Electronic Journal of Statistics
Vol. 6 (2012) 1967–1983
ISSN: 1935-7524
DOI: 10.1214/12-EJS736

Approximation of rejective sampling inclusion probabilities and application to high order correlations

Helène Boistard

Université de Toulouse
Toulouse, France
e-mail: helene@boistard.fr

Hendrik P. Lopuhaä

DIAM
Delft University of Technology
Delft, The Netherlands
e-mail: h.p.lopuhaa@tudelft.nl

and

Anne Ruiz-Gazen

Toulouse School of Economics
France
e-mail: anne.ruiz-gazen@tse-fr.eu

Abstract: This paper is devoted to rejective sampling. We provide an expansion of joint inclusion probabilities of any order in terms of the inclusion probabilities of order one, extending previous results by Hájek (1964) and Hájek (1981) and making the remainder term more precise. Following Hájek (1981), the proof is based on Edgeworth expansions. The main result is applied to derive bounds on higher order correlations, which are needed for the consistency and asymptotic normality of several complex estimators.

AMS 2000 subject classifications: Primary 62D05; secondary 60E10.
Keywords and phrases: Rejective Sampling, Poisson sampling, Edgeworth expansions, maximal entropy, Hermite polynomials.

Received July 2012.

Contents

1. Introduction

In a finite population of size N, sampling without replacement with unequal inclusion probabilities and fixed sample size is not straightforward, but there exist several sampling designs that satisfy these properties (see Brewer and Hanif (1983) for a review). Rejective sampling, which is also called maximum entropy sampling or conditional Poisson sampling, is one possibility, introduced by Hájek (1964). If n denotes the fixed sample size, the n units are drawn independently with probabilities that may vary from unit to unit and the samples in which all units are not distinct are rejected. In the particular case of equal drawing probabilities, rejective sampling coincides with simple random sampling without replacement. Rejective sampling with size n can also be regarded as Poisson sampling conditionally on the sample size being equal to n. The unconditional Poisson design can be easily implemented by drawing N independently distributed Bernoulli random variables with different probabilities of success, but it has the disadvantage of working with a random sample size. The conditional Poisson design can also be interpreted as a maximum entropy sampling design for a fixed sample size and a given set of first order inclusion probabilities.

Rejective sampling has been extensively studied in the literature. Hájek (1964, 1981) derives an approximation of the joint inclusion probabilities in terms of first order inclusion probabilities. By showing that the maximum entropy design belongs to a parametric exponential family, Chen, Dempster and Liu (1994) give a recursive expression of the joint inclusion probabilities and propose a new algorithm. This algorithm has been improved by Deville (2000), who gives another expression for the joint inclusion probabilities. Using the results in Chen, Dempster and Liu (1994), Qualité (2008) proves that the variance of the well-known unbiased Horvitz-Thompson estimator for rejective sampling is smaller than the variance of the Hansen-Hurvitz estimator for multinomial sampling. Several estimators of the variance of the Horvitz-Thompson estimator have also been proposed; see Matei and Tillé (2005) for a comparison by means of a large simulation study. The conditional Poisson sampling scheme is not only of interest in the survey sampling field, but also in the context of case-control studies or survival analysis, see Chen (2000).

The purpose of the present article is to generalize the result given in Hájek (1964) and Hájek (1981), obtained for the first and second order inclusion probabilities of rejective sampling, to inclusion probabilities of any order and also to provide a more precise remainder term. The proof of our result is along the lines of the proof by Hájek (1981) using Edgeworth expansions and leads to approximations that are valid when N, n and $N - n$ are large enough. One interesting application of our result is that it enables us to show that rejective

sampling satisfies the assumptions needed for the consistency and the asymptotic normality of some complex estimators, such as the ones defined in Breidt and Opsomer (2000), Breidt et al. (2007), Cardot et al. (2010) or Wang (2009). Such assumptions involve conditions on correlations up to order four, which are difficult to check for complex sampling designs that go beyond simple random sampling without replacement or Poisson sampling. Our result implies that the rejective sampling design also satisfies these conditions.

In the case-control context, Arratia, Goldstein and Langholz (2005) consider rejective sampling and also give approximations of higher order correlations. Their approach and the assumptions they need to derive their results are different from the ones we consider in the present paper. Instead of using Edgeworth expansions, they consider an expansion that involves the characteristic function. Their results are obtained using a condition, which is sufficient, but not necessary to derive our expansion. In view of this we provide an example of a rejective sampling design that does not satisfy the condition in Arratia, Goldstein and Langholz (2005), but does satisfy our weaker assumption. Moreover, Arratia *et al.* do not give an explicit approximation formula for higher order inclusion probabilities in rejective sampling, whereas we do provide such an approximation, which may be of interest in itself.

The paper is organized as follows: in Section 2 we introduce notations and state our main result which is Theorem 1. In Section 3, we apply this result and illustrate that rejective sampling satisfies conditions on higher order correlations imposed in the recent literature to derive several asymptotic results. Detailed proofs are provided in Section 4.

2. Notations and main result

In this paper, we use the first description of rejective sampling by Hájek (1981), namely as Poisson sampling conditionally on the sample size being equal to n. Let us denote $\mathcal{U}$ as the population of size N. Let $0 \leq p_1, p_2, \ldots, p_N \leq 1$ be a sequence of real numbers such that $p_1 + p_2 + \cdots + p_N = n$. The Poisson sampling design with parameters $p_1, p_2, \ldots, p_N$ is such that for any sample s, the probability of s is

$$P(s) = \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i).$$

The corresponding rejective sampling design is such that the probability of a sample s is

$$P_{RS}(s) = \begin{cases} c \displaystyle\prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i) & \text{if size } s = n, \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

where c is a constant such that $\sum_s P_{RS}(s) = 1$. We refer the reader to Hájek (1981) for more details.

The inclusion probabilities of order k under this sampling scheme are denoted as

$$\pi_{i_1, i_2, \ldots, i_k} = P_{RS}(i_1 \in s, i_2 \in s, \ldots, i_k \in s)$$

for any $\{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, N\}$. Our purpose is to obtain an expansion of inclusion probabilities of any order. Theorem 7.4 in Hájek (1981), see also Theorem 5.2 in Hájek (1964), provides such an expansion for inclusion probabilities of order two, i.e.,

$$\pi_{ij} = \pi_i \pi_j \left[1 - d^{-1}(1-\pi_i)(1-\pi_j) + o(d^{-1})\right], \quad \text{as } d \to \infty, \tag{2.2}$$

uniformly in i, j such that $1 \leq i \neq j \leq N$, where

$$d = \sum_{i=1}^{N} p_i(1-p_i). \tag{2.3}$$

We will obtain an extension of (2.2) and prove that a similar expansion holds for inclusion probabilities of higher order.

Our approach is along the lines of the method used in Hájek (1981). Consider Poisson sampling with parameters $p_1, p_2, \ldots, p_N$ and denote as P the corresponding probability measure on the set of samples under this sampling scheme. For $i = 1, 2, \ldots, N$, we denote as I_i the indicator of inclusion of unit i, that is

$$I_i = \mathbf{1}(i \in s) = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{otherwise.} \end{cases}$$

For every $i = 1, 2, \ldots, N$, the indicator I_i is a Bernoulli random variable with parameter p_i. Define

$$K = \text{size } s = I_1 + I_2 + \cdots + I_N. \tag{2.4}$$

Note that the expectation and the variance of K satisfy $\mathbb{E}_P(K) = n$ and $\mathbb{V}_P(K) = d$. By Bayes formula and by independence of the I_i's under Poisson sampling, the inclusion probability $\pi_{i_1, i_2, \ldots, i_k}$ can be written as

$$\begin{aligned} &\pi_{i_1, i_2, \ldots, i_k} \\ &= P\left(I_{i_1} = I_{i_2} = \cdots = I_{i_k} = 1 | K = n\right) \\ &= P(I_{i_1} = I_{i_2} = \cdots = I_{i_k} = 1) \frac{P(K = n | I_{i_1} = I_{i_2} = \cdots = I_{i_k} = 1)}{P(K = n)} \\ &= p_{i_1} p_{i_2} \cdots p_{i_k} \frac{P(K = n | I_{i_1} = I_{i_2} = \cdots = I_{i_k} = 1)}{P(K = n)}. \end{aligned} \tag{2.5}$$

The next step is to use Edgeworth expansions for the probabilities of K. This leads to the next lemma.

Lemma 1. *Consider Poisson sampling with parameters $p_1, p_2, \ldots, p_N$, such that $p_1 + p_2 + \cdots + p_N = n$ with corresponding probability measure P on the set of samples. Let d and K be defined in (2.3) and (2.4), respectively. Then, for all $A_k = \{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, N\}$, $k \geq 1$, it holds that if $d \to \infty$, then*

$$\begin{aligned} P(K = n) &= (2\pi d)^{-1/2} \left\{1 + c_1 d^{-1} + O\left(d^{-2}\right)\right\}, \\ P(K = n | I_{i_1} = \cdots = I_{i_k} = 1) &= (2\pi d)^{-1/2} \left\{1 + c_2 d^{-1} + O\left(d^{-2}\right)\right\}, \end{aligned}$$

where

$$\begin{aligned} c_1 &= \frac{1}{8}\left(1 - 6\overline{\overline{p(1-p)}}\right) - \frac{5}{24}\left(1 - 2\overline{\overline{p}}\right)^2, \\ c_2 &= \frac{1}{2}\left(B_2 - (B_1 - k)^2\right) - \frac{1}{2}(B_1 - k)\left(1 - 2\overline{\overline{p}}\right) + c_1, \end{aligned} \tag{2.6}$$

with

$$\begin{aligned} \overline{\overline{p}} &= d^{-1}\sum_{i=1}^{N} p_i^2(1-p_i), & B_1 &= \sum_{j\in A_k} p_j, \\ \overline{\overline{p(1-p)}} &= d^{-1}\sum_{i=1}^{N} p_i^2(1-p_i)^2, & B_2 &= \sum_{j\in A_k} p_j(1-p_j). \end{aligned} \tag{2.7}$$

The proof of the lemma is provided in Section 4. Let us now formulate our main result.

Theorem 1. *For $k \geq 1$, let $A_k = \{i_1, i_2, \ldots, i_k\} \subset \{1, \ldots, N\}$. Under rejective sampling (2.1), the following approximations hold as $d \to \infty$, where d is defined by (2.3).*

(i) For all $k \geq 2$,

$$\pi_{i_1,i_2,\ldots,i_k} = \pi_{i_1}\pi_{i_2}\cdots\pi_{i_k} \times \times \left(1 - d^{-1}\sum_{i,j\in A_k : i<j} (1-p_i)(1-p_j) + O(d^{-2})\right), \tag{2.8}$$

where $O(d^{-2})$ holds uniformly in $i_1, i_2, \ldots, i_k$.

(ii) For all $k \geq 2$,

$$\pi_{i_1,i_2,\ldots,i_k} = \pi_{i_1}\pi_{i_2}\cdots\pi_{i_k} \times \times \left(1 - d^{-1}\sum_{i,j\in A_k : i<j} (1-\pi_i)(1-\pi_j) + O(d^{-2})\right), \tag{2.9}$$

where $O(d^{-2})$ holds uniformly in $i_1, i_2, \ldots, i_k$.

Proof. From Lemma 1, we find

$$\frac{P(K=n \mid I_{i_1} = \cdots = I_{i_k} = 1)}{P(K=n)} = \frac{1 + c_2 d^{-1} + O(d^{-2})}{1 + c_1 d^{-1} + O(d^{-2})} = 1 + (c_2 - c_1)d^{-1} + O(d^{-2}).$$

Together with (2.5) it follows that for all $k \geq 1$,

$$\begin{aligned}
&\pi_{i_1,i_2,\ldots,i_k} \\
&= p_{i_1} p_{i_2} \cdots p_{i_k} \left\{1 + (c_2 - c_1) d^{-1} + O(d^{-2})\right\} \\
&= p_{i_1} p_{i_2} \cdots p_{i_k} \left\{1 + \frac{1}{2d} \sum_{j \in A_k} p_j (1 - p_j) - \frac{1}{2d} \left(\sum_{j \in A_k} p_j - k \right)^2 \right. \\
&\qquad \left. - \frac{1 - 2\overline{\overline{p}}}{2d} \left(\sum_{j \in A_k} p_j - k \right) + O(d^{-2}) \right\}.
\end{aligned} \tag{2.10}$$

Applying (2.10) to the case $k = 1$, yields that the first order inclusion probabilities satisfy

$$p_i = \pi_i \left(1 - d^{-1}(p_i - \overline{\overline{p}})(1 - p_i) + O(d^{-2})\right), \tag{2.11}$$

and as a consequence,

$$p_{i_1} p_{i_2} \cdots p_{i_k} = \pi_{i_1} \pi_{i_2} \cdots \pi_{i_k} \left\{ 1 - d^{-1} \sum_{j \in A_k} (p_j - \overline{\overline{p}})(1 - p_j) + O(d^{-2}) \right\}.$$

Combining this with (2.10) yields

$$\pi_{i_1,i_2,\ldots,i_k} = \pi_{i_1} \pi_{i_2} \cdots \pi_{i_k} \left\{1 + a d^{-1} + O(d^{-2})\right\}$$

where the contribution to terms of order d^{-1} is

$$\begin{aligned}
a &= \frac{1}{2} \sum_{j \in A_k} p_j (1 - p_j) - \frac{1}{2} \left(\sum_{j \in A_k} p_j - k \right)^2 - \frac{1 - 2\overline{\overline{p}}}{2} \left(\sum_{j \in A_k} p_j - k \right) \\
&\quad - \sum_{j \in A_k} (p_j - \overline{\overline{p}})(1 - p_j) \\
&= -\frac{1}{2} \sum_{j \in A_k} p_j (1 - p_j) - \frac{1}{2} \left(\sum_{j \in A_k} (1 - p_j) \right)^2 + \frac{1}{2} \left(\sum_{j \in A_k} (1 - p_j) \right) \\
&= \frac{1}{2} \sum_{j \in A_k} (1 - p_j)^2 - \frac{1}{2} \left(\sum_{j \in A_k} (1 - p_j) \right)^2 = - \sum_{i,j \in A_k : i < j} (1 - p_i)(1 - p_j).
\end{aligned}$$

This proves part (i). Part (ii) is deduced immediately from (i) and (2.11). □

3. Application: Bounds on higher order correlations under rejective sampling

Conditions on the order of higher order correlations, as $N \to \infty$, appear at several places in the literature, see e.g., Breidt and Opsomer (2000), Breidt et al.

(2007), Cardot et al. (2010) or Wang (2009), among others. Such conditions are used when studying asymptotic properties in survey sampling for general sampling designs, but they are difficult to check for more complex sampling designs, that go beyond simple random sampling without replacement. An attempt to provide simpler conditions for rejective sampling can be found in Arratia, Goldstein and Langholz (2005). They formulate some sort of asymptotic stability condition on inclusion frequencies that ensure bounds on general higher order correlations. The purpose of the present section is to explain how Theorem 1 can be used to establish several bounds on higher order correlations for the rejective sampling design. The bounds in Arratia, Goldstein and Langholz (2005) match with the ones that we find for correlations up to order four, which suffices for the conditions imposed in Breidt and Opsomer (2000); Breidt et al. (2007); Cardot et al. (2010); Wang (2009). However, in order to derive these bounds, we only need the simple requirement that

$$\limsup_{N\to\infty} \frac{N}{d} < \infty, \tag{3.1}$$

where d is defined in (2.3). Moreover, one can show that (3.1) is weaker than the asymptotic stability condition in Arratia, Goldstein and Langholz (2005) as detailed in Section 4.2.

Before we start a discussion on the assumptions on higher order correlations that appear for example in Breidt and Opsomer (2000); Breidt et al. (2007); Cardot et al. (2010); Wang (2009), first note that (3.1) necessarily yields that $d \to \infty$, which means that Theorem 1 holds. Moreover, condition (3.1) has a number of additional consequences, such as $n \geq d \to \infty$, $N - n \geq d \to \infty$, and

$$\limsup_{N\to\infty} \frac{N}{n} \leq \limsup_{N\to\infty} \frac{N}{d} < \infty. \tag{3.2}$$

A typical example of a condition on higher order correlations, is

$$\limsup_{N\to\infty} \; n \max_{(i,j)\in D_{2,N}} |\mathbb{E}_P(I_i - \pi_i)(I_j - \pi_j)| < \infty, \tag{3.3}$$

where for every integer $t \geq 1$:

$$D_{t,N} = \big\{(i_1, i_2, \ldots, i_t) : i_1, i_2, \ldots, i_t \text{ are all different and each } i_j \in \{1, 2, \ldots, N\}\big\}. \tag{3.4}$$

Condition (3.3) is one of the assumptions in Breidt and Opsomer (2000) among others. Since $\mathbb{E}_P(I_i - \pi_i)(I_j - \pi_j) = \pi_{ij} - \pi_i\pi_j$, condition (3.3) immediately follows from Theorem 1 and (3.2).

Interestingly, the simple representation of the second order correlations as a difference of second order inclusion probabilities and the product of single order inclusion probabilities can be generalized for correlations of higher order as detailed in the following lemma.

Lemma 2. *For any $k \geq 2$, let $A_k = \{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, N\}$. Then*

$$\mathbb{E}\left[\prod_{j=1}^{k}(I_{i_j} - \pi_{i_j})\right] = \sum_{m=2}^{k}(-1)^{k-m} \sum_{(i_1,\ldots,i_m)\in D_{m,k}} (\pi_{i_1,\ldots,i_m} - \pi_{i_1}\cdots\pi_{i_m})\,\pi_{i_{m+1}}\cdots\pi_{i_k}, \tag{3.5}$$

where $D_{m,k}$ is the set of distinct m-tuples in A_k and $\{i_{m+1}, \ldots, i_k\} = A_k \setminus \{i_1, \ldots, i_m\}$.

From this lemma, we can prove the following proposition that provides an expansion of higher order correlations for rejective sampling.

Proposition 1. *Consider a rejective sampling design. Then, for any $k \geq 3$ and any positive integers n_j, $j = 1, 2, \ldots, k$,*

$$\mathbb{E}\left[\prod_{j=1}^{k}(I_{i_j} - \pi_{i_j})^{n_j}\right] = O(d^{-2}) \tag{3.6}$$

as $d \to \infty$, where d is defined by (2.3).

The proofs of Lemma 2 and Proposition 1 are provided in Section 4.3.

Proposition 1 together with condition (3.2) imply that the following conditions that appear for example in Breidt and Opsomer (2000) are satisfied:

$$\begin{aligned} &\limsup_{N\to\infty} \frac{N^4}{n^2} \max_{(i,j,k,l)\in D_{4,N}} |\mathbb{E}(I_i - \pi_i)(I_j - \pi_j)(I_k - \pi_k)(I_l - \pi_l)| < \infty \\ &\limsup_{N\to\infty} \frac{N^3}{n^2} \max_{(i,j,k)\in D_{3,N}} \left|\mathbb{E}(I_i - \pi_i)^2(I_j - \pi_j)(I_k - \pi_k)\right| < \infty. \end{aligned} \tag{3.7}$$

Other conditions on higher order correlations, such as

$$\lim_{N\to\infty} \max_{(i,j,k,l)\in D_{4,N}} |\mathbb{E}(I_i I_j - \pi_{ij})(I_k I_l - \pi_{kl})| = 0, \tag{3.8}$$

that appears in Breidt and Opsomer (2000), can be treated in the same manner.

The conditions in Breidt et al. (2007) and Cardot et al. (2010) on higher order correlations are equivalent to the preceding ones. A stronger condition appears in Wang (2009): in assumption (A6) therein, the third condition is as follows:

$$\limsup_{N\to\infty} n^2 \max_{(i,j,k)\in D_{3,N}} \left|\mathbb{E}(I_i - \pi_i)^2(I_j - \pi_j)(I_k - \pi_k)\right| < \infty. \tag{3.9}$$

This is an easy consequence of Proposition 1 and of (3.1) which implies that $n^2 = O(d^2)$ as $N \to \infty$.

4. Proofs

4.1. Proof of Lemma 1

For the proof of Lemma 1, we use Edgeworth expansions for probabilities of sums of independent random variables, as given in Theorem 6.2 in Hájek (1981). Suppose $K = I_1 + I_2 \cdots + I_N$ is a sum of independent Bernoulli random variables with parameters $p_1, p_2, \ldots, p_N$, and let $d = \mathbb{V}(K)$. Then, for $0 \leq l \leq N$ and $m \geq 1$,

$$|P(K = l) - f_m(x)| = o(d^{-(m+1)/2}) \tag{4.1}$$

where $f_m(x)$ is the Edgeworth expansion of $P(K = l)$ up to order m, given by

$$f_m(x) = d^{-1/2}\phi(x)\left(1 + \sum_{j=1}^{m} P_j(x)\right), \quad \text{with } x = \frac{l - \mathbb{E}(K)}{d^{1/2}}, \tag{4.2}$$

where ϕ denotes the standard normal density and each P_j is a linear combination of (probabilistic) Hermite polynomials involving the cumulants of K. Recall that the Hermite polynomials are defined by

$$H_k(x) = (-1)^k \mathrm{e}^{x^2/2} \frac{\mathrm{d}^k}{\mathrm{d}x^k}\left[\mathrm{e}^{-x^2/2}\right] \tag{4.3}$$

for $k = 0, 1, 2, \ldots$ and that the cumulants of a random variable X are defined as the coefficients in the expansion of the logarithm of the moment-generating function, i.e., if

$$g(t) = \log \mathbb{E}(\mathrm{e}^{tX}) = \sum_{m=1}^{\infty} \kappa_m \frac{t^m}{m!},$$

the m-th cumulant is $\kappa_m = g^{(m)}(0)$.

In the following lemma, we provide a suitable expression for the polynomials P_j in (4.2).

Lemma 3. *The polynomials P_j in (4.2) can be expressed as:*

$$P_j(x) = d^{-j/2} \sum_{\{k_m\}} H_{j+2r}(x) \prod_{m=1}^{j} \frac{1}{k_m!} \frac{1}{((m+2)!)^{k_m}} \left(\frac{\kappa_{m+2}}{d}\right)^{k_m}, \tag{4.4}$$

where the sum is taken over all sets $\{k_m\}$ consisting of all non-negative integer solutions of

$$k_1 + 2k_2 + \cdots + jk_j = j, \tag{4.5}$$

and r is defined by $k_1 + k_2 + \cdots + k_j = r$, and where κ_m is the m-th cumulant of K and H_{j+2r} is the Hermite polynomial of degree $j + 2r$ as given in (4.3).

Proof. The proof relies on the Edgeworth expansion of $P(K=l)$, e.g., see (43) in Blinnikov and Moessner (1998),

$$P\left(K=l\right)=d^{-1/2}\phi(x)\left\{1+\sum_{j=1}^{\infty}d^{j/2}\sum_{\{k_m\}}H_{j+2r}(x)\frac{1}{k_m!}\prod_{m=1}^{j}\left(\frac{S_{m+2}}{(m+2)!}\right)^{k_m}\right\},$$

where $x=(l-\mathbb{E}_P(K))d^{-1/2}$ and $S_m=\kappa_m/d^{m-1}$. This means that

$$P_j(x)=d^{j/2}\sum_{\{k_m\}}H_{j+2r}(x)\prod_{m=1}^{j}\frac{1}{k_m!}\left(\frac{S_{m+2}}{(m+2)!}\right)^{k_m}.$$

Note that

$$\prod_{m=1}^{j}S_{m+2}^{k_m}=\prod_{m=1}^{j}\left(\frac{\kappa_{m+2}}{d^{m+1}}\right)^{k_m}=\prod_{m=1}^{j}\left(\frac{\kappa_{m+2}}{d}\right)^{k_m}\prod_{m=1}^{j}d^{-mk_m}=d^{-j}\prod_{m=1}^{j}\left(\frac{\kappa_{m+2}}{d}\right)^{k_m},$$

according to (4.5). This yields (4.4). □

The next lemma shows that the cumulants of the sum of independent Bernoulli variables are of the same order as the variance.

Lemma 4. *Let $K=I_1+I_2+\cdots+I_N$ be a sum of independent Bernoulli random variables with parameters $p_1,p_2,\ldots,p_N$. Let $d=\mathbb{V}(K)=\sum_{i=1}^{N}p_i(1-p_i)$. Then, for any positive integer m, we have $\kappa_m=O(d)$, as $d\to\infty$, uniformly in $p_1,p_2,\ldots,p_N$.*

Proof. The definition of cumulants implies that the m-th cumulant κ_m of the sum of independent Bernoulli random variables is equal to the sum of the m-th cumulants e_m of the individual Bernoulli variables. Moreover, we have the following recurrence relation between the cumulants of a single Bernoulli variable with parameter p:

$$e_{m+1}=p(1-p)\frac{\mathrm{d}}{\mathrm{d}p}e_m, \tag{4.6}$$

see for instance, example (c) in Section 4 in Khatri (1959). It is straightforward to see that $\kappa_1=p_1+p_2+\cdots+p_N$ and $\kappa_2=\sum_{i=1}^{N}p_i(1-p_i)=d$. It can be proved by induction, using (4.6), that $e_m=p(1-p)R_m(p)$, where R_m is a polynomial with degree less than or equal to $m-1$ and with coefficients depending only on m. Thus, $\kappa_m=dQ_m(p)$, where $Q_m(p)$ is of the form

$$Q_m(p)=\frac{\sum_{i=1}^{N}p_i(1-p_i)R_m(p_i)}{\sum_{i=1}^{N}p_i(1-p_i)}$$

and is bounded uniformly in $p_1,p_2\ldots,p_N$. This proves the lemma. □

Proof of Lemma 1. We use (4.1) with $m=4$. Because $\mathbb{E}_P(K)=n$, formula (4.2) is used with $x=0$. In order to determine the expressions of $P_j(0)$, for $j=$

$1, 2, 3, 4$, we use Lemma 3. It follows from (4.3) that the Hermite polynomials satisfy the following recurrence relationship

$$H_{k+1}(x) = -\mathrm{e}^{x^2/2}\frac{\mathrm{d}}{\mathrm{d}x}\left[H_k(x)\mathrm{e}^{-x^2/2}\right]. \tag{4.7}$$

By induction it follows from (4.7) that for any integer $j = 0, 1, \ldots$, the Hermite polynomials H_{2j} and H_{2j+1} are of the form

$$H_{2j}(x) = a_{0j} + a_{1j}x^2 + \cdots + a_{jj}x^{2j},$$
$$H_{2j+1}(x) = b_{1j}x + b_{2j}x^3 + \cdots + b_{jj}x^{2j+1}.$$

It follows that $H_{2j+1}(0) = 0$, for any integer j. Combining this with Lemmas 3 and 4, we can see that $P_{2j+1}(0) = 0$ and $P_{2j}(0) = O(d^{-j})$ for any integer j. Thus, $P_1(0) = P_3(0) = 0$ and $P_4(0) = O(d^{-2})$. Moreover,

$$P_2(0) = \frac{H_6(0)}{2!(3!)^2}\frac{\kappa_3^2}{d^3} + \frac{H_4(0)}{4!}\frac{\kappa_4}{d^2} = -\frac{15}{72}\frac{\kappa_3^2}{d^3} + \frac{3}{24}\frac{\kappa_4}{d^2}.$$

Finally, from (4.6) one can easily deduce that $\kappa_3 = d(1 - 2\overline{\overline{p}})$ and $\kappa_4 = d(1 - 6\overline{\overline{p(1-p)}})$. We then obtain:

$$\begin{aligned} P(K = n) &= d^{-1/2}\phi(0)\left\{1 + \sum_{j=1}^{4} P_j(0) + O\left(d^{-2}\right)\right\} \\ &= (2\pi d)^{-1/2}\left\{1 - \frac{5}{24}\left(1 - 2\overline{\overline{p}}\right)^2 d^{-1} + \frac{1}{8}\left(1 - 6\overline{\overline{p(1-p)}}\right)d^{-1} + O(d^{-2})\right\} \\ &= (2\pi d)^{-1/2}\left\{1 + c_1 d^{-1} + O\left(d^{-2}\right)\right\}. \end{aligned}$$

For the expansion of $P(K = n | I_{i_1} = \cdots = I_{i_k} = 1)$, let E_k denote the event $\{I_j = 1, \text{ for all } j \in A_k\}$ and define the random variable $\widetilde{K} = K \mid E_k$. Note that it can be written as the sum of independent Bernoulli's,

$$\widetilde{K} = \sum_{j \notin A_k} I_j + \sum_{j \in A_k} I_j^*$$

where $I_j^* = 1$. Thus, we can write an Edgeworth expansion for $\widetilde{K}$ as stated in (4.1). Since

$$\begin{aligned} \mathbb{E}(\widetilde{K}) &= \sum_{j \notin A_k} p_j + k = n + k - \sum_{j \in A_k} p_j = n + k - B_1, \\ \mathbb{V}(\widetilde{K}) &= \sum_{j \notin A_k} p_j(1 - p_j) = d - \sum_{j \in A_k} p_j(1 - p_j) = d - B_2, \end{aligned} \tag{4.8}$$

with $\tilde{d} = d - B_2$, the expansion is as follows:

$$P(\widetilde{K} = n) = \widetilde{d}^{-1/2}\phi(\widetilde{x})\left\{1 + \sum_{j=1}^{4} P_j^*(\widetilde{x})\right\} + o(\widetilde{d}^{-5/2}), \qquad \text{with } \widetilde{x} = \frac{n - \mathbb{E}(\widetilde{K})}{\widetilde{d}^{1/2}},$$

where the P_j^*'s are the polynomials given in (4.4) corresponding to $\widetilde{K}$.

Let us first compute an expansion for $\widetilde{d}^{-1/2}\phi(\widetilde{x})$. We start with the expansion of $\widetilde{d}^{-1/2}$:

$$\widetilde{d}^{-1/2} = (d - B_2)^{-1/2} = d^{-1/2}\left\{1 + \frac{1}{2}B_2 d^{-1} + O(d^{-2})\right\}. \tag{4.9}$$

Next, remark that

$$\widetilde{x} = (d - B_2)^{-1/2}(B_1 - k) = d^{-1/2}(B_1 - k)\left\{1 + \frac{1}{2}B_2 d^{-1} + O(d^{-2})\right\}, \tag{4.10}$$

so that

$$\phi(\widetilde{x}) = (2\pi)^{-1/2}\left\{1 - \frac{1}{2}\widetilde{x}^2 + O(\widetilde{x}^4)\right\} = (2\pi)^{-1/2}\left\{1 - \frac{1}{2}(B_1 - k)^2 d^{-1} + O(d^{-2})\right\}.$$

Together with (4.9), this gives

$$\widetilde{d}^{-1/2}\phi(\widetilde{x}) = (2\pi d)^{-1/2}\left\{1 + a_1 d^{-1} + O(d^{-2})\right\}, \tag{4.11}$$

where $a_1 = (B_2 - (B_1 - k)^2)/2$. Finally, we compute $P_j^*(\widetilde{x})$, for $j = 1, 2, 3, 4$. First, let us compute the third and fourth cumulants of $\widetilde{K}$. We find

$$\kappa_3^* = \kappa_3 - \sum_{A_k} p_j(1 - p_j)(1 - 2p_j) = \kappa_3 - B_3,$$
$$\kappa_4^* = \kappa_4 - \sum_{A_k} p_j(1 - p_j)(1 - 6p_j + 6p_j^2) = \kappa_4 - B_4,$$

for constants B_3 and B_4. Thus, by Lemmas 3 and 4 with (4.9) and (4.10),

$$\begin{aligned} P_1^*(\widetilde{x}) &= \frac{H_3(\widetilde{x})}{6}\frac{\kappa_3^*}{\widetilde{d}^{3/2}} = -\frac{1}{2}\frac{\kappa_3^*}{\widetilde{d}^{3/2}}\left(\widetilde{x} + O(\widetilde{x}^3)\right) \\ &= -\frac{1}{2}(B_1 - k)\left(1 - 2\overline{\overline{p}}\right)d^{-1}\left(1 + O(d^{-1})\right), \end{aligned}$$

and likewise

$$\begin{aligned} P_2^*(\widetilde{x}) &= \frac{H_6(\widetilde{x})}{72}\frac{(\kappa_3^*)^2}{\widetilde{d}^3} + \frac{H_4(\widetilde{x})}{24}\frac{\kappa_4^*}{\widetilde{d}^2} = \left(-\frac{5}{24}\frac{(\kappa_3^*)^2}{\widetilde{d}^3} + \frac{1}{8}\frac{\kappa_4^*}{\widetilde{d}^2}\right)\left(1 + O\left(\widetilde{x}^2\right)\right) \\ &= \left\{-\frac{5}{24}\left(1 - 2\overline{\overline{p}}\right)^2 + \frac{1}{8}\left(1 - 6\overline{\overline{p(1-p)}}\right)\right\}d^{-1}\left(1 + O(d^{-1})\right). \end{aligned}$$

Moreover, similarly to Lemma 4, one has $\kappa_m^* = O(d)$, for any positive integer m. Hence, for any integer j, $P_{2j}^*(\widetilde{x}) = O(d^{-j})$ and $P_{2j+1}^*(\widetilde{x}) = O(d^{-(j+1)})$, so that $P_3^*(\widetilde{x}) = O(d^{-2})$ and $P_4^*(\widetilde{x}) = O(d^{-2})$. It follows that

$$1 + \sum_{j=1}^{4} P_j^*(\widetilde{x}) = 1 + c_1^* d^{-1} + O(d^{-2}), \tag{4.12}$$

where

$$c_1^* = -\frac{1}{2}(B_1 - k)\left(1 - 2\overline{\overline{p}}\right) - \frac{5}{24}\left(1 - 2\overline{\overline{p}}\right)^2 + \frac{1}{8}\left(1 - 6\overline{\overline{p(1-p)}}\right)$$
$$= -\frac{1}{2}(B_1 - k)\left(1 - 2\overline{\overline{p}}\right) + c_1.$$

Combining (4.11) and (4.12) proves the lemma. □

4.2. Comparison with assumptions in Arratia et al.

In Arratia, Goldstein and Langholz (2005), the following condition is used for rejective sampling. For all $\delta \in (0,1)$, there exists $\epsilon \in (0,1)$, such that

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left\{\frac{\epsilon}{1+\epsilon} < p_i < \frac{1}{1+\epsilon}\right\} \geq 1 - \delta. \tag{4.13}$$

This condition implies our condition (3.1), because

$$d = \sum_{i=1}^{N} p_i(1-p_i) \geq N(1-\delta)\frac{\epsilon}{1+\epsilon}\left(1 - \frac{1}{1+\epsilon}\right) \geq N\lambda > 0,$$

where $\lambda = (1-\delta)(\epsilon/(1+\epsilon))^2 \in (0,1)$.

However, our condition is weaker, in the sense that we can construct an example which satisfies (3.1), but not (4.13). To this end, suppose that $n/N \to \gamma \in (0,1)$. Take $\delta \in (0,1)$, such that $0 < \gamma < 1 - \delta < 1$. Furthermore, choose $\alpha \in (0,1)$, such that $0 < \gamma < \alpha < 1 - \delta < 1$, and let $k = \alpha N$. Then define

$$p_1 = \cdots = p_k = \frac{\gamma}{\alpha} \quad \text{and} \quad p_{k+1} = \cdots = p_N = \delta_n = \frac{n/N - \gamma}{1-\alpha} \to 0.$$

First note that this choice is possible in rejective sampling, since

$$p_1 + \cdots + p_N = k \times \frac{\gamma}{\alpha} + (N-k) \times \delta_n = N\gamma + N(1-\alpha)\frac{n/N - \gamma}{1-\alpha} = n.$$

With these probabilities, condition (4.13) is not satisfied for any $\epsilon \in (0,1)$, because for N sufficiently large $p_{k+1} = \cdots = p_N < \epsilon/(1+\epsilon)$, so that

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left\{\frac{\epsilon}{1+\epsilon} < p_i < \frac{1}{1+\epsilon}\right\} \leq \frac{k}{N} = \alpha < 1 - \delta,$$

whereas condition (3.1) is fulfilled, as

$$\frac{d}{N} = \frac{n}{N} - \frac{1}{N}\sum_{i=1}^{N} p_i^2 = \frac{n}{N} - \frac{k}{N}\left(\frac{\gamma}{\alpha}\right)^2 - \frac{N-k}{N}\delta_n^2 = \frac{n}{N} - \frac{\gamma^2}{\alpha} - (1-\alpha)\delta_n^2 \to \gamma - \frac{\gamma^2}{\alpha} \geq \lambda$$

where $\lambda = (\gamma - \gamma^2/\alpha)/2 \in (0,1)$.

4.3. Proofs of Lemma 2 and Proposition 1

Proof of Lemma 2. We decompose the product in the following way:

$$\begin{aligned}
&\mathbb{E}\left[\prod_{j=1}^{k}(I_{i_j}-\pi_{i_j})\right]\\
&=\pi_{i_1}\pi_{i_2}\dots\pi_{i_k}(-1)^k+\mathbb{E}\left[\sum_{m=1}^{k}\sum_{D_{m,k}} I_{i_{j_1}}I_{i_{j_2}}\dots I_{i_{j_m}}\pi_{i_{j_{m+1}}}\dots\pi_{i_{j_k}}(-1)^{k-m}\right]\\
&=\pi_{i_1}\pi_{i_2}\dots\pi_{i_k}(-1)^k+\sum_{m=1}^{k}\sum_{D_{m,k}}\pi_{i_{j_1}i_{j_2}\dots i_{j_m}}\pi_{i_{j_{m+1}}}\dots\pi_{i_{j_k}}(-1)^{k-m}\\
&=\sum_{m=1}^{k}\sum_{D_{m,k}}\left(\pi_{i_{j_1}i_{j_2}\dots i_{j_m}}-\pi_{i_{j_1}}\pi_{i_{j_2}}\dots\pi_{i_{j_m}}\right)\pi_{i_{j_{m+1}}}\dots\pi_{i_{j_k}}(-1)^{k-m}\\
&\quad+\pi_{i_1}\pi_{i_2}\dots\pi_{i_k}(-1)^k+\sum_{m=1}^{k}\sum_{D_{m,k}}\pi_{i_{j_1}}\pi_{i_{j_2}}\dots\pi_{i_{j_m}}\pi_{i_{j_{m+1}}}\dots\pi_{i_{j_k}}(-1)^{k-m}.
\end{aligned}$$

The last two terms on the right hand side are equal to

$$\begin{aligned}
&\pi_{i_1}\pi_{i_2}\dots\pi_{i_k}(-1)^k+\sum_{m=1}^{k}\sum_{D_{m,k}}\pi_{i_1}\pi_{i_2}\dots\pi_{i_k}(-1)^{k-m}\\
&=\pi_{i_1}\pi_{i_2}\dots\pi_{i_k}(-1)^k+\pi_{i_1}\pi_{i_2}\dots\pi_{i_k}\sum_{m=1}^{k}\binom{k}{m}(-1)^{k-m}\\
&=\pi_{i_1}\pi_{i_2}\dots\pi_{i_k}\sum_{m=0}^{k}\binom{k}{m}(-1)^{k-m}=\pi_{i_1}\pi_{i_2}\dots\pi_{i_k}(1-1)^k=0.
\end{aligned}$$

□

Proof of Proposition 1. The proof is by induction on the powers n_j. We first prove that

$$\mathbb{E}\left[\prod_{j=1}^{k}(I_{i_j}-\pi_{i_j})\right]=O(d^{-2}), \tag{4.14}$$

for any $A_k=\{i_1,i_2,\dots,i_k\}\subset\{1,2,\dots,N\}$, with $3\le k\le N$ and then add an extra power one by one. From Lemma 2, we have that

$$\begin{aligned}
&\mathbb{E}\left[\prod_{j=1}^{k}(I_{i_j}-\pi_{i_j})\right]\\
&=\sum_{m=2}^{k}(-1)^{k-m}\sum_{(i_1,\dots,i_m)\in D_{m,k}}\left(\pi_{i_1,\dots,i_m}-\pi_{i_1}\cdots\pi_{i_m}\right)\pi_{i_{m+1}}\cdots\pi_{i_k},
\end{aligned}$$

where $\{i_{m+1},\dots,i_k\} = A_k \setminus \{i_1,\dots,i_m\}$. From Theorem 1, we have that

$$\pi_{i_1,\dots,i_m} - \pi_{i_1}\cdots\pi_{i_m} = -\pi_{i_1}\cdots\pi_{i_m} d^{-1} \sum_{i<j} (1-\pi_i)(1-\pi_j) + O(d^{-2}),$$

where the sum runs over all $i<j$, such that $i,j \in \{i_1,\dots,i_m\}$. This means that

$$\begin{aligned}&\mathbb{E}\left[\prod_{j=1}^{k}(I_{i_j}-\pi_{i_j})\right]\\ &= -d^{-1}\pi_{i_1}\cdots\pi_{i_k}\sum_{m=2}^{k}(-1)^{k-m}\sum_{(i_1,\dots,i_m)\in D_{m,k}}\sum_{i<j}(1-\pi_i)(1-\pi_j)+O(d^{-2}).\end{aligned}$$

For $2 \le m \le k$ fixed, consider the summation

$$\sum_{(i_1,\dots,i_m)\in D_{m,k}}\sum_{i<j}(1-\pi_i)(1-\pi_j).$$

The first summation is over all possible $(i_1,\dots,i_m) \in D_{m,k}$, which are all possible combinations of m different indices from $A_k = \{i_1,i_2,\dots,i_k\}$. From each such combination $i_1,\dots,i_m$, the second summation picks two different indices $i<j$ from the set $\{i_1,\dots,i_m\}$. This means that any combination of $(1-\pi_i)(1-\pi_j)$, with $\{i,j\} \subset A_k$ is possible. In fact, each such combination will appear several times and we only have to count how many times. Well, for a fixed combination (i,j), from the k possibilities A_k, we need to pick i and j, and for the $m-2$ remaining choices there are $k-2$ possibilities left. We conclude that each term $(1-\pi_i)(1-\pi_j)$, with $\{i,j\} \subset A_k$, appears $\binom{k-2}{m-2}$ times. Moreover, this holds for any $m = 2,3,\dots,k$. This means that

$$\begin{aligned}&\sum_{m=2}^{k}(-1)^{k-m}\sum_{(i_1,\dots,i_m)\in D_{m,k}}\sum_{i<j}(1-\pi_i)(1-\pi_j)\\ &= \sum_{\{i,j\}\subset A_k}(1-\pi_i)(1-\pi_j)\sum_{m=2}^{k}(-1)^{k-m}\binom{k-2}{m-2},\end{aligned}$$

where

$$\sum_{m=2}^{k}(-1)^{k-m}\binom{k-2}{m-2} = \sum_{n=0}^{k-2}(-1)^{k-2-n}\binom{k-2}{n} = (1-1)^{k-2} = 0.$$

We conclude that the coefficient of d^{-1} is zero, which proves (4.14).

Next, suppose that the expectation is of order $O(d^{-2})$ for all powers $1 \le m_j \le n_j$, and consider

$$\mathbb{E}\left[(I_{i_1}-\pi_{i_1})^{n_1+1}(I_{i_2}-\pi_{i_2})^{n_2}\cdots(I_{i_k}-\pi_{i_k})^{n_k}\right].$$

This can be written as

$$\begin{aligned}&\mathbb{E}\left[I_{i_1}(I_{i_1}-\pi_{i_1})^{n_1}(I_{i_2}-\pi_{i_2})^{n_2}\cdots(I_{i_k}-\pi_{i_k})^{n_k}\right]\\&\quad-\pi_{i_1}\mathbb{E}\left[(I_{i_1}-\pi_{i_1})^{n_1}(I_{i_2}-\pi_{i_2})^{n_2}\cdots(I_{i_k}-\pi_{i_k})^{n_k}\right]\\&=\mathbb{E}\left[I_{i_1}(I_{i_1}-\pi_{i_1})^{n_1}(I_{i_2}-\pi_{i_2})^{n_2}\cdots(I_{i_k}-\pi_{i_k})^{n_k}\right]+O(d^{-2})\end{aligned}$$

according to the induction hypothesis. Next, write

$$\begin{aligned}I_{i_1}(I_{i_1}-\pi_{i_1})^{n_1}&=(1-\pi_{i_1})I_{i_1}(I_{i_1}-\pi_{i_1})^{n_1-1}\\&=(1-\pi_{i_1})(I_{i_1}-\pi_{i_1})^{n_1}+(1-\pi_{i_1})\pi_{i_1}(I_{i_1}-\pi_{i_1})^{n_1-1}.\end{aligned}$$

When we insert this, we find

$$\begin{aligned}&\mathbb{E}\left[(I_{i_1}-\pi_{i_1})^{n_1+1}(I_{i_2}-\pi_{i_2})^{n_2}\cdots(I_{i_k}-\pi_{i_k})^{n_k}\right]\\&=(1-\pi_{i_1})\mathbb{E}\left[(I_{i_1}-\pi_{i_1})^{n_1}(I_{i_2}-\pi_{i_2})^{n_2}\cdots(I_{i_k}-\pi_{i_k})^{n_k}\right]\\&\quad+(1-\pi_{i_1})\pi_{i_1}\mathbb{E}\left[(I_{i_1}-\pi_{i_1})^{n_1-1}(I_{i_2}-\pi_{i_2})^{n_2}\cdots(I_{i_k}-\pi_{i_k})^{n_k}\right]+O(d^{-2})\\&=O(d^{-2})\end{aligned}$$

by applying the induction hypothesis.

□

Acknowledgements

The authors want to thank Guillaume Chauvet for helpful discussions.

References

Arratia, R., Goldstein, L. and Langholz, B. (2005). Local central limit theorems, the high-order correlations of rejective sampling and logistic likelihood asymptotics. *Ann. Statist.* **33** 871–914. MR2163162 (2006j:62094)

Blinnikov, S. and Moessner, R. (1998). Expansions for nearly Gaussian distributions. *Astron. Astrophys. Suppl. Ser.* **130** 193-205.

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regresssion estimators in survey sampling. *Ann. Statist.* **28** 1026–1053. MR1810918 (2001m:62012)

Breidt, F. J., Opsomer, J. D., Johnson, A. A. and Ranalli, M. G. (2007). Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology* **33** 35.

Brewer, K. R. W. and Hanif, M. (1983). *Sampling with unequal probabilities. Lecture Notes in Statistics* **15**. Springer-Verlag, New York. MR681289 (84i:62010)

Cardot, H., Chaouch, M., Goga, C. and Labruère, C. (2010). Properties of design-based functional principal components analysis. *J. Statist. Plann. Inference* **140** 75–91. MR2568123 (2010j:62166)

Chen, S. X. (2000). General properties and estimation of conditional Bernoulli models. *J. Multivariate Anal.* **74** 69–87. MR1790614 (2001i:62064)

Chen, X.-H., Dempster, A. P. and Liu, J. S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika* **81** 457–469. MR1311090 (96c:62022)

Deville, J.-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu Technical Report No. France, CREST-ENSAI.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Statist.* **35** 1491–1523. MR0178555 (31 ##2812)

Hájek, J. (1981). *Sampling from a finite population. Statistics: Textbooks and Monographs* **37**. Marcel Dekker Inc., New York. Edited by Václav Dupač, With a foreword by P. K. Sen. MR627744 (83d:62019)

Khatri, C. G. (1959). On certain properties of power-series distributions. *Biometrika* **46** 486–490. MR0109381 (22 ##267)

Matei, A. and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *J. Official Statistics* **21** 543–570.

Qualité, L. (2008). A comparison of conditional Poisson sampling versus unequal probability sampling with replacement. *J. Statist. Plann. Inference* **138** 1428–1432. MR2388021 (2009e:62044)

Wang, L. (2009). Single-index model-assisted estimation in survey sampling. *J. Nonparametr. Stat.* **21** 487–504. MR2571724 (2011c:62022)