The Annals of Probability
1977, Vol. 5, No. 6, 1028–1035

SOME ASYMPTOTIC RESULTS FOR OCCUPANCY PROBLEMS[1]

BY LARS HOLST

University of Wisconsin, Madison

Suppose n balls are placed into N cells with arbitrary probabilities. Limit distributions for the number of empty cells are considered when $N \to \infty$ and $n \to \infty$ in such a way that $n/N \to \infty$. Limit distributions for the number of balls to achieve exactly b empty cells are obtained for $N \to \infty$ with b fixed and for $b \to \infty$ with $b/N \to 0$.

1. Introduction. Suppose that balls are thrown independently of each other into N cells so that each ball has probability p_k of falling into the kth cell, $p_1 + \cdots + p_N = 1$. Let Y_n denote the number of empty cells after n throws and let T_b denote the throw on which for the first time exactly b cells remain empty, $0 \leqq b < N$. The symmetrical case $p_1 = \cdots = p_N = 1/N$ is discussed for example, in Feller (1968) under occupancy or waiting time problems. For an expository paper on these and related problems, see Kolchin and Chistyakov (1974).

Depending on how $b, n, N \to \infty$, different asymptotic distributions for Y_n and T_b can be obtained; see, for example, Holst (1971) and, for the symmetric case, Samuel-Cahn (1974). In this paper some remaining problems are investigated for the nonsymmetrical case.

To give precise meanings for the limits obtained, double sequences $(p_{kN})_N$, $(Y_{nN})_N$ are considered. But in order to simplify the notation the extra index N will usually be omitted.

2. A bounded number of empty cells. The following limit theorem for Y_n, the number of empty cells after n throws, was proved by Sevastyanov (1972).

THEOREM 1. *If the p's are such that*

$$\max_{1 \leqq k \leqq N} (1 - p_k)^n \to 0 \tag{2.1}$$

and

$$E(Y_n) = \sum_{k=1}^{N} (1 - p_k)^n \to m < \infty\,, \tag{2.2}$$

then

$$P(Y_n = y) \to m^y \cdot e^{-m}/y!\,, \tag{2.3}$$

or equivalently

$$Y_n \Rightarrow Po(m)\,, \qquad \textit{when} \quad N \to \infty\,. \tag{2.4}$$

Received April 8, 1976; revised February 24, 1977.

[1] Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

AMS 1970 *subject classifications.* Primary 60F05; Secondary 60C05.

Key words and phrases. Occupancy problems, coupon collectors problem, limit theorems.

REMARK. When the p's are equal an expression for $P(Y_n = y)$ can be obtained from which (2.3) can be derived by elementary methods; see Feller (1968). In this case (2.1) and (2.2) are replaced by

$$N \cdot \exp(-n/N) \to m < \infty . \tag{2.5}$$

For T_b, the number of balls until b empty cells remain, we have:

THEOREM 2. *If b is a fixed integer and for some fixed numbers C and D*

$$0 < C \leqq Np_k \leqq D < \infty , \qquad \text{for all} \quad k \quad \text{and} \quad N , \tag{2.6}$$

then, when $N \to \infty$,

$$\textstyle\sum_{k=1}^{N} (1 - p_k)^{T_b} \Rightarrow \frac{1}{2}\chi^2(2(b+1)) , \tag{2.7}$$

and

$$\textstyle\sum_{k=1}^{N} \exp(-T_b p_k) \Rightarrow \frac{1}{2}\chi^2(2(b+1)) . \tag{2.8}$$

Before we prove the theorem let us consider the functions

$$f(t) = f_N(t) = \textstyle\sum_{k=1}^{N} (1 - p_k)^t , \qquad t > 0 , \tag{2.9}$$

and

$$g(t) = g_N(t) = \textstyle\sum_{k=1}^{N} \exp(-tp_k) . \tag{2.10}$$

LEMMA 1. *If condition (2.6) is satisfied, $y > 0$ is a fixed number, and $t = t_N = t(y)$ is defined by the equation*

$$f(t) = y , \tag{2.11}$$

then

$$0 < C \leqq \liminf_{N\to\infty} N(\log N)/t_N \leqq \limsup_{N\to\infty} N(\log N)/t_N \leqq D < \infty , \tag{2.12}$$

and when $N \to \infty$

$$f([t]) \to y , \tag{2.13}$$

$$\max_{1\leqq k\leqq N} (1 - p_k)^{[t]} \to 0 , \tag{2.14}$$

$$g(t) \quad \text{and} \quad g([t]) \to y , \tag{2.15}$$

where $[t]$ denotes the integer part of t.

LEMMA 2. *If f is replaced by g and g by f in Lemma 1, then the same conclusions hold.*

PROOF OF LEMMA 1. From condition (2.6), it follows that

$$y = \textstyle\sum_{k=1}^{N} (1 - p_k)^t \geqq N \cdot (1 - D/N)^t . \tag{2.16}$$

Hence for $\varepsilon > 0$ and N sufficiently large,

$$\log y \geqq \log N - t \cdot (D + \varepsilon)/N , \tag{2.17}$$

and therefore

(2.18) $$D + \varepsilon = (D + \varepsilon)\lim_{N\to\infty} (1/(1 - \log y/\log N)) \geqq \limsup_{N\to\infty} N \log N/t_N ,$$

which proves the right inequality of (2.12).

To prove the left inequality of (2.12), note that by (2.6)

(2.19) $$\log y \leqq \log N - t \log (1 - C/N) \leqq \log N - tC/N .$$

From this it follows that

(2.20) $$C = C \lim_{N\to\infty} (1 - \log y/\log N)^{-1} \leqq \liminf_{N\to\infty} N \log N/t_N .$$

Using (2.6) and (2.11) we get

(2.21) $$(1 - D/N)^{-1} y \geqq f([t]) \geqq y ,$$

which proves (2.13).

Combining (2.6) and (2.12) shows that, for some $K > 0$ and N sufficiently large,

(2.22) $$\max (1 - p_k)^{[t]} \leqq (1 - C/N)^{[t]} \leqq (1 - C/N)^{KN \log N} \to 0 , \quad N \to \infty ,$$

proving (2.14).

From (2.6) and (2.12) it follows that for some constant K

(2.23) $$|1 - e^{-tp_k}/(1 - p_k)^t| \leqq K \cdot \log N/N ,$$

and therefore

(2.24) $$|f(t) - g(t)| \leqq \sum_1^N (1 - p_k)^t \cdot |1 - e^{-tp_k}/(1 - p_k)^t| \leqq K \sum_1^N (1 - p_k)^t (\log N)/N = Ky(\log N)/N \to 0 ,$$

which proves (2.15).

Proof of Lemma 2. The proof is essentially the same as that for Lemma 1.

Proof of Theorem 2. From the definitions it follows that

(2.25) $$Y_n \leqq b \Leftrightarrow T_b \leqq n ,$$

and therefore

(2.26) $$P(Y_n \leqq b) = P(T_b \leqq n) = P(f(T_b) \geqq f(n)) .$$

Let $y > 0$ be fixed and define $n = [t]$ with $t = t(y)$ as in Lemma 1. According to Lemma 1 the assumptions of Theorem 1 are satisfied. Hence

(2.27) $$P(f(T_b) \geqq y) = P(Y_n \leqq b) \to P(Y \leqq b) ,$$

where Y is $Po(y)$. Furthermore it is well-known that

(2.28) $$P(Y \leqq b) = P(\tfrac{1}{2}\chi^2(2(b + 1)) \geqq y) ;$$

(2.27) and (2.28) prove (2.7). From Lemma 2 the assertion (2.8) follows.

REMARK. When the p's are equal the theorem can be written

$$N \cdot (1 - 1/N)^{T_b} \Rightarrow \tfrac{1}{2}\chi^2(2(b+1)), \tag{2.29}$$

and therefore

$$T_b/N - \log N \Rightarrow \log(\tfrac{1}{2}\chi^2(2(b+1))). \tag{2.30}$$

This result was found by Baum and Billingsley (1965) using complicated calculations. From the result in Feller (1968) and the method of proof of Theorem 2, (2.29) and (2.30) follow. A consequence of (2.30) is

$$T_b/N \log N \to 1 \quad \text{in probability as} \quad N \to \infty. \tag{2.31}$$

Now (2.31) will be generalized. First introduce the distribution function

$$H_N(x) = \#(p_k;\, Np_k \leqq x)/N. \tag{2.32}$$

LEMMA 3. *If* $t = t_N = t(y)$ *is defined by*

$$g(t) = g_N(t_N) = y > 0, \tag{2.33}$$

and there exists a distribution function $H(x)$ *on* $[C, D]$ *such that*

$$H_N(x) \to H(x), \qquad N \to \infty, \tag{2.34}$$

and

$$0 < C = \inf\{x;\, H(x) > 0\}, \tag{2.35}$$

then for $1/C > \varepsilon > 0$,

$$g_N((\varepsilon + 1/C)(N \log N)) \to 0, \tag{2.36}$$

and

$$g_N((-\varepsilon + 1/C)(N \log N)) \to +\infty \tag{2.37}$$

as $N \to \infty$.

PROOF. From the definitions it follows that

$$\begin{aligned} 0 < y = g_N(t_N) &= N \cdot \textstyle\int_C^D \exp(-t_N x/N)\, dH_N(x) \\ &= \textstyle\int_C^D \exp((1 - t_N x/N \log N) \log N)\, dH_N(x). \end{aligned} \tag{2.38}$$

Consider

$$g_N((\varepsilon + 1/C)N \log N) = \textstyle\int_C^D \exp((1 - x(1+\varepsilon C)/C) \log N)\, dH_N(x). \tag{2.39}$$

Now, for $C \leqq x \leqq D$ it is true that $1 - x(1 + \varepsilon C)/C < 0$ and therefore the exponent in (2.39) is negative, so the integral tends to 0 when $N \to \infty$, which proves (2.36).

In a similar way (2.38) can be proved.

COROLLARY (to Theorem 2). *If the conditions* (2.34) *and* (2.35) *are satisfied, then*

$$T_b/N \log N \to 1/C \quad \textit{in probability as} \quad N \to \infty. \tag{2.40}$$

PROOF. Let $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ be given. Take a $\delta > 0$ so that

$$P(\tfrac{1}{2}\chi^2(2(b+1)) < \delta) < \varepsilon_2/2 . \tag{2.41}$$

For N sufficiently large it follows from Theorem 2 that

$$P(g_N(T_b) < \delta) < \varepsilon_2/2 \tag{2.42}$$

and from Lemma 3 that

$$g_N((\varepsilon_1 + 1/C)(N \log N)) < \delta . \tag{2.43}$$

Hence

$$\begin{aligned} P(T_b/N \log N > \varepsilon_1 + 1/C) &= P(g_N(T_b) < g_N((\varepsilon_1 + 1/C)(N \log N))) \\ &\leqq P(g_N(T_b) < \delta) < \varepsilon_2/2 . \end{aligned} \tag{2.44}$$

In a similar way we prove

$$P(T_b/N \log N < -\varepsilon_1 + 1/C) < \varepsilon_2/2 . \tag{2.45}$$

By (2.44) and (2.45) the assertion follows.

3. A small fraction of empty cells. As above, Y_n denotes the number of empty cells after n throws.

THEOREM 3. *If*

$$0 < C \leqq Np_k \leqq D < \infty \qquad \textit{for all} \quad k \quad \textit{and} \quad N , \tag{3.1}$$

$$n/N \to \infty , \tag{3.2}$$

$$f(n) = E(Y_n) = \textstyle\sum_{k=1}^{N} (1 - p_k)^n \to +\infty , \tag{3.3}$$

then when $n \to \infty$,

$$(Y_n - f(n))/(f(n))^{\frac{1}{2}} \Rightarrow N(0, 1) , \tag{3.4}$$

and

$$(Y_n - g(n))/(g(n))^{\frac{1}{2}} \Rightarrow N(0, 1) , \tag{3.5}$$

where

$$g(n) = \textstyle\sum_{k=1}^{N} \exp(-np_k) . \tag{3.6}$$

PROOF. From (3.1) and (3.3) it follows that

$$\textstyle\sum_1^N (1 - p_k)^n \leqq N \cdot (1 - C/N)^n \to +\infty ; \tag{3.7}$$

hence

$$n/N \log N = O(1) . \tag{3.8}$$

By (3.1), (3.2), and (3.8) there exists a constant K such that

$$|f(n) - g(n)| \leqq K \cdot (n/N) \cdot \exp(-Cn/N) \to 0 . \tag{3.9}$$

Hence it is sufficient to prove (3.5). This will be done using a technique similar to that of Karlin (1967).

Let $\{X(t)\}$ be a Poisson process with unit parameter and assume that at each event of the Poisson process we independently choose one of the N cells according to $\{p_k\}_{k=1}^N$ and place a ball in it. For $1 \leqq k \leqq N$ let

(3.10) $$W_k(t) = \text{number of balls in cell } k \text{ at time } t\,.$$

It is well known that for each t, $\{W_k(t)\}$ are independent random variables with $W_k(t)$ distributed as $Po(p_k t)$. Let

(3.11) $$\begin{aligned} I(y) &= 1 \quad \text{if } y = 0 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Define

(3.12) $$Y(t) = \sum_{k=1}^N I(W_k(t))\,.$$

Then $Y(t)$ is the number of empty cells at time t, and by (3.3) and (3.9)

(3.13) $$E(Y(n)) = g(n) = \sum_{k=1}^N e^{-np_k} \to \infty\,.$$

Thus, by the central limit theorem and the independence of the $\{W_k(t)\}$,

(3.14) $$(Y(n) - g(n))/g(n)^{\frac{1}{2}} \Rightarrow N(0, 1)\,.$$

We now need to show for all x

(3.15) $$|P((Y(n) - g(n))/g(n)^{\frac{1}{2}} \leqq x) - P((Y_n - g(n))/g(n)^{\frac{1}{2}} \leqq x)| \to 0$$

as $n \to \infty$. But

(3.16) $$P((Y(n) - g(n))/g(n)^{\frac{1}{2}} \leqq x) = \sum_{j=0}^\infty P((Y_j - g(n))/g(n)^{\frac{1}{2}} \leqq x)e^{-n} \cdot n^j/j!\,.$$

Let $\delta > 0$. Then by the central limit theorem there exists a $A > 0$ such that for all x and n sufficiently large,

(3.17) $$\begin{aligned} &|P((Y(n) - g(n))/g(n)^{\frac{1}{2}} \leqq x) \\ &\quad - \textstyle\sum_{|j-n| \leqq An^{\frac{1}{2}}} P((Y_j - g(n))/g(n)^{\frac{1}{2}} \leqq x)e^{-n} \cdot n^j/j!| < \delta\,. \end{aligned}$$

Let $\varepsilon > 0$ and suppose we can prove that

(3.18) $$\sup_{|j-n| < An^{\frac{1}{2}}} P(|Y_n - Y_j| > \varepsilon(g(n))^{\frac{1}{2}}) = o(1)\,.$$

It then will follow from (3.17) and (3.18) that

(3.19) $$\begin{aligned} &P((Y(n) - g(n))/g(n)^{\frac{1}{2}} < x - \varepsilon) - \delta + o(1) \\ &\quad \leqq P((Y_n - g(n))/g(n)^{\frac{1}{2}} \leqq x) \\ &\quad \leqq P((Y(n) - g(n))/g(n)^{\frac{1}{2}} \leqq x + \varepsilon) + \delta + o(1)\,, \end{aligned}$$

which in turn will establish the theorem by the continuity of the normal distribution. So it remains only to prove (3.18).

Markov's inequality yields

(3.20) $$P(|Y_n - Y_j| > \varepsilon(g(n))^{\frac{1}{2}}) \leqq E|Y_n - Y_j|/\varepsilon(g(n))^{\frac{1}{2}}\,.$$

Assume first that $j > n$, so $j = n + i$, $0 < i < An^{\frac{1}{2}}$. Since $Y_n \geqq Y_j$,

(3.21) $$\begin{aligned} E|Y_n - Y_{n+i}| &= E(Y_n) - E(Y_{n+i}) = \textstyle\sum_{k=1}^N (e^{-np_k} - e^{-(n+i)p_k}) \\ &\leqq \textstyle\sum_{k=1}^N e^{-np_k}(1 - \exp(-An^{\frac{1}{2}}p_k))\,. \end{aligned}$$

Since $\max_k p_k \leqq D/N$ and $n/N \log N = O(1)$, $n^{\frac{1}{2}} \max p_k \to 0$ and so for some constant D_2,

$$\sup_{0 \leqq i \leqq An^{\frac{1}{2}}} E|Y_n - Y_{n+i}| \leqq D_2 g(n) n^{\frac{1}{2}}/N . \tag{3.22}$$

Similarly

$$\sup_{0 \leqq i \leqq An^{\frac{1}{2}}} E|Y_n - Y_{n-i}| \leqq D_2 g(n) n^{\frac{1}{2}}/N . \tag{3.23}$$

Thus

$$\sup_{|k-n| < An^{\frac{1}{2}}} P(|Y_n - Y_k| > \varepsilon(g(n))^{\frac{1}{2}}) \leqq D_2(g(n)n)^{\frac{1}{2}}/\varepsilon N . \tag{3.24}$$

But $g(n) \leqq Ne^{-Cn/N}$ and $n/N \to \infty$, so the right-hand side of (3.24) converges to 0.

This proves (3.18) and completes the proof of the theorem.

4. The waiting time for a small fraction. As above let T_b denote the number of balls thrown until exactly $b = b_N$ cells remain empty. Let t_b be the unique solution of the equation

$$b = g(t_b) = \sum_{k=1}^N \exp(-t_b p_k) . \tag{4.1}$$

THEOREM 4. *If*

$$b_N \to +\infty , \tag{4.2}$$

$$b_N/N \to 0 , \tag{4.3}$$

as $N \to \infty$ *and*

$$0 < C \leqq Np_k \leqq D < \infty , \qquad \textit{for all} \quad k \quad \textit{and} \quad N , \tag{4.4}$$

then

$$b_N^{-\frac{1}{2}}(T_b - t_b) \sum_{k=1}^N p_k \exp(-t_b p_k) \Rightarrow N(0, 1) . \tag{4.5}$$

PROOF. From (4.1) and (4.4) it follows that

$$Cb/N \leqq \Delta = \sum_1^N p_k \exp(-t_b p_k) \leqq Db/N . \tag{4.6}$$

Thus for N sufficiently large

$$0 < C \leqq \Delta \cdot N/b \leqq D < \infty . \tag{4.7}$$

As in the proof of Theorem 2 the relation

$$P((T_b - t_b)\Delta/b^{\frac{1}{2}} \leqq x) = P(Y_n \leqq b) , \tag{4.8}$$

holds, where

$$n = [t_b + xb^{\frac{1}{2}}/\Delta] . \tag{4.9}$$

We have

$$\begin{aligned} g(n)(1 + o(1)) &= g(t_b + xb^{\frac{1}{2}}/\Delta) \\ &= \sum \exp(-t_b p_k) \cdot (1 - xp_k b^{\frac{1}{2}}/\Delta + O(1/b)) \\ &= b - x \cdot b^{\frac{1}{2}} + O(1) , \end{aligned} \tag{4.10}$$

and thus

$$g(n) \to +\infty , \tag{4.11}$$

and from (3.9)

$$f(n) \to +\infty . \tag{4.12}$$

Furthermore,

$$b = g(t_b) \geqq N \exp(-Dt_b/N) , \tag{4.13}$$

so by (4.3) we have

$$t_b/N \to +\infty , \tag{4.14}$$

and therefore

$$n/N \to +\infty . \tag{4.15}$$

Hence the assumptions of Theorem 3 are satisfied. Now (4.8) and (4.10) give

$$\begin{aligned} &P(T_b - t_b)\Delta/b^{\frac{1}{2}} \leqq x) \\ &\quad = P(Y_n \leqq b) = \Phi((b - g(n))/(g(n))^{\frac{1}{2}}) + o(1) \\ &\quad = \Phi((xb^{\frac{1}{2}} + O(1))/(b(1 + o(1)))^{\frac{1}{2}}) + o(1) \to \Phi(x) , \end{aligned} \tag{4.16}$$

where $\Phi(x)$ is the standardized normal distribution function. This proves the theorem.

Acknowledgment. I wish to thank Professor B. Harris for pointing out some errors in the original draft of this manuscript and the referee for giving an improved proof of Theorem 3.

REFERENCES

[1] Baum, L. E. and Billingsley, P. (1965). Asymptotic distributions for the coupon collector's problem. *Ann. Math. Statist.* **36** 1835–1839.

[2] Feller, W. (1968). *An Introduction to Probability Theory and its Applications* **1**, 3rd ed. Wiley, New York.

[3] Holst, L. (1971). Limit theorems for some occupancy and sequential occupancy problems. *Ann. Math. Statist.* **42** 1671–1680.

[4] Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *J. Math. Mech.* **17** 373–401.

[5] Kolchin, V. F. and Chistyakov, V. P. (1974). Combinatorial problems of probability theory. *Itogi Nauki i Tekhniki. Teoriya Veroyatnostei Matematicheskaya Statistika. Teoreticheskaya Kibernetika* **11** 5–45. (Translation in *J. Soviet Mathematics* **4** 217–243.)

[6] Samuel-Cahn, E. (1974). Asymptotic distributions for occupancy and waiting time problems with positive probability of falling through the cells. *Ann. Probability* **2** 515–521.

[7] Sevastyanov, B. A. (1972). Poisson limit law for a scheme of sums of dependent random variables. *Theor. Probability Appl.* **17** 695–699.

Department of Mathematics
Uppsala University
S-75252 Uppsala, Sweden