

On the number of overflowed urns and excess balls in an allocation model with limited urn capacity

M.V. Boutsikas¹ and M.V. Koutras

Department of Mathematics, University of Athens, Greece

Abstract: In the present paper we consider an equiprobable allocation urn model with m urns of limited capacity and study the distribution of the number of overflowed urns and the number of excess balls (i.e those assigned to already filled up urns) after n balls have been assigned. The exact distribution of the variables of interest are studied by a Markov chain imbedding method; a Poisson and compound Poisson approximation is discussed as well along with estimates of their rates of convergence. Certain limit laws as $n, m \rightarrow \infty$ are also established under appropriate conditions on n, m and urns' capacities.

AMS 1991 Subject classifications. Primary 60F05, 60C05; secondary 62E17, 60J10.

Key words and phrases: Limit Theorems, Negative association, Poisson approximation, Compound Poisson approximation, Rate of convergence, Sequential occupancy, Urn models, Generalized birthday problems, Markov chain imbedding, Kolmogorov distance.

1. Introduction

Many combinatorial problems in probability and statistics can be formulated and best understood by using appropriate urn models. Since the publication of the classical book by Johnson and Kotz (1977) on urn models, the theory and applications of the so called ball-in-urn distribution problems and the associated drawing schemes have received substantial attention from probabilists, statisticians and applied scientists. For an up to date review on the recent developments in this area the reader might wish to consult Kotz and Balakrishnan (1997).

In the present article we consider an urn model with cells of limited capacity and proceed to the investigation of two random variables (r.v.'s) which are closely related to a generalization of the classical birthday problem. More specifically, assume that n indistinguishable balls are distributed into m cells (urns) of limited capacity $k - 1$ ($k \geq 2$). Each ball is equally likely to be assigned to any of the m cells, but if the cell is full, the ball is placed in a spare (overflow) urn with unlimited capacity. The random variables of interest here, are

- a. The number U of urns that have «contributed» at least one ball to the spare urn, i.e. the number of different urns to whom at least k balls (out of n) were assigned.
- b. The total number W of balls placed in the spare urn (number of excess balls).

In the birthday problem (see e.g. Feller(1968) or Johnson and Kotz (1977)) one assumes that each person is equally likely to have any of the 365 days in the year as his or her birthday and asks for the probability that among n persons chosen at random, at least k have the same birthday. Although the literature related to the classical birthday problem ($k = 2$) is large (see for example Johnson and Kotz (1977) and references therein), there are relatively few papers dealing with the general case $k > 2$ (see McKinney (1966), Holst (1995), Menon and Indira (1990), Henze (1998) or Barbour, Holst and Janson (1992), Kotz and Balakrishnan

¹ Research supported by the National Scholarship Foundation of Greece.

(1997) and references therein). It is evident that the probability asked for in the birthday problem can be expressed as $P(U \neq 0)$ or $P(W \neq 0)$ by identifying days with cells ($m = 365$) and persons with balls in the urn model described earlier.

Another interesting application comes from the area of inventory control. Assume that in a retail store, m compartments (or shells) are used with a limited capacity of $k-1$ items for each compartment. If n customers select at random a compartment and leave the store whenever the chosen compartment is found empty, then U enumerates the compartments that caused at least one customer loss and W the number of lost customers. Similar problems could also be stated for multilevel parking places with limited capacity in each level, railway terminals being able to accommodate a restricted number of trains etc. Finally, the random variables U and W may be used to test the null hypothesis that a series of multistate outcomes follows a uniform distribution against several clustering alternatives.

In this article we present both exact and approximate formulae for the evaluation of the distribution of the random variables U and W . In Section 2 we introduce the necessary notations and present some results that facilitate the subsequent analysis. In Section 3 we illustrate how the exact distributions of U and W can be derived through a Markov chain imbedding technique. Finally, in Section 4 we establish some Poisson and compound Poisson approximations along with error estimates and exploit them to set up limit theorems for the distributions of U and W respectively.

2. Preliminaries.

Let $S_j, j = 1, 2, \dots, m$ denote the total number of balls assigned to urns $1, 2, \dots, m$. Manifestly, the number U of «*overflowed*» urns (urns to whom at least k balls were assigned) can be expressed as

$$U = \sum_{j=1}^m I_{[k, \infty)}(S_j) \quad (2.1)$$

where $I_A(\cdot)$ stands for the usual indicator function of set A . On the other hand, the number of balls placed in the spare urn is given by

$$W = \sum_{j=1}^m (S_j - k + 1) I_{[k, \infty)}(S_j). \quad (2.2)$$

Recently, Fu and Koutras (1994) and Fu (1996) developed a unified Markov chain imbedding technique for computing the exact distribution of enumerating random variables in sequences of binary or multistate trials. Koutras and Alexandrou (1995) suggested a refinement of this method which used multidimensional binomial type probability vectors and exploited it in the study of urn models associated to several drawing schemes. Their approach gave rise to exact compact formulae for the probability generating functions of the variables under investigation, in terms of certain matrices which characterize the enumeration scheme in use. Since we are going to employ this approach for the investigation of the exact distributions of U and W we deem it necessary to present here a brief outline of the Markov chain imbedding technique; for more details refer to Fu and Koutras (1994), Koutras and Alexandrou (1995) and Fu (1996).

Let V_n (n a non-negative integer) be an integer valued random variable (r.v.) and denote by $l_n = \max\{v: P(V_n = v) > 0\} < \infty$ its upper end point. V_n will be called *Markov chain imbeddable Variable of Binomial type* (MVB) if

1. there exists a Markov chain $\{Z_t, t \geq 0\}$ defined on a discrete state space Ω which can be partitioned as

$$\Omega = \bigcup_{v \geq 0} C_v, \quad C_v = \{c_{v,0}, c_{v,1}, \dots, c_{v,s-1}\}.$$

2. $P(Z_t \in C_w | Z_{t-1} \in C_v) = 0$ for all $w \neq v, v+1$ and $t \geq 1$.

3 The event $V_n = v$ is equivalent to $Z_n \in C_v$ and therefore the probability mass function of V_n is given by

$$P(V_n = v) = P(Z_n \in C_v), \quad v=1,2,\dots,l_n.$$

The distribution of a MVB is completely determined by the following three quantities:

- the initial probabilities

$$\boldsymbol{\pi}_v = (P(Z_0 = c_{v,0}), P(Z_0 = c_{v,1}), \dots, P(Z_0 = c_{v,s-1})), \quad v=1,2,\dots,l_n$$

- the *within states* one step transition matrix

$$\mathbf{A}_t(v) = (P(Z_t = c_{v,j} | Z_{t-1} = c_{v,i}))_{s \times s}, \quad v = 1, 2, \dots, l_n, \quad t \geq 1$$

- the *between states* one step transition matrix

$$\mathbf{B}_t(v) = (P(Z_t = c_{v+1,j} | Z_{t-1} = c_{v,i}))_{s \times s}, \quad v = 1, 2, \dots, l_n, \quad t \geq 1.$$

More specifically, as Koutras and Alexandrou (1995) indicated, if

$$\mathbf{f}_t(v) = (P(Z_t = c_{v,0}), P(Z_t = c_{v,1}), \dots, P(Z_t = c_{v,s-1}))$$

then the next recurrences hold true for all $1 \leq t \leq n$,

$$\mathbf{f}_t(0) = \mathbf{f}_{t-1}(0) \mathbf{A}_t(0)$$

$$\mathbf{f}_t(v) = \mathbf{f}_{t-1}(v) \mathbf{A}_t(v) + \mathbf{f}_{t-1}(v-1) \mathbf{B}_t(v-1), \quad 1 \leq v \leq l_n.$$

These relations, used in conjunction with the initial conditions $\mathbf{f}_0(v) = \boldsymbol{\pi}_v$, $0 \leq v \leq l_n$, offer a very simple computational scheme for the evaluation of the probability mass function of V_n through the formula

$$P(V_n = v) = \mathbf{f}_n(v) \cdot \mathbf{1}', \quad v=0,1,\dots,l_n$$

($\mathbf{1} = (1,1,\dots,1)$ is the row vector of $\tilde{\mathbf{N}}^s$ with all its entries being 1).

It is sufficient for our purposes and also of greater simplicity (especially for the statement of more compact formulae) to assume that $\boldsymbol{\pi}_v = \mathbf{0}$, $v \geq 1$ and $\boldsymbol{\pi}_0 \mathbf{1}' = 1$; this convention is in fact equivalent to the condition $P(V_0 = 0) = 1$.

The models where the MVB technique will be applied later, lead to transition matrices independent of v and t , i.e.

$$\mathbf{A}_t(v) = \mathbf{A}, \quad \mathbf{B}_t(v) = \mathbf{B} \quad \text{for all } t \geq 0 \text{ and } 1 \leq v \leq l_n.$$

In this case, the double generating function of V_n

$$\Phi(z, w) = \sum_{n=0}^{\infty} \sum_{v=0}^{l_n} P(V_n = v) z^v w^n$$

takes on the next compact form (c.f. Koutras and Alexandrou (1995))

$$\Phi(z, w) = \pi_0 (\mathbf{I} - w(\mathbf{A} + z\mathbf{B}))^{-1} \mathbf{1}'.$$

As mentioned in the introduction, besides the exact formulae, we shall provide Poisson approximations for the distribution of U and W , along with estimates of the error incurred by these approximations. To achieve this we are going to exploit the next result which has been recently reported by Boutsikas and Koutras (2000). The distance metric appearing there is the Kolmogorov distance, i.e.

$$d(\mathbb{L}(X), \mathbb{L}(Y)) = \sup_w |P(X \leq w) - P(Y \leq w)|.$$

Manifestly, a sequence of random variables X_n , $n=1, 2, \dots$ converges in distribution to Y if $d(\mathbb{L}(X_n), \mathbb{L}(Y))$ converges to 0.

Theorem 1. *Let X_1, X_2, \dots, X_m be associated or NA (negatively associated) integer valued r.v.'s with $E(|X_i|), E(|X_i X_j|) < \infty$, $i, j = 1, 2, \dots, m$, $i \neq j$. If X'_i are independent random variables such that X'_i is distributed according to the marginal distribution of X_i ($\mathbb{L}(X_i) = \mathbb{L}(X'_i)$) then*

$$d(\mathbb{L}(\sum_{i=1}^m X_i), \mathbb{L}(\sum_{i=1}^m X'_i)) \leq \left| \sum_{i < j} \text{Cov}(X_i, X_j) \right|.$$

Moreover, if X_1, X_2, \dots, X_m are NA non-negative r.v.'s, then

$$0 \leq \prod_{i=1}^m P(X_i = 0) - P(X_i = 0, i=1, 2, \dots, m) \leq - \sum_{i < j} \text{Cov}(X_i, X_j).$$

Another useful result pertaining to compound Poisson approximation is given by the following theorem. The notation $CP(\lambda, F)$ used there indicates the distribution of the sum of $Y = \sum_{i=1}^N Y_i$ where N is a Poisson r.v. with mean $E(N) = \lambda$, independent of Y_i and F is the c.d.f. of the i.i.d. r.v.'s Y_1, Y_2, \dots .

Theorem 2. *Let X_1, X_2, \dots, X_m be associated or NA identical integer-valued r.v.'s with finite moments $E(|X_i|), E(|X_i X_j|)$ for $i, j=1, 2, \dots, m$, $i \neq j$. Then*

$$d(\mathbb{L}(\sum_{a=1}^m X_a), CP(\lambda, F)) \leq \left| \sum_{a < \beta} \text{Cov}(X_a, X_\beta) \right| + m[P(X_1 \neq 0)]^2$$

where $\lambda = mP(X_1 \neq 0)$ and $F(x) = P(X_1 \leq x | X_1 \neq 0)$, $x \in \mathbb{N}$.

Proof. Let Y_1, Y_2, \dots be a sequence of independent r.v.'s following the zero-truncated distribution of X_i i.e. $P(Y_a = j) = P(X_1 = j | X_1 \neq 0)$, $j \in \mathbb{U}$ and $Y = \sum_{i=1}^N Y_i$ with N being a Poisson r.v. independent of Y_i and $E(N) = mP(X_1 \neq 0)$. Invoking Theorem 1 we immediately deduce

$$d\left(\mathbb{L}\left(\sum_{a=1}^m X_a\right), \mathbb{L}\left(\sum_{a=1}^m X'_a\right)\right) \leq \left| \sum_{a < \beta} \text{Cov}(X_a, X_\beta) \right|.$$

Observe next that $\sum_{a=1}^m X'_a = \sum_{a=1}^M Y_a$ where $M = \sum_{a=1}^m (1 - I_{\{0\}}(X'_a))$ is the number of the non-zero X'_a 's. Therefore,

$$d\left(\mathbb{L}\left(\sum_{a=1}^m X'_a\right), \mathbb{L}\left(\sum_{a=1}^N Y_a\right)\right) = d\left(\mathbb{L}\left(\sum_{a=1}^M Y_a\right), \mathbb{L}\left(\sum_{a=1}^N Y_a\right)\right) \leq d_{TV}(\mathbb{L}(M), \mathbb{L}(N))$$

$$\leq \sum_{a=1}^m [P(1 - I_{\{0\}}(X'_a)=1)]^2 = m[P(X_1 \neq 0)]^2$$

(d_{TV} denotes the total variation distance between two distributions; for the first inequality see Vellaisamy and Chandhuri (1996); the second is a special case of the well known fact that the total variation distance between a sum of independent binary r.v.'s Z_1, \dots, Z_m with $E(Z_i) = p_i$ and the Poisson distribution with mean $\sum_{i=1}^m p_i$ is upper bounded by $\sum_{i=1}^m p_i^2$, see e.g. Serfling (1978)). The proof is now easily completed by applying the triangle inequality. \square

Additional bounds between a sum of associated or NA r.v.'s and a compound Poisson distribution may be found in Boutsikas and Koutras (2000). For the definition and basic properties of associated and NA r.v.'s the interested reader might wish to consult Esary, Proschan and Walkup (1967) and Joag-Dev and Proschan (1983). Apparently, the lower and upper bounds for $P(\sum_{i=1}^m X_i \leq x)$ conferred from the inequalities of Theorems 1 and 2 may take negative values or values greater than 1 respectively. It goes without saying that in such cases 0, 1 will be used as a lower, upper bound for $P(\sum_{i=1}^m X_i \leq x)$ accordingly.

Before closing this section we state some auxiliary results that will be repeatedly used in the sequel and are of independent interest. The multinomial coefficients appearing below are defined as follows

$$\binom{a}{b_1, b_2, \dots, b_t} = \begin{cases} \frac{a!}{\prod_{i=1}^t b_i! (a - \sum_{i=1}^t b_i)!}, & a \geq 0, b_i \geq 0, \sum_{i=1}^t b_i \leq a \\ 0 & \text{if } a < 0 \text{ or some } b_i < 0 \text{ or } \sum_{i=1}^t b_i > a \end{cases}$$

Lemma 1. *The following identities are valid*

$$\begin{aligned} \text{a. } \sum_{i=k}^n (i-k+1) \binom{n}{i} p^i (1-p)^{n-i} &= \frac{n!}{(k-2)!(n-k)!} \int_0^p (p-x)x^{k-2}(1-x)^{n-k} dx, \quad 2 \leq k \leq n, p \in (0,1), \\ \text{b. } \sum_{i=k}^{n-k} \sum_{j=k}^{n-i} (i-k+1)(j-k+1) \binom{n}{i, j} p^{i+j} (1-2p)^{n-i-j} &= \\ &= \frac{n!}{(k-2)!^2 (n-2k)!} \int_0^p \int_0^p (p-x)(p-y)x^{k-2}y^{k-2}(1-x-y)^{n-2k} dx dy, \quad 2 \leq k \leq n/2, p \in (0,1/2). \end{aligned}$$

Proof. We recall first the following well known identity which expresses the tail probabilities of the multinomial distribution in terms of incomplete Dirichlet integrals (see for example Olkin and Sobel (1965) or Gradshteyn and Ryzhik(1980))

$$\begin{aligned} \sum_{i_1 \geq k_1} \sum_{i_2 \geq k_2} \dots \sum_{i_r \geq k_r} \binom{n}{i_1, i_2, \dots, i_r} p_1^{i_1} p_2^{i_2} \dots p_r^{i_r} (1-p_1-p_2-\dots-p_r)^{n-i_1-i_2-\dots-i_r} &= \\ &= \frac{n!}{\prod_{i=1}^r (k_i-1)!(n-\sum_{i=1}^r k_i)!} \int_0^{p_1} \int_0^{p_2} \dots \int_0^{p_r} \prod_{i=1}^r x_i^{k_i-1} (1-\sum_{i=1}^r x_i)^{n-\sum_{i=1}^r k_i} dx_r \dots dx_2 dx_1 \end{aligned} \quad (2.3)$$

($p_i > 0, \sum p_i < 1, r \geq 1, n - \sum k_i \geq 0$). Note that the above identity can also be considered as the multivariate analogue of the well known formula for order statistics (see e.g. Arnold and Balakrishnan (1989))

$$P(X_{k:n} \leq x) = \sum_{i=k}^n \binom{n}{i} F^i(x) (1-F(x))^{n-i} = \frac{n!}{(k-1)!(n-k)!} \int_0^{F(x)} t^{k-1} (1-t)^{n-k} dt.$$

Applying formula (2.3) for $r = 1$ we easily deduce

$$\sum_{i=k}^n (k-1) \binom{n}{i} p^i (1-p)^{n-i} = \frac{n!}{(k-2)!(n-k)!} \int_0^p x^{k-1} (1-x)^{n-k} dx.$$

Moreover, the incomplete first moment of the binomial distribution, by virtue of the same formula, takes on the form

$$\sum_{i=k}^n i \binom{n}{i} p^i (1-p)^{n-i} = np \sum_{i=k-1}^{n-1} \binom{n-1}{i} p^i (1-p)^{n-i-1} = \frac{n!p}{(k-2)!(n-k)!} \int_0^p x^{k-2} (1-x)^{n-k} dx.$$

This completes the proof of part (a). The proof of part (b) is easily performed by applying (2.3) for $r=2$. More precisely, the LHS of (b) can be decomposed in three parts for which the next alternative expressions emanate

- $$\sum_{i=k}^{n-k} \sum_{j=k}^{n-i} ij \binom{n}{i, j} p^{i+j} (1-2p)^{n-i-j} = n(n-1)p^2 \sum_{i \geq k-1} \sum_{j \geq k-1} \binom{n-2}{i, j} p^{i+j} (1-2p)^{n-2-i-j} =$$

$$= \frac{n!p^2}{(k-2)!^2 (n-2k)!} \int_0^p \int_0^p x^{k-2} y^{k-2} (1-x-y)^{n-2k} dx dy,$$

(joint incomplete second moment of multinomial distribution)
- $$(k-1)^2 \sum_{i=k}^{n-k} \sum_{j=k}^{n-i} \binom{n}{i, j} p^{i+j} (1-2p)^{n-i-j} = \frac{n!}{(k-2)!^2 (n-2k)!} \int_0^p \int_0^p x^{k-1} y^{k-1} (1-x-y)^{n-2k} dx dy,$$
- $$-2(k-1) \sum_{i=k}^{n-k} \sum_{j=k}^{n-i} i \binom{n}{i, j} p^{i+j} (1-2p)^{n-i-j} = -2n(k-1)p \sum_{i \geq k-1} \sum_{j \geq k} \binom{n-1}{i, j} p^{i+j} (1-2p)^{n-1-i-j}$$

$$= -\frac{2n!p}{(k-2)!^2 (n-2k)!} \int_0^p \int_0^p x^{k-2} y^{k-1} (1-x-y)^{n-2k} dx dy.$$

Summing up the RHS of the last three identities we easily arrive at the RHS of part (b). \square

Lemma 2. Let $F(x)$ denote the c.d.f. of the Erlang distribution with probability mass function

$$f(x) = \frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x}, \quad x > 0.$$

Then the next inequality holds true for all $x < k/\lambda$,

$$\frac{x f(x)}{k - \lambda x} \left(1 - \frac{\lambda x}{k(k+1)(1 - \lambda x/k)^2} \right) \leq F(x) \leq \frac{x f(x)}{k - \lambda x}.$$

Proof. It can be readily verified that the above inequality is equivalent to

$$0 \leq 1 - \frac{\int_0^a x^{k-1} e^{-x} dx}{\frac{a}{k-a} a^{k-1} e^{-a}} \leq \frac{a}{k(k+1)(1-a/k)^2} \quad \text{for } 0 < a < k, \quad k=1,2,\dots \quad (2.4)$$

Taking into account the well known formula for the incomplete gamma function (see e.g. Gradshteyn and Ryzhik(1980))

$$\gamma(k, a) = \int_0^a x^{k-1} e^{-x} dx = (k-1)! e^{-a} \sum_{i=k}^{\infty} \frac{a^i}{i!} \quad (2.5)$$

we obtain

$$\frac{\int_0^a x^{k-1} e^{-x} dx}{\frac{a}{k-a} a^{k-1} e^{-a}} = (1-c) \left(1 + \frac{k!}{(ck)^k} \sum_{i=k+1}^{\infty} \frac{(ck)^i}{i!} \right) \quad (2.6)$$

where $c = a/k \in (0,1)$. Observe next that

$$\frac{c}{1-c} - \frac{k!}{(ck)^k} \sum_{j=k+1}^{\infty} \frac{(ck)^j}{j!} = \sum_{i=1}^{\infty} \left(1 - \frac{k!k^i}{(i+k)!} \right) c^i = \sum_{i=1}^{\infty} \left(1 - \frac{k}{k+1} \frac{k}{k+2} \cdots \frac{k}{k+i} \right) c^i \geq 0$$

which, in view of the elementary inequality (see for example Billingsley (1986), page 367)

$$\left| \prod_{j=1}^s a_j - \prod_{j=1}^s b_j \right| \leq \sum_{j=1}^s |a_j - b_j|, \quad |a_j|, |b_j| \leq 1$$

reveals that

$$\begin{aligned} \frac{c}{1-c} - \frac{k!}{(ck)^k} \sum_{j=k+1}^{\infty} \frac{(ck)^j}{j!} &\leq \sum_{i=1}^{\infty} \left(\left(1 - \frac{k}{k+1} \right) + \left(1 - \frac{k}{k+2} \right) + \dots + \left(1 - \frac{k}{k+i} \right) \right) c^i \\ &\leq \sum_{i=1}^{\infty} \left(\frac{1}{k+1} + \frac{2}{k+1} + \dots + \frac{i}{k+1} \right) c^i = \frac{1}{2(k+1)} \sum_{i=1}^{\infty} i(i+1) c^i = \frac{c}{(k+1)(1-c)^3}. \end{aligned}$$

Hence

$$0 \leq \frac{c}{1-c} - \frac{k!}{(ck)^k} \sum_{j=k+1}^{\infty} \frac{(ck)^j}{j!} \leq \frac{c}{(k+1)(1-c)^3} \quad (2.7)$$

and the required inequality (2.4) results immediately by combining (2.7) and (2.6). \square

It is of interest to note that, when $x < k/\lambda$ is sufficiently small, the inequality described in Lemma 2 offers very tight bounds for the c.d.f. of the Erlang distribution. This point is illustrated in the next table where the bounds and the exact values of $F(x)$ have been provided for several k , λ and x .

Table 1: exact and approximate values for the c.d.f. of Erlang distribution

k	λ	x	lower bound	exact value	upper bound
50	100	0.1	1.85448 10^{-19}	1.85473 10^{-19}	1.86591 10^{-19}
		0.2	1.24399 10^{-8}	1.24589 10^{-8}	1.27170 10^{-8}
		0.3	0.000511594	0.000518891	0.000552196
		0.4	0.0538154	0.0703351	0.0885351
100	10	2.	3.48875 10^{-37}	3.48888 10^{-37}	3.49958 10^{-37}
		4.	1.20576 10^{-15}	1.20625 10^{-15}	1.21917 10^{-15}
		6.	1.47557 10^{-6}	1.48153 10^{-6}	1.53247 10^{-6}
		8.	0.0157968	0.0171083	0.0196973

For larger values of x a relatively simple approximation of $F(x)$ could be accomplished by exploiting the next formula (Gray, Thompson and McWilliams (1969))

$$\frac{1}{a^{k-1} e^{-a}} \int_a^{\infty} x^{k-1} e^{-x} dx \approx \frac{a}{a-k+1} \left(1 - \frac{k-1}{(a-k+1)^2 + 2a} \right),$$

which yields

$$F(x) \approx 1 - \frac{x f(x)}{\lambda x - k + 1} \left(1 - \frac{k-1}{(\lambda x - k + 1)^2 + 2\lambda x} \right), \quad x > k/\lambda.$$

(the labels in the left and above matrices \mathbf{A} , \mathbf{B} , refer to the values of (x_1, x_2)). The distribution of $U = U_n$ can now be easily captured through formula $P(U_n = v) = \mathbf{f}_n(v)\mathbf{1}'$, $v = 0, 1, 2$ with $\mathbf{f}_n(v)$ being computed by repeated application of the recursive formulae

$$\begin{aligned} \mathbf{f}_t(0) &= \mathbf{f}_{t-1}(0)\mathbf{A} \\ \mathbf{f}_t(v) &= \mathbf{f}_{t-1}(v)\mathbf{A} + \mathbf{f}_{t-1}(v-1)\mathbf{B}, \quad v = 1, 2. \end{aligned} \quad (3.1)$$

for $t = 1, 2, \dots, n$ ($\mathbf{f}_0(0) = \mathbf{e}_1 \in \tilde{\mathcal{N}}^9$, $\mathbf{f}_0(v) = \mathbf{0} \in \tilde{\mathcal{N}}^9$ for $v > 0$). In addition, the double generating function

$$\Phi(z, w) = \sum_{n=0}^{\infty} \sum_{v=0}^m P(U_n = v) z^v w^n = \boldsymbol{\pi}_0 (\mathbf{I} - w(\mathbf{A} + z\mathbf{B}))^{-1} \mathbf{1}'$$

can be effortlessly calculated as

$$\Phi(z, w) = 1 + w + \frac{w^2}{2} + w^2 \frac{2+2w-w^2}{(2-w)^2} z + \frac{w^4(3-w)}{2(1-w)(2-w)^2} z^2,$$

and the following (simple) generating functions ensue

$$\sum_{n=0}^{\infty} P(U_n = 0) w^n = 1 + w + \frac{w^2}{2}, \quad \sum_{n=0}^{\infty} P(U_n = 1) w^n = w^2 \frac{2+2w-w^2}{(2-w)^2}, \quad \sum_{n=0}^{\infty} P(U_n = 2) w^n = \frac{w^4(3-w)}{2(1-w)(2-w)^2}.$$

We shall now turn our attention to the r.v. W , which enumerates the balls placed in the spare urn. The upper end point of W is $l_n = \max\{v: P(W = v) > 0\} = n - k + 1$. Let us consider the state space $\Omega = \cup_{v \geq 0} C_v$, with $C_v = \{0, 1, \dots, k-1\}^m \times \{v\}$, ($|\Omega| = k^m(n-k+1)$) and introduce the r.v.'s Z_t , $t=0, 1, \dots$ as follows: $Z_t = (x_1, x_2, \dots, x_m, v)$ if and only if, after having distributed t balls into the urns, x_i balls have been placed in urn i , $i=1, 2, \dots, m$ and v balls have been forwarded to the spare urn (convention: $Z_0 = (0, 0, \dots, 0) \in \tilde{\mathcal{N}}^{m+1}$). It can now be readily verified that $\{Z_t, t \geq 0\}$ is a Markov chain and $P(Z_t \in C_w | Z_{t-1} \in C_v) = 0$ for all $w \neq v, v+1$ and $t \geq 1$, $P(W = v) = P(Z_n \in C_v)$, $v = 1, 2, \dots, l_n$. Thus, W is a MVB and its exact distribution can be easily computed in a recursive fashion on using

- the initial probabilities:

$$\boldsymbol{\pi}_0 = (P(Z_0 = (0, \dots, 0, 0)), P(Z_0 = (1, 0, \dots, 0, 0)), \dots, P(Z_0 = (k-1, \dots, k-1, 0))) = (1, 0, \dots, 0)$$

$$\boldsymbol{\pi}_v = (P(Z_0 = (0, \dots, 0, v)), P(Z_0 = (1, 0, \dots, 0, v)), \dots, P(Z_0 = (k-1, \dots, k-1, v))) = (0, 0, \dots, 0), \quad v > 0.$$

- the *within states* one step transition matrix $\mathbf{A}_t(v) = \mathbf{A} = (P(Z_t = c_{v,j} | Z_{t-1} = c_{v,i}))_{s \times s}$ whose entries are given by

$$P(Z_t = (x_1, x_2, \dots, x_m, v) + \mathbf{e}_i | Z_{t-1} = (x_1, x_2, \dots, x_m, v)) = I_{\{0, 1, \dots, k-2\}}(x_i) / m, \quad i=1, 2, \dots, m, (x_1, \dots, x_m, v) \in C_v$$

- the *between states* one step transition matrix $\mathbf{B}_t(v) = \mathbf{B} = (P(Z_t = c_{v+1,j} | Z_{t-1} = c_{v,i}))_{s \times s}$ whose entries are

$$P(Z_t = (x_1, x_2, \dots, x_m, v+1) | Z_{t-1} = (x_1, x_2, \dots, x_m, v)) = \sum_{i=1}^m I_{\{k-1\}}(x_i) / m, \quad (x_1, x_2, \dots, x_m, v) \in C_v.$$

The transition probabilities that are not described explicitly in the above statements vanish.

As an application, consider the case $k = 2$, $m = 3$. Then $\boldsymbol{\pi}_0 = (1, 0, \dots, 0) \in \tilde{\mathcal{N}}^8$, $\boldsymbol{\pi}_v = \mathbf{0} \in \tilde{\mathcal{N}}^8$, and matrices \mathbf{A} , \mathbf{B} read

$$\begin{array}{c}
(0,0,0)(0,1,0)(0,0,1)(0,1,1)(1,0,0)(1,1,0)(1,0,1)(1,1,1) \\
\mathbf{A} = \begin{bmatrix}
(0,0,0) & 0 & \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 \\
(0,1,0) & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\
(0,0,1) & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\
(0,1,1) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} \\
(1,0,0) & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\
(1,1,0) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} \\
(1,0,1) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} \\
(1,1,1) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\end{array}
\quad
\begin{array}{c}
(0,0,0)(0,1,0)(0,0,1)(0,1,1)(1,0,0)(1,1,0)(1,0,1)(1,1,1) \\
\mathbf{B} = \begin{bmatrix}
(0,0,0) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
(0,1,0) & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\
(0,0,1) & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\
(0,1,1) & 0 & 0 & 0 & \frac{2}{3} & 0 & 0 & 0 & 0 \\
(1,0,0) & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\
(1,1,0) & 0 & 0 & 0 & 0 & 0 & \frac{2}{3} & 0 & 0 \\
(1,0,1) & 0 & 0 & 0 & 0 & 0 & 0 & \frac{2}{3} & 0 \\
(1,1,1) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\end{array}$$

(the labels in the left and above matrices \mathbf{A} , \mathbf{B} , refer to the values of (x_1, x_2, x_3)). Launching the recursive scheme (3.1) with initial conditions $\mathbf{f}_0(0) = \mathbf{e}_1 \in \tilde{\mathcal{N}}^8$, $\mathbf{f}_0(v) = \mathbf{0} \in \tilde{\mathcal{N}}^8$ for $v > 0$, we can effortlessly compute $\mathbf{f}_n(v)$, $v = 1, 2, \dots, I_n$ and thereof the exact distribution of $W = W_n$ by the formula $P(W_n = v) = \mathbf{f}_n(v)\mathbf{1}'$. Finally, the corresponding double generating function reads

$$\Phi(z, w) = \sum_{n=0}^{\infty} \sum_{v=0}^m P(W_n = v) z^n w^v = 1 + \frac{9w + (6-15z)w^2 + (2-6z+6z^2)w^3}{(1-wz)(9-9wz+2w^2z^2)}.$$

An interesting feature of the aforementioned approach is that by a trivial adjustment of the transition matrices \mathbf{A} , \mathbf{B} we can also encompass the case of having different probabilities p_1, p_2, \dots, p_m of assigning a ball to the m cells (non-equiprobable scheme) or even more generally the case where the assignment of the i -th ball depends on the previous (one or more) assignments in a Markovian fashion.

However, it should be mentioned that the Markov chain imbedding method elucidated above becomes unwieldy for large values of k and m (even for the equiprobable case) because of the exponential increase of the dimensions of \mathbf{A} , \mathbf{B} (it is $(k+1)^m$ for U and k^m for W). This remark brings up the question whether one could establish neat and computationally tractable approximations for large values of k , n and m . In the next section we are going to address this problem.

4. Approximations and limit theorems

In the present section we deal with the problem of approximating the distribution of the r.v.'s U and W by appropriate Poisson distributions. The main tools to accomplish that are Theorems 1, 2 of Section 2 and representations (2.1), (2.2) of U and W in terms of the total number S_j of balls assigned to urn j , $j=1, 2, \dots, m$.

In all that follows we shall use p for the probability of assigning a ball to a specific urn i.e. $p=1/m$.

It is clear that the random vector (S_1, S_2, \dots, S_m) follows a multinomial distribution with parameters n and (p, p, \dots, p) . This ascertains that S_1, S_2, \dots, S_m are NA (see Joag-Dev and Proschan(1983)) and since the r.v.'s $I_j = I_{[k, \infty)}(S_j)$, $j = 1, 2, \dots, m$ are non decreasing functions of $\{S_1, S_2, \dots, S_m\}$ defined on the disjoint subsets $\{S_j\}$, $j = 1, 2, \dots, m$, we conclude that I_1, I_2, \dots, I_m are also NA . A direct application of Theorem 1 yields

$$d(L(\sum_{j=1}^m I_j), L(\sum_{j=1}^m I'_j)) \leq -\sum_{i < j} Cov(I_i, I_j) \quad (4.1)$$

Table 3. Comparison between our binomial type bounds, improved Bonferroni bounds and Chen-Stein bounds for the generalized birthday problem ($m=365$)

n	k	lower bounds			Exact value	upper bounds		
		Improved Bonferroni	Chen-Stein	Binomial type		Binomial type	Chen-Stein	Improved Bonferroni
10	2	.11586	.11469	.11443	.11695	.11717	.11745	.12296
10	3	.00087	.00087	.0008874	.0008877	.0008877	.00090	.00090
15	3	.00324	.00324	.0033265	.0033294	.0033294	.00341	.00341
23	2	.45387	.48397	.48712	.50730	.52350	.51680	.69126
23	3	.01212	.01210	.01268	.01271	.01271	.01323	.01326
40	2	.68842	.82085	.86461	.89123	1	.94474	1
40	3	.06017	.06011	.06644	.06689	.06692	.07187	.07396
60	3	.16387	.16836	.20437	.20723	.20787	.22950	.25625
88	2	.91490	.78682	.99989	.99999	1	1	1
88	3	.34040	.32147	.49989	.51107	.52120	.57888	.82143
88	4	.02298	.02126	.03913	.03925	.03925	.04697	.04782
100	3	.40369	.33796	.63097	.64586	.66921	.73325	1
120	4	.06364	.02860	.12293	.12380	.12391	.15675	.16847
150	3	.65589	0	.95408	.96477	1	1	1
187	3	.76415	0	.99624	.99815	1	1	1
187	4	.21467	0	.49438	.50269	.51020	.66123	1
187	5	.02023	0	.06507	.06530	.06532	.09724	.10145
250	5	.05870	0	.22340	.22535	.22587	.36281	.43922

Let us return again to the original general ball-in-urn distribution problem and inequality (4.2). Should one be interested in a Poisson approximation for $L(U)$ with the same mean (instead of the binomial approximation suggested by (4.2)) he could make use of the triangle inequality in conjunction with (4.2) to get

$$d(L(U), Po\left(m \sum_{s=k}^n \binom{n}{s} p^s (1-p)^{n-s}\right)) \leq UB + m \left(\sum_{s=k}^n \binom{n}{s} p^s (1-p)^{n-s} \right)^2 \quad (4.4)$$

Results of similar flavor have also been formulated by Barbour, Holst and Janson (1992) with the aid of the celebrated Chen-Stein method.

Our next task is to investigate the limiting behavior of the distribution of U as $n, m \rightarrow \infty$. Two different cases will be considered for k : in the first k will be increasing with no bound ($k \rightarrow \infty$) as n and m increase whilst in the second k will be fixed.

Theorem 3. If $n, m, k \rightarrow \infty$ such that

$$\frac{n}{km} = c + o\left(\frac{1}{k}\right), \quad c \in (0,1) \quad \text{and} \quad \frac{m}{1-c} \frac{e^{-ck} (ck)^k}{k!} \rightarrow \lambda < \infty, \quad (4.5)$$

then the limiting distribution of U is Poisson with expected value (parameter) λ . Moreover, the rate of convergence of $L(U)$ to $Po(\lambda)$ is at least $O(k^{1/2} (ce^{1-c})^k)$.

Proof. The expected value of U equals, by virtue of (2.3) for $r=1$,

$$E(U) = mP(S_1 \geq k) = m \sum_{s=k}^n \binom{n}{s} p^s (1-p)^{n-s} = \frac{mn!}{(k-1)!(n-k)!} \int_0^p x^{k-1} (1-x)^{n-k} dx.$$

With the aid of

$$e^{-x} (1-x)^x < 1-x < e^{-x}, \quad 0 < x < 1 \quad (4.6)$$

(which results immediately from the elementary inequality $x < -\ln(1-x) < x/(1-x)$, $0 < x < 1$) the integral of the RHS can be bounded above and below as follows

$$(1-p)^{p(n-k)} \int_0^p x^{k-1} e^{-x(n-k)} dx \leq \int_0^p x^{k-1} e^{-x(n-k)} (1-x)^{x(n-k)} dx \leq \int_0^p x^{k-1} (1-x)^{n-k} dx \leq \int_0^p x^{k-1} e^{-x(n-k)} dx.$$

Writing the last outcome as

$$\frac{(1-p)^{p(n-k)}}{(n-k)^k} \int_0^{p(n-k)} x^{k-1} e^{-x} dx \leq \int_0^p x^{k-1} (1-x)^{n-k} dx \leq \frac{1}{(n-k)^k} \int_0^{p(n-k)} x^{k-1} e^{-x} dx \quad (4.7)$$

and making use of (2.4) for $a=p(n-k)$ we obtain

$$\frac{mn!(1-p)^{p(n-k)}}{(k-1)!(n-k)!} \frac{p^k e^{-p(n-k)}}{k-p(n-k)} \left(1 - \frac{p(n-k)}{k(k+1)(1-p(n-k)/k)^2} \right) \leq E(U) \leq \frac{mn!}{(k-1)!(n-k)!} \frac{p^k e^{-p(n-k)}}{k-p(n-k)} \quad (4.8)$$

for sufficiently large n, m, k such that $p(n-k)/k < 1$ (this is always feasible because, under the assumptions made, we have $\lim p(n-k)/k = c < 1$). It is now easy to verify that, due to conditions (4.5), both lower and upper bounds for $E(U)$ converge to λ and consequently $\lim E(U) = \lambda$. This secures that the Poisson distribution appearing in the LHS of (4.4) converges to $Po(\lambda)$, whilst the rightmost term of the upper bound (i.e. $(E(U))^2/m$) converges to 0. We shall now prove that UB as stated in (4.3) converges to 0 thereof obtaining the desired limiting result.

Let us start by writing UB in terms of appropriate integral expressions. Exploiting (2.3) for $r = 1, 2$ we get

$$\begin{aligned} UB &= \frac{m(m-1)}{2} \left(\left(\frac{n!}{(k-1)!(n-k)!} \right)^2 \int_0^p \int_0^p x^{k-1} y^{k-1} (1-x)^{n-k} (1-y)^{n-k} dx dy - \right. \\ &\quad \left. - \frac{n!}{(k-1)!^2 (n-2k)!} \int_0^p \int_0^p x^{k-1} y^{k-1} (1-x-y)^{n-2k} dx dy \right) \\ &= \frac{m(m-1)}{2} \left(\frac{n!}{(k-1)!(n-k)!} \right)^2 \int_0^p \int_0^p x^{k-1} y^{k-1} A(x, y) dx dy, \quad n \geq 2k \end{aligned}$$

where

$$\begin{aligned} A(x, y) &= (1-x)^{n-k} (1-y)^{n-k} - \frac{((n-k)!)^2}{n!(n-2k)!} (1-x-y)^{n-2k} \\ &\leq (1-x)^{n-k} (1-y)^{n-k} - \left(1 - \frac{k}{n-k+1} \right)^k (1-x-y)^{n-k}. \end{aligned} \quad (4.9)$$

Recalling (4.6) we gain the inequality

$$\begin{aligned} A(x, y) &\leq e^{-(x+y)(n-k)} \left(1 - \left(1 - \frac{k}{n-k+1} \right)^k (1-x-y)^{(x+y)(n-k)} \right) \\ &\leq e^{-(x+y)(n-k)} \left(1 - \left(1 - \frac{k}{n-k+1} \right)^k (1-2p)^{2p(n-k)} \right) \end{aligned} \quad (4.10)$$

which leads to the following upper bound for UB

$$UB \leq \frac{m(m-1)}{2} \left(\frac{n!}{(k-1)!(n-k)!} \right)^2 \left(1 - \left(1 - \frac{k}{n-k+1} \right)^k (1-2p)^{2p(n-k)} \right) \int_0^p \int_0^p x^{k-1} y^{k-1} e^{-(x+y)(n-k)} dx dy.$$

Observe next that the double integral simplifies to

$$\left(\int_0^p x^{k-1} e^{-x(n-k)} dx \right)^2 = \left(\frac{1}{(n-k)^k} \int_0^{p(n-k)} x^{k-1} e^{-x} dx \right)^2 \leq \left(\frac{p^k e^{-p(n-k)}}{k-p(n-k)} \right)^2$$

with the inequality holding true for sufficiently large n, m, k such that $p(n-k)/k < 1$ (to verify that, apply (2.4) for $a = p(n-k)$). It is now straightforward that

$$\begin{aligned} UB &\leq \frac{1}{2} \left(1 - \left(1 - \frac{k}{n-k+1} \right)^k (1-2p)^{2p(n-k)} \right) \left(\frac{mn! p^k e^{-p(n-k)}}{(k-1)!(n-k)!(k-p(n-k))} \right)^2 \\ &\leq \frac{1}{2} \left(1 - \left(1 - \frac{k}{n-k+1} \right)^k (1-2p)^{2p(n-k)} \right) \left(\frac{m e^{-np} (np)^k}{k!(1-p(n-k)/k)} \right)^2 e^{2pk} \equiv UB' \end{aligned} \quad (4.11)$$

Under the assumptions (4.5), the above bound converges to 0. More specifically we have

$$UB' \sim \frac{1}{2} \left(4c + \frac{1}{c} \right) \frac{k}{m} \cdot \lambda^2 \cdot 1 \sim \left(4c + \frac{1}{c} \right) \frac{\lambda (ce^{1-c})^k}{2(1-c)} \sqrt{\frac{k}{2\pi}} = O(k^{1/2} (ce^{1-c})^k) \rightarrow 0$$

(Note that $ce^{1-c} < 1$ for $c \in (0, 1)$. The symbol \sim indicates that the ratio of the two sides tends to 1). An upper bound for the rate of convergence of $L(U)$ to $Po(\lambda)$ is offered by

$$UB' + mP(S_1 \geq 0)^2 = UB' + \frac{E(U)^2}{m} \sim \left(4c + \frac{1}{c} \right) \frac{\lambda (ce^{1-c})^k}{2(1-c)} \sqrt{\frac{k}{2\pi}} + \frac{\lambda^2}{m} = O(k^{1/2} (ce^{1-c})^k) \rightarrow 0$$

This completes the proof. \square

We are now going to state a second limit result which covers the case of fixed k .

Theorem 4. *If k is fixed and $n, m \rightarrow \infty$ such that*

$$m \frac{(n/m)^k}{k!} \rightarrow \lambda < \infty \quad (4.12)$$

then the limiting distribution of U is Poisson with expected value (parameter) λ . Moreover, the rate of convergence of $L(U)$ to $Po(\lambda)$ is at least $O(1/m^{1-1/k}) = O(1/n)$.

Proof. It is couched on essentially the same considerations as the ones used in obtaining the proof of Theorem 3. Condition (4.12) implies that $\lim p(n-k)/k = \lim n/(km) = 0$ and choosing n, m sufficiently large so that $p(n-k)/k < 1$ we can secure the validity of the inequalities (4.8) and (4.11). It can now be easily verified that, under assumption (4.12),

a. the lower and upper bounds in (4.8) converge again to λ as $n, m \rightarrow \infty$; therefore $\lim E(U) = \lambda$.

b. UB' of (4.11) converges to 0; more specifically,

$$UB' \sim \frac{1}{2} \left(\frac{k^2}{n} + 4 \frac{n}{m^2} \right) \left(\frac{n^k}{k! m^{k-1}} \right)^2 \sim \frac{1}{2} \left(\frac{k^2}{n} + 4 \frac{n}{m^2} \right) \lambda^2 = O(1/m^{1-1/k})$$

c. the last term in (4.4) i.e. $m(P(U \geq k))^2 = E^2(U)/m$ is of order $O(1/m)$.

Using (a), (b), (c) in conjunction with (4.4) we immediately yield the desired limit result. \square

Recently Henze(1998) proved, using an entirely different technique, that, when $m \rightarrow \infty$ and $n = [m^{1-1/k}t]$, U converges to a Poisson distribution with parameter $\lambda = t^k/k!$. This can also be established by exploiting Theorem 4 since in that case we have

$$\frac{n^k}{k!m^{k-1}} = \frac{[m^{1-1/k}t]^k}{k!m^{k-1}} \rightarrow \frac{t^k}{k!} = \lambda.$$

Let us now turn our attention to the limiting behavior of W , the number of balls placed in the spare (overflow) urn. Observe first that

$$W_j = (S_j - k + 1)I_{[k, \infty)}(S_j), \quad j=1, 2, \dots, m,$$

are NA as non decreasing functions of the set of NA r.v.'s $\{S_1, S_2, \dots, S_m\}$ defined on the disjoint subsets $\{S_j\}, j=1, 2, \dots, m$. Apply next Theorem 2 to gain the inequality

$$d(L(W), CP(\mu, F)) = d(L(\sum_{j=1}^m W_j), CP(\mu, F)) \leq -\sum_{i < j} Cov(W_i, W_j) + mP(W_1 \neq 0)^2. \quad (4.13)$$

where F is the cdf of the zero truncated distribution of W_1 with probability mass function

$$P(W_1 = i | W_1 \neq 0) = P(S_1 = i + k - 1 | S_1 \geq k) = \frac{\binom{n}{i+k-1} p^{i+k-1} (1-p)^{n-i-k+1}}{\sum_{s=k}^n \binom{n}{s} p^s (1-p)^{n-s}}, \quad i=1, 2, \dots \quad (4.14)$$

and

$$\mu = mP(W_1 \neq 0) = mP(S_1 \geq k) = E(U) = m \sum_{s=k}^n \binom{n}{s} p^s (1-p)^{n-s} \quad (4.15)$$

The terms in the RHS of (4.13) take the form

$$\begin{aligned} UB_1 &= -\sum_{j=1}^m \sum_{i=1}^{j-1} Cov(W_i, W_j) = \frac{m(m-1)}{2} (E(W_1)E(W_2) - E(W_1 W_2)) = \\ &= \frac{m(m-1)}{2} \left(\left(\sum_{i=k}^n (i-k+1) \binom{n}{i} p^i (1-p)^{n-i} \right)^2 - \sum_{i=k}^{n-k} \sum_{j=k}^{n-i} (i-k+1)(j-k+1) \binom{n}{i, j} p^{i+j} (1-2p)^{n-i-j} \right) \end{aligned} \quad (4.16)$$

and

$$UB_2 = mP(W_1 \neq 0)^2 = mP(S_1 \geq k)^2 = \mu^2/m$$

respectively.

We are now equipped with all necessary machinery to establish a limiting result for the distribution of W under the same conditions on n, m, k as those used in Theorem 3. The limiting law turns out to be a Polya-Aeppli distribution $PA(\lambda, c)$ i.e. a $CP(\lambda, F)$ with F being a geometric distribution with parameter c . The corresponding pdf is given by

$$PA(\lambda, c)\{0\} = e^{-\lambda}, \quad PA(\lambda, c)\{u\} = e^{-\lambda} c^u \sum_{j=1}^u \binom{u-1}{j-1} \frac{(\lambda(1-c)/c)^j}{j!}, \quad u=1, 2, \dots$$

Theorem 5. *If $k, n, m \rightarrow \infty$ such that*

$$\frac{n}{km} = c + o\left(\frac{1}{k}\right), \quad c \in (0, 1) \quad \text{and} \quad \frac{m}{1-c} \frac{e^{-ck} (ck)^k}{k!} \rightarrow \lambda < \infty, \quad (4.17)$$

then the limiting distribution of W is a Polya-Aeppli (compound Poisson) distribution with parameters λ and c . Moreover, the rate of convergence is at least $O(k^{1/2} (ce^{1-c})^k)$.

Proof. From the proof of Theorem 3 it is clear that, under assumptions (4.17), $\mu = E(U)$ of (4.15) converges to λ and UB_2 converges to 0. Manifestly, the proof of our target limiting result would be complete if in addition we succeeded to verify that

- a. the limiting distribution F is geometric with parameter (success probability) c ,
- b. the quantity UB_1 (and therefore the upper bound $UB_1 + UB_2$) converges to 0.

For part (a) apply inequality (4.8) for the sum in the denominator of (4.14) to bound $P(W_1 = i | W_1 \neq 0)$ as follows

$$\begin{aligned} \frac{(k-1)!(n-k)!p^{i-1}(1-p)^{n-i-k+1}(k-p(n-k))}{(i+k-1)!(n-i-k+1)!e^{-p(n-k)}} &\leq P(W_1 = i | W_1 \neq 0) \leq \\ &\leq \frac{(k-1)!(n-k)!p^{i-1}(1-p)^{n-i-k+1}(k-p(n-k))}{(i+k-1)!(n-i-k+1)!(1-p)^{p(n-k)} e^{-p(n-k)}} \left(1 - \frac{p(n-k)}{k(k+1)(1-p(n-k)/k)^2}\right)^{-1} \end{aligned}$$

(for sufficiently large n, m, k such that $p(n-k)/k < 1$). It is rather straightforward that, under assumptions (4.17), both upper and lower bounds stated above converge to $c^{i-1}(1-c)$. For example, the asymptotic behavior of the lower bound can be successively established by

$$\frac{(k-1)!(n-k)!p^{i-1}(1-p)^{n-i-k+1}(k-p(n-k))}{(i+k-1)!(n-i-k+1)!e^{-p(n-k)}} \sim \frac{n^{i-1}p^{i-1}(1-p)^{n-i-k+1}(1-c)}{k^{i-1}e^{-p(n-k)}} \sim c^{i-1}(1-c)$$

whereas similar reasoning applies for the upper bound as well.

Let us now switch to assertion (b). Replacing the sums in (4.16) by the integral expressions provided by Lemma 1, we get

$$UB_1 = \frac{m(m-1)}{2} \left(\frac{n!}{(k-2)!(n-k)!} \right)^2 \int_0^p \int_0^p (p-x)(p-y)x^{k-2}y^{k-2}A(x,y)dx dy$$

with $A(x,y)$ given as in (4.9). Employing once more inequality (4.10) we deduce

$$UB_1 \leq \frac{1}{2} \left(\frac{mn!}{(k-2)!(n-k)!} \right)^2 \left(1 - \left(1 - \frac{k}{n-k+1} \right)^k (1-2p)^{2p(n-k)} \right) \left(\int_0^p (p-x)x^{k-2}e^{-x(n-k)}dx \right)^2. \quad (4.18)$$

We shall try next to construct an appropriate upper bound for

$$\int_0^p (p-x)x^{k-2}e^{-x(n-k)}dx = \frac{p^k}{k-1} e^{-p(n-k)} - \left(1 - p \frac{n-k}{k-1} \right) \frac{1}{(n-k)^k} \gamma(k, (n-k)p).$$

Exploiting the lower bound inferred from (2.4) for $\gamma(k,a)$ the next inequality ensues

$$\int_0^p (p-x)x^{k-2}e^{-x(n-k)}dx \leq \frac{p^k e^{-p(n-k)}}{k-1} \left(1 - \left(1 - \frac{1}{k-p(n-k)} \right) \left(1 - \frac{p(n-k)}{k(k+1)(1-p(n-k)/k)^2} \right) \right). \quad (4.19)$$

Combining (4.18), (4.19) we obtain

$$UB_1 \leq \frac{e^{2pk}}{2} \left(1 - \left(1 - \frac{k}{n-k+1} \right)^k (1-2p)^{2p(n-k)} \right) \left(\frac{me^{-np} (np)^k}{k! \left(1 - \frac{p(n-k)}{k} \right)} \right)^2 \left(1 + \frac{\frac{p(n-k)}{k+1}}{1 - \frac{p(n-k)}{k}} - \frac{\frac{p(n-k)}{k+1}}{k \left(1 - \frac{p(n-k)}{k} \right)^2} \right)^2. \quad (4.20)$$

Straightforward calculations on the last bound reveal that (under assumptions (4.17)) it converges to 0. More specifically, the asymptotic rate of convergence of it is

$$\frac{1}{2} \left(4c + \frac{1}{c} \right) \frac{k}{m} \cdot \lambda^2 \cdot \left(1 + \frac{c}{1-c} \right)^2 \sim \left(4c + \frac{1}{c} \right) \frac{\lambda (ce^{1-c})^k}{2(1-c)^3} \sqrt{\frac{k}{2\pi}} = O(k^{1/2} (ce^{1-c})^k) \rightarrow 0$$

and an upper bound for the rate of convergence of UB_1+UB_2 to 0 is offered by

$$\left(4c + \frac{1}{c} \right) \frac{\lambda (ce^{1-c})^k}{2(1-c)^3} \sqrt{\frac{k}{2\pi}} + \frac{\lambda^2}{m} = O(k^{1/2} (ce^{1-c})^k) \rightarrow 0$$

This completes the proof. \square

In closing this section we give some tables illustrating the quality of the Poisson approximations established in Theorems 3-5. In Tables 4 and 5 several large values of n and m were chosen and the cumulative distribution function $P(U \leq i)$ of U (calculated by simulation, 10^5 iterations), along with the c.d.f. of $Po(E(U)) = Po(\lambda_0)$ and $Po(\lim E(U)) = Po(\lambda)$ were evaluated at $i = 0, 1, 2, \dots$. Upper bounds UB_U for the errors incurred when approximating $P(U \leq i)$ by the c.d.f. of $Po(\lambda_0)$ are also provided. To be more precise, the values displayed above the main body of the tables were computed by formulae

$$\lambda_0 = m \sum_{s=k}^n \binom{n}{s} p^s (1-p)^{n-s}, \quad \lambda = \frac{m}{1-c} \frac{e^{-ck} (ck)^k}{k!}, \quad c = \frac{n}{km},$$

$$UB_U = \frac{m(m-1)}{2} \left(\left(\sum_{i=k}^n \binom{n}{s} p^s (1-p)^{n-s} \right)^2 - \sum_{i=k}^{n-k} \sum_{j=k}^{n-i} \binom{n}{i,j} p^{i+j} (1-2p)^{n-i-j} \right) + m \left(\sum_{s=k}^n \binom{n}{s} p^s (1-p)^{n-s} \right)^2,$$

(c.f. (4.3), (4.4))

$$UB_U^* = \frac{1}{2} \left(1 - \left(1 - \frac{k}{n-k+1} \right)^k (1-2p)^{2p(n-k)} \right) \left(\frac{me^{-np} (np)^{k-1}}{(k-1)!} \right)^2 e^{2pk} \left(\frac{np/k}{1 - p(n-k)/k} \right)^2 + m \left(\frac{n!}{(k-1)!(n-k)!} \frac{p^k e^{-p(n-k)}}{k - p(n-k)} \right)^2$$

(c.f. (4.11) and the upper bound in (4.8)); note that UB_U^* is a weaker upper bound than UB_U ($UB_U \leq UB_U^*$) but it is easier to evaluate. In Table 4, which serves as an illustration for Theorem 3, large values of k were chosen whereas in Table 5, which elucidates Theorem's 4 performance, a variety of small values was assigned to k ; in this case the c.d.f. of $Po(\lambda_2)$, $\lambda_2 = m(n/m)^k/k!$ was also incorporated in the table for comparison reasons. Finally, in Table 6 we have evaluated the exact cumulative distribution function $P(W \leq i)$ of W (by simulation, 10^5 iterations), along with the corresponding values of the $CP(\mu, F)$ distribution used in formula (4.13) and the Polya-Aeppli distribution described in Theorem 5. The error estimates (upper bounds) UB_W , UB_W^* for the approximation of $P(W \leq i)$ by the c.d.f. $CP(\mu, F)$ were calculated through formulae

$$UB_W = \frac{m(m-1)}{2} \left(\left(\sum_{i=k}^n (i-k+1) \binom{n}{i} p^i (1-p)^{n-i} \right)^2 - \sum_{i=k}^{n-k} \sum_{j=k}^{n-i} (i-k+1)(j-k+1) \binom{n}{i,j} p^{i+j} (1-2p)^{n-i-j} \right) + m \left(\sum_{s=k}^n \binom{n}{s} p^s (1-p)^{n-s} \right)^2$$

(c.f. (4.13) and (4.16)),

$$UB_W^* = \frac{e^{2pk}}{2} \left(1 - \left(1 - \frac{k}{n-k+1} \right)^k (1-2p)^{2p(n-k)} \right) \left(\frac{me^{-np} (np)^k}{k!} \right)^2 \left(\frac{1}{1 - \frac{p(n-k)}{k}} + \frac{\frac{p(n-k)}{k+1}}{\left(1 - \frac{p(n-k)}{k} \right)^2} - \frac{\frac{p(n-k)}{k+1}}{k \left(1 - \frac{p(n-k)}{k} \right)^3} \right)^2 + m \left(\frac{n!}{(k-1)!(n-k)!} \frac{p^k e^{-p(n-k)}}{k-p(n-k)} \right)^2$$

(c.f. (4.20) and upper bound in (4.8)); note once more that UB_W^* is worse than UB_U but is computationally more tractable.

Table 4. Exact and approximate c.d.f. of U for large k .

	$k=20, n=9000, m=1000$			$k=20, n=80000, m=10000$			$k=30, n=74000, m=5000$			$k=30, n=140000, m=10000$		
	$UB_U=.00940, UB_U^*=.0517$ $\lambda_0=1.049, \lambda=1.121, c=.45$			$UB_U=.0069, UB_U^*=.0295$ $\lambda_0=2.527, \lambda=2.650, c=.4$			$UB_U=.0055, UB_U^*=.0387$ $\lambda_0=1.687, \lambda=1.782, c=.493$			$UB_U=.0020, UB_U^*=.0124$ $\lambda_0=1.357, \lambda=1.423, c=.467$		
i	$Po(\lambda)$	$Po(\lambda_0)$	Exact value	$Po(\lambda)$	$Po(\lambda_0)$	Exact value	$Po(\lambda)$	$Po(\lambda_0)$	Exact value	$Po(\lambda)$	$Po(\lambda_0)$	Exact value
0	.3259	.3503	.3460	.0707	.0799	.0816	.1683	.1852	.1857	.2411	.2575	.2590
1	.6913	.7177	.7161	.2580	.2818	.2847	.4683	.4974	.4972	.5841	.6069	.6087
2	.8960	.9105	.9125	.5061	.5369	.5393	.7355	.7608	.7602	.8280	.8439	.8453
3	.9722	.9779	.9783	.7252	.7517	.7553	.8942	.9088	.9092	.9437	.9510	.9523
4	.9941	.9955	.9959	.8703	.8875	.8916	.9649	.9712	.9713	.9848	.9874	.9884
5	.9989	.9992	.9993	.9472	.9561	.9580	.9901	.9923	.9923	.9965	.9973	.9974
6				.9812	.9850	.9856	.9976	.9982	.9980	.9993	.9995	.9995

Table 5. Exact and approximate c.d.f. of U for small k .

	$k=3, n=100, m=600$				$k=3, n=1000, m=10000$				$k=5, n=3 \cdot 10^7, m=5 \cdot 10^8$			$k=5, n=20000, m=50000$		
	$UB_U=.0067, UB_U^*=.0081$ $\lambda_0=.3980, \lambda=.4149,$ $\lambda_2=.4630, c=.05556$				$UB_U=.0104, UB_U^*=.0112$ $\lambda_0=1.542, \lambda=1.560,$ $\lambda_2=1.667, c=.0333$				$UB_U^*=3.99 \cdot 10^{-6}$ $\lambda_0=3.082, \lambda=3.088,$ $\lambda_2=3.24, c=.0120$			$UB_U=.0053, UB_U^*=.0064$ $\lambda_0=3.061, \lambda=3.109,$ $\lambda_2=4.267, c=.080$		
i	$Po(\lambda)$	$Po(\lambda_2)$	$Po(\lambda_0)$	Exact value	$Po(\lambda)$	$Po(\lambda_2)$	$Po(\lambda_0)$	Exact value	$Po(\lambda)$	$Po(\lambda_2)$	$Po(\lambda_0)$	$Po(\lambda)$	$Po(\lambda_2)$	$Po(\lambda_0)$
0	.6604	.6294	.6716	.6645	.2101	.1889	.2139	.2113	.0456	.0392	.0459	.0447	.0140	.0468
1	.9344	.9208	.9390	.9410	.5379	.5037	.5438	.5406	.1863	.1661	.1872	.1835	.0739	.1902
2	.9913	.9883	.9922	.9938	.7936	.7660	.7982	.7978	.4037	.3716	.4051	.3993	.2016	.4097
3	.9991	.9987	.9992	.9994	.9266	.9117	.9290	.9297	.6274	.5936	.6289	.6229	.3832	.6336
4					.9784	.9725	.9794	.9804	.8002	.7735	.8013	.7967	.5769	.8049
5					.9946	.9927	.9949	.9950	.9069	.8900	.9076	.9047	.7422	.9098
6									.9618	.9529	.9622	.9607	.8597	.9633

Table 6. Exact and approximate c.d.f. of W .

	$k=20, n=9000, m=1000$			$k=20, n=80000, m=10000$			$k=30, n=74000, m=5000$			$k=30, n=140000, m=10000$		
	$UB_W=.0271, UB_W^*=.148,$ $\lambda_0=1.049, \lambda=1.121, c=.45$			$UB_W=.0175, UB_W^*=.0727$ $\lambda_0=2.527, \lambda=2.650, c=.4$			$UB_W=.0188, UB_W^*=.1352,$ $\lambda_0=1.687, \lambda=1.782, c=.493$			$UB_W=.0064, UB_W^*=.0393,$ $\lambda_0=1.357, \lambda=1.423, c=.467$		
i	Polya-Aeppli	Comp. Poisson	Exact value	Polya-Aeppli	Comp. Poisson	Exact value	Polya-Aeppli	Comp. Poisson	Exact value	Polya-Aeppli	Comp. Poisson	Exact value
0	.3259	.3503	.3457	.0707	.0799	.0816	.1683	.1852	.1857	.2411	.2575	.2625
1	.5268	.5650	.5627	.1831	.2068	.2082	.3203	.3521	.3531	.4240	.4527	.4590
2	.6792	.7228	.7200	.3173	.3559	.3595	.4639	.5070	.5105	.5788	.6148	.6211
3	.7884	.8302	.8295	.4541	.5036	.5093	.5892	.6383	.6420	.7009	.7389	.7447
4	.8635	.8994	.8999	.5797	.6344	.6386	.6927	.7427	.7442	.7927	.8285	.8339
5	.9136	.9420	.9428	.6870	.7412	.7462	.7748	.8217	.8209	.8593	.8902	.8944
6	.9462	.9673	.9676	.7735	.8231	.8277	.8379	.8793	.8777	.9061	.9314	.9345
7	.9669	.9820	.9824	.8402	.8827	.8867	.8851	.9199	.9180	.9383	.9579	.9605
8				.8898	.9243	.9273	.9197	.9478	.9460	.9600	.9747	.9766
9				.9255	.9524	.9543	.9445	.9666	.9654	.9744	.9850	.9860

References

- ARNOLD, B.C. AND N. BALAKRISHNAN (1989). *Relations, Bounds and Approximations for Order Statistics*. Lecture Notes in Statistics. Springer-Verlag.
- BARBOUR, A.D., L. HOLST AND S. JANSON (1992). *Poisson Approximation*. Clarendon Press, Oxford.
- BILLINGSLEY, P. (1986). *Probability & Measure*. 2ed, Wiley, New York.
- BOUTSIKAS, M.V. AND M.V. KOUTRAS (1999). A bound for the distribution of the sum of discrete associated or NA random variables. *Annals of Applied Probability*. (to appear)
- ESARY, J.D., F. PROSCHAN AND D. WALKUP (1967). Association of random variables with applications. *Annals of Mathematical Statistics* **38**, 1466-1474.
- FELLER, W. (1968). *An Introduction to Probability Theory and its Applications*, Vol 1, 3rd ed. Wiley, N.Y.
- FU, J.C. (1996). Distribution theory of run and patterns associated with a sequence of multi-state trials. *Statistica Sinica* **6**, 957-974.
- FU, J.C. AND M.V. KOUTRAS (1994). Distribution theory of runs: A Markov chain approach. *Journal of the American Statistical Association* **89**, 1050-1058.
- GRADSHTEYN, I.S. AND I.M. RYZHIK (1980). *Tables of Integrals, Series and Products*. Academic Press, N.Y.
- GRAY, H.L., R.W. THOMPSON AND G.V. MCWILLIAMS (1969). A new approximation for the chi-square integral. *Mathematics of computation*, **23**, 85-89.
- HENZE, N. (1998). A Poisson limit law for a generalized birthday problem. *Statistics and Probability Letters* **39**, 333-336.
- HOLST, L. (1995). The general birthday problem. *Random Structures and Algorithms* **6**, 201-208.
- JOAG-DEV, K. AND F. PROSCHAN (1983). Negative association of random variables with applications. *The Annals of Statistics* **11**, 286-295.
- JOHNSON, N.L. AND S. KOTZ (1977). *Urn models and their applications*, New York: John Wiley and Sons.
- KOTZ, S. AND N. BALAKRISHNAN (1997). Advances in urn models during the past two decades. In *Advances in Computational methods and Applications to Probability and Statistics* (Edr N.Balakrishnan), Birkhauser, Boston, 203-257.
- KOUNIAS S. (1995). Poisson approximation and Bonferroni bounds for the probability of the union of events. *Int. J. Math. & Stat. Sci.* **4**, 43-52.
- KOUTRAS, M.V. AND V.A. ALEXANDROU (1995). Runs, scans and urn model distributions: a unified Markov chain approach. *Ann. Inst. Statist. Math.* **47**, 743-766.
- MCKINNEY, E. (1966). Generalized birthday problem. *Amer. Math. Monthly* **73**, 385-387.
- MENON, V.V. AND N.K. INDIRA (1990). A Poisson approximation in an urn model with indistinguishable balls, *Journal of Statistical Planning and Inference* **26**, 93-101.
- OLKIN, I. AND M. SOBEL (1965). Integral expressions for tail probabilities of the multinomial and the negative multinomial distribution. *Biometrika* **52**, 167-179.
- SERFLING, R. J. (1978). Some elementary results on Poisson approximations in a sequence of Bernoulli trials. *SIAM Rev.* **20**, 567-579.
- VELLAISAMY, P. AND B. CHAUDHURI (1996). Poisson and Compound Poisson approximations for random sums of random variables. *Journal of Applied Probability* **33**. 127-137.