Journal of Statistical Planning and Inference 140 (2010) 2204–2212

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Competition between discrete random variables, with applications to occupancy problems

Julia Eaton [a], Anant P. Godbole [b,*], Betsy Sinclair [c]

[a] Department of Mathematics, University of Washington, Seattle, WA 98195, USA
[b] Department of Mathematics, East Tennessee State University, Johnson City, TN 37614, USA
[c] Department of Political Science, University of Chicago, Chicago, IL 60637, USA

ARTICLE INFO

Available online 20 January 2010

Keywords:
Occupancy problems
Poisson approximations

ABSTRACT

Consider n players whose "scores" are independent and identically distributed values $\{X_i\}_{i=1}^n$ from some discrete distribution F. We pay special attention to the cases where (i) F is geometric with parameter $p \to 0$ and (ii) F is uniform on $\{1, 2, \ldots, N\}$; the latter case clearly corresponds to the classical occupancy problem. The quantities of interest to us are, first, the U-statistic W which counts the number of "ties" between pairs i, j; second, the univariate statistic Y_r, which counts the number of strict r-way ties between contestants, i.e., episodes of the form $X_{i_1} = X_{i_2} = \cdots = X_{i_r}; X_j \neq X_{i_1}; j \neq i_1, i_2, \ldots, i_r$; and, last but not least, the multivariate vector $Z_{AB}=(Y_A, Y_{A+1},\ldots,Y_B)$. We provide Poisson approximations for the distributions of W, Y_r and Z_{AB} under some general conditions. New results on the joint distribution of cell counts in the occupancy problem are derived as a corollary.

1. Introduction

In this paper we hope to shed new light on an old problem, studied extensively in, e.g., Barbour et al. (1992) and Kolchin et al. (1978). Consider n players whose "scores" are independent and identically distributed values $\{X_i\}_{i=1}^n$ from some discrete distribution F. We consider the case of general distributions F but pay special attention to the cases where (i) F is geometric with parameter $p \to 0$ and (ii) F is uniform on $\{1, 2, \ldots, N\}$; the latter case corresponds to the classical occupancy problem. The quantities of interest to us are

- the U-statistic W which counts the number of "ties" between pairs i, j (with $X_a=X_b=X_c=X_d$, for example, leading to a contribution of $\binom{4}{2} = 6$ to the value of W);
- the univariate statistic Y_r which counts the number of strict r-way ties between contestants, i.e., episodes of the form $X_i=x$ for some x iff $i \in A$, $|A| = r$; and
- the multivariate vector $Z_{AB}=(Y_A,Y_{A+1},\ldots,Y_B)$.

We provide Poisson approximations for the distributions of W, Y_r and Z_{AB} under some general conditions. New results on the joint distribution of cell counts in the occupancy problem are derived as a corollary.

Consider the following elementary problem from Warner and Kline (1999): "Two players use a coin that lands heads with probability p to play a game that consists of a sequence of rounds. In each round, the first player tosses the coin until a head appears. Then the second player tosses the coin until a head appears. If the players have the same number of flips in a round, the round is declared a tie and another round is played. If not, the player with the larger number of flips wins the

* Corresponding author.
E-mail address: godbolea@etsu.edu (A.P. Godbole).

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.01.016

game. Rounds are played successively until one of the two players wins the game." Readers are asked to find the expected number of rounds; the expected value of the total number of flips; and the probability distribution of the difference between the number of flips made by players 1 and 2 in a given round. We briefly mention the solution for the first two of these questions: the probability of a two person tie is clearly

$$\sum_{x=1}^{\infty}(1-p)^{2x-2}p^2=\frac{p}{2-p},$$

so that $\mathbb{E}(R)$, the expected number of rounds is given by

$$\mathbb{E}(R)=\sum_{x=1}^{\infty}x(\mathbb{P}(\text{tie}))^{x-1}(1-\mathbb{P}(\text{tie}))=\sum_{x=1}^{\infty}x\left(\frac{p}{2-p}\right)^{x-1}\left(1-\frac{p}{2-p}\right)=\frac{2-p}{2-2p},$$

so that Wald's lemma yields for $\mathbb{E}(F)$, the expected total number of flips,

$$\mathbb{E}(F)=\mathbb{E}(F/R)\mathbb{E}(R)=\frac{2-p}{2-2p}\,\mathbb{E}(F/R)=\frac{2(2-p)}{p(2-2p)},$$

since the expected number $\mathbb{E}(F/R)$ of flips per round is clearly $2/p$. Computations for a three-person game, not mentioned in Warner and Kline (1999), are similar, but we need to lay down some rules as follows: three players each flip a p-coin until heads is flipped. The player with the highest number of flips wins unless there are ties between two or more players, in which case we repeat the process. That is, the value of each of the three geometric variables in question must be unique. We next compute the probability of a two- or three-way tie; the expected number of rounds; and the expected number of flips for $n=3$—to convince the reader that the situation rapidly becomes quite complicated as n increases. [The authors had a lively discussion with Lloyd Douglas, NSF Program Officer, about the following "real-life" application of the n-person model with $p\to 0$. We wish to rank n of the greatest free-throw shooters (or slam dunkers, or, etc.) in the National Basketball Association. The players each shoot free throws until they miss—conditional on the fact that no two players miss on the same attempt. Rankings are then awarded in the obvious fashion.]

With three players, whose outcomes in each round are denoted by the independent Geometric(p) random variables A, B and C, there are $3!=6$ ways for a round to end with no two players tied, and seven ways for a round to end with some tie between players—one is a three-way tie (i.e., $A=B=C$); there are $\binom{3}{1}=3$ ways for events both of the form $A>B=C$ and $A=B>C$ to occur. Note that

$$\mathbb{P}(A=B=C)=p^3+p^3(1-p)^3+p^3(1-p)^6\ldots=\sum_{x=1}^{\infty}p^3(1-p)^{3x-3}=\frac{p^2}{3-3p+p^2},$$

while the table below

Case	A	B	C
1	TH, TTH, …	H	H
2	TTH, TTTH, …	TH	TH
3	TTTH, TTTTH, …	TTH	TTH
4	TTTTH, TTTTTH, …	TTTH	TTTH
⋮	⋮	⋮	⋮

reveals that

$$\mathbb{P}(A>B=C)=p^3\sum_{m=1}^{\infty}(1-p)^m\sum_{i=0}^{m-1}(1-p)^{2i}=\frac{p(1-p)}{3-3p+p^2}.$$

Finally, we observe from the table

Case	A	B	C
1	TH	TH	H
2	TTH	TTH	H, TH
3	TTTH	TTTH	H, TH, TTH
4	TTTTH	TTTTH	H, TH, TTH, TTTH
⋮	⋮	⋮	⋮

that

$$\mathbb{P}(A=B>C)=p^3\sum_{m=1}^{\infty}(1-p)^{2m}\sum_{i=0}^{m-1}(1-p)^i=\frac{p(1-p)^2}{(2-p)(3-3p+p^2)},$$

which leads to

$$\mathbb{P}(\text{tie})=\mathbb{P}(A=B=C)+3\mathbb{P}(A>B=C)+3\mathbb{P}(A=B>C)=\frac{5p^3-13p^2+9p}{(2-p)(3-3p+p^2)},$$

and hence as before to

$$\mathbb{E}(R) = \frac{1}{1-\frac{5p^3-13p^2+9p}{(2-p)(3-3p+p^2)}}$$

and

$$\mathbb{E}(F) = \mathbb{E}(F/R)\mathbb{E}(R) = \frac{3}{p}\left(\frac{1}{1-\frac{5p^3-13p^2+9p}{(2-p)(3-3p+p^2)}}\right).$$

Competitions of the kind discussed above are best formulated in the more general context of occupancy models as follows: n balls are independently thrown into an infinite array of boxes so that any ball hits the j th box with probability p_j. Let X_j be the number of balls in box j. Then, with $p_j=(1-p)^{j-1}p$, we have the game inspired by Warner and Kline (1999) ending iff $X_j \le 1\ \forall j$. Extremal versions of such questions have arisen in the literature before, often with surprising results. Motivated by a question, posed by Carl Pomerance and arising in an additive number theory context, Athreya and Fidkowski (2000) proved that the probability π_n that the highest numbered non-empty box has exactly one ball in it converges to a constant (which is shown to be one) iff $\lim_{n\to\infty} p_n/\sum_{j=n}^{\infty} p_j = 0$. This is a condition that is not satisfied by, e.g., the sequence $p_n=1/2^n$ for which, quite interestingly, the limit superior and the limit inferior of the sequence π_n differ in the *fourth* decimal place. These results had been independently obtained a few years earlier by Baryshnikov et al. (1995), Eisenberg et al. (1993), Eisenberg and Stengle (1996) and also by Bruss and O'Cinneide (1990). The comprehensive paper of Móri (2000) is most relevant too: here it is proven that given a double sequence of integer valued random variables, i.i.d. within rows, and letting $\mu(n)$ denote the multiplicity of the maximal value in the n th row, the limiting distribution of $\mu(n)$ does not exist in the ordinary sense—but that the intriguing empirical type a.s. limit result

$$\lim_{t\to\infty}\frac{1}{\log t}\sum_{n=1}^{t}\frac{1}{n}I(\mu(n)=m) = \frac{r^m}{m\log\left(\frac{1}{1-r}\right)}, \quad m=1,2,\ldots$$

holds, where r is a parameter that depends on the distribution. The whole field appears to be extraordinarily rich with known facts and tantalizing possibilities.

Results of the kind described above indicate that the cell counts for the n person (Geometric) coin game are unlikely to behave in an asymptotically smooth way if $p=p_n \nrightarrow 0$. This fact is exhibited in Theorem 1, where we study the distribution of the number W of pairs of equalities in the n person game, with $W=0$ corresponding to the end of a "round" in the sense of Warner and Kline (1999), and show that a good Poisson approximation is obtained if $np\to 0$ (Geometric distribution) or $n/N\to 0$ (Uniform distribution). Theorem 2 concerns itself with the distribution of the number Y_r of strict r-way ties (=the number of boxes with exactly r balls) and Theorem 3 with a multivariate generalization of Theorem 2. The approximating distribution is Poisson (Theorem 2) or a product of independent Poisson variates (Theorem 3). We note, moreover, that we were able to prove a result such as Theorem 3 relatively easily *probably* due to the approach taken—we use as a counter the event that r *specific* balls go into the same urn, rather than the conventional approach (e.g., Barbour et al., 1992, Section 6.2) of counting the number of urns with r balls. See also Arratia et al. (1989, 1999, 2000, 2003) and Kolchin et al. (1978).

2. Results

For any two discrete random variables T and U, let $d_{\mathrm{TV}}(\mathcal{L}(T),\mathcal{L}(U))$ denote the usual total variation distance between their distributions $\mathcal{L}(T)$ and $\mathcal{L}(U)$, i.e.,

$$d_{\mathrm{TV}}(\mathcal{L}(T),\mathcal{L}(U)) = \sup_{A\subseteq\mathbb{Z}^+}|\mathbb{P}(T\in A)-\mathbb{P}(U\in A)|.$$

Theorem 1. *Let $\{X_j\}_{j=1}^n$ be an integer valued sequence of i.i.d. random variables with $\mathbb{P}(X_1=i)=p_i$, and consider the U-statistic*

$$W = \sum_{K=1}^{\binom{n}{2}} I_K,$$

where, with $\mathcal{K}$ denoting the Kth 2-subset of $\{1,2,\ldots,n\}$, $I_K=1$ if $X_i=X_j; i,j\in\mathcal{K}$ ($I_K=0$ otherwise). Let $\mathrm{Po}(\lambda)$ be the Poisson distribution with parameter $\lambda=\mathbb{E}(W)$. Then

$$d_{\mathrm{TV}}(\mathcal{L}(W),\mathrm{Po}(\lambda)) \le 2n\pi + \frac{2n\rho}{\pi},$$

where $\pi=\mathbb{P}(X_1=X_2)$ and $\rho=\mathbb{P}(X_1=X_2=X_3)$.

Proof. The proof is an elementary application of, e.g., Theorem 2.C.5 in Barbour et al. (1992), which yields with $\lambda = \mathbb{E}(W) = \binom{n}{2}\pi$

$$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \le \frac{1-e^{-\lambda}}{\lambda}\left(\sum_K \mathbb{P}^2(I_K=1) + \sum_K \sum_{\{L:K\cap L\neq\emptyset\}} \{\mathbb{E}(I_K I_L)+\pi^2\}\right) \le \pi + \frac{2(n-2)\rho}{\pi} + 2(n-2)\pi \le 2n\pi + \frac{2n\rho}{\pi},$$

as asserted. □

In Theorem 1, if the variables are uniform over $\{1,2,\ldots,N\}$, then $\pi = 1/N$; $\rho = 1/N^2$, so that $d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \le 4n/N \to 0$ if $N \gg n$, where, throughout this paper we write for $f_n, g_n \ge 0$, $f_n \ll g_n$ (or $g_n \gg f_n$) if $f_n/g_n \to 0$ as $n\to\infty$. If the variables are Geometric(p), then the discussion in Section 1 yields $\pi = p/(2-p)$ and $\rho = p^2/(3-3p+p^2)$, so that we get $d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \le 2np/(2-p) + 2np(2-p)/(3-3p+p^2) \le 6np \to 0$ if $p \ll 1/n$. For the n person game discussed in Section 1, we thus get

$$\mathbb{P}(W=0) = \mathbb{P}(\text{no ties}) = \exp\{-(n(n-1)p)/(2(2-p))\} \pm 6np = \exp\{-\lambda\} \pm 6np,$$

$$\mathbb{E}(R) = \frac{1}{\mathbb{P}(W=0)} = \frac{1}{e^{-\lambda} \pm 6np},$$

and

$$\mathbb{E}(F) = \frac{n}{p}\mathbb{E}(R).$$

The random variable W, while providing us with some insight, does not yield the level of detail that we desire. For this reason, we turn our attention next to the variable Y_r that counts the "number of strict r-way ties," or, in other words, the "number of boxes with exactly r balls." The development that follows is alternative to that provided, say, in Barbour et al. (1992), Theorems 6.C, 6.E, and particularly 6.F, though we do not make too many comparisons between our results and those of Barbour et al. (1992), since our main focus will be on the multivariate Theorem 3; the strategy of looking at specific sets of r players is what sets our method apart.

Letting as before $\{X_j\}_{j=1}^n$ be an integer valued sequence of i.i.d. random variables with $\mathbb{P}(X_1 = i) = p_i$, we denote by

$$\pi = \sum_{x=1}^{\infty} p_x^r (1-p_x)^{n-r}$$

the probability that a specific set of r players are involved in a strict tie, and thus

$$\lambda = \binom{n}{r}\pi$$

is the expected number of boxes with exactly r balls. Throughout this paper we will employ, as in the previous sentence, the dual analogies of "balls in boxes" and "ties between contestants." It may be readily verified that $\pi = (1/N^{r-1})(1-N^{-1})^{n-r}$ in the uniform case, and that in the geometric case $\pi = \sum_{x=1}^{\infty}(1-p)^{rx-r}p^r(1-(1-p)^{x-1}p)^{n-r}$ may be estimated as follows; the estimates may be seen to be tight provided that $p\to 0$. First we have

$$\pi = \sum_{x=1}^{\infty}(1-p)^{rx-r}p^r(1-(1-p)^{x-1}p)^{n-r} \ge \sum_{x=1}^{\infty}(1-p)^{rx-r}p^r[1-(n-r)(1-p)^{x-1}p] = \sum_{x=1}^{\infty}(1-p)^{rx-r}p^r - (n-r)\sum_{x=1}^{\infty}(1-p)^{rx-r}p^r(1-p)^{x-1}p \ge \frac{p^{r-1}}{r} - \frac{2(n-r)p^r}{(r+1)(2-rp)},$$

where the above inequalities follow since $(1-p)^r \ge 1-rp$ and $(1-p)^{r+1} \le 1-(r+1)p+[(r+1)rp^2]/2$. Next note that

$$\begin{aligned}\pi &= \sum_{x=1}^{\infty}(1-p)^{rx-r}p^r(1-(1-p)^{x-1}p)^{n-r} \le \sum_{x=1}^{\infty}(1-p)^{rx-r}p^r \\ &\quad\times\left(1-(n-r)(1-p)^{x-1}p + \frac{(n-r)(n-r-1)}{2}(1-p)^{2x-2}p^2\right) \\ &= \frac{p^r}{1-(1-p)^r} - \frac{(n-r)p^{r+1}}{1-(1-p)^{r+1}} + \frac{(n-r)(n-r-1)}{2}\frac{p^{r+2}}{1-(1-p)^{r+2}} \\ &\le \frac{2p^{r-1}}{r(2-(r-1)p)} - \frac{(n-r)p^r}{r+1} + \frac{(n-r)^2p^{r+1}}{(r+2)(2-(r+1)p)},\end{aligned}$$

so that in the geometric case, $\pi \sim p^{r-1}/r$ provided that $np^{(r+1)/2}\to 0$; $rp\to 0$.

We shall use the coupling approach as in Barbour et al. (1992) to show that $\mathcal{L}(Y_r)$ may be closely approximated by a Poisson distribution with the same mean. We need to first find, given a sum $\sum_{j=1}^n I_j$ of indicator variables, a sequence $\{J_{ij}\}$ of indicator variables, defined on the same probability space as the I_j's, so that for each j,

$$\mathcal{L}(J_{1j}, J_{2j}, \ldots, J_{nj}) = \mathcal{L}(I_1, I_2, \ldots, I_n | I_j = 1). \tag{1}$$

Good error bounds on a Poisson approximation are obtained if the J_{ji}'s are chosen in a fashion that makes them "not too far apart" from the I_j's. We proceed in a manner similar to that in Theorem 6.F in Barbour et al. (1992), but the coupling we use

is conditional, thus imparting a different flavor to the argument: we have

$$Y_r = \sum_{j=1}^{\binom{n}{r}} I_j,$$

where I_j=1 if the j th r-set is engaged in a strict tie. Now we define the indicator variable I_{jx} as being one if and only if I_j=1 and the members of the j th r-set all have "value" x. Now we proceed as follows: If I_{jx}=1, we "do nothing", setting J_i=I_i for all i. If, however, I_{jx}=0, we move all members of the j th r set into the x th box (some of these might of course have occupied the x th box to begin with), while ejecting all its "illegal" occupants and moving each of these independently with probability $p_k/(1-p_x)$ to box k; $k\neq x$. Finally we set J_i=J_{ijx}=1 if the i th r set is involved in a strict tie *after* this interchange. We need to verify that (1) holds in the modified form

$$\mathcal{L}(J_{1jx}, J_{2jx}, \ldots, J_{\binom{n}{r}jx}) = \mathcal{L}(I_1, I_2, \ldots, I_{\binom{n}{r}} | I_{jx} = 1), \tag{2}$$

while this may be viewed as being "obvious," we provide a proof next. To show that (2) holds, it clearly suffices to show that any configuration (or sample point) corresponding to the members of the j th r-set being "the only occupants of the x th box" is equally likely under both the conditional and unconditional models in (2). This strategy will achieve more, in fact, since we will not have to verify a condition similar to (2) when we move on to the multivariate case.

We let a_i denote the score of the i th player not in the r-clique in question ($a_i\neq x$), and b_i the score of the i th player in the r-clique, so that b_i=x. Now,

$$\mathbb{P}(\text{configuration} | I_{jx} = 1) = \frac{\mathbb{P}(\text{configuration})}{\mathbb{P}(I_{jx}=1)} = \frac{\mathbb{P}(a_1, a_2, \ldots, a_{n-r}, b_{n-r+1}, \ldots, b_n)}{p_x^r (1-p_x)^{n-r}} = \frac{\mathbb{P}(a_1)\mathbb{P}(a_2)\ldots\mathbb{P}(a_{n-r})}{(1-p_x)^{n-r}}.$$

Note also that the probability of the configuration under the coupled model is given by

$$\sum_{l=0}^{r} \binom{r}{l} p_x^l [1-p_x]^{r-l} \sum_{S\subseteq\{1,\ldots,n-r\}} p_x^{|S|} \prod_{j\in\{1,2,\ldots,n-r\}\setminus S} \mathbb{P}(a_j) \prod_{j\in S} \frac{\mathbb{P}(a_j)}{1-p_x}. \tag{3}$$

Now

$$\sum_{S\subseteq\{1,\ldots,n-r\}} p_x^{|S|} \prod_{j\in\{1,2,\ldots,n-r\}\setminus S} \mathbb{P}(a_j) \prod_{j\in S} \frac{\mathbb{P}(a_j)}{1-p_x} = \left(\frac{1}{1-p_x}\right)^{n-r} \prod_{j=1}^{n-r} \mathbb{P}(a_j),$$

which shows that (3) yields the same expression as before. This proves the claim.

Now Theorem 2.B in Barbour et al. (1992) leads to the following inequality:

$$d_{\mathrm{TV}}(\mathcal{L}(Y_r), \mathrm{Po}(\lambda)) \leq \left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_j \sum_x \mathbb{P}(I_{jx}=1) \left\{ \mathbb{P}(I_j=1) + \sum_{i\neq j} \mathbb{P}(I_i \neq J_{ijx}) \right\} \leq \pi + \left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_j \sum_x \mathbb{P}(I_{jx}=1) \sum_{i\neq j} \mathbb{P}(I_i \neq J_{ijx}), \tag{4}$$

where $\lambda = \mathbb{E}(Y_r)$, $\pi = \mathbb{P}(I_j=1)$, and the coupled sequence $\{J_i\} = \{J_{ijx}\}$ satisfies (2) for each j and x. Consider first the case $\mathbb{P}(I_i=0, J_{ijx}=1)$, which is clearly impossible when $|i\cap j|\geq 1$, and which we shall call Case I. We thus have for $|i\cap j| = 0$,

$$\sum_{i\neq j} \mathbb{P}(I_i=0, J_{ijx}=1) = \binom{n-r}{r} \sum_{y\neq x} \mathbb{P}(I_i=0, J_{ijx}=1, y), \tag{5}$$

where the summand $\mathbb{P}(I_i=0, J_{ijx}=1, y)$ represents the probability that the i th r-set is not engaged in a strict r-way tie before the coupling, but is part of such a tie with common value y after the coupling. Thus, relation (5) yields

$$\sum_{i\neq j} \mathbb{P}(I_i=0, J_{ijx}=1) = \sum_{i\neq j} \sum_{y\neq x} \{\mathbb{P}(J_{ijx}=1, y) - \mathbb{P}(I_i=1, J_{ijx}=1, y)\} = \binom{n-r}{r} \sum_{y\neq x} \mu_y,$$

where

$$\begin{aligned}\mu_y &= \sum_{q=0}^{r} \binom{r}{q} p_y^q p_x^{r-q} \sum_{s=0}^{n-2r} \binom{n-2r}{s} p_x^s (1-p_x-p_y)^{n-2r-s} \left(\frac{p_y}{1-p_x}\right)^{r-q} \cdot \left(\frac{1-p_x-p_y}{1-p_x}\right)^s - (1-p_y)^r p_y^r \sum_{s=0}^{n-2r} \binom{n-2r}{s} p_x^s (1-p_x-p_y)^{n-2r-s} \left(\frac{1-p_x-p_y}{1-p_x}\right)^s \\ &= p_y^r \left(1-\frac{p_y}{1-p_x}\right)^{n-2r} \left\{ \left(\frac{1}{1-p_x}\right)^r - (1-p_y)^r \right\}. \end{aligned} \tag{6}$$

We now check to see the nature of the bound (6) in the uniform and geometric cases: when the balls are distributed uniformly in N boxes, we see that (6) leads to

$$\sum_{i\neq j} \mathbb{P}(I_i=0, J_{ijx}=1) = \binom{n-r}{r} \sum_{y\neq x} \mu_y \leq \binom{n}{r} \cdot N \cdot \frac{1}{N^r} \left(1-\frac{1}{N}\right)^{n-2r} \left\{ \left(\frac{N}{N-1}\right)^r - \left(\frac{N-1}{N}\right)^r \right\} = \lambda \left(1-\frac{1}{N}\right)^{-2r} \left\{ 1 - \left(1-\frac{1}{N}\right)^{2r} \right\} \leq \frac{2r}{N} \exp\{2r/(N-1)\}\lambda, \tag{7}$$

while in the geometric case we have

$$\begin{aligned}\sum_{i\neq j}\mathbb{P}(I_i=0,J_{ijx}=1)&=\binom{n-r}{r}\sum_{y\neq x}p_y^r\left(1-\frac{p_y}{1-p_x}\right)^{n-2r}\left\{\frac{1}{(1-p^x)^r}-(1-p_y)^r\right\}\\&\leq\binom{n}{r}\sum_{y=1}^{\infty}p_y^r(1-p_y)^{n-r}(1-p_y)^{-r}((1-p_x)^{-r}-(1-p_y)^r)\leq\lambda\max_{x,y}\{(1-p_x)^{-r}(1-p_y)^{-r}-1\}\\&\leq\lambda\left(\frac{1}{1-p}\right)^{2r}\{1-(1-p)^{2r}\}\leq\lambda\exp\{2rp/(1-p)\}2rp.\end{aligned}\tag{8}$$

We now consider the case where $I_i=1$ and $J_{ijx}=0$ (Case II). We clearly have $\mathbb{P}(I_i=1,J_{ijx}=0)=\mathbb{P}(I_i=1)$ for $|i\cap j|\geq 1$ (Case II′), so we obtain

$$\sum_{|i\cap j|\geq 1}\mathbb{P}(I_i=1,J_{ijx}=0)\leq r\binom{n-1}{r-1}\pi=\frac{r^2}{n}\lambda.\tag{9}$$

Next assume (Case II″) that $|i\cap j|=0$, and we seek to estimate the probability $\mathbb{P}(I_i=1,J_{ijx}=0)$. If $y=x$, we bound $\mathbb{P}(I_i=1,J_{ijx}=0)$ by π_x, so that

$$\sum_{|i\cap j|=0}\mathbb{P}(I_{ix}=1,J_{ijx}=0)\leq\binom{n-r}{r}\pi_x.\tag{10}$$

If, on the other hand, $y\neq x$, then we have

$$\sum_{|i\cap j|=0}\mathbb{P}(I_i=1,J_{ijx}=0)=\binom{n-r}{r}\sum_{y\neq x}\mathbb{P}(I_{iy}=1,J_{ijx}=0)\leq\binom{n-r}{r}\sum_{y\neq x}\sum_{q\geq 1}\binom{n-2r}{q}p_x^qp_y^r\frac{qp_y}{1-p_x};$$

the above equation follows since in order for $I_{iy}=1,J_{ijx}=0$ to occur, we must have at least one of the q "bad" balls present in urn x land in urn y and thus "spoil" the fact that $I_{iy}=1$. We thus get

$$\sum_{|i\cap j|=0}\mathbb{P}(I_i=1,J_{ijx}=0)\leq\binom{n-r}{r}\sum_{y\neq x}\frac{p_y^{r+1}}{1-p_x}\sum_{q\geq 1}\binom{n-2r}{q}qp_x^q\leq\binom{n-r}{r}n\frac{p_x}{1-p_x}(1+p_x)^{n-2r-1}\sum_{y\neq x}p_y^{r+1}\leq\binom{n-r}{r}n\frac{p_x}{1-p_x}e^{np_x}\sum_{y\neq x}p_y^{r+1}.\tag{11}$$

Now (11) reduces in the uniform case to

$$\binom{n-r}{r}\frac{n}{N-1}e^{n/N}\frac{N-1}{N^{r+1}}\leq\frac{n}{N^2}\lambda\frac{e^{n/N}}{\left(1-\frac{1}{N}\right)^{n-r}}\leq\frac{n}{N^2}\lambda e^{n/N}\exp\{(n-r)/(N-1)\}\leq\frac{n}{N^2}\lambda e^{2n/(N-1)}\tag{12}$$

and is bounded in the geometric case by

$$\binom{n-r}{r}\frac{np^{r+2}}{1-p}\sum_{y\neq x}(1-p)^{(y-1)(r+1)}\leq\lambda np^2e^{np}(1+o(1)),\tag{13}$$

provided that $np^{(r+1)/2}\to 0, rp\to 0$. Eqs. (4), (6), (9), (10) and (11) now yield

$$\begin{aligned}d_{\mathrm{TV}}(\mathcal{L}(Y_r),\mathrm{Po}(\lambda))&\leq\pi+\left(\frac{1-e^{-\lambda}}{\lambda}\right)\sum_j\sum_x\mathbb{P}(I_{jx}=1)\sum_{i\neq j}\mathbb{P}(I_i\neq J_{ijx})\\&\leq\pi+\left(1\wedge\frac{1}{\lambda}\right)\sum_j\sum_x\mathbb{P}(I_{jx}=1)\times\left[\binom{n-r}{r}\sum_{y\neq x}p_y^r\left(1-\frac{p_y}{1-p_x}\right)^{n-2r}\left\{\left(\frac{1}{1-p_x}\right)^r-(1-p_y)^r\right\}+\frac{r^2}{n}\lambda\right.\\&\left.+\binom{n-r}{r}\pi_x+\binom{n-r}{r}n\frac{p_x}{1-p_x}e^{np_x}\sum_{y\neq x}p_y^{r+1}\right].\end{aligned}\tag{14}$$

We next evaluate (14) in the uniform case: Eqs. (4), (7), (9), (10) and (12) give

$$\begin{aligned}d_{\mathrm{TV}}(\mathcal{L}(Y_r),\mathrm{Po}(\lambda))&\leq\pi+\left(1\wedge\frac{1}{\lambda}\right)\sum_j\sum_x\mathbb{P}(I_{jx}=1)\sum_{i\neq j}\mathbb{P}(I_i\neq J_{ijx})\\&\leq\pi+\left(1\wedge\frac{1}{\lambda}\right)\lambda\times\left\{\frac{2r}{N}\lambda\exp\{2r/(N-1)\}+\frac{r^2}{n}\lambda+\binom{n-r}{r}\frac{\pi}{N}+\frac{n}{N^2}\lambda\exp\{2n/(N-1)\}\right\}\\&=\pi+(\lambda\wedge\lambda^2)\times\left\{\frac{2r}{N}\exp\{2r/(N-1)\}+\frac{r^2}{n}+\frac{1}{N}+\frac{n}{N^2}\exp\{2n/(N-1)\}\right\}.\end{aligned}\tag{15}$$

We compare (15) with Eq. (6.2.18) in Barbour et al. (1992), which yields the upper bound

$$d_{\mathrm{TV}}(\mathcal{L}(Y_r),\mathrm{Po}(\lambda))\leq(\lambda\wedge\lambda^2)\left\{\frac{1}{N}+\frac{6n}{N^2}+\frac{6r^2}{n}\right\};$$

it is evident that (15) provides a better estimate if

$$\frac{2r}{N}\exp\{2r/(N-1)\} \le \frac{5r^2}{n} + (6-\exp\{2n/(N-1)\})\frac{n}{N^2},$$

which is a condition that holds under a wide range of circumstances, and certainly if $n/N \to 0$. Now in the geometric case, Eqs. (4), (8), (9), (10), and (13) reveal that (14) reduces as follows:

$$\begin{aligned} d_{\mathrm{TV}}(\mathcal{L}(Y_r), \mathrm{Po}(\lambda)) &\le \pi + \left(1 \wedge \frac{1}{\lambda}\right) \sum_j \sum_x \mathbb{P}(I_{jx} = 1) \times \left(2\lambda rp \exp\{2rp/(1-p)\} + \frac{r^2}{n}\lambda + \binom{n}{r}\pi_x + \lambda np^2\right) \\ &\le \pi + (\lambda \wedge \lambda^2)\left\{2rp\exp\{2rp/(1-p)\} + \frac{r^2}{n} + rp + np^2 e^{np}\right\}. \end{aligned} \tag{16}$$

We have thus proved the following result.

Theorem 2. *Let $\{X_j\}_{j=1}^n$ be an integer valued sequence of i.i.d. random variables with $\mathbb{P}(X_1 = i) = p_i$. Define Y_r to be the number of strict r-way ties between these random variables. Then the total variation distance between $\mathcal{L}(Y_r)$ and a Poisson distribution with the same mean is given by* (14). *This expression reduces to the one in Eq.* (15) *when the distribution of the X_i's is uniform on $\{1, 2, \ldots, N\}$ and to the expression in Eq.* (16) *when $X_1 \sim \mathrm{Geo}(p)$.*

For the rest of the paper we will, for simplicity, restrict our attention to the classical occupancy problem of n balls in N boxes, assuming furthermore that $n/N \to 0$. The goal is to obtain a multivariate Poisson approximation for the vector $Z_{AB} = \{Y_A, Y_{A+1}, \ldots, Y_B\}$, for suitably restricted A and B, and where the approximating Poisson vector consists of independent components. First consider the quantities $\{\lambda_a : A \le a \le B\}$. Since

$$\lambda_a = \binom{n}{a}\left(\frac{1}{N}\right)^{a-1}\left(1-\frac{1}{N}\right)^{n-a} \sim \frac{1}{\sqrt{2\pi a}}\mathrm{N}\left(\frac{ne}{Na}\right)^a \exp\{(n-a)/N\} \sim \frac{1}{\sqrt{2\pi a}}\mathrm{N}\left(\frac{ne}{Na}\right)^a (1+\mathrm{o}(1)),$$

it follows, due to the fact that $n/N \to 0$, that λ_a is monotone decreasing in a. Suppose that $\lambda_A < \infty$ for some finite A. It then follows that the approximating Poisson distribution for Y_{A+1} would have mean close to zero, making our agenda somewhat uninteresting. We shall assume therefore that $\lambda_A \to \infty$ as $n, N \to \infty$. Choices of the parameters that make this occur might be, e.g., $n = N^\alpha$; $\alpha < 1$, when $\mathbb{E}(Y_a) \to 0$ for all $a \ge A_0$, or, more interestingly, $n = N/\log N$ in which case the threshold A_0 would tend to infinity with N. We thus seek values of A and B for which we get an "interesting" multivariate Poisson approximation for the ensemble $(Y_A, \ldots, Y_B)$. Using the notation suggested in the proof of Theorem 2, Theorem 10.J of Barbour et al. (1992) yields

$$d_{\mathrm{TV}}\left(\mathcal{L}(Y_A, \ldots, Y_B), \prod_{a=A}^B \mathrm{Po}(\lambda_a)\right) \le \sum_{a=A}^B \sum_{j=1}^{\binom{n}{a}} \sum_{x=1}^N \mathbb{P}(I_{ajx} = 1)\left\{\mathbb{P}(I_{aj} = 1) + \sum_{biy \ne ajx} \mathbb{P}(I_{biy} \ne J_{biy})\right\}, \tag{17}$$

where the last sum does not include the case $bi{=}aj$. Correspondingly, we let T_1, T_2, T_3 denote the quantities

$$\sum_{a=A}^B \sum_{j=1}^{\binom{n}{a}} \sum_{x=1}^N \mathbb{P}(I_{ajx} = 1)\mathbb{P}(I_{aj} = 1), \sum_{a=A}^B \sum_{j=1}^{\binom{n}{a}} \sum_{x=1}^N \mathbb{P}(I_{ajx} = 1) \sum_{iy \ne jx} \mathbb{P}(I_{aiy} \ne J_{aiy})$$

and

$$\sum_{a=A}^B \sum_{j=1}^{\binom{n}{a}} \sum_{x=1}^N \mathbb{P}(I_{ajx} = 1) \sum_{b \ne a} \sum_{i=1}^{\binom{n}{b}} \sum_{y=1}^N \mathbb{P}(I_{biy} \ne J_{biy}),$$

respectively; we need to compute the sum $T_1 + T_2 + T_3$. First, we see that

$$T_1 = \sum_a \sum_j \mathbb{P}^2(I_{aj} = 1) = \sum_a \binom{n}{a}\frac{1}{N^{2a-2}}\left(1-\frac{1}{N}\right)^{2n-2a} \le N^2 \sum_{a=A}^B \left(\frac{ne}{aN^2}\right)^a \le N^2 \sum_{a=2}^B \left(\frac{ne}{2N^2}\right)^a = \frac{e^2}{4}\frac{n^2}{N^2}(1+\mathrm{o}(1)) \to 0 \tag{18}$$

for each A, B. The computation of T_2 follows as in the proof of Theorem 2. The first component, T_{21} is, by (7), given by

$$T_{21} = \sum_{a=A}^B \sum_j \sum_x \mathbb{P}(I_{ajx} = 1) \times \frac{2a}{N}\exp\{2a/(N-1)\}\lambda_a \le \sum_a \lambda_a^2 \frac{2a}{N}(1+\mathrm{o}(1)). \tag{19}$$

Under what circumstances might the bound in (19) tend to zero? Let us pause to consider this question before continuing. If $n = \sqrt{N\log N}$ and $A{=}2$, then $\lambda_A \sim (\log N)/2$, the error bound of Theorem 2 is of magnitude $\sqrt{\log N/N}$, and the bound in (19) *does* approach zero. However in this case $\lambda_3 \to 0$ so we are able to derive little useful beyond a Poisson approximation for Y_2, the number of "days" with exactly two "birthdays". If $n{=}N^{0.9}$, then $\lambda_a \sim N^{1-0.1a} \to \infty$ for all $a{=}2,3,\ldots,9$ but the summands in (19), asymptotically equal to $N^{1-0.2a}$, tend to zero only if $a{=}6,7,8,9$. We thus have a potential multivariate approximation for (Y_6, Y_7, Y_8, Y_9). Finally, let $n = N/\log N$. In this case, $\lambda_a \sim (e/a\log N)^a \cdot N$, and, with $a = \log N/(2\log\log N)$,

for example, we see that

$$\lambda_a \sim \left(\frac{e}{a\log N}\right)^a \cdot N = \left(\frac{2e\log\log N}{\log^2 N}\right)^{\log N/2\log\log N} \cdot (e^{2\log\log N})^{\log N/2\log\log N} = (2e\log\log N)^{\log N/2\log\log N} \to \infty,$$

while with $a = \log N/[(4-\varepsilon)\log\log N]$ (we use $a = \log N/(3\log\log N)$ below) we have

$$\frac{\lambda_a^2}{N} \sim \left(\frac{e}{a\log N}\right)^{2a} \cdot N = \left(\frac{3e\log\log N}{\log^{1/2} N}\right)^{2\log N/3\log\log N},$$

which leads, with $A = \log N/(3\log\log N)$ and $B = \log N/(2\log\log N)$, to

$$T_{21} \le 2(B-A)\frac{\lambda_A^2}{N} \le \frac{\log N}{3\log\log N}\left(\frac{3e\log\log N}{\log^{1/2} N}\right)^{2\log N/3\log\log N} \to 0;$$

we thus have a potential Poisson approximation for the vector $(Y_A, \ldots, Y_B)$. Next note that the term T_{22} that corresponds to (9) is given by

$$T_{22} = \sum_{a=A}^{B}\sum_{j=1}^{\binom{n}{a}}\sum_{x=1}^{N}\mathbb{P}(I_{ajx}=1)\frac{a^2}{n}\lambda_a = \sum_a \frac{a^2}{n}\lambda_a^2. \tag{20}$$

Finally we combine the two remaining terms (10) and (12), as reflected in (15), to get

$$T_{23} = \sum_{a=A}^{B}\sum_{j=1}^{\binom{n}{a}}\sum_{x=1}^{N}\mathbb{P}(I_{ajx}=1)\left(\frac{\binom{n}{a}\pi_a}{N} + \frac{n}{N^2}\lambda_a \exp\{2n/(N-1)\}\right) = \sum_a \frac{\lambda_a^2}{N} + \frac{n}{N^2}\lambda_a^2(1+o(1)). \tag{21}$$

Turning to a computation of T_3, we first observe that for $a \ne b$ and $|i\cap j| \ge 1$, it is impossible for $I_{biy}=0$, $J_{biy}=1$ to occur. Accordingly, as in the calculation leading up to (6) we see that

$$\begin{aligned}
&\sum_b\sum_i\sum_y \mathbb{P}(I_{biy}=0, J_{biy}=1)\\
&= \sum_b\sum_i\sum_y \{\mathbb{P}(J_{biy}=1) - \mathbb{P}(I_{biy}=1, J_{biy}=1)\}\\
&= \sum_b \binom{n-a}{b}\sum_y\left\{\sum_{q=0}^{b}\binom{b}{q}\left(\frac{1}{N}\right)^q\left(\frac{1}{N}\right)^{b-q}\right.\\
&\times\sum_s\binom{n-a-b}{s}\left(\frac{1}{N}\right)^s\left(1-\frac{2}{N}\right)^{n-a-b-s}\left(\frac{1}{N-1}\right)^{b-q}\left(\frac{N-2}{N-1}\right)^s\\
&\left.-\left(\frac{N-1}{N}\right)^a\left(\frac{1}{N}\right)^b\sum_s\binom{n-a-b}{s}\left(\frac{1}{N}\right)^s\left(1-\frac{2}{N}\right)^{n-a-b-s}\left(\frac{N-2}{N-1}\right)^s\right\}\\
&= \sum_b \binom{n-a}{b} N\cdot\times\left\{\left(\frac{1}{N-1}\right)^b\left(\frac{N-2}{N-1}\right)^{n-a-b} - \left(\frac{N-1}{N}\right)^a\left(\frac{1}{N}\right)^b\left(\frac{N-2}{N-1}\right)^{n-a-b}\right\}\\
&\le \sum_b \lambda_b\frac{(a+b)}{N}(1+o(1)).
\end{aligned} \tag{22}$$

As in the univariate case, $\mathbb{P}(I_{biy}=1, J_{biy}=0) = \mathbb{P}(I_{biy}=1)$ if $|i\cap j| \ge 1$. Hence

$$\sum_b\sum_{|i\cap j|\ge 1}\sum_y \mathbb{P}(I_{biy}=1, J_{biy}=0) \le \sum_b a\binom{n-1}{b-1}\pi_b = \sum_b \frac{ab}{n}\lambda_b. \tag{23}$$

If, however, $|i\cap j| = 0$, then

$$\sum_b\sum_{|i\cap j|=0}\mathbb{P}(I_{bix}=1, J_{bix}=0) \le \sum_b\binom{n-a}{b}\pi_{bx}, \tag{24}$$

and, being rather crude with the final estimation

$$\begin{aligned}
&\sum_b\sum_{|i\cap j|=0}\sum_{y\ne x}\mathbb{P}(I_{biy}=1, J_{biy}=0)\\
&\le \sum_b\binom{n-a}{b}\frac{1}{N^b}\sum_y\sum_{q\ge 1}\binom{n-a-b}{q}\frac{1}{N^q}\frac{q}{N-1}\\
&\le \sum_b\binom{n}{b}\frac{n}{N^{b+1}}\sum_{q\ge 1}\binom{n-1}{q-1}\frac{1}{N^{q-1}}(1+o(1))
\end{aligned}$$

$$\leq \sum_b \frac{n}{N^2}\lambda_b(1+o(1)). \tag{25}$$

Collecting Eqs. (18)–(25), we see that the following holds:

Theorem 3. *When n balls are randomly assigned to N boxes, where $n \ll N$, the joint distribution of the multivariate vector $(Y_A,\ldots,Y_B)$ of exact box counts may be approximated by a Poisson vector with independent components. More specifically,*

$$d_{\mathrm{TV}}\left(\mathcal{L}(Y_A,\ldots,Y_B), \prod_{a=A}^{B} \mathrm{Po}(\lambda_a)\right) \leq \varepsilon_{n,N,A,B},$$

where

$$\lambda_a = \mathbb{E}(Y_a) = \binom{n}{a}\frac{1}{N^{a-1}}\left(1-\frac{1}{N}\right)^{n-a}$$

and $\varepsilon_{n,N,A,B}$ is of magnitude

$$\sum_{a=A}^{B}\lambda_a^2\left(\frac{2a}{N}+\frac{a^2}{n}+\frac{1}{N}+\frac{n}{N^2}\right)+\sum_a \lambda_a\left(\sum_{b\neq a}\lambda_b\left(\frac{(a+b)}{N}+\frac{ab}{n}+\frac{1}{N}+\frac{n}{N^2}\right)\right).$$

In addition, an application, e.g., of Theorem 10.K in Barbour et al. (1992) may provide slight improvements in the above, through a partial reinstatement of the so-called "magic factor".

Acknowledgments

The research of all three authors was supported by NSF Grants DMS-0049015 and DMS-0552730, and was conducted at East Tennessee State University, when Eaton and Sinclair were undergraduate students at Rochester University and the University of Redlands, respectively. The research of Godbole was further supported by JHU's Acheson J. Duncan Fund for the Advancement of Research in Statistics.

References

Arratia, R., Barbour, A., Tavaré, S., 1999. On Poisson–Dirichlet limits for random decomposable combinatorial structures. Combin. Probab. Comput. 8, 193–208.
Arratia, R., Barbour, A., Tavaré, S., 2000. Limits of logarithmic combinatorial structures. Ann. Probab. 28, 1620–1644.
Arratia, R., Barbour, A., Tavaré, S., 2003. Logarithmic Combinatorial Structures: A Probabilistic Approach. European Mathematical Society, Zürich.
Arratia, R., Goldstein, L., Gordon, L., 1989. Two moments suffice for Poisson approximation: the Chen–Stein method. Ann. Probab. 17, 9–25.
Athreya, J., Fidkowski, L., 2000. Number theory, balls in boxes, and the asymptotic uniqueness of maximal discrete order statistics, Electron. J. Combin. Number Theory 0, Paper A3, available at ⟨http://www.integers-ejcnt.org/vol0.html⟩.
Barbour, A., Holst, L., Janson, S., 1992. Poisson Approximation. Oxford University Press, Oxford.
Baryshnikov, Y., Eisenberg, B., Stengle, G., 1995. A necessary and sufficient condition for the existence of the limiting probability of a tie for first place. Statist. Probab. Lett. 23, 203–209.
Bruss, F., O'Cinneide, C., 1990. On the maximum and its uniqueness for geometric random samples. J. Appl. Probab. 22, 598–610.
Eisenberg, B., Stengle, G., Strang, G., 1993. The asymptotic probability of a tie for first place. Ann. Appl. Probab. 3, 731–745.
Eisenberg, B., Stengle, G., 1996. Minimizing the probability of a tie for first place. J. Math. Anal. Appl. 198, 458–472.
Kolchin, V., Sevast'yanov, B., Chistyakov, V., 1978. Random Allocations. Winston, Washington, DC.
Móri, T., 2000. On the multiplicity of the sample maximum and the longest head run. Period. Math. Hungar. 41, 195–212.
Warner, B., Kline, B., 1999. Problem number 662. College Mathematics Journal 30, 407.