

A POISSON APPROXIMATION FOR AN OCCUPANCY PROBLEM WITH COLLISIONS

TOSHIO NAKATA,* *Fukuoka University of Education*

Abstract

We study collision probabilities concerning the simple balls-and-bins problem developed by Wendl (2003). In this article we give the factorial moment of the number of collisions. Moreover, we obtain a Poisson approximation for the number of collisions using the Chen–Stein method.

Keywords: Collision probability; occupancy problem; Poisson approximation; Chen–Stein method

2000 Mathematics Subject Classification: Primary 60C05
Secondary 90D60

1. Introduction

There is an enormous body of literature on occupancy problems which are focused on combinatorial arguments and limit probabilities. In regard to these problems the reader should see standard textbooks by Johnson and Kotz [8] and Kolchin *et al.* [10]. Wendl [14] proposed the following natural occupancy problem, keeping effective implementations in mind:

Throw m black balls, denoted by A , and n white balls, denoted by B , into t bins. What is the probability that the number of bins containing both colors, that is, the number of collisions between A and B , is k ?

Wendl referred to some attractive applications, for example, collisions of airborne planes, celestial objects, and transportations. In particular, he concretely calculated numerical probabilities concerning a DNA clone mapping problem (see [14] and [15]). On the other hand, Wendl's occupancy problem is regarded as a random coloring problem on a complete bipartite graph constructed by vertex sets A and B . Namely, consider that all the vertices of A and B are independently colored from a uniform distribution on $\{1, \dots, t\}$. Then we would like to know the probability that the number of maximal subtrees constructed by the same color vertices is k . Krivelevich and Nachmias [11] studied a variant of the problem, that is, the asymptotic probability of the existence of a proper coloring for a complete bipartite graph with the same number of vertices on each side.

Let us return to Wendl's occupancy problem. We denote by $X = X(m, n, t)$ the number of collisions between A and B . Wendl [14] gave an exact expression for the no collisions probability, $P(X = 0)$, formulated via graph theory. Moreover, Nakata [12] gave not only $P(X = 0)$ but also $P(X = k)$ for $0 \leq k \leq \min\{m, n, t\}$ using the method of *enumerating surjections* with *Stirling numbers of the second kind* (see [5, Section II.3.1]). On the other

Received 21 November 2007; revision received 1 May 2008.

* Postal address: Department of Mathematics, Fukuoka University of Education, Akama-Bunkuyomachi, Munakata, Fukuoka, 811-4192, Japan. Email address: nakata@fukuoka-edu.ac.jp

hand, Wendl [15] studied another extension of $P(X = 0)$ using the *match number* between A and B denoted by $\sum_{l=1}^t \min\{\alpha_l, \beta_l\}$, where α_l and β_l are the number of respective balls of A and B that exist in the l th bin. In this article we develop some combinatorial arguments to give the explicit form of a (descending) factorial moment of X , whose definition is given in, for example, [8, p. 51].

Using the factorial moment, we study asymptotic probabilities of the number of collisions. We are especially interested in the *Poisson law of small numbers* for X . Intuitively, if $E(X)$ is small, X should be approximately Poisson distributed with parameter $E(X)$. Note that X is *not* a sum of independent and identically distributed random variables. Therefore, it is difficult to effectively use a standard technique of characteristic functions (Fourier transformations). So in this article the Poisson convergence is shown by the *moment method* (see [10, Section I.1]). Indeed, we estimate the limit of the factorial moment of X . The traditional approach to the Poisson paradigm is called *Brun’s sieve* (see [1, Theorem 8.3.1] and [13, Proposition 10.1.1]). Moreover, we obtain an error bound for the Poisson approximation using the *Chen–Stein method*, which is applicable to a sum of weak dependent random variables (see [2], [3], [6], and [13]). It gives a substantial bound between the actual distribution and the target Poisson distribution. Determining the number of empty bins remaining when balls are thrown randomly into some bins is a typical example of a Poisson approximation; see [6, Example 4.12 (14)] and [13, Example 10.2 (B)]. The use of the Chen–Stein method in this article is based on these examples. Moreover, we give a numerical example of the error bounds.

The plan of the article is as follows. In Section 2 we state the factorial moment of X (Theorem 2). Moreover, in Subsection 3.1 we show the Poisson law of small numbers (Theorem 3). Finally, in Subsection 3.2 we use the Chen–Stein method to obtain an error bound for the Poisson approximation (Theorem 4).

2. Combinatorial results

We begin by introducing the following combinatorial notation. First, let $\left\{ \begin{matrix} n \\ k \end{matrix} \right\}$ denote Stirling numbers of the second kind, which is inductively defined as follows: let $\left\{ \begin{matrix} 0 \\ 0 \end{matrix} \right\} = 1$,

$$\left\{ \begin{matrix} n \\ 0 \end{matrix} \right\} = \left\{ \begin{matrix} 0 \\ k \end{matrix} \right\} = 0$$

for $n, k \neq 0$, and

$$\left\{ \begin{matrix} n \\ k \end{matrix} \right\} = \left\{ \begin{matrix} n - 1 \\ k - 1 \end{matrix} \right\} + k \left\{ \begin{matrix} n - 1 \\ k \end{matrix} \right\} \quad \text{for } n, k \in \mathbb{Z}.$$

Various notation for Stirling numbers of the second kind exists. We use the notation $\left\{ \begin{matrix} n \\ k \end{matrix} \right\}$ proposed in [9], since its usage is now widespread. Second, let $(t)_k$ denote

$$\binom{t}{k} k! = t(t - 1) \cdots (t - k + 1)$$

for nonnegative integers $t, k \geq 0$. Note that $(t)_0 = 1$ for $t \geq 0$ and $(t)_k = 0$ for $t < k$.

Recall that the integer-valued random variable X denotes the number of collisions between A and B . Wendl [14] and Nakata [12] investigated the k collisions probability $P(X = k)$ for any $0 \leq k \leq \min\{m, n, t\}$.

Theorem 1. ([12], [14].) *Throw m black balls and n white balls into t bins. Then the probability of k collisions, $P(X = k)$, is, for $0 \leq k \leq \min\{m, n, t\}$,*

$$P(X = k) = \frac{1}{t^{m+n}} \sum_{i=k}^m \sum_{j=k}^n \binom{m}{i} \binom{n}{j} \frac{(i)_k (j)_k (t)_{i+j-k}}{k!}.$$

Note that $P(X = k) = 0$ if $k > \min\{m, n, t\}$ is trivial. Now we show the factorial moment of X .

Theorem 2. *The factorial moment of X is the following: for $1 \leq l \leq t$,*

$$E((X)_l) = (t)_l \left(\sum_{i=0}^l \binom{l}{i} (-1)^i \left(1 - \frac{i}{t}\right)^m \right) \left(\sum_{j=0}^l \binom{l}{j} (-1)^j \left(1 - \frac{j}{t}\right)^n \right). \tag{1}$$

In particular, the expectation is

$$E(X) = t \left(1 - \left(1 - \frac{1}{t}\right)^m\right) \left(1 - \left(1 - \frac{1}{t}\right)^n\right). \tag{2}$$

We use two combinatorial lemmas in Subsection 2.1 to obtain the factorial moment, (1). After proving (1), we point out that X is represented by a sum of dependent random variables to recheck (2).

2.1. Proof of (1)

We prepare two combinatorial lemmas to calculate the factorial moment. It is known that

$$\sum_{j \geq 0} \binom{n}{j} (t)_j = t^n \quad \text{for } n \geq 1 \text{ and } t > 0 \tag{3}$$

(see [9, Section 2.6]). The following lemma generalizes (3) and [12, Lemma 1].

Lemma 1. *For any $l = 1, \dots, \min\{n, t\}$, we have*

$$\sum_{j \geq 0} \binom{n}{j} (t)_j (j)_l = (t)_l \sum_{j \geq 0} \binom{l}{j} (-1)^j (t - j)^n. \tag{4}$$

Proof. It is well known that

$$\binom{t}{k} \binom{s}{k}^{-1} = \binom{s-k}{t-k} \binom{s}{t}^{-1} \tag{5}$$

(see [9, Equation (20)]). Using (5), we deduce that (4) is equivalent to

$$\sum_{j \geq 0} \binom{n}{j} \binom{t-l}{j-l} j! = \sum_{j=0}^l \binom{l}{j} (-1)^j (t-j)^n. \tag{6}$$

Therefore, we check (6) by induction with respect to l . If $l = 0$, both sides of (6) are equal to t^n by (3). Now assume that (6) holds for l . Then the right-hand side of (6), for $l + 1$, is

$$\begin{aligned} & \sum_{j=0}^{l+1} \binom{l+1}{j} (-1)^j (t-j)^n \\ &= \sum_{j=1}^{l+1} \binom{l}{j-1} (-1)^j (t-j)^n + \sum_{j=0}^l \binom{l}{j} (-1)^j (t-j)^n \\ &= - \sum_{i=0}^l \binom{l}{i} (-1)^i (t-1-i)^n + \sum_{j=0}^l \binom{l}{j} (-1)^j (t-j)^n \\ &= - \sum_{j \geq 0} \begin{Bmatrix} n \\ j \end{Bmatrix} \binom{t-1-l}{j-l} j! + \sum_{j \geq 0} \begin{Bmatrix} n \\ j \end{Bmatrix} \binom{t-l}{j-l} j! \\ &= \sum_{j \geq 0} \begin{Bmatrix} n \\ j \end{Bmatrix} \left(\binom{t-l}{j-l} - \binom{t-l-1}{j-l} \right) j!, \end{aligned}$$

which is equal to the left-hand side of (6).

Note that if $l = t$ for (4) then we have

$$\sum_{j \geq 0} \begin{Bmatrix} n \\ j \end{Bmatrix} (t)_j (j)_t = t! \begin{Bmatrix} n \\ t \end{Bmatrix}.$$

We now state the second lemma, which generalizes [12, Lemma 2].

Lemma 2. For any $0 \leq l \leq \min\{i, j\} \leq \max\{i, j\} \leq t$,

$$\sum_{k=0}^{\min\{i,j\}} \binom{i}{k} \binom{j}{k} \binom{t}{i+j-k} \binom{k}{l} / \binom{i+j}{k} = \binom{i}{l} \binom{j}{l} \binom{t}{i} \binom{t}{j} / \binom{i+j}{i} \binom{t}{l}.$$

Proof. Using (5), we have

$$\binom{i}{k} \binom{i+j}{k}^{-1} = \binom{i+j-k}{i-k} \binom{i+j}{i}^{-1} = \binom{i+j-k}{j} \binom{i+j}{i}^{-1}$$

and $\binom{i}{j} \binom{t}{i}^{-1} = \binom{t-l}{i-l} \binom{t}{i}^{-1}$. Therefore, Lemma 2 is equivalent to

$$\sum_{k=0}^{\min\{i,j\}} \binom{i+j-k}{j} \binom{j}{k} \binom{t}{i+j-k} \binom{k}{l} = \binom{t-l}{i-l} \binom{j}{l} \binom{t}{j}.$$

Multiplying both sides of the above equation by $(\binom{j}{l} \binom{t}{j})^{-1}$ and applying (5) again, we obtain

$$\sum_{k=0}^{\min\{i,j\}} \binom{t-j}{i-k} \binom{j-l}{k-l} = \binom{t-l}{i-l},$$

that is, the Vandermonde formula (see [9, Equation (21)]).

Let us return to the proof of (1). By Lemma 2, it is easy to see that

$$\sum_k (k)_l \frac{(i)_k (j)_k (t)_{i+j-k}}{k!} = \sum_{k \geq l} \frac{(i)_k (j)_k (t)_{i+j-k}}{(k-l)!} = \frac{(i)_l (j)_l (t)_i (t)_j}{(t)_l}. \tag{7}$$

Under these preliminaries, we show the factorial moment of X :

$$\begin{aligned} E((X)_l) &= \sum_k (k)_l P(X = k) \\ &= \frac{1}{t^{m+n}} \sum_{k \geq l} \sum_{i=k}^m \sum_{j=k}^n \left\{ \begin{matrix} m \\ i \end{matrix} \right\} \left\{ \begin{matrix} n \\ j \end{matrix} \right\} \frac{(i)_k (j)_k (t)_{i+j-k}}{(k-l)!} \\ &= \frac{1}{t^{m+n}} \sum_{i=0}^m \sum_{j=0}^n \left\{ \begin{matrix} m \\ i \end{matrix} \right\} \left\{ \begin{matrix} n \\ j \end{matrix} \right\} \sum_{k \geq l} \frac{(i)_k (j)_k (t)_{i+j-k}}{(k-l)!} \\ &= \frac{1}{t^{m+n}} \sum_{i=0}^m \sum_{j=0}^n \left\{ \begin{matrix} m \\ i \end{matrix} \right\} \left\{ \begin{matrix} n \\ j \end{matrix} \right\} \frac{(i)_l (j)_l (t)_i (t)_j}{(t)_l} \\ &= (t)_l \sum_{i=0}^l \sum_{j=0}^l \binom{l}{i} \binom{l}{j} (-1)^{i+j} \left(1 - \frac{i}{t}\right)^m \left(1 - \frac{j}{t}\right)^n. \end{aligned}$$

The fourth equality holds by (7). The last equality holds by Lemma 1. This completes the proof of (1).

2.2. Proof of (2)

Now we give another approach to show (2). The number of collisions X may be written as $X = \sum_{i=1}^t \xi_i$, where

$$\xi_i = \begin{cases} 1 & \text{if the collision occurs in the } i\text{th bin,} \\ 0 & \text{if the collision does not occur in the } i\text{th bin.} \end{cases} \tag{8}$$

Note that the $\{\xi_i\}$ are *not* independent. However, since the actions of throwing a ball are independent,

$$P(\xi_i = 1) = P(A(i) \cap B(i)) = P(A(i)) P(B(i)) = \left(1 - \left(1 - \frac{1}{t}\right)^m\right) \left(1 - \left(1 - \frac{1}{t}\right)^n\right),$$

where $A(i)$ and $B(i)$ denote the events that there exist some balls from A and B in the i th bin, respectively. Hence, we have

$$E(X) = \sum_{i=1}^t E(\xi_i) = \sum_{i=1}^t P(\xi_i = 1) = t \left(1 - \left(1 - \frac{1}{t}\right)^m\right) \left(1 - \left(1 - \frac{1}{t}\right)^n\right).$$

3. Asymptotic results

First we directly show the asymptotic properties of $E(X(m, n, t))$ for sufficiently large m , n , and t . For convenience, we use the following notation.

- $a_n = \omega(b_n)$ is defined by $\lim_{n \rightarrow \infty} a_n/b_n = \infty$.

- $a_n = O(b_n)$ is defined by $\sup_n a_n/b_n < \infty$.
- $a_n = o(b_n)$ is defined by $\lim_{n \rightarrow \infty} a_n/b_n = 0$.
- $a_n \sim b_n$ is defined by $\lim_{n \rightarrow \infty} a_n/b_n = 1$.

By (2), the following claims hold.

1. If $\min\{m, n\} = \omega(t)$ then $E(X(m, n, t))/t \sim 1$.
2. If $\min\{m, n\} = o(t)$ then $E(X(m, n, t))/t = o(1)$.
3. In particular, if $m \sim \alpha t$ and $n \sim \beta t$ for $\alpha, \beta > 0$ then

$$\frac{E(X(m, n, t))}{t} \sim (1 - e^{-\alpha})(1 - e^{-\beta}). \tag{9}$$

For a more detailed discussion of (9), we state the weak law of large numbers for X/t under some strong conditions in the following proposition.

Proposition 1. *If $m \sim \alpha t$ and $n \sim \beta t$ for $\alpha, \beta > 0$ then*

$$\lim_{t \rightarrow \infty} \frac{X}{t} = (1 - e^{-\alpha})(1 - e^{-\beta}) \text{ in probability.}$$

Proof. The proof is based on a trivial estimate of $\text{var}(X)$ derived from $E((X)_2)$. Since $\text{var}(X) = E(X^2) - (E(X))^2$, we obtain

$$\begin{aligned} \text{var}(X) &= t^2((2a_1 - 1)(b_1^2 - b_2) + (2b_1 - 1)(a_1^2 - a_2) - (a_1^2 b_1^2 - a_2 b_2)) \\ &\quad - t((2a_1 - 1)(b_1 - b_2) + (2b_1 - 1)(a_1 - a_2) - (a_1 b_1 - a_2 b_2)), \end{aligned}$$

where

$$a_1 = \left(1 - \frac{1}{t}\right)^m, \quad a_2 = \left(1 - \frac{2}{t}\right)^m, \quad b_1 = \left(1 - \frac{1}{t}\right)^n, \quad b_2 = \left(1 - \frac{2}{t}\right)^n.$$

Assume that $m \sim \alpha t$ and $n \sim \beta t$. Then we have

$$\lim_{t \rightarrow \infty} a_1 = e^{-\alpha}, \quad \lim_{t \rightarrow \infty} a_2 = e^{-2\alpha}, \quad \lim_{t \rightarrow \infty} b_1 = e^{-\beta}, \quad \lim_{t \rightarrow \infty} b_2 = e^{-2\beta}.$$

Therefore, $\lim_{t \rightarrow \infty} \text{var}(X)/t^2 = 0$. Since $\text{var}(X/t) = \text{var}(X)/t^2 \rightarrow 0$ as $t \rightarrow \infty$, we see that

$$\lim_{t \rightarrow \infty} \frac{X}{t} = (1 - e^{-\alpha})(1 - e^{-\beta}) \text{ in probability,}$$

by [4, Theorem 5.4].

Under some strong conditions, we give a Poisson convergence of X in Subsection 3.1 and the error bounds concerning the convergence in Subsection 3.2.

3.1. The Poisson law of small numbers

Before stating Poisson limit theorems, we introduce some notation. Let $\mathcal{L}(X)$ denote the distribution of the random variable X . We define the *total variation distance* between two distributions as

$$d_{TV}(\mathcal{L}(X), \mathcal{L}(Y)) = \frac{1}{2} \sum_{k \geq 0} |\mathbb{P}(X = k) - \mathbb{P}(Y = k)|.$$

Moreover, we define $\lim_{t \rightarrow \infty} \mathcal{L}(X_t) \stackrel{D}{=} \mathcal{L}(Y)$ if $\lim_{t \rightarrow \infty} \mathbb{E}(f(X_t)) = \mathbb{E}(f(Y))$ for all bounded continuous functions f , where ‘ $\stackrel{D}{=}$ ’ denotes equality in distribution. Let $\text{Poi}(\gamma)$ denote a Poisson distribution with parameter $\gamma > 0$, that is, Y is distributed with $\text{Poi}(\gamma)$ if

$$\mathbb{P}(Y = k) = e^{-\gamma} \frac{\gamma^k}{k!} \quad \text{for } k = 0, 1, \dots$$

Theorem 3. *If $\max\{m, n\} = o(t)$ and $mn \sim \gamma t$ for $\gamma > 0$ then the following claims hold:*

$$\lim_{t \rightarrow \infty} \mathbb{E}(X(m, n, t)) = \gamma, \tag{10}$$

$$\lim_{t \rightarrow \infty} \mathcal{L}(X(m, n, t)) \stackrel{D}{=} \text{Poi}(\gamma). \tag{11}$$

Proof. First we show that (10) holds. Since $\max\{m, n\} = o(t)$, we have, for $i, j = 0, \dots, l$,

$$\left(1 - \frac{i}{t}\right)^m = 1 - \frac{im}{t} + O\left(\frac{m^2}{t^2}\right) \sim e^{-mi/t}, \quad \left(1 - \frac{j}{t}\right)^n = 1 - \frac{jn}{t} + O\left(\frac{n^2}{t^2}\right) \sim e^{-nj/t}.$$

Therefore, the following holds:

$$\begin{aligned} \mathbb{E}(X) &= t \left(1 - \left(1 - \frac{1}{t}\right)^m\right) \left(1 - \left(1 - \frac{1}{t}\right)^n\right) \\ &= t \left(\frac{m}{t} - O\left(\frac{m^2}{t^2}\right)\right) \left(\frac{n}{t} - O\left(\frac{n^2}{t^2}\right)\right) \\ &= \frac{mn}{t} - O\left(\frac{m+n}{t^2}\right) \\ &\rightarrow \gamma \quad \text{as } t \rightarrow \infty. \end{aligned}$$

Next we show that (11) holds. The proof is essentially based on the convergence of the moments (see [10, Theorem I.1.2]). Since each ordinary moment can be represented by a linear combination of factorial moments, we may check the convergence of the factorial moments instead of the ordinary moments. Indeed, we have

$$\mathbb{E}(X^l) = \sum_{j=0}^l \begin{Bmatrix} l \\ j \end{Bmatrix} \mathbb{E}((X)_j), \tag{12}$$

because of applying (3) to the random variable X .

It is known that the l th factorial moment of the Poisson random variable is

$$\mathbb{E}((Y)_l) = \gamma^l, \tag{13}$$

where Y is distributed with $\text{Poi}(\gamma)$ (see [8, Equation (2.13)]). Hence, we show that the l th factorial moment of X converges to γ^l . Since l is fixed, we have

$$\begin{aligned} E((X)_l) &= (t)_l \left(\sum_{i=0}^l \binom{l}{i} (-1)^i \left(1 - \frac{i}{t}\right)^m \right) \left(\sum_{j=0}^l \binom{l}{j} (-1)^j \left(1 - \frac{j}{t}\right)^n \right) \\ &\sim t^l \left(\sum_{i=0}^l \binom{l}{i} (-1)^i e^{-mi/t} \right) \left(\sum_{j=0}^l \binom{l}{j} (-1)^j e^{-nj/t} \right) \\ &= (t(1 - e^{-m/t})(1 - e^{-n/t}))^l \\ &= \left(t \left(\frac{m}{t} - O\left(\frac{m^2}{t^2}\right) \right) \left(\frac{n}{t} - O\left(\frac{n^2}{t^2}\right) \right) \right)^l \\ &\rightarrow \gamma^l \quad \text{as } t \rightarrow \infty. \end{aligned}$$

Additionally, according to [4, Theorem 3.12] we have to check the following sufficient condition to show the convergence in distribution:

$$\limsup_{l \rightarrow \infty} \frac{E(Y^{2l})^{1/2l}}{2l} < \infty. \tag{14}$$

By (3), (12), and (13), the $2l$ th ordinary moment of Y is

$$E(Y^{2l}) = \sum_{j=0}^{2l} \left\{ \begin{matrix} 2l \\ j \end{matrix} \right\} E((Y)_j) = \sum_{j=0}^{2l} \left\{ \begin{matrix} 2l \\ j \end{matrix} \right\} \gamma^j \sim \gamma^{2l}.$$

Therefore, the limit of (14) is 0. Hence, the theorem is proved.

3.2. An error bound given by the Chen–Stein method

In Subsection 3.1 we only examined the Poisson convergence of X without considering any error bounds. In this subsection we obtain an error bound for the Poisson approximation using the Chen–Stein method. Before stating the result, we introduce some notation. Let λ be the expectation of X , that is,

$$\lambda = E(X) = E\left(\sum_{i=1}^t \xi_i\right) = t \left(1 - \left(1 - \frac{1}{t}\right)^m\right) \left(1 - \left(1 - \frac{1}{t}\right)^n\right). \tag{15}$$

Moreover, define

$$\Delta = \frac{1}{2} \sum_{i \neq j} E(\xi_i \xi_j), \tag{16}$$

where the sum is over unordered pairs for $1 \leq i \neq j \leq t$. Note that the $\{\xi_i\}$ are not independent and that Δ is a tool for measuring the ‘dependence’ of the $\{\xi_i\}$ (see [1, p. 116] and [7, p. 31]). Then we have

$$E((X)_2) = 2\Delta. \tag{17}$$

In fact,

$$E(X^2) = E\left(\left(\sum_{i=1}^t \xi_i\right)\left(\sum_{j=1}^t \xi_j\right)\right) = \sum_{i=1}^t E(\xi_i^2) + \sum_{i \neq j} E(\xi_i \xi_j) = \lambda + 2\Delta,$$

TABLE 1: Expectations and error bounds for $m = n = 10$ evaluated using (18).

t	λ	Error bound
10	4.242 19	0.675 58
50	1.673 11	0.294 16
100	0.914 27	0.152 47
200	0.478 04	0.042 523
500	0.196 43	0.007 269 7
1000	0.099 10	0.001 858 2

because of $\xi_i^2 = \xi_i$. Accordingly, by (1) we have

$$\Delta = \binom{t}{2} \left(1 - 2 \left(1 - \frac{1}{t} \right)^m + \left(1 - \frac{2}{t} \right)^m \right) \left(1 - 2 \left(1 - \frac{1}{t} \right)^n + \left(1 - \frac{2}{t} \right)^n \right).$$

Using λ and Δ , we state the following theorem.

Theorem 4. *For λ and Δ defined by (15) and (16), respectively, we have*

$$d_{\text{TV}}(\mathcal{L}(X(m, n, t)), \text{Poi}(\lambda)) \leq \min \left\{ 1, \frac{1}{\lambda} \right\} (\lambda^2 - 2\Delta). \tag{18}$$

If t is large enough, the Poisson approximation is better fit; see, for example, Table 1, in which we present the expectations and the error bounds for $10 \leq t \leq 1000$, that is, the right-hand side of (18) for $m = n = 10$.

Recall that the dependent Bernoulli random variables $\{\xi_i\}_{i=1}^t$ with respective parameters $p_i = (1 - (1 - 1/t)^m)(1 - (1 - 1/t)^n)$ are defined in (8). We now state Chen–Stein’s theorem [6, Section 4.12], [13, Theorem 10.2.3], [3, Theorem 6.B] for the Poisson approximation of $X = \sum_{i=1}^t \xi_i$.

Theorem 5. (Chen–Stein theorem.) *If there exist random variables $\{V_i\}$ coupling with X satisfying*

$$P(V_i = k - 1) = P(X = k \mid \xi_i = 1) \quad \text{for } i = 1, \dots, t, \tag{19}$$

then

$$d_{\text{TV}}(\mathcal{L}(X), \text{Poi}(\lambda)) \leq \min \left\{ 1, \frac{1}{\lambda} \right\} \sum_{i=1}^t p_i E|X - V_i|,$$

where $\lambda = E(X) = \sum_{i=1}^t p_i$. Moreover, if $X \geq V_i$ for every i then

$$d_{\text{TV}}(\mathcal{L}(X), \text{Poi}(\lambda)) \leq \min \left\{ 1, \frac{1}{\lambda} \right\} (\lambda - \text{var}(X)). \tag{20}$$

Proof of Theorem 4. Following Theorem 5, we construct a sequence V_1, \dots, V_t of random variables satisfying (19). If the collision occurs in the i th bin, we set $V_i = X - 1$. Suppose that the collision does not occur in the i th bin. Then we determine V_i referring to [3, Theorem 6.B]. Let η_j^A and η_j^B denote the number of balls from A and B , respectively, in the j th bin for $j = 1, \dots, t$. Since the collision does not occur in the i th bin, we have $\eta_i^A \eta_i^B = 0$. Let $\tilde{\eta}_i^A$ denote a random variable having distribution $\mathcal{L}(\eta_i^A \mid \eta_i^A \geq 1)$ and being independent of

$\{\eta_j^A\}_{j=1}^t$. Note that $\tilde{\eta}_i^A \geq 1$ with probability 1. We define $\tilde{\eta}_i^B$ similarly, only $\tilde{\eta}_i^B$ is independent of $\{\eta_j^A\}_{j=1}^t$, $\{\eta_j^B\}_{j=1}^t$, and $\{\tilde{\eta}_j^A\}_{j=1}^t$. If $\eta_i^A = 0$ then draw $\tilde{\eta}_i^A$ balls from A by simple random sampling without replacement from the bins different from the i th bin, and put these balls into the i th bin. Similarly, if $\eta_i^B = 0$ then draw $\tilde{\eta}_i^B$ balls from B by simple random sampling without replacement from the bins different from the i th bin, and put these balls into the i th bin. After these procedures, let V_i define the number of these $t - 1$ bins in which collisions occur. Then (19) holds and, furthermore, $V_i \leq X$. Thus, we can use (20). By (17) we have

$$\lambda - \text{var}(X) = E(X) - (E(X^2) - E(X)^2) = \lambda^2 - E((X)_2) = \lambda^2 - 2\Delta.$$

Hence, inserting it into (20), we have (18).

Acknowledgement

The author would like to thank an anonymous referee for several improvements in the presentation of this paper.

References

- [1] ALON, N. AND SPENCER, J. (2000). *The Probabilistic Method*, 2nd edn. John Wiley, New York.
- [2] ARRATIA, R., GOLDSTEIN, L. AND GORDON, L. (1990). Poisson approximation and the Chen–Stein method. *Statist. Sci.* **5**, 403–434.
- [3] BARBOUR, A. D., HOLST, L. AND JANSON, S. (1992). *Poisson Approximation*. Oxford University Press.
- [4] DURRETT R. (1996). *Probability: Theory and Examples*, 2nd edn. Duxbury Press, Belmont, CA.
- [5] FLAJOLET, P. AND SEDGEWICK, R. (2008). *Analytic Combinatorics*. Cambridge University Press.
- [6] GRIMMETT, G. AND STIRZAKER, D. (2001). *Probability and Random Processes*, 3rd edn. Oxford University Press.
- [7] JANSON, S., ŁUCZAK, T. AND RUCIŃSKI, A. (2000). *Random Graphs*. John Wiley, New York.
- [8] JOHNSON, N. L. AND KOTZ, S. (1977). *Urn Models and Their Application*. John Wiley, New York.
- [9] KNUTH, D. (1973). *The Art of Computer Programming, Fundamental Algorithms*, Vol. 1, 3rd edn. Addison-Wesley, Reading, MA.
- [10] KOLCHIN, F., SEVASTYANOV, A. AND CHISTYAKOV, P. (1978). *Random Allocations*. John Wiley, New York.
- [11] KRIVELEVICH, M., AND NACHMIAS, A. (2006). Coloring complete bipartite graphs from random lists. *Random Structures Algorithms* **29**, 436–449.
- [12] NAKATA, T. (2008). Collision probability for an occupancy problem. To appear in *Statist. Prob. Lett.*
- [13] ROSS, S. (1996). *Stochastic Processes*, 2nd edn. John Wiley, New York.
- [14] WENDL, M. (2003). Collision probability between sets of random variables. *Statist. Prob. Lett.* **64**, 249–254.
- [15] WENDL, M. (2005). Probabilistic assessment of clone overlaps in DNA fingerprint mapping via a priori models. *J. Comput. Biol.* **12**, 283–297.