this collection. Then,

$$Y_r \sim Hyp(N_r - n_r; M_r, N_r) \quad \text{and} \quad P(X_r = j) = P(Y_r = M_r - j), \tag{3.10}$$

for all $r \in \mathbb{N}$ and $j \in \mathbb{Z}$. Hence,

$$Var(Y_r) = Var(X_r) \quad \text{and} \quad P(X_r \leq k) = P(Y_r \geq M_r - k).$$

Thus, for each $x \in \mathbb{R}$,

$$\begin{aligned}
& P\left(\frac{X_r - n_r p_r}{\sigma_r} \leq x\right) \\
&= P(X_r \leq \lfloor n_r p_r + x\sigma_r \rfloor) \\
&= P(Y_r \geq M_r - \lfloor n_r p_r + x\sigma_r \rfloor) \\
&= P\left(\frac{Y_r - (N_r - n_r)p_r}{\sigma_r} \geq \frac{M_r - \lfloor n_r p_r + x\sigma_r \rfloor - (N_r - n_r)p_r}{\sigma_r}\right) \\
&= P(\tilde{Y}_r \geq \check{x}_r) \qquad \text{(say)},
\end{aligned}$$

where $\tilde{Y}_r = \frac{Y_r - (N_r - n_r)p_r}{\sigma_r}$ and $\check{x}_r = \frac{M_r - \lfloor n_r p_r + x\sigma_r \rfloor - (N_r - n_r)p_r}{\sigma_r}$. Note that

$$\check{x}_r < \frac{1}{\sigma_r}\left[N_r p_r - (n_r p_r + x\sigma_r - 1) - N_r p_r + n_r p_r\right] = -x + \sigma_r^{-1}$$

and similarly, $\check{x}_r \geq -x$. Hence, this implies,

$$P(\tilde{Y}_r < \check{x}_r) \leq P(\tilde{Y}_r \leq \check{x}_r) \leq P(\tilde{Y}_r \leq -x + \sigma_r^{-1})$$

and

$$P(\tilde{Y}_r < \check{x}_r) \geq P(\tilde{Y}_r < -x) \geq P(\tilde{Y}_r \leq -x - \sigma_r^{-1}).$$

Now using the above identity and inequalities, we have

$$\begin{aligned}\Delta_r^*(x) &= |P(\tilde{Y}_r \geq \check{x}_r) - (1-\Phi(-x))| = |\Phi(-x) - P(\tilde{Y}_r < \check{x}_r)| \\ &\leq \max_{y\in A}|P(\tilde{Y}_r \leq y) - \Phi(y)| + \max_{y\in A}|\Phi(-x) - \Phi(y)|, \end{aligned} \tag{3.11}$$

where $A = \{-x - \sigma_r^{-1}, -x + \sigma_r^{-1}\}$. By repeating the arguments leading to (3.9), it follows that there exists $r_1 \in \mathbb{N}$ such that for all $r \geq r_1$ with $(1 - f_r) \leq \frac{1}{2}$,

$$\sup_{x\in\mathbf{R}}|P(\tilde{Y}_r \leq x) - \Phi(x)| \leq \epsilon. \tag{3.12}$$

Hence, (2.4) now follows from (2.5), (3.9), (3.11) and (3.12), with $W \sim N(0,1)$. In particular, if (2.5) holds, then one must have $\mu = 0$ and $\sigma = 1$.

Conversely, suppose that (2.4) holds for some $\mu \in \mathbb{R}$ and $\sigma \in (0,\infty)$. Then, for any sequences $\{a_r\}_{r\geq 1}$,$\{b_r\}_{r\geq 1} \subset \mathbb{R}$ with $a_r < b_r$ for all $r \geq 1$,

$$\left|P\left(a_r < \frac{X_r - n_r p_r}{\sigma_r} \leq b_r\right) - P(a_r < W \leq b_r)\right| \leq 2\Delta_r \to 0 \quad \text{as} \quad r \to \infty. \tag{3.13}$$

If possible, suppose that $\sigma_r < 1$ infinitely often. Then, we can pick $a_r, b_r \in [-1,1]$ such that for all such r, $a_r - b_r = 1$ and $\frac{\lfloor n_r p_r\rfloor - n_r p_r}{\sigma_r} < a_r < b_r < \frac{\lfloor n_r p_r\rfloor + 1 - n_r p_r}{\sigma_r}$. Then,

$$P\left(a_r < \frac{X_r - n_r p_r}{\sigma_r} \leq b_r\right) = 0$$

but

$$P\left(a_r < W \leq b_r\right) \geq \inf\{P(a < W \leq b) : a, b \in [-1,1], b - a = 1\} > 0,$$

infinitely often. This contradicts (3.13). Hence, we may suppose that $\sigma_r \geq 1$ for all but finitely many r's. Now define $a_r = \frac{\lfloor n_r p_r\rfloor - n_r p_r + \frac{1}{3}}{\sigma_r}$ and $b_r = \frac{\lfloor n_r p_r\rfloor - n_r p_r + \frac{2}{3}}{\sigma_r}$. Since $P(X_r \in \{0, 1, \ldots, n_r\}) = 1$,

$$P\left(a_r < \frac{X_r - n_r p_r}{\sigma_r} \leq b_r\right) = P\left(\lfloor n_r p_r\rfloor + \frac{1}{3} < X_r \leq \lfloor n_r p_r\rfloor + \frac{2}{3}\right) = 0.$$

Next using the definitions of a_r, b_r, and the fact that '$x - 1 < \lfloor x\rfloor \leq x$ for all $x \in \mathbb{R}$', we get

$$-\frac{2}{3\sigma_r} < a_r < b_r \leq \frac{2}{3\sigma_r}, \quad r \geq 1. \tag{3.14}$$

By (3.13) and (3.14), it follows that

$$\begin{aligned} &\frac{1}{3\sigma_r}\min\{\phi_\sigma(x-\mu) : |x| \leq \frac{2}{3\sigma_r}\} \\ \leq\ & \int_{a_r}^{b_r}\phi_\sigma(x-\mu)dx = P(a_r < W \leq b_r) \\ =\ & \left|P\left(a_r < \frac{X_r - n_r p_r}{\sigma_r} \leq b_r\right) - P(a_r < W \leq b_r)\right| \to 0 \quad \text{as} \quad r \to \infty. \end{aligned}$$

As a result, $\sigma_r \to \infty$ as $r \to \infty$, and (2.5) holds. This completes the proof of the theorem. □

Proof of Theorem 2.2. Let $r \in \mathbb{N}$ be an integer such that $\sigma_r \delta_r > 1$. *First, suppose that* $f_r \leq \frac{1}{2}$. Consider the case $x \leq 0$. For $k = 0, 1, \ldots, n_r$, let $\tilde{x}_k \equiv \tilde{x}_{k,r} = \frac{k-np}{\sigma}$, and define

$$\begin{aligned} K_{0,r} &= \sup\{k \in \mathbb{Z}_+ : \tilde{x}_k \leq 0\},\ K_{1,r} = \inf\{k \in \mathbb{Z}_+ : \tilde{x}_k \geq -1\}, \\ K_{2,r} &= \inf\{k \in \mathbb{Z}_+ : \tilde{x}_k \geq -\delta_r \sigma_r\} \text{ and } J_{x,r} = \lfloor np + x\sigma \rfloor, x \in \mathbb{R}, \end{aligned}$$

where $\delta_r \in (0, \frac{1}{2}]$ is as in (4.2). For notational simplicity, we drop the subscript r from the indices $\tilde{x}_{k,r}, K_{0,r}, K_{1,r}, K_{2,r}$ and $J_{x,r}$. Note that by definition $K_1 - 1 < n_r p_r - \sigma_r \leq K_1$, $K_2 - 1 < n_r p_r - \delta_r \sigma_r^2 \leq K_2$, $\tilde{x}_{j,r} \in [-1, 0]$ for all $K_1 \leq j \leq K_0$ and $\tilde{x}_{j,r} \in [-\delta_r \sigma_r, -1)$ for all $K_2 \leq j < K_1$. Hence, for any $x \in [-\delta_r \sigma_r, 0]$,

$$\begin{aligned} |\Delta_r^*(x)| &\leq P(X_r < K_2) + \sum_{j=K_2}^{J_x} \left| P(X_r = j) - \frac{\phi(\tilde{x}_{j,r})}{\sigma_r} \right| + \left| \sum_{j=K_2}^{J_x} \frac{\phi(\tilde{x}_{j,r})}{\sigma_r} - \Phi(x) \right| \\ &= I_{1,r} + I_{2,r}(x) + I_{3,r}(x), \qquad \text{say.} \end{aligned} \tag{3.15}$$

By Chebyshev's inequality, noting that $K_2 - 1 < n_r p_r - \delta_r \sigma_r^2 \leq K_2$, we have

$$\begin{aligned} I_{1,r} &\equiv P(X_r \leq K_2 - 1) \leq P\left(\left| \frac{X_r - n_r p_r}{\sigma_r} \right| \geq \left| \frac{K_2 - n_r p_r - 1}{\sigma_r} \right| \right) \\ &\leq \frac{Var(X_r)}{(K_2 - 1 - n_r p_r)^2} \leq \frac{N_r \sigma_r^2}{N_r - 1} (\delta_r \sigma_r^2)^{-2} \\ &\leq \frac{2}{\delta_r^2 \sigma_r^2}. \end{aligned} \tag{3.16}$$

Next, consider $I_{2,r}(x)$ for $x \in [-\delta_r \sigma_r, -1)$. Note that for $x < -1$, $\frac{J_x - n_r p_r}{\sigma_r} \leq x < -1$. Hence $J_x < K_1$ and $\tilde{x}_{j,r} < -1$ for all $j < J_x$. From Lemma 3.1, writing $R_r^*(j) \equiv R_{n_r}^*(j)$, we get

$$|R_r^*(j)| \leq \frac{1}{6\sigma_r^2(1-\delta_r)} + \left[\frac{|\tilde{x}_{j,r}|^2}{2\sigma_r} + \frac{|\tilde{x}_{j,r}|^2}{\sigma_r^2} \left\{ \frac{1}{4} + \frac{2\delta_r}{(1-\delta_r)^3} \right\} + \frac{|\tilde{x}_{j,r}|^3}{2\sigma_r} A_r \right] \tag{3.17}$$

for all $K_2 \leq j < K_1$, where $A_r = a_{1,r}\left(1 + \frac{4(1+\delta_r)}{(1-\delta_r)^3}\right)$ and $a_{1,r} = \frac{f_r + 4}{4(1-f_r)}$. It is easy to verify that $\delta_r \leq \frac{1}{20}$ and $\delta_r A_r < .59$ for all r satisfying $\delta_r \sigma_r > 1$. Hence

$$\begin{aligned} |R_r^*(j)| &\leq (0.2)\sigma_r^{-2} + \frac{\tilde{x}_{j,r}^2}{2} \left[\frac{1}{\sigma_r} + \frac{2}{\sigma_r^2}(0.3667) + \delta_r A_r \right] \\ &\leq (0.2)\sigma_r^{-2} + \frac{\tilde{x}_{j,r}^2}{2} \left[\min\{0.86, \frac{6}{5\sigma_r} + 0.59\} \right]. \end{aligned} \tag{3.18}$$

Now, from (3.17), for all $K_2 \leq j < K_1$,

$$|R_r^*(j)| \leq (0.2)\sigma_r^{-2} + |\tilde{x}_{j,r}|^3 \left| \left[\frac{1}{2\sigma_r} + \frac{1}{\sigma_r^2}(0.3667) + \frac{3a_{1,r}}{\sigma_r} \right] \leq 4|\tilde{x}_{j,r}|^3 \frac{a_{1,r}}{\sigma_r}. \tag{3.19}$$

Next note that for any $a \in (0, \infty)$, the function $g(y; a) = y^3 \exp(-ay)$, $y \in [0, \infty)$, is increasing on $[0, \sqrt{3/2a}]$, and decreasing on $(\sqrt{3/2a}, \infty)$. Hence, by Lemmas 3.1

and 3.2, (3.18) and (3.19), with $c = .07$, we have

$$\begin{aligned} I_{2,r}(x) &\leq \sum_{j=K_2}^{J_x} \left| \frac{\phi(\tilde{x}_{j,r})}{\sigma_r} \exp(R_r^*(j)) - \frac{\phi(\tilde{x}_{j,r})}{\sigma_r} \right| \\ &\leq \frac{1}{\sigma_r} \sum_{j=K_2}^{J_x} \phi(\tilde{x}_{j,r}) |R_r^*(j)| \exp(|R_r^*(j)|) \\ &\leq \frac{4a_{1,r}}{\sqrt{2\pi}\sigma_r^2} \exp(\sigma_r^{-2}) \sum_{j=K_2}^{J_x} |\tilde{x}_{j,r}|^3 \exp(-c\tilde{x}_{j,r}^2) \leq \frac{C}{\sigma_r}. \end{aligned} \tag{3.20}$$

Also, noting that $|R_r^*(j)| \leq \frac{1}{\sigma_r} + \left[(.43)\tilde{x}_{j,r}^2\right] \wedge \frac{4a_{1,r}}{\sigma_r}$ for all $K_1 \leq j \leq K_0$ and $K_0 - K_1 \leq \sigma_r$, by Lemma 3.1, it follows that

$$\begin{aligned} &\sum_{j=K_1}^{K_0} \left| P(X = j) - \frac{1}{\sigma_r}\phi(\tilde{x}_{j,r}) \right| \leq \sum_{j=K_1}^{K_0} \exp\left(-\frac{\tilde{x}_{j,r}^2}{2}\right) |R_r^*(j)| \frac{\exp(|R_r^*(j)|)}{\sqrt{2\pi}\sigma_r}. \\ &\leq (K_0 - K_1)\exp(\sigma_r^{-1}) \frac{5a_{1,r}}{\sqrt{2\pi}\sigma_r^2} \leq \frac{C}{\sigma_r}. \end{aligned} \tag{3.21}$$

Thus, the bound (3.20) on $I_{2,r}(x)$ holds for all $x \in [-\delta_r\sigma_r, 0]$. Next note that by definition, $\tilde{x}_{J_x,r} \leq x$ and $\tilde{x}_{K_2,r} \leq -\delta_r\sigma_r + \sigma_r^{-1}$. Hence, for $x \in [-\delta_r\sigma_r, 0]$, by Lemma 3.3,

$$\begin{aligned} I_{3,r}(x) &\leq \left| \frac{1}{\sigma_r} \sum_{j=K_2}^{J_x} \phi(\tilde{x}_{j,r}) - \int_{\tilde{x}_{K_2,r}-(2\sigma_r)^{-1}}^{\tilde{x}_{J_x,r}+(2\sigma_r)^{-1}} \phi(y)dy \right| \\ &\qquad + \left| \Phi(x) - \Phi\left(\tilde{x}_{J_x,r} + \frac{1}{(2\sigma_r)}\right) \right| + \Phi\left(\tilde{x}_{K_2,r} - \frac{1}{(2\sigma_r)}\right) \\ &\leq \frac{1}{12\sigma_r^2} \left[\int_{-\infty}^{x+\frac{1}{2\sigma_r}} |\phi''(y)|dy + 5\max\{|\phi''(y)| : -\infty < y < x + \frac{1}{2\sigma_r}\} \right] \\ &\qquad + \Phi\left(x + \frac{1}{2\sigma_r}\right) - \Phi\left(x - \frac{1}{2\sigma_r}\right) + \Phi(-\delta_r\sigma_r + \frac{1}{2\sigma_r}) \\ &\leq \frac{C}{\sigma_r}. \end{aligned} \tag{3.22}$$

Since $\sup_{-\infty \leq x \leq -\delta_r\sigma_r} |\Delta_r^*(x)| \leq P(X_r \leq K_2 - 1) + \Phi(-\delta_r\sigma_r) \leq \frac{C}{\delta_r^2\sigma_r^2}$,

$$\sup_{x \in (-\infty, 0]} |\Delta_r^*(x)| \leq C/\sigma_r \quad \text{for} \quad f_r \leq 1/2. \tag{3.23}$$

To establish the upper bound for $x \geq 0$ and $f_r \leq \frac{1}{2}$, define $V_r = n_r - X_r$, $r \in \mathbb{N}$. Note that V_r has a Hypergeometric distribution with parameters $n_r, N_r - M_r, N_r$. Further, $[X_r - n_r p_r]/\sigma_r = -[V_r - n_r q_r]/\sigma_r$ for all $r \in \mathbb{N}$. Hence, the desired upper bound on the right tails of $[X_r - n_r p_r]/\sigma_r$ can be obtained by repeating the arguments above with X_r replaced by V_r, and p_r replaced by q_r for any r such

that $\delta_r\sigma_r > 1$. This, together with (3.23) proves the upper bound in Theorem 2.2 for all r satisfying $f_r \le \frac{1}{2}$ and $\delta_r f_r > 1$. The proof of the upper bound in (2.9) for '$f_r \in [\frac{1}{2}, 1)$ and $x \in \mathbb{R}$' follows by replacing the above arguments with X_r, f_r replaced by $Y_r, 1 - f_r$, respectively, and using the bound (3.10) and (3.11).

To establish the lower bound in (2.9), write $x_r^* = (\lfloor n_r p_r \rfloor - n_r p_r)/\sigma_r$. Clearly, $\liminf_{r\to\infty} \sigma_r \Delta_r \ge \liminf_{r\to\infty} \sigma_r \left|\Delta_r^*(x_r^*)\right| \equiv C_0$, say. If $C_0 > 0$, then $\Delta_r > \frac{C_0}{2\sigma_r}$ for all but finitely many r's, and the lower bound holds. On the other hand, if $C_0 = 0$, then, using the fact that $\sigma_r^{-1}(X_r - n_r p_r)$ is a lattice random variable with maximal span σ_r^{-1}, we get

$$\begin{aligned} &\liminf_{r\to\infty} \sigma_r \Delta_r \ge \liminf_{r\to\infty} \sigma_r \left|\Delta_r^*(x_r^* + [2\sigma_r]^{-1})\right| \\ = \; &\liminf_{r\to\infty} \sigma_r \left|\Delta_r^*(x_r^*) + \Phi(x_r^*) - \Phi(x_r^* + [2\sigma_r]^{-1})\right| = \phi(0)/2 > 0. \end{aligned}$$

This completes the proof of Theorem 2.2.

References

1. Babu, G.J. and Singh, K. (1985). Edgeworth expansions for sampling without replacement from finite populations. *Journal of Multivariate Analysis* **17** 261-278. MR0813236 (87h:62027)
2. Bloznelis, M. (1999). A Berry-Esseen bound for finite population Student's statistic. *Annals of Probability* **27** 2089-2108. MR1742903 (2000m:62004)
3. Bloznelis, M. and Götze, F. (2000). An Edgeworth expansion for finite-population U-statistics. *Bernoulli* **6** 729-760. MR1777694 (2001k:62010)
4. Blyth, C. R. and Staudte, R. G. (1997). Hypothesis estimates and acceptability profiles for 2×2 contingency tables. *Journal of the American Statistical Association* **92** 694-699. MR1467859
5. Burstein, H. (1975). Finite population correction for binomial confidence limits. *Journal of the American Statistical Association* **70** 67-69.
6. Erdos, P. and Renyi, A. (1959). On the Central Limit Theorem for samples from a finite population. *Magyar Tudoanyos Akademia Budapest Matematikai Kutato Intezet Koezlemenyei*, Trudy Publications **4** 49-57. MR0107294 (21:6019)
7. Feller, W. (1971). *An introduction to probability theory and its applications. Volume* I. Wiley, New York, NY. MR0270403 (42:5292)
8. Hajek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magyar Tudoanyos Akademia Budapest Matematikai Kutato Intezet Koezlemenyei*, Trudy Publications **5** 361-374. MR0125612 (23:A2911)
9. Lahiri, S. N., Chatterjee, A. and Maiti, T. (2004). A Sub-Gaussian Berry-Esseen Theorem for the Hypergeometric probabilities. *Preprint # 2004-21* (posted at *http://seabiscuit.stat.iastate.edu/departmental/preprint/preprint.html#2004*), Iowa State University, IA (also, posted as *math.PR/0602276* at *http://arxiv.org*).
10. Nicholson, W. L. (1956). On the Normal approximation to the Hyprgeometric distribution. *Annals of Mathematical Statistics* **27** 471-483. MR0087246 (19:326c)
11. Madow, W. G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Annals of Mathematical Statistics* **19** 535-545. MR0029136 (10:554a)
12. Patel, J. K. and Samaranayake, V. A. (1991). Prediction intervals for some discrete distributions. *Journal of Quality Technology* **23** 270-278.
13. Seber, G. A. F. (1970). The effects of trap response on tag recapture estimates. *Biometrics* **26** 13-22.
14. Sohn, S. Y. (1997). Accelerated life-tests for intermittent destructive inspection, with logistic failure-distribution. *IEEE Transactions on Reliability* **46** 122-1295.
15. Wendell, J. P. and Schmee, J. (1996). Exact inference for proportions from a stratified finite population. *Journal of the American Statistical Association* **91** 825-830. MR1395749 (97a:62022)
16. Wittes, J. T. (1972). On the bias and estimated variance of Chapman's two-sample capture-recapture population estimate. *Biometrics* **28** 592-597.

Department of Statistics, Iowa State University, Ames, Iowa 50011
Current address: Department of Statistics, Texas A&M University, College Station, Texas 77843
E-mail address: snlahiri@iastate.edu

Department of Statistics, Iowa State University, Ames, Iowa 50011
Current address: Department of Statistics, Texas A&M University, College Station, Texas 77843
E-mail address: cha@iastate.edu