

# On the number of unobserved and observed categories when sampling from a multivariate hypergeometric population

Sungsu Kim<sup>a</sup> and Chong Jin Park<sup>b</sup>

<sup>a</sup>*University of Louisiana Lafayette*

<sup>b</sup>*San Diego State University*

**Abstract.** Consider taking a random sample of size  $n$  from a finite population that consists of  $N$  categories with  $M_i$  copies in the  $i$ th category for  $i = 1, \dots, N$ . Each observed unit in a sample is presumed to have a probability  $1 - p$  ( $0 < p < 1$ ) of getting lost from the sample. Let  $S$  denote the number of categories not observed in the sample and  $S_j$  denote the number of categories where  $j$  samples are observed for  $j = 1, \dots, n$ . In this paper, the probability distribution and factorial moments of  $S$  and  $S_j$  are studied. A matrix inversion algorithm is used in order to facilitate numerical computations in obtaining the probabilities and factorial moments. A couple of examples of the problem considered in this paper may include a filing or storage process, where objects are randomly assigned to files or storage bins, and from time to time, objects may be missing or have disappeared, species as categories in a capture-recapture problem, or DNA sequence study.

## 1 Introduction

Consider taking a random sample of size  $n$  from a finite population with  $N$  categories with  $M_i$  copies in the  $i$ th category for  $i = 1, \dots, N$ . Let  $S$  denote the number of unobserved categories in the sample and  $S_j$  denote the number of categories where  $j$  samples are observed for  $j = 1, \dots, n$ . Each observed sample is presumed to have a probability  $1 - p$ ,  $0 < p < 1$ , of being lost. In this paper, the probability distribution and factorial moments of  $S$  and  $S_j$  are studied and we use a matrix inversion method to facilitate numerical computations using the relationship between probabilities and factorial moments. Several examples of the problem are as follows: a filing or storage process, where objects are randomly assigned to files or storage bins, and from time to time, objects may be missing or have disappeared; species as categories in a capture-recapture problem; a DNA sequence study, where categories are created by combinations of A, C, G and T in a sequence. For example, when taking a sample of sequence with size 10, there are  $4^{10}$  categories.

The random variable  $S$  explored in this article originates in the context of the classical occupancy problem. In Feller (1957), the probability that exactly  $S$  categories are remain empty is given in the context of a multinomial sampling, and

---

*Key words and phrases.* Factorial moments, matrix inversion method, multivariate hypergeometric distribution, multinomial distribution, occupancy problem, Sterling's number of the second kind.

Received February 2015; accepted November 2016.

the example of finding  $S$  days of the year which are not birthdays in a village of 1900 people is mentioned as an application. Samuel-Cahn (1974) discussed the number of observed categories  $N - S$ , and derived the mean and variance of  $N - S$ . Korwar (1988) considers a general framework applied to various classes of multivariate distributions. The probability distribution of the number of observed classes sampling from a multivariate hypergeometric population was derived in Walton (1986). However, the method used in this paper is different from his work, and our method is more general in that we discuss the case when the proportion of number of copies in  $i$ th category tends to be  $p_i$  as  $M_i$  and  $N$  tend to be infinity. To our knowledge, a study about the probability distribution of  $S_j$  has not appeared in literature. Various asymptotic results were produced in Feller (1957), Harris et al. (1987), Park (1972, 1981) and Samuel-Cahn (1974). For instance, Feller (1957) and Samuel-Cahn (1974) proved that  $S$  is approximately a Poisson random variable for large  $n$  and  $N$ , and Park (1972, 1981) and Harris et al. (1987) proved that  $S$  is asymptotically a normal random variable.

Our paper is composed of the following sections. Section 1 is an **Introduction**. In Section 2 and 3, we present the results for  $S$  in cases of having equal number of copies among  $N$  categories and of unequal number of copies among  $N$  categories, respectively. In Section 4, we present the results for  $S_j$ , followed by numerical examples and concluding remarks in Section 5 and 6, respectively.

## 2 Results for $S$ with equal number of copies in each categories

### 2.1 Case when $M$ is a finite number

In this section, we begin with the following lemma.

**Lemma 1.** *Assume that the number of copies in each category is equal to  $M$ . Let  $H_{n,M}(k)$  denote the number of ways exactly  $k$  out of  $N$  categories are occupied. Then,*

$$H_{n,M}(k) = \sum_{j=0}^{k-1} (-1)^j \binom{k}{j} \binom{M(k-j)}{n} I_{[M(k-j) \geq n]}.$$

**Proof.** Let  $X_i = 0$  if category  $i$  is empty and  $X_i = 1$  if category  $i$  is occupied. Hence, we need  $|\cap_{i=1}^l \{X_i = 1\}|$ , where  $|\{X_i = 1\}| = \binom{M}{n}$ . This can be calculated by using the inclusion-exclusion principle.  $\square$

Using Lemma 1, we have the following results.

**Proposition 1.** *The probability distribution of  $S$  is given by*

$$P(S = s) = \frac{\binom{N}{s} H_{n,M}(N-s)}{\binom{NM}{n}}.$$

Let  $X_i$  denote the number of units observed in the  $i$ th category for  $i = 1, \dots, N$ . Let  $Y_i = 1$  for  $X_i = 0$  and otherwise,  $Y_i = 0$ . Then,  $S = \sum_{i=1}^N Y_i$  denotes the number of unobserved categories, and  $N - S$  denotes the number of distinct categories included in a sample of size  $n$ . Now, let  $T_n(k)$  denote the number of possible combination of a subset  $k$  of  $N$ , that is,  $T_n(k) = \binom{N}{k} \binom{Mk}{n}$  for  $k = 1, \dots, N$ . Since  $S = \sum_{i=1}^N Y_i$ ,  $E(S) = N \frac{T_n(N-1)}{T_n(N)}$ .

**Proposition 2.** *The  $k$ th factorial moment of  $S$  is given by*

$$E(S^{(k)}) = k! \frac{T_n(N-k)}{T_n(N)} = \frac{N!}{(N-k)!} \frac{\binom{M(N-k)}{n}}{\binom{MN}{n}}.$$

For example,

$$E(S) = \frac{\binom{N}{1} \binom{M(N-1)}{n}}{\binom{MN}{n}} \quad \text{and} \quad E(S^{(2)}) = 2! \frac{\binom{N}{2} \binom{M(N-2)}{n}}{\binom{MN}{n}}.$$

**Theorem 2.1.** *The probability distribution of  $S$  is given by*

$$P(S=s) = \frac{1}{s!} \sum_{j=0}^{N-s-1} (-1)^j \frac{E(S^{(s+j)})}{j!}.$$

**Proof.**

$$P(S=s) = \frac{\binom{N}{s} H_{n,M}(N-s)}{\binom{NM}{n}} = \frac{\binom{N}{s} \sum_{j=0}^{N-s-1} (-1)^j \binom{N-s}{j} \binom{M(N-s-j)}{n}}{\binom{NM}{n}}$$

and

$$\frac{E(S^{(s+j)})}{s!j!} = (s+j)! \frac{\binom{N}{s+j} \binom{M(N-s-j)}{n}}{j! \binom{NM}{n}} = \frac{\binom{N}{s} \binom{N-s}{j} \binom{M(N-s-j)}{n}}{\binom{NM}{n}}. \quad \square$$

Equivalently, we write

$$P(S=s) = \sum_{k=s}^{N-1} (-1)^{k-s} \binom{k}{s} \frac{E(S^{(k)})}{k!}.$$

Note that we can write as  $P = CT$ , where  $P$  is a  $N \times 1$  vector with  $P(S=s)$  as an element,  $T$  is a  $N \times 1$  vector with  $\frac{E(S^{(k)})}{k!}$  as an element for  $k = 0, \dots, N-1$ ,

and  $C$  is  $N \times N$  coefficients matrix shown below.

$$\begin{pmatrix} \binom{0}{0} & (-1)^1 \binom{1}{1} & \binom{2}{2} & \cdots & (-1)^{N-2} \binom{N-2}{N-2} & (-1)^{N-1} \binom{N-1}{N-1} \\ 0 & \binom{1}{0} & (-1)^1 \binom{2}{1} & \cdots & (-1)^{N-3} \binom{N-2}{N-3} & (-1)^{N-2} \binom{N-1}{N-2} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & \ddots & \binom{N-2}{0} & (-1)^1 \binom{N-1}{1} \\ 0 & 0 & 0 & \cdots & 0 & \binom{N-1}{0} \end{pmatrix}.$$

When  $N = 5$ ,  $C$  is given by

$$\begin{pmatrix} 1 & -1 & 1 & -1 & 1 \\ 0 & 1 & -2 & 3 & -4 \\ 0 & 0 & 1 & -3 & 6 \\ 0 & 0 & 0 & 1 & -4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Note that the  $X$ th column of  $C$  is the coefficient of  $A^{X-1-k} B^k$  in the binomial expansion of  $(A - B)^{X-1} = \sum_{k=0}^{X-1} (-1)^{X-1-k} \binom{X-1}{k} A^{X-1-k} B^k$  for  $X = 1, \dots, N$ . Now it can be verified that  $P(S \leq s) = ACE$ , where  $E$  is the vector of factorial moments and  $A$  is the lower triangular matrix with 1's. The following lemma gives the key instrument for computing the factorial moments.

**Lemma 2.** *The upper triangular Pascal's matrix has the inverse that is the same matrix, except with every other sub-diagonal multiplied by  $-1$ .*

For the case of  $N = 5$ ,

$$\begin{pmatrix} 1 & -1 & 1 & -1 & 1 \\ 0 & 1 & -2 & 3 & -4 \\ 0 & 0 & 1 & -3 & 6 \\ 0 & 0 & 0 & 1 & -4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 3 & 6 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

**Theorem 2.2.** *The factorial moment of  $S$  is given by the*

$$E(S^{(k)}) = k! \sum_{s=k}^{N-1} \binom{s}{k} \binom{N}{s} P(S=s).$$

**Theorem 2.3.** Assume that an observed unit has a probability  $1 - p$  of being lost. Then, the  $k$ th factorial moment of  $S$  is given by

$$E(S^{(k)}) = \sum_{t=0}^n k! \frac{\binom{N}{k} \binom{M(N-k)}{t}}{\binom{MN}{t}} \binom{n}{t} p^t (1-p)^{n-t}.$$

**Proof.**

$$\begin{aligned} E(S^{(k)}) &= \sum_{t=0}^n E(S^{(k)}|t) \binom{n}{t} p^t (1-p)^{n-t} = \sum_{t=0}^n k! \frac{T_t(N-k)}{T_t(N)} \binom{n}{t} p^t (1-p)^{n-t} \\ &= \sum_{t=0}^n k! \frac{\binom{N}{k} \binom{M(N-k)}{t}}{\binom{MN}{t}} \binom{n}{t} p^t (1-p)^{n-t}. \end{aligned}$$

□

## 2.2 Asymptotic results for $S$ as $M$ tends to infinity

From Theorem 2.1, we get the following result.

**Proposition 3.** As  $M$  tends to infinity, the factorial moment of  $S$  is given by

$$E(S^{(k)}) \xrightarrow{M \rightarrow \infty} k! \binom{N}{k} \left(1 - \frac{k}{N}\right)^n.$$

**Proposition 4.** As  $M$  tends to infinity, the probability distribution of  $S$  is given by

$$P(S=s) \xrightarrow{M \rightarrow \infty} \frac{N!}{(N-s)!N^n} S(n, N-s),$$

where  $S(\cdot)$  denotes the Sterling's number of the second kind.

**Proof.**

$$\begin{aligned} P(S=s) &\xrightarrow{M \rightarrow \infty} \sum_{k=s}^{N-1} (-1)^{k-s} \binom{k}{s} \binom{N}{k} \left(1 - \frac{k}{N}\right)^n \\ &= \frac{\binom{N}{s}}{N^n} \sum_{k=s}^{N-1} (-1)^{k-s} \binom{N-s}{k-s} (N-k)^n. \end{aligned}$$

Now, substitute  $q$  for  $k - s$ ,

$$= \frac{\binom{N}{s}}{N^n} \sum_{q=0}^{N-s-1} (-1)^q \binom{N-s}{q} (N-s-q)^n.$$

□

From Theorem 2.2, we get the following result.

**Proposition 5.** For  $M \rightarrow \infty$ ,

$$E(S^{(k)}) \xrightarrow{M \rightarrow \infty} N^{(k)} \left(1 - \frac{kp}{N}\right)^n.$$

Using the relationship between the probability generating function and the factorial moment generating function, we establish the following proposition.

**Proposition 6.** For  $M \rightarrow \infty$ , the asymptotic probability distribution of  $S$  is given by

$$P(S = s) \xrightarrow{M \rightarrow \infty} \sum_{k=s}^{N-1} (-1)^{k-s} \binom{k}{s} \binom{N}{k} \left(1 - \frac{kp}{N}\right)^n.$$

### 3 Results for $S$ with unequal number of copies in each categories

#### 3.1 Case when $M_i$ 's are finite numbers

Suppose there are  $N$  categories and the  $i$ th category consists of  $M_i$  copies.

**Lemma 3.** Let  $H^n(k)$  denote the number of ways  $(N - k)$  categories is unobserved for a given subset of  $k$  out of  $N$  categories. Then, it can be calculated using the recursive formula shown below.

$$H^n(k) = \sum_{j=0}^{k-1} (-1)^j \sum_{i_1, \dots, i_{k-j}} \binom{\sum_{u=1}^{k-j} M_{i_u}}{n} I_{[\sum_{u=1}^{k-j} M_{i_u} \geq n]},$$

where the summation is over all possible subsets of size  $k - j$ .

**Proof.** First, note that  $\sum_{i_1, \dots, i_{k-j}} \binom{\sum_{u=1}^{k-j} M_{i_u}}{n} I_{[\sum_{u=1}^{k-j} M_{i_u} \geq n]}$  is the number of ways at least  $j$  categories are unobserved among a given set of  $k$  categories. Denote it by  $H^{n*}(j)$ . By using the inclusion-exclusion principle, we have the following recursive formula

$$H^n(k) = \sum_{j=0}^{k-1} (-1)^j H^{n*}(j).$$

□

Using Lemma 3, the following theorem is established.

**Theorem 3.1.** When the number of copies in  $N$  categories are not all the same, the probability distribution of  $S$  is given by

$$P(S = s) = \frac{\binom{N}{s} \sum_{j=0}^{N-s-1} (-1)^j \sum_{i_1, \dots, i_{N-s-j}} \binom{\sum_{u=1}^{N-s-j} M_{i_u}}{n}}{\binom{\sum_{i=1}^N M_i}{n}}.$$

**Theorem 3.2.** *The  $k$ th factorial moment of  $S$  is given by*

$$E(S^{(k)}) = k! \sum_{i_1, \dots, i_k} \left( 1 - \frac{\sum_{q=1}^k M_{i_q}}{\sum_{i=1}^N M_i} \right)^n$$

### 3.2 Asymptotic results as $M_i$ 's tend to infinity

Denote  $\frac{M_i}{\sum_{i=1}^N M_i} = p_i$  as  $M_i$ 's tend to infinity and  $\sum_{i=1}^N p_i = 1$ .

**Proposition 7.** *As  $M_i$ 's tend to infinity,*

$$E(S^{(k)}) \longrightarrow k! \sum_{i_1, \dots, i_k} \left( 1 - \sum_{q=1}^k p_{i_q} \right)^n,$$

where the summation is over all possible subsets of size  $k$ .

Using the relationship between the probability generating function and the factorial moment generating function, we establish the following proposition.

**Proposition 8.** *The asymptotic probability distribution of  $S$  is given by*

$$P(S = s) \longrightarrow \sum_{k=s}^{N-1} (-1)^{k-s} \binom{k}{s} T^n(k) = \sum_{k=s}^{N-1} (-1)^{k-s} \binom{k}{s} \sum_{i_1, \dots, i_k} \left( 1 - \sum_{q=1}^k p_{i_q} \right)^n.$$

**Proposition 9.** *Assume that an observed unit has a probability  $1 - p$  of being lost. Then*

$$E(S^{(k)}) \longrightarrow k! \sum_{i_1, \dots, i_k} \left( 1 - p \sum_{q=1}^k p_{i_q} \right)^n.$$

**Proof.**

$$\begin{aligned} E(S^{(k)}) &\longrightarrow \sum_{t=0}^n E(S^{(k)} | t) \binom{n}{t} p^t (1-p)^{n-t} \\ &= k! \sum_{i_1, \dots, i_k} \sum_{t=0}^n \binom{n}{t} \left( p - p \sum_{q=1}^k p_{i_q} \right)^t (1-p)^{n-t}. \end{aligned}$$

□

Let  $B = \sum_{j=1}^N Z_j$ , where  $Z_j = 1$  if  $j$ th category is observed in a sample of size  $n$ , and  $Z_j = 0$  otherwise, for  $j = 1, \dots, N$ . Then,  $B$  is the sum of  $N$  dependent Bernoulli random variables  $Z_j$ , where  $P(Z_j = 1) = 1 - (1 - p_j)^n$ . Hence, the asymptotic distribution of  $S (= N - B)$  is given in the following theorem.

**Theorem 3.3.** As  $N$  and  $M_i$ 's tend to infinity,

$$\frac{S - E(S)}{(\text{Var}(S))^{1/2}} \longrightarrow N(0, 1),$$

where

$$\begin{aligned} E(S) &= \sum_{j=1}^N (1 - p_j)^n \quad \text{and} \\ \text{Var}(S) &= \sum_{j=1}^N [(1 - p_j)^n (1 - (1 - p_j)^n)] \\ &\quad + \sum_{i \neq k}^N [(1 - p_i - p_k)^n - (1 - p_i)^n (1 - p_k)^n]. \end{aligned}$$

#### 4 Asymptotic results for $S_j$

Let  $f_{ij} = 1(0)$  if  $x_i = j (\neq j)$ , where  $x_i$  denotes the frequency counts of  $i$ th category. Then,  $S_j = \sum_i^N f_{ij}$  denotes the number of categories with  $j$  samples with  $\sum_j^N S_j = N$  and  $\sum_j^n j S_j = n$ , and the number of possible sequences  $(S_0, S_1, \dots, S_N)$  is  $\binom{N+n-1}{n}$ . Denote  $\frac{M_i}{\sum_{i=1}^N M_i} \longrightarrow p_i$  as  $M_i \longrightarrow \infty$  for  $i = 1, \dots, N$  and  $\sum_{i=1}^N p_i = 1$ .

Note that we get

$$\begin{aligned} E(S_j) &\longrightarrow \sum_i^N \binom{n}{j} p_i^j (1 - p_i)^{n-j} \quad \text{and} \\ E(S_j^{(2)}) &\longrightarrow \sum_{i_1, i_2} \frac{n!}{j! j! (n-2j)!} p_{i_1}^j p_{i_2}^j (1 - p_{i_1} - p_{i_2})^{n-2j}, \end{aligned}$$

since  $E(f_{ij}) \longrightarrow \binom{n}{j} p_i^j (1 - p_i)^{n-j}$  and  $E(f_{i_1 j} f_{i_2 j}) \longrightarrow \frac{n!}{j! j! (n-2j)!} p_{i_1}^j p_{i_2}^j (1 - p_{i_1} - p_{i_2})^{n-2j}$ . Hence, the following proposition follows.

**Proposition 10.**

$$E(S_j^{(k)}) \longrightarrow \sum_{i_1, \dots, i_k} \frac{n!}{(j!)^k (n - kj)!} \prod_{m=1}^k p_{i_m}^j (1 - \sum_{m=1}^k p_{i_m})^{n-kj}.$$

Similarly to the case for  $S$ , we obtain the probabilities of  $S_j$  as shown in the following matrix notation,

$$P_j = C T_j,$$



where  $P_j$  is  $(N + 1) \times 1$  vector of  $[P(S_j = 0), \dots, P(S_j = N)]'$ ,  $C$  is the  $(N + 1) \times (N + 1)$  matrix and  $T_j$  is  $(N + 1) \times 1$  vector of  $[1, T_j(1), \dots, T_j(N)]'$ , where  $T_j(k) = \frac{E(S_j^{(k)})}{k!}$ .

**Proposition 11.** *Suppose that an observed unit has a probability  $1 - p$  of being lost, the  $k$ th factorial moment of  $S$  is written as*

$$E(S_j^{(k)}) \longrightarrow \binom{N}{k} \frac{n!}{(j!)^k (n - kj)!} \prod_{m=1}^k (pp_{i_m})^j (1 - p \sum_{m=1}^k p_{i_m})^{n-kj}.$$

**Proof.**

$$\begin{aligned} E(S_j^{(k)}) &\longrightarrow \sum_{t=0}^n E(S_j^{(k)} | t) \binom{n}{t} p^t (1 - p)^{n-t} \\ &= \sum_{t=0}^n \sum_{i_1, \dots, i_k} \frac{t!}{(j!)^k (t - kj)!} \prod_{m=1}^k p_{i_m}^j (1 - \sum_{m=1}^k p_{i_m}^j)^{t-kj} \binom{n}{t} p^t (1 - p)^{n-t} \\ &= \sum_{i_1, \dots, i_k} \frac{n!}{(j!)^k (n - kj)!} \prod_{m=1}^k (pp_{i_m})^j (1 - p \sum_{m=1}^k p_{i_m})^{n-kj}. \end{aligned} \quad \square$$

## 5 Numerical examples

In this section, we present simple numerical exercises using the proposed algorithms. In the following example, we have 4 categories having unequal number of copise, where  $p_1 = 0.1$ ,  $p_2 = 0.2$ ,  $p_3 = 0.3$  and  $p_4 = 0.4$ , and the sample size is given by 5.

### 5.1 Case for $S$

In this exercise, we use

$$W_4 = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 0 & 1 & -2 & 3 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Using Proposition 8, we compute

$$T_1(1) = E(S_1) = 1.164,$$

$$T_1(2) = \frac{E(S_1^{(2)})}{2!} = 0.20175,$$

$$T_1(3) = \frac{E(S_1^{(3)})}{3!} = 0.013.$$

The probabilities of  $S$  by

$$\begin{pmatrix} P(S=0) \\ P(S=1) \\ P(S=2) \\ P(S=3) \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 0 & 1 & -2 & 3 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1.164 \\ 0.20175 \\ 0.013 \end{pmatrix} = \begin{pmatrix} 0.02475 \\ 0.7995 \\ 0.16275 \\ 0.013 \end{pmatrix}.$$

## 5.2 Case for $S_1$

In this exercise, we have

$$W_5 = \begin{pmatrix} 1 & -1 & 1 & -1 & 1 \\ 0 & 1 & -2 & 3 & -4 \\ 0 & 0 & 1 & -3 & 6 \\ 0 & 0 & 0 & 1 & -4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Using Proposition 10, the factorial moments are given by

$$\begin{aligned} E(S_1) &= 0.2714, & \frac{E(S_1^{(2)})}{2!} &= 0.0151, \\ \frac{E(S_1^{(3)})}{3!} &= 0.0024 & \text{ and } & \frac{E(S_1^{(4)})}{4!} = 0. \end{aligned}$$

We compute the probabilities of  $S_1$  as shown below.

$$\begin{pmatrix} P(S_1=0) \\ P(S_1=1) \\ P(S_1=2) \\ P(S_1=3) \\ P(S_1=4) \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 & -1 & 1 \\ 0 & 1 & -2 & 3 & -4 \\ 0 & 0 & 1 & -3 & 6 \\ 0 & 0 & 0 & 1 & -4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0.2714 \\ 0.0151 \\ 0.0024 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.7413 \\ 0.2484 \\ 0.0079 \\ 0.0024 \\ 0 \end{pmatrix}$$

## 6 Concluding remarks

We investigate the problem of sampling from a multivariate hypergeometric distribution. We propose the probability distribution and factorial moments of the number of unobserved categories and the number of categories with  $j$  samples for  $j = 1, \dots, n$ . The asymptotic results are also provided in this paper. Some of applications of the proposed results include: a filing or storage process, where objects are randomly assigned to files or storage bins, and from time to time, objects may be missing or have disappeared, species of insects as categories in a Biological application, or a DNA experimental study. We hope that our readers find this paper useful for theoretical developments and practical applications in their researches.

## References

- Feller, W. (1957). *An Introduction to Probability Theory and Its Applications*. New York: Wiley. [MR0088081](#)
- Harris, B., Marden, M. and Park, C. J. (1987). The distribution of the number of empty cells in a generalized random allocation scheme. *Annals of Discrete Mathematics* **33**, 77–90. [MR0930485](#)
- Korwar, R. M. (1988). On the observed number of classes from multivariate power series and hypergeometric distributions. *Sankhya* **50**, 39–59. [MR1056251](#)
- Park, C. J. (1972). A note on the classical occupancy problem. *Annals of Mathematical Statistics* **43**, 1698–1701. [MR0346870](#)
- Park, C. J. (1981). On the distribution of the number of unobserved elements when  $m$  samples of size  $n$  are drawn from a finite population. *Communications in Statistics, Theory and Methods* **10**, 371–383.
- Samuel-Cahn, E. (1974). Asymptotic distributions for occupancy and waiting time problems with positive probability of falling through the cells. *Annals of Probability* **2**, 515–521. [MR0365668](#)
- Walton, G. S. (1986). The number of observed classes from a multiple hypergeometric distributions. *Journal of American Statistical Association* **81**, 169–171. [MR0830577](#)

University of Louisiana at Lafayette  
104 E. University Circle  
Lafayette, Louisiana 70503  
USA  
E-mail: [dr.sungsu@gmail.com](mailto:dr.sungsu@gmail.com)

San Diego State University  
550 Campanile Dr  
San Diego, California 92182  
USA  
E-mail: [chongjin.park@mail.sdsu.edu](mailto:chongjin.park@mail.sdsu.edu)