Asymptotics of multivariate contingency tables with fixed marginals

Quan Zhou[a]

[a] *Department of Statistics, Rice University, Houston, Texas 77005, U.S.A.*

Abstract

We consider the asymptotic distribution of a cell in a $2 \times \cdots \times 2$ contingency table as the fixed marginal totals tend to infinity. The asymptotic order of the cell variance is derived and a useful diagnostic is given for determining whether the cell has a Poisson limit or a Gaussian limit. There are three forms of Poisson convergence. The exact form is shown to be determined by the growth rates of the two smallest marginal totals. The results are generalized to contingency tables with arbitrary sizes and are further complemented with concrete examples.

Keywords: coupon collector's problem, negative association, negative relation, random allocation, Stein-Chen's method

1. Introduction

This work considers the asymptotic distribution of a cell in a $2 \times \cdots \times 2$ contingency table as the fixed marginal totals tend to infinity. The literature on this problem has been documented under various names: "the coupon collector problem," "capture-recapture," "the committee problem," "matrix occupancy," "random allocation," and "allocation by complexes" (Barbour et al., 1992, Sec. 6.4). The reader is encouraged to consult Holst (1986) and Stadje (1990) for historical accounting of these problems.

The present work borrows the framework and terminology of the coupon-collector problem. Consider n distinct coupons and m coupon collectors operating independently and let the ith collector collect a_i distinct coupons. Let $\mathcal{C} = \{1, 2, \dots, m\}$ denote the set of the collectors. For each set $\mathcal{C}' \subseteq \mathcal{C}$, we are interested in the number of coupons that are collected by $\mathcal{C}'$ and by no others. These counts may be summarized in an m-way $2 \times 2 \times \cdots \times 2$ contingency table. Let $X_{\boldsymbol{v}}$ denote the count in the cell $\boldsymbol{v} = (v_1, \dots, v_m)$, where $v_i \in \{1, 2\}$ and $v_i = 1$ indicates that a coupon is collected by collector i. This contingency table must satisfy $\sum_{v_i=1} X_{\boldsymbol{v}} = a_i$ and $\sum_{v_i=2} X_{\boldsymbol{v}} = n - a_i$, for $i = 1, \dots, m$, where the marginal total a_i is treated as fixed. For the case of two collectors, the 2×2 contingency table is shown in Table 1.

		Collector 2	
		Collected	Not collected
Collector 1	Collected	$X_{(1,1)}$	$X_{(1,2)}$
	Not collected	$X_{(2,1)}$	$X_{(2,2)}$

Table 1: The contingency table for $m = 2$. The cell counts must satisfy $X_{(1,1)} + X_{(1,2)} = a_1$ and $X_{(1,1)} + X_{(2,1)} = a_2$. When we have a third collector, we can construct a $2 \times 2 \times 2$ contingency table by splitting each cell in the above table into two, according to whether the coupon is collected by collector 3.

We consider the distribution of an arbitrary cell under the following asymptotic conditions:

(A1) $n \to \infty$;

(A2) $a_i = a_i(n) \to \infty$ and $n - a_i \to \infty$ for $i = 1, \dots, m$;

(A3) $1 \le a_1 \le a_2 \le \cdots \le a_m \le n - 1$;

(A4) $a_i/n \to \alpha_i \in [0, 1]$ for $i = 1, \dots, m$.

Under (A1)–(A4), each cell can be treated equivalently up to relabelling of rows and columns. Therefore, without loss of generality, it suffices to consider one cell. Henceforth our analysis shall concern the cell $X_{\mathbf{1}}$, where $\mathbf{1} = (1, \ldots, 1)$, i.e. the number of the coupons that are collected by all collectors.

To the best of our knowledge, the first complete analysis of all the possible asymptotic limits of $X_{\mathbf{1}}$ is due to Vatutin and Mikhailov (1983). The authors showed that $X_{\mathbf{1}}$ has either a normal or a Poisson limit depending on whether $\mathrm{Var}(X_{\mathbf{1}})$ converges (see Theorem 1 below). This was accomplished by verifying that its generating function has only real roots (see also Kou and Ying, 1996). Alternative proofs for this problem and its variants are given in Kolchin et al. (1978, Chap. VII), Holst (1980), Mitwalli (2002), Harris (1989), and Cekanavicius et al. (2000). See Smythe (2011) for an extension to the case in which $a_1, \ldots, a_m$ are random. See Lareida et al. (2017) for a more recent application of these results.

Theorem 1 (Vatutin and Mikhailov (1983)). *Under the asymptotic assumptions* (A1)–(A4)*, if* $\mathrm{Var}(X_{\mathbf{1}}) \to \infty$*,* $X_{\mathbf{1}}^* \equiv (X_{\mathbf{1}} - \mathbb{E}(X_{\mathbf{1}}))/\sqrt{\mathrm{Var}(X_{\mathbf{1}})} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$*; if* $\mathrm{Var}(X_{\mathbf{1}}) \to \rho < \infty$*,* $X_{\mathbf{1}}$ *has a Poisson limit in the sense that there exists a sequence of constants* C_n *such that* $X_{\mathbf{1}} + C_n \xrightarrow{\mathcal{D}} \mathrm{Pois}(\rho)$ *or* $-X_{\mathbf{1}} + C_n \xrightarrow{\mathcal{D}} \mathrm{Pois}(\rho)$*. (If* $\rho = 0$*,* $\mathrm{Pois}(0)$ *refers to the degenerate distribution* δ_0*.)*

In Section 2, we calculate the asymptotic order of $\mathrm{Var}(X_{\mathbf{1}})$. This provides a useful diagnostic for determining whether the limiting distribution of $X_{\mathbf{1}}$ given by Theorem 1 is normal or Poisson. In Section 3, we show that the exact form of Poisson convergence is determined only by a_1 and a_2. Section 4 generalizes the results of Sections 2 and 3 to contingency tables of arbitrary size.

2. Asymptotics of the cell variance

Mathematical induction will be used to prove most of our key results. The induction setup is described as follows. In lieu of considering an m-way contingency table, we consider a sequence of contingency tables, each of which has a grand total count of n. The kth table records the coupon counts for the first k coupon collectors and we use $X_{\mathbf{1}}^{(k)}$ to denote the number of the coupons that are collected by each of the first k collectors (whether the coupon is collected by the other collectors is not considered.) When we use induction to prove a statement regarding $X_{\mathbf{1}}$, we always start from checking the statement for $X_{\mathbf{1}}^{(2)}$ and then proceed to prove it for $X_{\mathbf{1}}^{(k)}$, for $k = 2, \ldots, m$.

Let E_k and V_k denote the expectation and the variance of $X_{\mathbf{1}}^{(k)}$. Clearly, $E_1 = a_1$ and $V_1 = 0$, and for $k = 2, 3, \ldots$,

$$E_k = n \prod_{i=1}^{k} \frac{a_i}{n} = \frac{a_k}{n} E_{k-1},$$

Since $X_{\mathbf{1}}^{(2)}$ follows a hypergeometric distribution,

$$V_2 = \frac{a_1 a_2 (n - a_1)(n - a_2)}{n^2(n-1)}. \tag{1}$$

We proceed to derive a recursive characterization of V_k.

Lemma 1. *For* $k = 2, 3, \dots,$

$$V_k = \frac{a_k(n - a_k)E_{k-1}(n - E_{k-1})}{n^2(n-1)} + \frac{a_k(a_k - 1)}{n(n-1)} V_{k-1}. \tag{2}$$

Remark 1. *The formula* (2) *decomposes* V_k *into two additive components. The first component is the variance of a cell from a* 2×2 *contingency table with fixed marginal totals* a_k *and* E_{k-1}*. The second component captures the variation of* $X_{\mathbf{1}}^{(k-1)}$*, which is 0 if* $V_{k-1} \to 0$*. If* $\alpha_k = 1$*, the second component converges to* V_{k-1}*. See Darroch (1958) for a closed-form expression for* V_k*.*

Proof. By the law of total variance, we express $\mathrm{Var}(X_{\mathbf{1}}^{(k)})$ as

$$V_k = \mathbb{E}\left(\mathrm{Var}(X_{\mathbf{1}}^{(k)} \mid X_{\mathbf{1}}^{(k-1)})\right) + \mathrm{Var}\left(\mathbb{E}(X_{\mathbf{1}}^{(k)} \mid X_{\mathbf{1}}^{(k-1)})\right).$$

After conditioning on $X_{\mathbf{1}}^{(k-1)}$, $X_{\mathbf{1}}^{(k)}$ is a hypergeometric random variable and thus we obtain

$$\begin{aligned}
\mathbb{E}(X_{\mathbf{1}}^{(k)} \mid X_{\mathbf{1}}^{(k-1)}) &= a_k X_{\mathbf{1}}^{(k-1)}/n, \\
\mathrm{Var}(X_{\mathbf{1}}^{(k)} \mid X_{\mathbf{1}}^{(k-1)}) &= a_k(n - a_k) X_{\mathbf{1}}^{(k-1)}(n - X_{\mathbf{1}}^{(k-1)})/n^2(n-1).
\end{aligned}$$

Routine calculations using $\mathbb{E}(X_{\mathbf{1}}^{(k-1)})^2 = V_{k-1} + E_{k-1}^2$ yield (2). □

Lemma 1 will be important for proving a series of asymptotic results for our problem. Our first asymptotic result regards the asymptotic order of $\mathrm{Var}(X_{\mathbf{1}})$. Let $\sim$ denote the asymptotic equivalence, i.e., $x_n \sim y_n$ if $\lim_{n\to\infty} x_n/y_n = 1$. Let $\asymp$ denote that two positive sequences have the same asymptotic order, i.e., $x_n \asymp y_n$ if both $\limsup_{n\to\infty} x_n/y_n$ and $\liminf_{n\to\infty} x_n/y_n$ are finite and strictly positive. Hence, $x_n \sim y_n$ is a special case of $x_n \asymp y_n$.

Theorem 2 (order of $\mathrm{Var}(X_{\mathbf{1}})$)**.** *Under the assumptions* (A1)–(A4)*, the asymptotic order of* $\mathrm{Var}(X_{\mathbf{1}})$ *is*

$$\mathrm{Var}(X_{\mathbf{1}}) \asymp \frac{(n - a_1)(n - a_2)}{n} \prod_{i=1}^{m} \frac{a_i}{n} = \left(1 - \frac{a_1}{n}\right)\left(1 - \frac{a_2}{n}\right) \mathbb{E}(X_{\mathbf{1}}).$$

Remark 2. *The claim is not true if* $m \to \infty$*. For example, let* $a_i = n - 1$ *for* $i = 1, \dots, m$ *and* $m^2/n \to 2\lambda$*, we have* $m - n + X_{\mathbf{1}} \xrightarrow{\mathcal{D}} \mathrm{Pois}(\lambda)$ *and thus* $\mathrm{Var}(X_{\mathbf{1}}) \to \lambda$*. This is in fact the classical birthday problem (Arratia et al., 1989; Diaconis and Holmes, 2002; DasGupta, 2005).*

Proof. We prove by induction on $X_{\mathbf{1}}^{(k)}$. That is, we aim to prove that

$$\mathrm{Var}(X_{\mathbf{1}}^{(k)}) = V_k \asymp \frac{(n-a_1)(n-a_2)}{n}\prod_{i=1}^{k}\frac{a_i}{n} = \left(1-\frac{a_1}{n}\right)\left(1-\frac{a_2}{n}\right)E_k, \tag{3}$$

for $k = 2, \ldots, m$. By (1), the claim holds trivially for $X_{\mathbf{1}}^{(2)}$. We now suppose the above claim holds for $X_{\mathbf{1}}^{(k-1)}$ $(k \geq 3)$ and consider $X_{\mathbf{1}}^{(k)}$.

The first subcase to consider is $\alpha_k = \lim a_k/n = 0$. In this subcase, by assumption (A3), $\alpha_i = 0$ for $i \leq k$. Hence,

$$\left(1-\frac{a_1}{n}\right)\left(1-\frac{a_2}{n}\right) \sim 1.$$

Since $E_{k-1} \leq a_1$, $E_{k-1}/n \to 0$. Hence, the first component of V_k in (2) is

$$\frac{a_k(n-a_k)E_{k-1}(n-E_{k-1})}{n^2(n-1)} \sim \frac{a_k E_{k-1}}{n} = E_k. \tag{4}$$

According to the induction assumption, $V_{k-1} \asymp E_{k-1}$ and thus (4) has the same order as $a_k V_{k-1}/n$. Since $\alpha_k = 0$, the second component of V_k in (2) has a strictly smaller order. Hence, the order of V_k is determined by its first component, which is asymptotically equal to E_k by (4), and thus (3) holds.

The second subcase we consider is $\alpha_k \in (0, 1]$. By the induction assumption, $V_{k-1} \asymp (n-a_1)(n-a_2)E_{k-1}/n^2$. Hence,

$$\frac{a_k(n-a_k)E_{k-1}(n-E_{k-1})}{n^2(n-1)} \asymp \frac{(n-a_k)(n-E_{k-1})V_{k-1}}{(n-a_1)(n-a_2)} < \frac{(n-E_{k-1})V_{k-1}}{n-a_1}.$$

However, since $n - a_1 \leq n - X_{\mathbf{1}}^{(k-1)} \leq (k-1)(n-a_1)$ and k is finite, we have $n - E_{k-1} \asymp n - a_1$. Thus the first component of V_k in (2) has the same or a smaller order than V_{k-1}. Since $\alpha_k > 0$ implies that the second component of V_k in (2) has the same asymptotic order as V_{k-1},

$$V_k \asymp V_{k-1} \asymp \left(1-\frac{a_1}{n}\right)\left(1-\frac{a_2}{n}\right)E_{k-1} \asymp \left(1-\frac{a_1}{n}\right)\left(1-\frac{a_2}{n}\right)E_k.$$

This completes the proof. □

By Theorem 1, the limiting distribution of $X_{\mathbf{1}}$ is fully determined by the convergence of the sequence $n^{-(m+1)}(n-a_1)(n-a_2)a_1\cdots a_m$. If it converges to zero, $X_{\mathbf{1}}$ converges in probability to some constant; if it converges to some finite nonzero constant, $X_{\mathbf{1}}$ has a Poisson limit. The following corollary shows that $X_{\mathbf{1}}$ has a Poisson limit only when $\alpha_1, \alpha_2 \in \{0, 1\}$.

Corollary 1. *Under the assumptions* (A1)–(A4), $\mathrm{Var}(X_{\mathbf{1}})$ *may converge to a finite constant only if* $\alpha_1, \alpha_2 \in \{0, 1\}$ *where* $\alpha_i = \lim a_i/n$. *This condition is necessary but not sufficient.*

Proof. By assumption (A2), $a_i(n-a_i)/n \to \infty$ for every i. Hence, according to (1), the claim holds for $\mathrm{Var}(X_{\mathbf{1}}^{(2)})$. Now consider $\mathrm{Var}(X_{\mathbf{1}}^{(k)})$ with $k \geq 3$. By assumption (A3), if α_1 or α_2 is in $(0,1)$, we have $\alpha_k > 0$. By Theorem 2, this implies that $\mathrm{Var}(X_{\mathbf{1}}^{(k)})$ has the same order as $\mathrm{Var}(X_{\mathbf{1}}^{(k-1)})$ and thus diverges. To see this condition is not sufficient, consider $\mathrm{Var}(X_{\mathbf{1}}^{(2)})$ for $a_1 = a_2 = n^{2/3}$, which implies $\alpha_1 = \alpha_2 = 0$. A direct calculation using (1) gives $\mathrm{Var}(X_{\mathbf{1}}^{(2)}) \sim n^{1/3}$. $\square$

3. Poisson convergence

Consider the simplest case of two coupon collectors and the associated 2×2 contingency table. If $\mathrm{Var}(X_{\mathbf{1}}) \to 0$, the variance of any other cell must also tend towards zero since there is only one degree of freedom when the marginal totals are fixed. It is straightforward to see that $X_{\mathbf{1}}$ should have three different "limits". First, if $\alpha_1 = \alpha_2 = 0$, we have $X_{\mathbf{1}} \to 0$. Second, if $\alpha_1 = 0, \alpha_2 = 1$, then $X_{(1,2)} = a_1 - X_{\mathbf{1}} \to 0$, i.e. every coupon collected by the first collector would also be collected by the second. Third, if $\alpha_1 = \alpha_2 = 1$, then $X_{(2,2)} = X_{\mathbf{1}} + n - a_1 - a_2 \to 0$, i.e. no coupon would be missed by both collectors.

For the Poisson convergence of m-way $2 \times \cdots \times 2$ contingency table, it still suffices to consider the above three scenarios.

Lemma 2. *Under the assumptions* (A1)–(A4)*, for* $m \geq 2$*,*

(i) $a_1/n \to 0$, $a_2/n \to 0$*:* $\mathbb{E}(X_{\mathbf{1}}) \sim \mathrm{Var}(X_{\mathbf{1}})$*;*

(ii) $a_1/n \to 0$, $a_2/n \to 1$*:* $\mathbb{E}(a_1 - X_{\mathbf{1}}) \sim \mathrm{Var}(X_{\mathbf{1}})$*;*

(iii) $a_1/n \to 1$, $a_2/n \to 1$*:* $\mathbb{E}(X_{\mathbf{1}} + (m-1)n - \sum_{i=1}^{m} a_i) \sim \mathrm{Var}(X_{\mathbf{1}})$.

Remark 3. *No assumption about the convergence of* $\mathrm{Var}(X_{\mathbf{1}})$ *is needed.*

Proof. Just like we did in the proof for Theorem 2, we prove each case separately by induction on the sequence $X_{\mathbf{1}}^{(2)}, \ldots, X_{\mathbf{1}}^{(m)}$.

Case (i): We use induction to prove that, given $a_1/n \to 0$, $a_2/n \to 0$, we have $E_k \sim V_k$ for $k = 2, \ldots, m$. For $k = 2$, the statement can be verified immediately using (1). Next, we assume $E_{k-1} \sim V_{k-1}$ and consider E_k and V_k. Note that $\alpha_1 = \lim a_1/n = 0$ implies $E_j/n \to 0$ for every j since $E_j < a_1$. By the induction assumption and the identity $E_k = a_k E_{k-1}/n$, for the two components of V_k in (2) we have

$$\frac{a_k(n-a_k)E_{k-1}(n-E_{k-1})}{n^2(n-1)} \sim \frac{E_k(n-a_k)}{n-1}, \quad \frac{a_k(a_k-1)}{n(n-1)}V_{k-1} \sim \frac{E_k(a_k-1)}{n-1}.$$

It thus follows that $V_k = E_k + o(E_k)$.

Case (ii): We use induction to prove that, given $a_1/n \to 0$, $a_2/n \to 1$, we have $a_1 - E_k \sim V_k$ for $k = 2, \ldots, m$. Again, for $k = 2$, the statement is immediate by (1). For the induction step, observe that $(n - a_k)E_{k-1}/n = E_{k-1} - E_k$. Hence, assuming $a_1 - E_{k-1} \sim V_k - 1$, which is the induction assumption, and using the fact that $\alpha_k = 1$ and $E_{k-1}/n \to 0$, we obtain

$$\frac{a_k(n-a_k)E_{k-1}(n-E_{k-1})}{n^2(n-1)} \sim E_{k-1} - E_k, \quad \frac{a_k(a_k-1)}{n(n-1)} V_{k-1} \sim a_1 - E_{k-1}.$$

Since both terms are always positive, by (2), we arrive at $V_k = a_1 - E_k + o(V_k)$.

Case (iii): We use induction to prove that, given $a_1/n \to 1$, $a_2/n \to 1$, we have $E_k + (k-1)n - \sum_{i=1}^{k} a_i \sim V_k$, for $k = 2, \ldots, m$. For $k = 2$, the statement follows from (1). The induction argument is very similar to that of case (ii). We need only observe that $E_{k-1}/n \to 1$ and $(n - a_k)(n - E_{k-1})/n = E_k - E_{k-1} + n - a_k$. $\square$

To establish the Poisson convergence of $X_{\mathbf{1}}$, we use Stein-Chen's method for "negatively associated" and "negatively related" random variables, the definitions of which are given below.

Definition 1 (Joag-Dev and Proschan (1983))**.** *Random variables $Y_1, \ldots, Y_N$ are said to be negatively associated if for every pair of disjoint subsets $A_1, A_2 \subseteq \{1, 2, \ldots, N\}$ and any nondecreasing functions f_1 and f_2, we have*

$$\mathrm{Cov}(f_1(Y_i, i \in A_1), f_2(Y_j, j \in A_2)) \leq 0.$$

Definition 2 (Erhardsson (2005))**.** *Bernoulli random variables $Y_1, \ldots, Y_N$ are said to be negatively related if for each $i \in \{1, 2, \ldots, N\}$ and any nondecreasing function $f : \{0,1\}^{N-1} \mapsto \{0,1\}$, we have*

$$\mathbb{E}[f(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_N) \mid Y_i = 1] \leq \mathbb{E}[f(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_N)].$$

In particular, negatively associated Bernoulli random variables are negatively related (Barbour et al., 1992, Theorem 2.I). We will first show that $X_{\mathbf{1}}$, $a_1 - X_{\mathbf{1}}$ and $X_{\mathbf{1}} + (m-1)n - \sum_{i=1}^{m} a_i$ can be decomposed into sums of negatively associated random variables. For $m = 2$, all the three random variables follow hypergeometric distribution, and the negative association property of hypergeometric random variables has been well studied (Joag-Dev and Proschan, 1983; Daly et al., 2012). Here we prove the general case $m \geq 2$.

Lemma 3. *$X_{\mathbf{1}}$ and $a_1 - X_{\mathbf{1}}$ can be written as sums of negatively related Bernoulli random variables. $X_{\mathbf{1}} + (m-1)n - \sum_{i=1}^{m} a_i$ can be written as a sum of non-negative integer-valued negatively associated random variables.*

Proof. For $X_{\mathbf{1}}$, the statement was proven in Barbour and Holst (1989) via coupling methods. Here we use another method, which works for all three random variables. Let I_{ij} ($i = 1, \dots, m$, $j = 1, \dots, n$) be a Bernoulli random variable such that $I_{ij} = 1$ if coupon j is collected by the ith collector. Let $J_{ij} = 1 - I_{ij}$. For each i, $\{I_{ij} : j = 1, \dots, n\}$ and $\{J_{ij} : j = 1, \dots, n\}$ are sets of negatively related random variables (Joag-Dev and Proschan, 1983, Theorem 2.11). The three random variables can be decomposed as

$$\begin{aligned} X_{\mathbf{1}} &= \sum_{j=1}^{n} Y_j, & Y_j &\equiv \min(I_{1j}, I_{2j}, \dots, I_{mj}), \\ a_1 - X_{\mathbf{1}} &= \sum_{j=1}^{n} Y'_j, & Y'_j &\equiv I_{1j} \max(J_{2j}, \dots, J_{mj}), \\ X_{\mathbf{1}} + (m-1)n - \sum_{i=1}^{m} a_i &= \sum_{j=1}^{n} Y''_j, & Y''_j &\equiv (-1 + \textstyle\sum_{i=1}^{m} J_{ij}) \vee 0. \end{aligned} \tag{5}$$

All the three functions, Y_j, Y'_j and Y''_j are nondecreasing. Applying Property P6 and Property P7 of Joag-Dev and Proschan (1983) and using the independence assumption of the collectors, we see that $\{Y_j\}, \{Y'_j\}$ and $\{Y''_j\}$ are sets of negatively associated random variables. Furthermore, $\{Y_j\}$ and $\{Y'_j\}$ are negatively related since they are indicator random variables. $\square$

For a sum of negatively related random variables, Stein-Chen's method allows us to establish the Poisson convergence by simply comparing the first two moments.

Theorem 3 (Poisson convergence of $X_{\mathbf{1}}$)**.** *Under the assumptions* (A1)–(A4), $X_{\mathbf{1}}$ *has a Poisson limit if and only if* $\mathrm{Var}(X_{\mathbf{1}}) \to \rho \in [0, \infty)$. *(*$\mathrm{Pois}(0)$ *refers to the degenerate distribution* δ_0*.) Let* $\mathrm{Pois}(\rho)$ *denote the Poisson distribution with parameter* ρ*. There are only three possible subcases:*

(i) $a_1/n \to 0$, $a_2/n \to 0$*:* $X_{\mathbf{1}} \xrightarrow{\mathcal{D}} \mathrm{Pois}(\rho)$*.*

(ii) $a_1/n \to 0$, $a_2/n \to 1$*:* $a_1 - X_{\mathbf{1}} \xrightarrow{\mathcal{D}} \mathrm{Pois}(\rho)$*;*

(iii) $a_1/n \to 1$, $a_2/n \to 1$*:* $X_{\mathbf{1}} + (m-1)n - \sum_{i=1}^{m} a_i \xrightarrow{\mathcal{D}} \mathrm{Pois}(\rho)$*.*

Proof. We need only prove sufficiency. By Corollary 1, the convergence of $\mathrm{Var}(X_{\mathbf{1}})$ requires $\alpha_1, \alpha_2 \in \{0, 1\}$. Since, by assumption (A3), $a_1 \le a_2$, Theorem 3 includes all the possible subcases where $\mathrm{Var}(X_{\mathbf{1}})$ converges. By Barbour et al. (1992, Corollary 2.C.2), if a random variable Z is a sum of negatively related Bernoulli random variables,

$$||\mathcal{L}(Z) - \mathrm{Pois}(\mathbb{E}(Z))||_{\mathrm{TV}} < 1 - \mathrm{Var}(Z)/\mathbb{E}(Z),$$

where $||\cdot||_{\mathrm{TV}}$ denotes the total variation distance. Thus the Poisson convergence for case (i) and (ii) immediately follows from Lemma 2 and Lemma 3.

We now turn to case (iii). To simplify notation, let $W \equiv X_{\mathbf{1}} + (m-1)n - \sum_{i=1}^m a_i$. Recall the decomposition $W = \sum_{j=1}^n Y_j''$ given in (5). Let $\theta \equiv \mathbb{E}(W)$ and $p \equiv \theta^{-1} \sum_{j=1}^n \mathbb{P}(Y_j'' = 1)$. By Daly and Johnson (2017, Corollary 4.2),

$$||\mathcal{L}(W) - \text{Pois}(\theta)||_{\text{TV}} \leq 1 + \theta + (1-2p)\left(\frac{\text{Var}(X_{\mathbf{1}})}{\theta} + \theta\right). \tag{6}$$

By construction, for $k \geq 1$, $\mathbb{P}(Y_j'' = k)$ is the probability that coupon j is not collected by exactly $k+1$ collectors. Using the fact that $a_i/n \to 1$ for each i, we can show that for each $k' \geq 2$, $\mathbb{P}(Y_j'' = k')/\mathbb{P}(Y_j'' = 1) \to 0$. This further implies that $\mathbb{E}(Y_j'') \sim \mathbb{P}(Y_j'' = 1)$ and thus $p \to 1$. Plugging this into (6) and using Lemma 2, we obtain $||\mathcal{L}(W) - \text{Pois}(\theta)||_{\text{TV}} \leq o(\theta) = o(1)$, which concludes the proof. □

4. Contingency tables with arbitrary sizes

We now extend our results to a general m-way contingency table with size $r_1 \times r_2 \times \cdots \times r_m$. We use $\tilde{X}_{\boldsymbol{v}}$ to denote a cell in the general contingency table with position $\boldsymbol{v} = (v_1, v_2, \ldots, v_m)$. The grand total of all the cells is n. The marginal totals are fixed and are denoted by $b_i(j)$ $(i = 1, \ldots, m$ and $j = 1, \ldots, r_i)$ which satisfy

$$b_i(j) = \sum_{v_i = j} \tilde{X}_{\boldsymbol{v}}, \qquad \sum_{j=1}^{r_i} b_i(j) = n. \tag{7}$$

Note that the coupon collector's problem is a special case of the above with $r_i = 2$, $b_i(1) = a_i$ and $b_i(2) = n - a_i$ for each i. To study the asymptotic distribution of $\tilde{X}_{\boldsymbol{v}}$, we return to the coupon collector's model specified in Section 1 and set $a_i = b_i(v_i)$. Then $\tilde{X}_{\boldsymbol{v}}$ has the same distribution as $X_{\mathbf{1}}$ in the coupon collector model and its asymptotic distribution can be determined by Theorem 2 (after reordering $a_1, \ldots, a_m$).

We conclude the present work with two examples. First, consider a three-way contingency table with $r_1 = 3, r_2 = r_3 = 2$. The marginals are given by $\boldsymbol{b}_1 = (n^{1/4}, n^{1/2}, n - n^{1/4} - n^{1/2})$, $\boldsymbol{b}_2 = (n^{1/2}, n - n^{1/2})$ and $\boldsymbol{b}_3 = (n^{1/2}, n - n^{1/2})$ where $\boldsymbol{b}_i = (b_i(1), \ldots, b_i(r_i))$. The limiting distributions of all the cells are given in Table 2. Using Theorem 1 and Lemma 2, each cell can be verified easily. It is also straightforward to check that all the marginal constraints are satisfied. Second, consider a three-way contingency table with the same size, same marginals $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$, but $\boldsymbol{b}_3 = (n/2, n/2)$. The limiting distributions of all the cells are given in Table 3. Now two thirds of the cells have normal limits and the variances of these cells are calculated manually.

$\tilde{X}_{ij1}$	$j=1$	$j=2$
$i=1$	$\tilde{X}_{111} \xrightarrow{\mathcal{P}} 0$	$\tilde{X}_{121} \xrightarrow{\mathcal{P}} 0$
$i=2$	$\tilde{X}_{211} \xrightarrow{\mathcal{P}} 0$	$\tilde{X}_{221} \xrightarrow{\mathcal{D}} \text{Pois}(1)$
$i=3$	$\tilde{X}_{311} \xrightarrow{\mathcal{D}} \text{Pois}(1)$	$n^{1/2} - \tilde{X}_{321} \xrightarrow{\mathcal{D}} \text{Pois}(2)$

$\tilde{X}_{ij2}$	$j=1$	$j=2$
$i=1$	$\tilde{X}_{112} \xrightarrow{\mathcal{P}} 0$	$n^{1/4} - \tilde{X}_{122} \xrightarrow{\mathcal{P}} 0$
$i=2$	$\tilde{X}_{212} \xrightarrow{\mathcal{D}} \text{Pois}(1)$	$n^{1/2} - \tilde{X}_{222} \xrightarrow{\mathcal{D}} \text{Pois}(2)$
$i=3$	$n^{1/2} - \tilde{X}_{312} \xrightarrow{\mathcal{D}} \text{Pois}(2)$	$\tilde{X}_{322} - n + n^{1/4} + 3n^{1/2} \xrightarrow{\mathcal{D}} \text{Pois}(3)$

Table 2: Example 1. The asymptotic distribution of a $3 \times 2 \times 2$ contingency table with fixed marginals: $\boldsymbol{b}_1 = (n^{1/4}, n^{1/2}, n - n^{1/4} - n^{1/2})$, $\boldsymbol{b}_2 = (n^{1/2}, n - n^{1/2})$ and $\boldsymbol{b}_3 = (n^{1/2}, n - n^{1/2})$ where $b_i(j)$ is defined in (7).

$\tilde{X}_{ij2}$	$j=1$	$j=2$
$i=1$	$\tilde{X}_{112} \xrightarrow{\mathcal{P}} 0$	$2n^{-1/8}(\tilde{X}_{122} - n^{1/4}/2) \xrightarrow{\mathcal{D}} \mathcal{N}^*$
$i=2$	$\tilde{X}_{212} \xrightarrow{\mathcal{D}} \text{Pois}(1/2)$	$2n^{-1/4}(\tilde{X}_{222} - n^{1/2}/2) \xrightarrow{\mathcal{D}} \mathcal{N}^*$
$i=3$	$2n^{-1/4}(\tilde{X}_{312} - n^{1/2}/2) \xrightarrow{\mathcal{D}} \mathcal{N}^*$	$\dfrac{\tilde{X}_{322} - n/2 + n^{1/2} + n^{1/4}/2}{n^{1/4}/\sqrt{2}} \xrightarrow{\mathcal{D}} \mathcal{N}^*$

Table 3: Example 2. The asymptotic distribution of a $3 \times 2 \times 2$ contingency table with fixed marginals: $\boldsymbol{b}_1 = (n^{1/4}, n^{1/2}, n - n^{1/4} - n^{1/2})$, $\boldsymbol{b}_2 = (n^{1/2}, n - n^{1/2})$ and $\boldsymbol{b}_3 = (n/2, n/2)$ where $b_i(j)$ is defined in (7). $\mathcal{N}^*$ denotes the standard normal distribution. Note that for any i, j, $\tilde{X}_{ij2}$ has the same distribution as $\tilde{X}_{ij1}$.

Acknowledgments

The author is grateful to Dr. Philip Ernst and Dr. Larry Brown for their helpful discussions.

References

Richard Arratia, Larry Goldstein, and Louis Gordon. Two moments suffice for poisson approximations: the chen-stein method. *The Annals of Probability*, pages 9–25, 1989.

A. D. Barbour and Lars Holst. Some applications of the stein-chen method for proving poisson convergence. *Advances in Applied Probability*, 21(1):74–90, 1989.

Andrew D. Barbour, Lars Holst, and Svante Janson. *Poisson approximation.* Clarendon Press Oxford, 1992.

Vydas Cekanavicius, Julius Kruopis, et al. Signed poisson approximation: a possible alternative to normal and poisson laws. *Bernoulli*, 6(4):591–606, 2000.

Fraser Daly and Oliver Johnson. Relaxation of monotone coupling conditions: Poisson approximation and beyond. *arXiv preprint arXiv:1706.04064*, 2017.

Fraser Daly, Claude Lefèvre, and Sergey Utev. Stein's method and stochastic orderings. *Advances in Applied Probability*, 44(2):343–372, 2012.

John N. Darroch. The multiple-recapture census: I. estimation of a closed population. *Biometrika*, 45(3/4):343–359, 1958.

Anirban DasGupta. The matching, birthday and the strong birthday problem: a contemporary review. *Journal of Statistical Planning and Inference*, 130(1): 377–389, 2005.

Persi Diaconis and Susan Holmes. A bayesian peek into feller volume i. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 820–841, 2002.

Torkel Erhardsson. Steins method for poisson and compound poisson. *'An Introduction to Stein's Method'*, 4:61, 2005.

Bernard Harris. Poisson limits for generalized random allocation problems. *Statistics & Probability Letters*, 8(2):123–127, 1989.

Lars Holst. On matrix occupancy, committee, and capture-recapture problems. *Scandinavian Journal of Statistics*, pages 139–146, 1980.

Lars Holst. On birthday, collectors', occupancy and other classical urn problems. *International Statistical Review/Revue Internationale de Statistique*, pages 15–27, 1986.

Kumar Joag-Dev and Frank Proschan. Negative association of random variables with applications. *The Annals of Statistics*, pages 286–295, 1983.

Valentin Fedorovich Kolchin, Boris Aleksandrovich Sevastyanov, and Vladimir Pavlovich Chistyakov. Random allocations. 1978.

S. G. Kou and Z. Ying. Asymptotics for a 2× 2 table with fixed margins. *Statistica Sinica*, pages 809–829, 1996.

Andri Lareida, Tobias Hoßfeld, and Burkhard Stiller. The bittorrent peer collector problem. In *Integrated Network and Service Management (IM), 2017 IFIP/IEEE Symposium on*, pages 449–455. IEEE, 2017.

Saleh M. Mitwalli. An occupancy problem with group drawings of different sizes. *Mathematica Slovaca*, 52(2):235–242, 2002.

R. T. Smythe. Generalized coupon collection: the superlinear case. *Journal of Applied Probability*, 48(1):189–199, 2011.

Wolfgang Stadje. The collector's problem with group drawings. *Advances in Applied Probability*, 22(4):866–882, 1990.

V. A. Vatutin and V. G. Mikhailov. Limit theorems for the number of empty cells in an equiprobable scheme for group allocation of particles. *Theory of Probability & Its Applications*, 27(4):734–743, 1983.