Mathematica Slovaca

Saleh M. Mitwalli

An occupancy problem with group drawings of different sizes

Mathematica Slovaca, Vol. 52 (2002), No. 2, 235--242

Persistent URL: http://dml.cz/dmlcz/132830

Math. Slovaca, **52** (2002), No. 2, 235–242

Mathematica Slovaca

AN OCCUPANCY PROBLEM WITH GROUP DRAWINGS OF DIFFERENT SIZES

SALEH M. MITWALLI

(*Communicated by Gejza Wimmer*)

ABSTRACT. We consider the classical occupancy problem with group drawings of different sizes. We derive the exact distribution of the number of balls occurring in all sampled groups and its asymptotic behavior. Furthermore, the number of drawings until at most k balls are observed in all samples is studied.

1. Introduction

Consider an urn containing s different balls. Groups of m_i balls ($m_i \le s$, $m_1 \ne m_2 \ne \cdots$) are drawn successively with replacement; in each drawing all $\binom{s}{m_i}$, $i = 1, 2, \dots$, possible samples of balls are equally likely. Let $U_j(s, m_1, m_2, \dots, m_n, n)$ be the number of balls occurring exactly j times in the first n drawings. Problems concerning the distribution of $\{U_j(s, m_1, m_2, \dots \dots, m_n, n) : j = 0, 1, \dots, n,\ n \in \mathbb{N}\}$ or of functions of this sequence are called *occupancy problems*. Their history can be traced back to DeMoivre, Laplace, and Euler (see Stadje (1990)). For the case $m_1 = m_2 = \cdots = m_n = 1$ we refer to Holst (1986) for a survey and a unifying approach. For arbitrary $m_1 = m_2 = \cdots = m_n = l > 1$ in particular $U_0(s, l, n)$, the number of balls that have not shown up in n drawings, has been extensively studied. Harris et al. (1987) derive its finite and asymptotic distribution (in their model each sampled ball has a probability $q \in [0, 1]$ of "disappearing" after being drawn). Asymptotic results on $U_0(s, l, n)$ can be found e.g. in Mikhailov (1980) and Park (1981). If $s = 2l$, the exact probability distributions of $U_0(s, l, n)$ and $U_n(s, l, n)$ are equal. In this note we consider $U(s, m_1, m_2, \dots, m_n, n) = U_n(s, m_1, m_2, \dots, m_n)$, the number of balls occurring in *all* n sampled groups, which is a generalization of Stadje et al. (1998). We determine its exact distribution and its asymptotic behavior for fixed n, as

2000 Mathematics Subject Classification: Primary 60E05, 62E15.
Keywords: occupancy problem, waiting times.

$s, m_1, m_2, \dots, m_n \longrightarrow \infty$. If $\frac{s}{m}$, $m = m_1 = m_2 = \cdots = m_n$, is fixed, $U(s, m, n)$ is asymptotically normal, while for fixed $s \prod_{i=1}^{n} \frac{m_i}{s}$ it tends to a Poisson distribution. The functional limit theorem for exchangeable random variables and the method of moments are used in the proofs.

2. The exact distribution

We derive formulas for the probability distribution of $U(s, m_1, m_2, \dots, m_n, n)$ and its factorial moments. Let $(a)_b = \prod_{i=0}^{b-1} (a - i)$ for $a, b \in \mathbb{Z}_+$ and $\mu_k = \mathbf{E}\Big(\big(U(s, m_1, m_2, \dots, m_n, n)\big)_k\Big)$.

THEOREM (1). *We have*

$$\mathbf{P}\big(U(s, m_1, m_2, \dots, m_n, n) = k\big) = \binom{s}{k} \sum_{i=0}^{\min\{m_1,\dots,m_n\}-k} (-1)^{-i} \binom{s-k}{i} \prod_{l=1}^{n} \left[\frac{\binom{s-k-i}{m_l-k-i}}{\binom{s}{m_l}}\right], \tag{1}$$

$$k = 0, 1, \dots, \min\{m_1, m_2, \dots, m_n\},$$

$$\mu_k = (s)_k \prod_{i=1}^{n} \left(\frac{(m_i)_k}{(s)_k}\right). \tag{2}$$

P r o o f. We use the inclusion-exclusion formula for the events B_j "j is in all n sampled groups", $j = 1, 2, \dots, s$. We obtain

$$\mathbf{P}(\text{exactly } k \text{ of } B_1, \dots, B_s \text{ occur}) = \sum_{i=0}^{s-k} (-1)^i \binom{k+i}{i} \sum_{1 \le j_1 < \cdots < j_{k+i} \le s} \mathbf{P}\big(B_{j_1} \cap \cdots \cap B_{j_{k+i}}\big). \tag{3}$$

Since the n drawings are independent,

$$\mathbf{P}(B_{j_1} \cap \cdots \cap B_{j_l}) = \begin{cases} \prod_{i=1}^{n} \left[\binom{s-l}{m_i-l} / \binom{s}{m_i}\right] & \text{if } l \in \{0, \dots, \min\{m_1, m_2, \dots, m_n\}\} \\ 0 & \text{if } l > \min\{m_1, m_2, \dots, m_n\}. \end{cases} \tag{4}$$

It follows from (3) and (4) that

$$\mathbf{P}\big(U(s, m_1, m_2, \dots, m_n, n) = k\big) = \sum_{i=0}^{\min\{m_1,\dots,m_n\}-k} (-1)^i \binom{k+i}{i} \binom{s}{k+i} \prod_{j=1}^{n} \left[\frac{\binom{s-k-i}{m_j-k}}{\binom{s}{m_j}}\right]$$

which, after a slight simplification, yields (1).

To prove (2), note that the conditional distribution of $U(s,m_1,m_2,\dots,m_n,n)$ given $U(s,m_1,m_2,\dots,m_{n-1},n-1)$ is hypergeometric. Setting $U(n) = U(s,m_1,m_2,\dots,m_n,n)$ we have

$$\mathbf{P}\big(U(n)=k \mid U(n-1)=j\big) = \binom{j}{k}\frac{\binom{s-j}{m_n-k}}{\binom{s}{m_n}}. \tag{5}$$

Thus,

$$\mathbf{E}\Big(\big(U(n)\big)_k \mid U(n-1)=j\Big) = (j)_k(m_n)_k/(s)_k \tag{6}$$

(see Johnson and Kotz [4; p. 81]) and, by (6),

$$\begin{aligned}\mathbf{E}\Big(\big(U(n)\big)_k\Big) &= [(m_n)_k/(s)_k]\mathbf{E}\Big(\big(U(n-1)\big)_k\Big)\\ &= \prod_{i=2}^{n}[(m_i)_k/(s)_k]\mathbf{E}\Big(\big(U(1)\big)_k\Big)\\ &= \prod_{i=2}^{n}[(m_i)_k/(s)_k](m_1)_k\,.\end{aligned}$$

□

3. Asymptotic results

For obtaining asymptotic results on $U(s,m_1,m_2,\dots,m_n,n)$ the following theorem is crucial.

THEOREM (2). *Let $s \longrightarrow \infty$ and $m_i \longrightarrow \infty$ for all $i=1,\dots,n$ such that $\lambda_i = \lim_{s\to\infty}\frac{m_i}{s}$ exists and $\lambda_i \in (0,1)$. Let $\gamma_n = \prod_{i=1}^{n}\lambda_i^{\,2}\Big(\sum_{i=1}^{n}(1-\lambda_i^{-1})-1\Big)+\prod_{i=1}^{n}\lambda_i$. Then*

$$\gamma_n = \lim_{s\to\infty} s^{-1}\mathbf{Var}\big(U(s,m_1,m_2,\dots,m_n,n)\big) \tag{7}$$

and for every $\varepsilon>0$,

$$s^{1-\varepsilon}\Big(s^{-1}U(s,m_1,m_2,\dots,m_n,n)-\prod_{i=1}^{n}(m_i/s)\Big) \longrightarrow 0 \quad \text{in } L_2, \quad \text{as } s\longrightarrow\infty. \tag{8}$$

Proof. It follows from (2) that

$$\mathbf{E}\big(U(s,m_1,m_2,\dots,m_n,n)\big) = s\prod_{i=1}^{n}(m_i/s) \tag{9}$$

and

$$\begin{aligned}
&\mathbf{Var}(U(s, m_1, m_2, \dots, m_n, n)) \\
&= \mu_2 + \mu_1 - \mu_1^2 \\
&= \frac{\prod_{i=1}^{n} m_i(m_i - 1)}{s^{n-1}(s-1)^{n-1}} + \frac{\prod_{i=1}^{n} m_i}{s^{n-1}} - \frac{\prod_{i=1}^{n} m^2}{s^{2n-2}} \\
&= s\left[\prod_{i=1}^{n}\left(\frac{m_i}{s}\right)^2 (s-1)\prod_{i=1}^{n}\frac{1-m_i^{-1}}{1-s^{-1}} + \prod_{i=1}^{n}\frac{m_i}{s} - s\prod_{i=1}^{n}\left(\frac{m_i}{s}\right)^2\right].
\end{aligned} \tag{10}$$

Since

$$\lim_{s\to\infty} s\prod_{i=1}^{n}\left(\left(\frac{1-m_i^{-1}}{1-s^{-1}}\right) - 1\right) = \sum_{i=1}^{n}(1-\lambda_i^{-1}),$$

equation (10) yields

$$s^{-1}\mathbf{Var}(U(s, m_1, m_2, \dots, m_n, n)) \longrightarrow \prod_{i=1}^{n}\lambda_i{}^2\sum_{i=1}^{n}(1-\lambda_i^{-1}) - \prod_{i=1}^{n}\lambda_i^2 + \prod_{i=1}^{n}\lambda_i,$$

proving (7). Assertion (8) now follows from $\mathbf{E}\Big(\Big[s^{-1}U(s, m_1, m_2, \dots, m_n, n) - \prod_{i=1}^{n}(m_i/s)\Big]^2\Big) = \mathbf{Var}(s^{-1}U(s, m_1, m_2, \dots, m_n, n)) \implies s^{-1}\gamma_n$, as $s \longrightarrow \infty$. □

THEOREM (3). *Under the conditions in Theorem (2), if $\lambda = \lambda_1 = \dots = \lambda_n$, let $m_1 = m_2 = \dots = m_n = m = m(s)$ as function of s and $U(s, m_1, m_2, \dots, m_n\ n) = U(s, m, n)$. The asymptotic behavior of $U(s, m, n)$ is*

$$s^{-1/2}\big(U(s, m, n) - (m^n/s^{n-1})\big) \longrightarrow N(0, \alpha_n) \quad \textit{in distribution, as} \quad s \longrightarrow \infty \tag{11}$$

where $\alpha_n = \lambda^{2n}\big(n(1-\lambda^{-1}) - 1\big) + \lambda^n = \lim_{s\to\infty} s^{-1}\mathbf{Var}(U(s, m, n))$.

Proof. First consider the case $n = 2$. Let $(X_1^{(s)}, \dots, X_s^{(s)})$ be an equidistributed random permutation of the numbers $1, \dots, s$. A short reflection shows that the random variable

$$\mathbf{V}_s = \#\{i \in \{1, \dots, m\} : X_i^{(s)} \le m\}$$

has the same distribution of $U(s, m, 2)$. Define the number $a^{(s)}(i)$ for $s \in \mathbb{N}$ and $i = 1, \dots, s$ by

$$a^{(s)}(i) = \begin{cases} \big[(s-m)/ms\big]^{1/2}, & i = 1, \dots, m, \\ -\big[m/(s(s-m))\big]^{1/2}, & i = m+1, \dots, s \end{cases}$$

and introduce the $D[0,1]$-valued process $\mathbf{V}^{(s)} = \big(\mathbf{V}^{(s)}(t)\big)_{t\in[0,1]}$ by setting

$$\mathbf{V}^{(s)}(t) = \sum_{i=1}^{[st]} a^{(s)}\big(X_i^{(s)}\big)\,, \qquad t \in [0,1]\,.$$

Since

$$\sum_{i=1}^{s} a^{(s)}(i) = 0\,, \qquad \sum_{i=1}^{s} \big(a^{(s)}(i)\big)^2 = 1$$

and

$$\max_{1\le i\le s} \big|a^{(s)}(i)\big| \longrightarrow 0 \qquad \text{as} \quad s \longrightarrow \infty\,,$$

we can use a functional limit theorem for exchangeable random variables (see Aldous (1985)) and conclude that $\mathbf{V}^{(s)}$ converges in distribution to a Brownian bridge $\mathbf{W}^0 = \big(\mathbf{W}^0(t)\big)_{t\in[0,1]}$. This yields

$$\mathbf{V}^{(s)}\big(\tfrac{m}{s}\big) \xrightarrow{D} \mathbf{W}^0(\lambda) \qquad \text{as} \quad s \longrightarrow \infty\,. \tag{12}$$

But a short computation shows that

$$\begin{aligned} \mathbf{V}^{(s)}\big(\tfrac{m}{s}\big) &= \big(\tfrac{s-m}{sm}\big)^{1/2}\mathbf{V}_s - \Big(\tfrac{m}{s(s-m)}\Big)^{1/2}(m - \mathbf{V}_s) \\ &= \Big(\mathbf{V}_s - s\big(\tfrac{m}{s}\big)^2\Big)\Big/\Big(\tfrac{m}{s}\big(1-\tfrac{m}{s}\big)s\Big)^{1/2} \\ &\overset{D}{=} \Big(U(s,m,n) - s\big(\tfrac{m}{s}\big)^2\Big)\Big/\Big(\tfrac{m}{s}\big(1-\tfrac{m}{s}\big)s\Big)^{1/2}. \end{aligned} \tag{13}$$

As $\mathbf{W}^0(\lambda)$ is $N\big(0,\lambda(1-\lambda)\big)$-distributed, $m/s \longrightarrow \lambda$, and $\alpha_2 = \lambda^2(1-\lambda)^2$, we obtain assertion (11) for $n = 2$ from (13) and (12).

We can now proceed by complete induction. Suppose that the theorem holds for some $n \ge 2$. Then let $\mathbf{Y}_s$ be a random variable that has the same distribution as $U(s,m,n)$ and is independent of $\big(X_1^{(s)},\dots,X_s^{(s)}\big)$ and thus also of the process $\big(\mathbf{V}^{(s)}(t)\big)_{t\in[0,1]}$. Then as $\mathbf{V}^{(s)} \xrightarrow{D} \mathbf{W}^0$, the induction hypothesis implies that

$$\bigg(\mathbf{V}^{(s)}(t) + \Big(\tfrac{m}{s(s-m)}\Big)^{1/2}\big[\mathbf{Y}_s - s(m/s)^n\big]\bigg)_{t\in[0,1]} \xrightarrow{D} \big(\mathbf{W}^0(t) + \mathbf{Z}\big)_{t\in[0,1]}\,, \tag{14}$$

where $\mathbf{Z}$ is $N\big(0,\frac{\lambda}{1-\lambda}\alpha_n\big)$-distributed and independent of $\mathbf{W}^0$. Let us take for t in (14) the random time $t = \mathbf{Y}_s/s$. By Theorem (2), $\mathbf{Y}_s/s \longrightarrow \lambda^n$ in L_2 so that

$$\mathbf{V}^{(s)}\Big(\tfrac{\mathbf{Y}_s}{s}\Big) + \Big(\tfrac{m}{s(s-m)}\Big)^{1/2}\big[\mathbf{Y}_s - s(m/s)^n\big] \xrightarrow{D} \mathbf{W}^0(\lambda^n) + \mathbf{Z}\,. \tag{15}$$

The limiting distribution in (15) is

$$N\left(0, \lambda^n(1-\lambda^n)\right) * N\left(0, \tfrac{\lambda}{1-\lambda}\alpha_n\right) = N\left(0, \tfrac{\alpha_{n+1}}{\lambda(1-\lambda)}\right).$$

To evaluate the left-hand side of (15), let

$$\tilde{\mathbf{V}}_s = \#\{i:\ 1 \le i \le \mathbf{Y}_s\,,\ X_i^{(s)} \le \mathbf{Y}_s\}.$$

Then

$$\begin{aligned} \mathbf{V}^{(s)}\left(\tfrac{\mathbf{Y}_s}{s}\right) &= \sum_{i=1}^{\mathbf{Y}_s} a^{(s)}\left(X_i^{(s)}\right) \\ &= \tilde{\mathbf{V}}_s\left(\tfrac{s-m}{sm}\right)^{1/2} - \left(\mathbf{Y}_s - \tilde{\mathbf{V}}_s\right)\left(\tfrac{m}{s(s-m)}\right)^{1/2} \end{aligned}$$

so that

$$\begin{aligned} &\mathbf{V}^{(s)}\left(\tfrac{\mathbf{Y}_s}{s}\right) + \left(\tfrac{m}{s(s-m)}\right)^{1/2}\left[\mathbf{Y}_s - s\left(\tfrac{m}{s}\right)^n\right] \\ &= \tilde{\mathbf{V}}_s\left(\left(\tfrac{s-m}{sm}\right)^{1/2} + \left(\tfrac{m}{s(s-m)}\right)^{1/2}\right) - \left(\tfrac{m}{s(s-m)}\right)^{1/2} s\left(\tfrac{m}{s}\right)^n \\ &= \frac{s^{-1/2}\left(\tilde{\mathbf{V}}_s - s\left(\frac{m}{s}\right)^{n+1}\right)}{\left(\frac{m}{s}\left(1-\frac{m}{s}\right)\right)^{1/2}}. \end{aligned} \tag{16}$$

Now note that by (5),

$$\tilde{\mathbf{V}}_s \overset{D}{=} \#\{i:\ 1 \le i \le \mathbf{Y}_s\,,\ X_i^{(s)} \le \mathbf{Y}_s\} \overset{D}{=} U(s,m,n+1). \tag{17}$$

Combining (15), (16) and (17) we obtain (11) for $n+1$ instead of n. The theorem is proved. □

THEOREM (4). *Assume that* $s\prod_{i=1}^{n}\left(\frac{m_i}{s}\right) \longrightarrow \alpha$ *as* $s \longrightarrow \infty$, *for some* $\alpha > 0$ *and some* $n \in \mathbb{N}$.

Then $U(s, m_1, m_2, \dots, m_n, n)$ *converges to the Poisson distribution* π_α *with mean* α.

Proof. By (2), the factorial moments μ_k of $U(s, m_1, m_2, \dots, m_n, n)$ converge to the corresponding factorial moments α^k of π_α. Thus, the moments of $U(s, m_1, m_2, \dots, m_n, n)$ also converge to the corresponding moments of π_α. Since π_α is the only distribution on $\mathbb{R}$ with this moment sequence, $U(s, m_1, m_2, \dots, m_n, n)$ must converge to π_α in distribution. □

4. The waiting times

The number of balls occurring in all n sampled groups $U(s, m_1, m_2, \dots, m_n, n)$, $n = 1, 2, \dots$, may be considered as a system observed at time points n. The system will be said *to be in state* i if and only if $U(s, m_1, m_2, \dots, m_n, n) = i$, $i = 0, 1, \dots, \min\{m_1, m_2, \dots, m_n\}$. The transition probabilities are

$$\begin{aligned} p_{ij}(n) &= P\Big(U(s, m_1, m_2, \dots, m_n, n) = j \mid U(s, m_1, m_2, \dots, m_{n-1}, n-1) = i\Big) \\ &= \frac{\binom{i}{j}\binom{s-i}{m_n - j}}{\binom{s}{m_n}}, \qquad j = 0, 1, \dots, i, \end{aligned}$$

and the initial distribution $P\big(U(s, m_1, m_2, \dots, m_n, 0) = m_1\big) = 1$, and $P\big(U(s, m_1, m_2, \dots, m_n, 0) = i\big) = 0$, $i = 0, 1, \dots, m_1 - 1$. This completely defines a discrete nonhomogeneous Markov chain with states $0, 1, \dots, \min\{m_1, m_2, \dots, m_n\}$. We denote by τ_k (the waiting time) the random variable representing the number of drawings necessary, after which at most k elements will be observed in all drawings. The probability distribution of τ_k is

$$\begin{aligned} &\mathbf{P}(\tau_k = n+1) \\ &= \sum_{h=k+1}^{m} \binom{s}{h} \prod_{l=1}^{n-1} \binom{s}{m_l}^{-1} \sum_{i=0}^{m-h} (-1)^i \binom{s-h}{i} \cdot \\ &\qquad \cdot \prod_{j=1}^{n-1} \binom{s-h-i}{m_j - h - i} \left[1 - \binom{s}{m_n}^{-1} \binom{s-h-i}{m_n - h - i}\right] \end{aligned}$$

where the random variables τ_k and $U(s, m_1, m_2, \dots, m_n, n)$ are related to each other in an obvious way for the distribution function

$$\mathbf{P}(\tau_k > n) = \mathbf{P}\big(U(s, m_1, m_2, \dots, m_n, n) > k\big)$$

and, by (1), the result follows.

The absorbtion time into state zero.

Denote by τ the first entrance time into state zero, i.e., $\tau = \min\{n : U(s, m_1, m_2, \dots, m_n, n) = 0\}$. Then

$$\mathbf{P}(\tau = n) = \sum_{j=1}^{m} (-1)^j \binom{s}{j} \left[\prod_{r=1}^{n-1} \frac{\binom{s-j}{m_r - j}}{\binom{s}{m_r}}\right] \left[\frac{\binom{s-j}{m_n - j}}{\binom{s}{m_n}} - 1\right]$$

using the facts that

$$\begin{aligned} &\mathbf{P}(\tau = n) \\ &= \mathbf{P}(\tau \le n) - \mathbf{P}(\tau \le n-1) \\ &= \mathbf{P}\big(U(s, m_1, m_2, \dots, m_n, n) = 0\big) - \mathbf{P}\big(U(s, m_1, m_2, \dots, m_n, n-1) = 0\big) \end{aligned}$$

and

$$\mathbf{P}\big(U(s,m_1,m_2,\dots,m_n,n)=0\big)=\sum_{j=0}^{m}(-1)^j\binom{s}{j}\left[\prod_{r=1}^{n}\frac{\binom{s-j}{m_r-j}}{\binom{s}{m_n}}\right]$$

the result follows. Also, the pth factorial moments of τ are

$$\mathbf{E}\big((\tau)_p\big)=p!\prod_{r=1}^{p-1}\frac{\binom{s-j}{m_r-j}}{\binom{s}{m_r}}\sum_{j=1}^{m}(-1)^{j+1}\binom{s}{j}\sum_{n=0}^{\infty}\prod_{i=p}^{n+p-1}\frac{\binom{s-j}{m_i-j}}{\binom{s}{m_i}}\left[\frac{\binom{s-j}{m_{p+n}-j}}{\binom{s}{m_{p+n}}-1}\right].$$

REFERENCES

[1] ALDOUS, D. J.: *Exchangeability and related topics.* In: Ecole d'Ete de Probabilites de Saint-Flour XIII – 1983 (P. L. Hennequin, ed.), Lecture Notes in Math. 1117, Springer, New York, 1985, pp. 1–199.

[2] HARRIS, B.- MARDEN, M.—PARK, C. J.: *The distribution of the number of empty cells in a generalized random allocation scheme.* In: Random Graphs'85 (M. Karonski, Z. Palka, eds.), Ann. Discrete Math. 33, North-Holland, Amsterdam, 1987, pp. 77–90.

[3] HOLST, L.: *On birthday, collector's, occupance, committee and other classical urn problems*, Internat. Statist. Rev. **54** (1986), 15–27.

[4] JOHNSON, N. L.—KOTZ, S.: *Urn Models and Their Application*, Wiley, New York, 1977.

[5] KOLCHIN, V. F.—SEVASTYANOV, B. A.—CHISTYAKOV, V. P.: *Random Allocations*, Winston, Washington D.C., 1978.

[6] MIKHAILOV, V. G.: *Asymptotic normality of the number of empty cells for group allocation of particles*, Theory Probab. Appl. **25** (1980), 82 90.

[7] PARK, C. J.: *On the distribution of the number of unobserved elements when m-samples of size n are drawn from a finite population*, Comm. Statist. Theory Methods **10** (1981), 371 383.

[8] STADJE, W.: *The Collector's problem with group drawings*, Adv. Appl. Probab. **22** (1990), 866–882.

[9] STADJE, W.—MITWALLI, S.: *An occupancy problem with group drawings I.* In: The 23rd International Conference for Statistics, Computer Science, and its Applications, Cairo, Egypt, May 9–14, 1998.

Received December 28, 2000
Revised June 25, 2001

Department of Mathematics
Faculty of Science at Damiatta
Mansoura University
P.O. 34517(3) New Damiatta
EGYPT