Commentationes Mathematicae Universitatis Carolinae

Zuzana Prášková

A note on Sampford-Durbin sampling

Commentationes Mathematicae Universitatis Carolinae, Vol. 31 (1990), No. 2, 367--372

Persistent URL: http://dml.cz/dmlcz/106866

Comment.Math.Univ.Carolinae 31,2(1990)367–372

A note on Sampford–Durbin sampling

ZUZANA PRÁŠKOVÁ

Abstract. In the present paper, a convergence of the sample sum to Poisson distribution is studied in the case that Sampford–Durbin sampling from a finite population is used. The results are close to those obtained by the author for the rejective sampling.

Keywords: Sampford–Durbin sampling, rejective sampling, sample sum, Poisson distribution

Classification: 60F05,62D05

1. Introduction. It is known that Sampford–Durbin sampling from a finite population is a modification of the rejective sampling which yields exact values of including probabilities. Víšek [5] developed the asymptotic normality of Horvitz–Thompson estimator of the population total and Prášková [3] obtained the rate of convergence. In this paper we shall study a convergence of the sample sum to the Poisson distribution.

Let us recall some definitions. Consider a population U of N units which can be identified with the numbers $1, \dots, N$, and a sample $s \subset U$.

Rejective sampling of size n with parameters $p_1, \dots, p_N$ is defined by probabilities

$$\begin{aligned} R(s) &= C \prod_{j \in s} p_j \prod_{j \notin s} (1 - p_j) \quad \text{if } K(s) = n \\ &= 0 \qquad \text{otherwise,} \end{aligned}$$

where $K(s)$ denotes the size of the sample s and C is a positive constant such that $\sum_{s \subset U} R(s) = 1$. We suppose that $0 < p_j < 1$ for all $j = 1, \dots, N$, and $\sum_{j=1}^{N} p_j = n$. The parameters $p_1, \dots, p_N$ can be controlled.

The probabilities of inclusion $\pi_j(R) = \sum_{s \ni j} R(s)$ satisfy the asymptotic relation

$$\pi_j(R) = \kappa_j = p_j \left[1 - \frac{(\bar{\bar{p}} - p_j)(1 - p_j)}{d(p)} + o\left(d^{-1}(p)\right) \right], \tag{1}$$

where

$$d(p) = \sum_{j=1}^{N} p_j (1 - p_j), \tag{2}$$

$$\bar{\bar{p}} = \sum_{j=1}^{N} p_j^2(1-p_j)\,(d(p))^{-1} \tag{3}$$

and $d(p)o\left(d^{-1}(p)\right) \longrightarrow 0$ if $d(p) \longrightarrow \infty$ (see Hájek [2], Theorem 7.3).

Sampford–Durbin sampling of size n with parameters $p_1,\dots,p_N$ is defined by probabilities

$$\begin{aligned} Q(s) &= C^* \prod_{i\in s} p_i(1-p_i)^{-1} \sum_{j\in s}(1-p_j) \quad \text{if } K(s)=n \\ &= 0 \qquad \text{otherwise,} \end{aligned}$$

where C^* is a constant such that $\sum_{s\subset U} Q(s) = 1$ and the parameters p_j, $1\le j\le N$, satisfy the conditions formulated above for the rejective sampling. It can be shown that in this case the including probabilities are

$$\pi_j(Q) = \sum_{s\ni j} Q(s) = p_j$$

(see Hájek [2], Chapt. 8, for details).

In our next considerations R and Q will denote probability measures generated by the rejective and Sampford–Durbin sampling, respectively. Further, c will stand for a positive constant the value of which can change in different formulas, or even in different places of the same formula. Let y_j denote the value of a characteristic y on the unit j, $1\le j\le N$, and S be the sample sum, $S=\sum_{j\in s} y_j$. Denote by f, h the characteristic function of S with respect to Q, R, respectively, i.e.

$$f(t) = E_Q e^{itS}, \qquad h(t) = E_R e^{itS}.$$

Lemma. *Let us consider the rejective sampling of size n with parameters $p_1,\dots,p_N$ and its Sampford–Durbin modification with the same parameters. Suppose that $\max_{1\le j\le N} p_j < \frac{1}{2}$. Then there exists a constant c such that for n sufficiently large and for all t*

$$|f(t)-h(t)| < cn^{-\frac{1}{2}}. \tag{4}$$

Proof : When using the same arguments as in the proof of Lemma 3 in [3], we obtain that the inequality

$$|f(t)-h(t)| \le (\operatorname{var}_R V)^{\frac{1}{2}}$$

holds for all t, the random variable V being defined by (3.1) in the quoted paper. Further, from (3.1), (3.3) and (3.4) in the same paper we get

$$|f(t)-h(t)| \le$$

$$\leq \left[\frac{1}{2} \sum_i \sum_{j \neq i} (p_i - p_j)^2 (\kappa_i \kappa_j - \kappa_{ij}) \left(\sum_{j=1}^{N} p_j (1 - \kappa_j) \right)^{-2} \right]^{\frac{1}{2}}, \tag{5}$$

where κ_i, κ_{ij} are including probabilities of the unit i, respectively of the units i, j in the rejective sampling.
Now, it can be easily checked that $d(p) \longrightarrow \infty$ if $n \longrightarrow \infty$ because $\max p_j < \frac{1}{2}$ and $\sum p_j = n$. Consequently, (5.10) in Hájek [1] implies that $d(\kappa) = \sum_{j=1}^{N} \kappa_j (1 - \kappa_j) \longrightarrow \infty$ and therefore

$$\kappa_i \kappa_j - \kappa_{ij} = \kappa_i \kappa_j (1 - \kappa_i)(1 - \kappa_j) d^{-1}(\kappa) \left[1 + o(1)\right],$$

where $o(1) \longrightarrow 0$ if $n \longrightarrow \infty$ (see Hájek [1], Theorem 5.2).
Thus we have

$$\sum_i \sum_{j \neq i} (p_i - p_j)^2 (\kappa_i \kappa_j - \kappa_{ij}) \leq (1 + o(1)) \sum_{j=1}^{N} p_j^2 \kappa_j (1 - \kappa_j) \tag{6}$$

$$\leq c \sum_{j=1}^{N} p_j^2 \leq c \sum_{j=1}^{N} p_j = cn.$$

It remains to estimate $\sum_{j=1}^{N} p_j (1 - \kappa_j)$. From (1) we get

$$1 - \kappa_j = (1 - p_j) \left[1 + \frac{(\bar{\bar{p}} - p_j) p_j}{d(p)} + o\left(d^{-1}(p)\right) \right].$$

Obviously, $0 < \bar{\bar{p}} < 1$. As $\max p_j < \frac{1}{2}$, it can be easily shown that there exists a constant $c > 0$ such that for n sufficiently large

$$1 - \kappa_j \geq c(1 - p_j) \qquad \text{for all } j = 1, \ldots, N.$$

Thus

$$\sum_{j=1}^{N} p_j (1 - \kappa_j) \geq c \sum_{j=1}^{N} p_j (1 - p_j) \geq \frac{c}{2} \sum_{j=1}^{N} p_j = cn$$

which together with (5) and (6) completes the proof. ■

2. Convergence to the Poisson distribution.

Now we shall suppose that $y_1, \ldots, y_N$ are nonnegative integers. Denote by T the random variable having Poisson distribution with the parameter $a = \sum_{j=1}^{N} y_j p_j$. In

fact, a can depend on n and N. It can be easily shown that $a = E_Q S$.
Our aim is to estimate the distances

$$\rho_1(S,T) = \sup_x |F(x) - G(x)|,$$

$$\rho_2(S,T) = \sup_k |Q(S = k) - P(T = k)|,$$

where $F(x) = Q(S \le x)$ is the distribution function of S in the Sampford–Durbin sampling and G is the distribution function of the random variable T. Let g stand for the characteristic function of T.
Denote $q_j = 1 - p_j$ and put

$$\begin{aligned}
A &= \sum_{j=1}^{N} y_j p_j (q_j - p_j)^{-1}, \\
B_1 &= \sum_{j=1}^{N} y_j p_j q_j (q_j - p_j)^{-2}, \\
B_2 &= \sum_{j=1}^{N} \left[y_j(y_j - 1)p_j + (y_j p_j)^2 \right] (q_j - p_j)^{-2}, \\
d &= d(p).
\end{aligned}$$

Theorem 1. *Suppose that* $\max p_j < \frac{1}{2}$. *Then there exists a constant* c *such that for* n *sufficiently large*

$$\rho_1 \le c \left[n^{-\frac{1}{2}} \log n + n^{-\frac{1}{2}} a + e^{2(A+a)} (B_1 d^{-\frac{1}{2}} a^{-\frac{1}{2}} + B_2 a^{-1} \right], \tag{7}$$

$$\rho_2 \le c \left[n^{-\frac{1}{2}} + e^{2(A+a)} (B_1 d^{-\frac{1}{2}} a^{-1} + B_2 a^{-1}) \right]. \tag{8}$$

PROOF : First we estimate ρ_2. According to Lemma 1.1 in [4] we get

$$\rho_2 \le \frac{1}{2\pi} \int_{-\pi}^{\pi} |f(t) - g(t)|\, dt \le \frac{1}{2\pi} \int_{-\pi}^{\pi} |f(t) - h(t)|\, dt +$$

$$+ \frac{1}{2\pi} \int_{-\pi}^{\pi} |h(t) - g(t)|\, dt.$$

Now, utilizing (4) to estimate the first integral and applying the proof of Theorem 2.1 in [4] to the second integral, we easily obtain (8).
To obtain (7) we make use of Lemma 1.1 in [4] again. We have

$$\rho_1 \le \frac{1}{4\pi} \int_{-\pi}^{\pi} |f(t) - g(t)| \, |\sin \frac{t}{2}|^{-1}\, dt$$

$$\leq \frac{1}{4\pi}\int_{-\pi}^{\pi}|f(t)-h(t)|\,|\sin\frac{t}{2}|^{-1}\,dt+\frac{1}{4\pi}\int_{-\pi}^{\pi}|h(t)-g(t)|\,|\sin\frac{t}{2}|^{-1}\,dt.$$

The second integral can be estimated similarly as in the proof of Theorem 2.1 in [4]. Thus, it remains to estimate the integral

$$J=\frac{1}{4\pi}\int_{-\pi}^{\pi}|f(t)-h(t)|\,|\sin\frac{t}{2}|^{-1}\,dt=\frac{1}{4\pi}\left(\int_{-\pi}^{-n^{-\frac{1}{2}}}+\int_{-n^{-\frac{1}{2}}}^{n^{-\frac{1}{2}}}+\int_{n^{-\frac{1}{2}}}^{\pi}\right)$$

$$=J_1+J_2+J_3.$$

When utilizing (4) and the inequality $|\sin\frac{t}{2}|\geq\frac{2}{\pi}|t|$ valid for $0\leq|t|\leq\frac{\pi}{2}$, we find that J_1 and J_3 are bounded by $cn^{-\frac{1}{2}}\log n$. Since $f(0)=h(0)=1$, we have

$$J_2\leq\frac{1}{8}\left[\int_{|t|\leq n^{-\frac{1}{2}}}|f(t)-f(0)|\,|t|^{-1}\,dt+\int_{|t|\leq n^{-\frac{1}{2}}}|h(t)-h(0)|\,|t|^{-1}\,dt\right].$$

Further, we have for all t

$$\left|\frac{d}{dt}f(t)\right|\leq E_Q|S|=E_QS=\sum_{j=1}^{N}y_jp_j$$

and similarly

$$\left|\frac{d}{dt}h(t)\right|\leq E_RS=\sum_{j=1}^{N}y_j\kappa_j.$$

Thus, making use of the mean value theorem, we get

$$J_2\leq cn^{-\frac{1}{2}}\left(\sum_{j=1}^{N}y_jp_j+\sum_{j=1}^{N}y_j\kappa_j\right).$$

From (1) it follows that for n sufficiently large $0<\kappa_j<2p_j$ for all j, and thus we can conclude that

$$J\leq c\left(n^{-\frac{1}{2}}\sum_{j=1}^{N}y_jp_j+n^{-\frac{1}{2}}\log n\right).$$

∎

Let us consider a sequence of populations U_ν consisting of N_ν units with characteristics $y_{\nu_1},\ldots,y_{\nu_N}$ and a sequence of Sampford–Durbin samplings of sizes n_ν with parameters $p_{\nu_1},\ldots,p_{\nu N_\nu}$. Suppose that $n_\nu\longrightarrow\infty$, $N_\nu\longrightarrow\infty$ as $\nu\longrightarrow\infty$ and $\rho_{1\nu},\rho_{2\nu}$ refer to the ν-th experiment.

Theorem 2. *Suppose that*

$$\max_{1\le j\le N_\nu} p_{\nu_j} \le \frac{1}{4},$$

$$\lim_{\nu\longrightarrow\infty} \sum_{j=1}^{N_\nu} p_{\nu_j} y_{\nu_j} = \lambda > 0,$$

$$\lim_{\nu\longrightarrow\infty} \max_{1\le j\le N_\nu} p_{\nu_j} y_{\nu_j} = 0,$$

$$\lim_{\nu\longrightarrow\infty} \sum_{j=1}^{N_\nu} p_{\nu_j} y_{\nu_j} (y_{\nu_j} - 1) = 0.$$

Then

$$\lim_{\nu\longrightarrow\infty} \rho_{1\nu} = 0,$$

$$\lim_{\nu\longrightarrow\infty} \rho_{2\nu} = 0.$$

Proof follows immediately from (7), (8) and the proof of Theorem 2.2 in [4].

References

[1] Hájek J., *Asymptotic theory of rejective sampling with varying probabilities from a finite population*, Ann. Math. Statist. **35** (1964), 1491–1523.

[2] Hájek J., *Sampling from a Finite Population*, (Edited by V. Dupač), Dekker, New York (1981).

[3] Prášková Z., *On the rate of convergence in Sampford–Durbin sampling from a finite population*, Statistics & Decision **2** (1984), 339–350.

[4] Prášková Z., *On the convergence to the Poisson distribution in rejective sampling from a finite population*, In: Proc. of the 5th Pannonian Symp. on Math. Stat., W. Grossmann et al (eds), Akadémiai Kiadó, Budapest (1987).

[5] Víšek, J.A., *Asymptotic distribution of simple estimate for rejective, Sampford and successive sampling*, In: Contribution to Statistics, J. Jurečková (ed), Prague, Academia (1979).

Charles University, Department of Probability and Statistics, Sokolovská 83, 186 00 Prague 8, Czechoslovakia

(*Received* July 17, 1989)