Large sample distribution of the sample total in a generalized rejective sampling scheme

S. Rao Jammalamadaka, G. Michaletzky * and P. Todorovic

Statistics and Applied Probability Program, University of California, Santa Barbara, CA 93106, USA

Received March 1990

Abstract: The weak convergence of a sample sum, in a generalized rejective sampling from a finite population, to a Poisson and Normal distribution is discussed. The generalization consists in assuming that the elements of the population are random variables, rather than fixed values.

1. Introduction

Consider a finite population of \( N \) units, \( \{Y_j\}_j^N \), where the value of the \( j \)th unit, \( Y_j \), is a non-negative integer. From this, a sample of size \( n \) is drawn according to a rejective sampling plan with parameters \( p_1, \ldots, p_N \), with \( \sum_j^n p_j = n \) (see Hajek, 1981, Chapter 7). Here \( p_j \) denotes the probability that the \( j \)th population unit is included in the sample — this event being represented by the indicator variable, \( I_j \). Let \( S_{N_n} \) be the sample sum, then clearly \( \mathbb{P}(S_{N_n}) = \mathbb{P}(\sum_j^n Y_j I_j | \sum_j^n I_j = n) \), where \( \{I_j\}_j^N \) is a sequence of independent Bernoulli r.v.'s with \( \mathbb{P}(I_j) = p_j \).

In a recent paper (Praskova, 1985) the weak convergence of \( S_{N_n} \) to a Poisson r.v. as \( N, n \to \infty \), was discussed in some detail. In this note, we extend and generalize the results of Praskova (1985). First, we show that the Poisson convergence still holds when \( \{Y_j\}_j^N \) are non-negative independent integer valued random variables (r.v.'s), independent of \( \{I_j\}_j^N \), such that \( \mathbb{E}(Y_j) < \infty \), \( k = 1, 2 \). The randomness of \( Y_j \) covers cases such as multistage sampling where \( Y_j \) is the value corresponding to the \( j \)th primary stage unit. Second, under mild regularity assumptions on \( \{Y_j\}_j^N \), we also investigate the weak convergence of the standardized sample sum to a normal distribution.

2. Poisson convergence of the sample sum

Set \( P_{N_n} = \mathbb{P}(\sum_j^n I_j = n) \), \( f_j(t) = \mathbb{E}(e^{itY_j}) \) and

\[
\varphi_{N_n}(t) = \mathbb{E}\left(\exp\left(i \sum_1^N Y_j I_j \right) \mid \sum_1^n I_j = n\right).
\]

From a simple argument (see e.g. Holst, 1979, Theorem 1) we have:

\[
\varphi_{N_n}(t) = (2\pi P_{N_n})^{-1} \int_{-\pi}^{\pi} e^{-2\pi i n} \left[ \prod_{j=1}^{N} \mathbb{E}\left(\exp\{i(tY_j + s) I_j\} \right) \right] ds. \tag{2.1}
\]

* Permanent Address: Department of Mathematics, Eötvös University, Budapest, Hungary.
Let $Z_N$ be a Poisson r.v. with $E\{Z_N\} = \sum_{i=1}^{N} E(Y_i) / p_j$, then its characteristic function $g_N(t)$ is

$$g_N(t) = \exp \left( (e^{it} - 1) \sum_{j=1}^{N} p_j E(Y_j) \right). \quad (2.2)$$

Our aim is to evaluate $|\varphi_{N,j}(t) - g_N(t)|$, which we do in Proposition 2.1. Throughout this paper, we assume that $p_j < 0.5$. Set $q_j = 1 - p_j$, then

$$E\left\{ \exp \left\{ i(tY_j + s) I_j \right\} \right\} = (q_j + p_j e^{i\phi} f_j(t)). \quad (2.3)$$

This and (2.1) yield

$$\varphi_{N,j}(t) = (2\pi P_N)^{-1} \int_{-\pi}^{\pi} e^{-is} \left\{ \prod_{j=1}^{N} \left( q_j + p_j \ e^{i\phi} \right) \right\} \left( q_j(s) + p_j(s) f_j(t) \right) \ ds \quad (2.4)$$

where $q_j(s) = q_j / (q_j + p_j \ e^{i\phi})$ and $p_j(s) = 1 - q_j(s)$. Since

$$1 = (2\pi P_N)^{-1} \int_{-\pi}^{\pi} e^{-is} E\left\{ \exp \left\{ is \sum_{j=1}^{N} I_j \right\} \right\} \ ds$$

it follows, using a well known identity for products, that

$$\varphi_{N,j}(t) - g_N(t) = (2\pi P_N)^{-1} \int_{-\pi}^{\pi} e^{-is} \left\{ \prod_{j=1}^{N} \left( q_j + p_j \ e^{i\phi} \right) \right\} \left( q_j(s) + p_j(s) f_j(t) \right) \ ds$$

$$- (2\pi P_N)^{-1} \int_{-\pi}^{\pi} e^{-is} \left\{ \prod_{j=1}^{N} \left( q_j + p_j \ e^{i\phi} \right) \right\} \left( q_j(s) + p_j(s) f_j(t) - e^{p_j(e^{i\phi} - 1)E(Y_j)} \right) \ ds. \quad (2.5)$$

To evaluate the absolute value of the difference in (2.5), the following lemma is needed.

**Lemma 2.1.**

$$\left| q_j(s) + p_j(s) f_j(t) - e^{p_j(e^{i\phi} - 1)E(Y_j)} \right|$$

$$\leq |e^{it} - 1|^2 \left[ \frac{p_j E(Y_j)}{2(q_j - p_j)} + \frac{2p_j q_j |\sin \frac{s}{2} E(Y_j)}{(q_j - p_j) |e^{it} - 1|} + \frac{1}{2} \left( p_j E(Y_j) \right)^2 \right]. \quad (2.6)$$
Proof. First, write the left hand side of (2.6) as follows:

\[
|p_j(s)[f_j(s) - 1 - E(Y_j)(e^{ist} - 1)] + (p_j(s) - p_j)(e^{ist} - 1)E(Y_j) - (e^{p_j(e^{ist})E(Y_j)} - 1 - p_j(e^{ist} - 1)E(Y_j))|.
\]

(2.7)

Clearly,

\[
|p_j(s)| \leq \frac{p_j}{q_j - p_j}, \quad |p_j(s) - p_j| \leq \frac{2p_jq_j}{q_j - p_j} \sin \frac{s}{2},
\]

\[
|f_j(s) - 1 - E(Y_j)(e^{ist} - 1)| = \left| \sum_{k=0}^{\infty} e^{ikt}P\{Y_j = k\} - 1 - (e^{ist} - 1) \sum_{k=0}^{\infty} kP\{Y_j = k\} \right|
\]

\[
\leq |e^{ist} - 1| \sum_{k=0}^{\infty} \left| \frac{e^{ikt} - 1}{e^{ist} - 1} - k \right| P\{Y_j = k\}
\]

\[
\leq \frac{1}{2} |e^{ist} - 1|^2 E(Y_j)(Y_j - 1).
\]

Finally,

\[
|e^{p_j(e^{ist})E(Y_j)} - 1 - p_j(e^{ist} - 1)E(Y_j)| \leq \frac{1}{2} |e^{ist} - 1|^2 (p_jE(Y_j))^2.
\]

This proves the lemma. 0

Set

\[
A = \sum_{j=1}^{N} \frac{p_jE(Y_j)}{q_j - p_j}, \quad B = \sum_{j=1}^{N} \left[ \frac{p_jE(Y_j)(Y_j - 1)}{q_j - p_j} + (p_jE(Y_j))^2 \right],
\]

\[
C = \sum_{j=1}^{N} \frac{p_jq_j}{q_j - p_j} E\{Y_j\}, \quad d = \sum_{j=1}^{N} p_jq_j.
\]

Then we have:

Proposition 2.1.

\[
|\Phi_N(t) - g_N(t)| \leq \alpha e^{|e^{ist} - 1|^A} \cdot \{\beta |e^{ist} - 1|^B + \gamma |e^{ist} - 1|^D e^{-1/2C}\}
\]

where \(\alpha, \beta\) and \(\gamma\) are positive constants.

Proof. First, we have the following three inequalities:

(i) \(|q_j(s) + p_j(s)f_j(t)| = |1 + (f_j(t) - 1)p_j(s)| ≤ 1 + |e^{ist} - 1| p_j(s)E(Y_j) ≤ \exp\left\{ p_jE(Y_j) |e^{ist} - 1|/(q_j - p_j) \right\} \).

(ii) \(|e^{p_k(e^{ist})E(Y_k)}| ≤ e^{p_k|e^{ist} - 1|E(Y_k)/(q_k - p_k)}\).

(iii) \(|q_j + p_j e^{ist}| ≤ e^{-2p_jq_j \sin^2(s/2)}\).

This and (2.5) yield:

\[
|\Phi_N(t) - g_N(t)|
\]

\[
\leq (2\pi P_N)^{-1} e^{|e^{ist} - 1|^A} \int_{-\pi}^{\pi} e^{-2d \sin^2(s/2)} \left[ \frac{1}{2} B |e^{ist} - 1|^2 + C |e^{ist} - 1| \sin \frac{s}{2} \right] ds.
\]

465
Since
\[ \int_0^\pi e^{-2d \sin^2(s/2)} \, ds \leq \left( \frac{1}{2} \pi \right)^{3/2} d^{-1/2}, \quad \int_0^\pi e^{-2d \sin^2(s/2)} \sin s \, ds \leq \frac{(1 - e^{-2d})/d}{d}, \]
and \( P_{N_d} \geq k \cdot d^{1/2} \), where \( k > 0 \) is a constant, the proposition follows. \( \Box \)

An important consequence of the proposition is the Poisson convergence of the sample sum, stated formally in Corollary 2.1. Set \( \alpha = \alpha(N) \) is assumed to be a bounded sequence
\[ \alpha^{-1} = \min_{1 \leq j \leq N} (q_j - p_j). \]
Then clearly
\[ A \leq \alpha \sum_{1}^{N} p_j E \{ Y_j \}, \]
\[ B \leq \alpha \sum_{1}^{N} p_j E \{ Y_j (Y_j - 1) \} + \max_{1 \leq k \leq N} (p_k E \{ Y_k \}) \sum_{1}^{N} p_j E \{ Y_j \}, \]
\[ C \leq \alpha \sum_{1}^{N} p_j E \{ Y_j \}. \]
In addition, since \( \sum_{1}^{N} p_j = n \) and \( p_j < 0.5 \) it follows that \( d \geq 2n \).

**Corollary 2.1.** If we set \( I_j = I_{N_j}, \ p_j = p_{N_j}, \) and assume that as \( N \to \infty, \ n \to \infty, \)
\[ \max_{1 \leq j \leq N} p_{N_j} E \{ Y_j \} \to 0, \quad \sum_{1}^{N} p_{N_j} E \{ Y_j (Y_j - 1) \} \to 0 \]
and
\[ \sum_{1}^{N} p_{N_j} E \{ Y_j \} \to \lambda, \]
then \( q_{N_j}(t) \to g(t), \) where \( g(t) = e^{(e^t - 1)\lambda}. \) This implies that
\[ \mathcal{L} \left( \sum_{1}^{N} Y_j I_{N_j} \right) \to \mathcal{L}(Z) \]
where \( Z \) is a Poisson r.v. with \( E(Z) = \lambda. \) \( \Box \)

### 3. Convergence to a normal distribution

The result of this section can be formulated as:

**Proposition 3.1.** Let \( \{ Y_j \}_1^\infty \) be independent real-valued r.v.'s such that \( E \{ Y_j \} = 0 \) and \( \text{Var} \{ Y_j \} = 1, \ j = 1, 2, \ldots, \) independent of \( \{ I_j \}_1^\infty. \) Then
\[ S_{N_j} / \sqrt{n} \to Z \]
where $Z \sim N(0,1)$, or equivalently $\Psi_{N}(t) = \phi(t)$, uniformly where

$$
\Psi_{N}(t) = E \left( \exp \left( \frac{it}{\sqrt{n}} \sum_{1}^{N} Y_j I_j \middle| \sum_{1}^{N} I_j = n \right) \right) \quad \text{and} \quad \phi(t) = e^{-t^2/2}.
$$

**Proof.** The method of proof is the one used in the previous proposition. Write

$$
\Psi_{N}(t) = (2\pi P_{N})^{-1} \int_{-\pi}^{\pi} e^{-isn} \prod_{1}^{N} \left( q_j + p_j e^{is} \right) ds.
$$

Then, we have

$$
\Psi_{N}(t) - \phi(t) = (2\pi P_{N})^{-1} \int_{-\pi}^{\pi} e^{-isn} \left[ \prod_{1}^{N} \left( q_j + p_j e^{is} \right) \right]
$$

$$
\times \left[ \prod_{1}^{N} \left( q_j(s) + p_j(s) f_j \left( \frac{t}{\sqrt{n}} \right) \right) - \prod_{1}^{N} e^{-\left( t^2/2n \right) p_j} \right] ds
$$

$$
= (2\pi P_{N,n})^{-1} \int_{-\pi}^{\pi} e^{-isn} \left[ \prod_{1}^{N} \left( q_j + p_j e^{is} \right) \right]
$$

$$
\times \left( \sum_{j=1}^{N} \left[ \prod_{k=1}^{j-1} \left( q_k(s) + p_k(s) f_k \left( \frac{t}{\sqrt{n}} \right) \right) \right] \prod_{k=j+1}^{N} e^{-\left( t^2/2n \right) p_k} \right)
$$

$$
\times \left( q_j(s) + p_j(s) f_j \left( \frac{t}{\sqrt{n}} \right) - e^{-\left( t^2/2n \right) p_j} \right) ds.
$$

But

$$
\left| q_j(s) + p_j(s) f_j \left( \frac{t}{\sqrt{n}} \right) - e^{-\left( t^2/2n \right) p_j} \right|
$$

$$
= \left| p_j(s) \left( f_j \left( \frac{t}{\sqrt{n}} \right) - 1 \right) + (1 - e^{-\left( t^2/2n \right) p_j}) \right|
$$

$$
= \left| p_j(s) \left( f_j \left( \frac{t}{\sqrt{n}} \right) \right) + (p_j - p_j(s)) \frac{t^2}{2n} + 1 - \frac{t^2}{2n} p_j - e^{-\left( t^2/2n \right) p_j} \right|
$$

$$
\leq \frac{p_j}{q_j - p_j} o \left( \frac{1}{n} \right) + \frac{2p_j q_j \sin \frac{3}{2} s}{q_j - p_j} \frac{t^2}{2n} + o \left( \frac{1}{n} \right) \frac{t^2}{2} p_j,
$$

$$
\left| q_k(s) + p_k(s) f_k \left( \frac{t}{\sqrt{n}} \right) \right| = \left| 1 + p_k(s) \left( f_k \left( \frac{t}{\sqrt{n}} \right) - 1 \right) \right|
$$

$$
\leq \exp \left( \left| p_k(s) \right| \left| f_k \left( \frac{t}{\sqrt{n}} \right) - 1 \right| \right) \leq \exp \left( \frac{p_k}{q_k - p_k} o \left( \frac{t^2}{2n} \right) \right),
$$

and

$$
\left| q_j + p_j e^{is} \right| \leq \exp \left( 2p_j q_j \sin \frac{3}{2} s \right).
$$
From this, we obtain that

\[ \left| \Psi_{N,n}(t) - \phi(t) \right| \leq \left( 2 \pi P_{N,n} \right)^{-1} \int_{-\pi}^{\pi} e^{-\left( \sum_{j=1}^{N} q_j p_j \right) \sin^2(s/2) - \left( t^2/2n \right) \sum_{j=1}^{N} \left( p_{i,j} - p_{i,k} \right)} + o(1) \]

\[ \times \left[ o(1) + \frac{t^2}{2n} \left( \sum_{j=1}^{N} q_j p_j \right) \sin \frac{1}{2}s + o(1) \right] ds \]

\[ \leq d^{-1/2} \left[ \frac{o(1)}{\sqrt{d}} + \frac{t^2}{2n} \right] = o(1) + \frac{c}{\sqrt{d}} \to 0 \]

which proves the assertion. \( \Box \)

References

