The Annals of Probability
1974, Vol. 2, No. 3, 515-521

ASYMPTOTIC DISTRIBUTIONS FOR OCCUPANCY AND WAITING TIME PROBLEMS WITH POSITIVE PROBABILITY OF FALLING THROUGH THE CELLS

BY ESTER SAMUEL-CAHN

Hebrew University

Consider N cells into which balls are being dropped independently, in such a way that the cells are equiprobable, and each ball has probability $p_N > 0$ of staying in the cell. Let $W_N(p_N, k_N)$ denote the waiting time until $k_N + 1$ cells are occupied, and let $S_N(p_N, j_N)$ denote the number of distinct cells occupied after j_N balls have been dropped. The full characterization of the limiting distributions of these two random variables is obtained, depending upon the joint behaviour of p_N, k_N and p_N, j_N respectively, as $N \to \infty$. The limit distributions obtained are the negative binomial, binomial, Poisson, chi-square and normal distributions.

1. Introduction and Summary. Consider N cells into which balls are being dropped independently, in such a way that the cells are equiprobable, but each ball has probability $p_N > 0$ of staying in, and probability $1 - p_N$ of falling through the cell. Let $W_N(p_N, k)$ denote the minimal number of balls needed in order that $k + 1$ cells be occupied, and let $S_N(p_N, j)$ denote the number of distinct cells occupied after j balls have been dropped. The above mentioned model has been considered in a recent paper by Park (1972), who proves that for fixed p and under certain conditions on the behavior of j_N, $S_N(p_N, j_N)$ has an asymptotic normal distribution, as $N \to \infty$. In the present paper we give a complete characterization of the limiting behavior of $W_N(p_N, k_N)$ and $S_N(p_N, j_N)$, which depends upon the joint behavior of p_N, k_N and p_N, j_N, respectively as functions of N, when $N \to \infty$.

To abbreviate, we use the following notations: $NB(k, p)$, $B(j, p)$, $Po(\lambda)$, $\chi^2_{(k)}$, $N(0, 1)$ denote the negative binomial, binomial, Poisson, chi-square and Normal distributions, with corresponding parameters. In particular, $Po(0)$ is the degenerate distribution with unit mass at 0. $\to_{\mathcal{L}}$ denotes convergence in law. Let $N \to \infty$. We have

THEOREM 1. (i) *If* $k_N = k$ *fixed, and* $p_N \to p > 0$ *then* $W_N(p_N, k) \to_{\mathcal{L}} NB(k + 1, p)$.

(ii) *If* $p_N \to 0$ *then* $2p_N W_N(p_N, k) \to_{\mathcal{L}} \chi^2_{(2(k+1))}$.

Theorems 2 to 4 are under the assumption $k_N \to \infty$. Set $\lim k_N(1 - p_N) = \mu$, $\lim k_N^2/N = \lambda$.

THEOREM 2. *If* $0 \leqq \mu + \lambda < \infty$ *then* $\{W_N(p_N, k_N) - (k_N + 1)\} \to_{\mathcal{L}} Po(\mu + \lambda/2)$.

Received February 21, 1973; revised September 24, 1973.
AMS 1970 *subject classifications.* 60F05.
Key words and phrases. Asymptotic distribution, occupancy problem, waiting time.

THEOREM 3. *If* $\mu + \lambda = \infty$ *and* $N - k_N \to \infty$ *then* $\{W_N(p_N, k_N) - EW_N(p_N, k_N)\}\{\text{Var } W_N(p_N, k_N)\}^{-\frac{1}{2}} \to_{\mathcal{L}} N(0, 1)$.

THEOREM 4. *If* $k_N = N - b$, $b > 0$ *constant, then*

$$\exp\{-W_N(p_N, k_N)p_N/N + \log(2N)\} \to_{\mathcal{L}} \chi^2_{(2b)} .$$

REMARK TO THEOREMS. It is easily checked that the various conditions on k_N and p_N in Theorems 1 to 4, essentially take care of all possible relationships between k_N and p_N, and hence, in each case, the conditions stated are not only sufficient, but also necessary, for the corresponding result to obtain. (This statement is accurate only if one is willing to look at the corresponding subsequences of k_N and p_N, as one can, without loss of generality.)

Clearly, for $p_N \equiv 1$, one is back in the classical coupon collector's problem, and for this particular case our Theorems 2 to 4 are identical with Theorems 1 to 4 of Baum and Billingsley (1965). Our proof of Theorem 2 is similar to the proofs of the corresponding theorems in Baum and Billingsley (1965). Since our proofs of Theorems 3 and 4 are different, and presumably simpler than those of Baum and Billingsley, they may be of interest even in the simple case $p_N \equiv 1$.

The random variables $W_N(p_N, k)$ and $S_N(p_N, j)$ are related through the obvious relationship

$$W_N(p_N, k) > j \Leftrightarrow S_N(p_N, j) < k + 1 , \tag{1.1}$$

which yields a relationship between the asymptotic distributions of the two random variables. We have

THEOREM 1*. (i) *If* $j_N = j$ *fixed, and* $p_N \to p$ *then* $S_N(p_N, j) \to_{\mathcal{L}} B(j, p)$.
(ii) *If* $j_N \to \infty$ *and* $0 \leqq \lim j_N p_N = \nu < \infty$ *then* $S_N(p_N, j_N) \to_{\mathcal{L}} Po(\nu)$.

Theorems 2* to 4* are under the assumption $j_N \to \infty$. Set $\lim j_N(1 - p_N) = \mu^*$, $\lim j_N^2/N = \lambda^*$.

THEOREM 2*. *If* $0 \leqq \mu^* + \lambda^* < \infty$ *then* $j_N - S_N(p_N, j_N) \to_{\mathcal{L}} Po(\mu^* + \lambda^*/2)$.

THEOREM 3*. *If*

$$V_N^2(j_N) = Ne^{-p_N j_N/N}\{1 - e^{-p_N j_N/N}(1 + p_N^2 j_N/N)\} \to \infty . \tag{1.2}$$

then $\{S_N(p_N, j_N) - ES_N(p_N, j_N)\}\{\text{Var } S_N(p_N, j_N)\}^{-\frac{1}{2}} \to_{\mathcal{L}} N(0, 1)$.

THEOREM 4*. *If* $p_N j_N/N - \log N \to \rho$, $-\infty < \rho \leqq \infty$, *then* $\{N - S_N(p_N, j_N)\} \to_{\mathcal{L}} Po(e^{-\rho})$.

REMARK TO THEOREM 3*. Condition (1.2) can (without loss of generality) be split into the following three possibilities.

(i) $p_N j_N/N \to 0$ and $p_N j_N\{1 - p_N + (p_N - \frac{1}{2})p_N j_N/N\} \to \infty$ (i.e. $p_N j_N \to \infty$ and also $j_N(1 - p_N) \to \infty$ or $j_N^2/N \to \infty$).
(ii) $p_N j_N/N \to c$, $0 < c < \infty$.
(iii) $p_N j_N/N \to \infty$ and $p_N j_N/N - \log N \to -\infty$.

REMARK TO THEOREMS*. Again it is easily checked that the various conditions on j_N and p_N in Theorems 1* to 4* essentially take care of all possible relationships between j_N and p_N, and hence, in each case the conditions stated are essentially necessary and sufficient for the corresponding results to obtain.

For the case $p_N \equiv 1$ Theorems 2* to 4* are well-known theorems for the classical occupancy problem. See e.g. Rényi (1962). In his recent paper, Park (1972) proves Theorem 3* under more restrictive conditions than our (1.2). His proof essentially uses the method of moments, and is entirely different from the proof given here.

Results related to the results obtained here are obtained in a recent paper, Samuel (1973), which considers a different version of the occupancy problem. In Samuel (1973) the underlying model is, that a cell, once occupied, becomes magnetized, and at each subsequent drop of a ball is $\gamma > 0$ times as likely to receive the ball, than each empty cell.

2. Proof of Theorems 1 and 1*. It is easily seen that

$$W_N(p_N, k) = X_0(N) + \cdots + X_k(N) \tag{2.1}$$

where $X_r(N)$ are independent random variables and $X_r(N)$ denotes the waiting time from the time exactly r cells are occupied until and including the time the $r+$ 1st cell becomes occupied. Thus $X_r(N)$ is a geometric random variable, i.e.,

$$P(X_r(N) = u) = p_r(N)q_r^{u-1}(N) \qquad u = 1, \cdots, \tag{2.2}$$

where $p_r(N) = (N - r)p_N/N$ and $q_r(N) = 1 - p_r(N)$.
Now under the conditions of Theorem 1, if $p_N \to p > 0$ then also $p_r(N) \to p > 0$ and the result follows trivially. Similarly, for fixed j the result in Theorem 1* is trivial, (with the obvious interpretations for $p_N \to 0$ and $p_N \to 1$). The characteristic function of $X_r(N)$ is

$$E(e^{itX_r(N)}) = p_r(N)e^{it}/(1 - q_r(N)e^{it}) . \tag{2.3}$$

To obtain the second part of Theorem 1, notice that

$$E(e^{it2p_N W_N(p_N,k)}) = e^{it2(k+1)p_N} \prod_{r=0}^{k} p_r(N)/(1 - q_r(N)e^{it2p_N}) . \tag{2.4}$$

Now expanding the denominator of each term in the product, it is easily seen that each term tends to $(1 - 2it)^{-1}$ as $p_N \to 0$, and thus (2.4) tends to $(1 - 2it)^{-(k+1)}$ and the theorem follows.

Let $F_{Po(\lambda)}(x)$ and $F_{\chi^2_{(s)}}(x)$ denote the cumulative Poisson and chi-square distributions with corresponding parameters. Then it is well known and easily established through integration by parts, that

$$F_{Po(\lambda)}(s - 1) = 1 - F_{\chi^2_{(2s)}}(2\lambda) . \tag{2.5}$$

Notice that $j_N \to \infty$ and $p_N j_N \to \nu < \infty$ imply $p_N \to 0$. The second part of Theorem 1* thus follows from Theorem 1, (1.1) and (2.5).

3. Proof of Theorems 2 and 2*. Notice that the assumption $\mu < \infty$ and

$k_N \to \infty$ implies $p_N \to 1$. By (2.1) and (2.3) the characteristic function $M_{k_N}(t)$ of $W_N(p_N, k_N) - (k_N + 1)$ is

$$(3.1) \qquad M_{k_N}(t) = \prod_{r=0}^{k_N} p_r(N)/(1 - q_r(N)e^{it})$$
$$= {p_N}^{k_N+1} \prod_{r=0}^{k_N} \left(1 - \frac{r}{N}\right) \Big/ \left\{1 - (1 - p_N)e^{it} - \frac{r}{N} p_N e^{it}\right\} .$$

Now

$$(3.2) \qquad 1 + z = \exp(z + \theta z^2) \qquad \text{if} \quad |z| \leqq \tfrac{1}{2}, \qquad \text{where} \quad |\theta| \leqq 1$$

and z and θ are real or complex numbers. We shall use the same notation θ for possibly different numbers satisfying $|\theta| \leqq 1$. Since $p_N \to 1$ and $k_N/N \to 0$ then for t fixed and N sufficiently large we can use (3.2) for each term in the numerator and denominator in the product of the right hand side of (3.1), to obtain

$$(3.3) \qquad M_{k_N}(t) = {p_N}^{k_N+1} \exp\left\{(k_N + 1)(1 - p_N)e^{it} + (p_N e^{it} - 1) \sum_{r=0}^{k_N} \frac{r}{N} + \varepsilon_N\right\}$$

where $\varepsilon_N = \theta \sum_{r=0}^{k_N} r^2/N^2 + \theta e^{2it} \sum_{r=0}^{k_N} ((1 - p_N) + (r/N)p_N)^2$. It is easy to verify that the assumptions of the Theorems imply $\lim \varepsilon_N = 0$, $\sum_{r=0}^{k_N} r/N \to \lambda/2$ $(k_N + 1)(1 - p_N) \to \mu$ and ${p_N}^{k_N+1} \to e^{-\mu}$. Thus $M_{k_N}(t) \to \exp\{(\mu + \lambda/2)(e^{it} - 1)\}$, and Theorem 2 follows.

For $k = 0, 1, \cdots$ fixed, (1.1) implies $P(j_N - S_N(p_N, j_N) \leqq k) = P(W_N(p_N, j_N - k - 1) - (j_N - k) \leqq k)$. Since under Theorem 2* $(j_N - k - 1)(1 - p_N) \to \mu^*$ and $(j_N - k - 1)^2/N \to \lambda^*$ Theorem 2* follows from Theorem 2.

4. Proof of Theorems 3 and 3*. We shall use representation (2.1) to show that under the conditions of Theorem 3 the summands in (2.1) satisfy the Lyapunov condition of the central limit theorem. Notice that $EX_r(N) = 1/p_r(N)$, $\operatorname{Var} X_r(N) = q_r(N)/{p_r}^2(N)$. To abbreviate notation set $k_N/N = \alpha_N$, $1 - \alpha_N = \beta_N$. Then

$$(4.1) \qquad \sigma^2(k_N) = \operatorname{Var} W(p_N, k_N) = (N^2/{p_N}^2) \sum_{u=N-k_N}^{N} u^{-2} - (N/p_N) \sum_{u=N-k_N}^{N} u^{-1}$$
$$= \{1 - \beta_N + p_N \beta_N \log \beta_N\}\{N/(\beta_N {p_N}^2)\} + 6\theta/(\beta_N p_N)^2 ,$$

where the right-hand side is obtained by simple approximations of the sums. The conditions of Theorem 3 imply $\sigma^2(k_N) \to \infty$.

To verify the Lyapunov condition, notice that if X is geometrically distributed with parameter p, then

$$E|X - p^{-1}|^3 < \sum_{r<p^{-1}} (p^{-1} - 1)^3 pq^{r-1} + \sum_{r \geqq p^{-1}} (r - 1)^3 pq^{r-1}$$
$$< q^3/p^3 + qEX^3 < 7q/p^3 ,$$

since $EX^3 = (1 + 4q + q^2)/p^3$. If $q > \frac{1}{4}$, $7q/p^3 < 10(\operatorname{Var} X)^{\frac{3}{2}}$ and if $q \leqq \frac{1}{4}$, $7q/p^3 < 10 \operatorname{Var} X$. Thus $E|X - p^{-1}|^3 < 10(\operatorname{Var} X + (\operatorname{Var} X)^{\frac{3}{2}})$, and hence

$$(4.2) \qquad \sum_{r=0}^{k_N} E|X_r(N) - p_r(N)^{-1}|^3/\sigma^3(k_N) < 10\{\sigma(k_N)^{-1} + (\operatorname{Var} X_{k_N}(N))^{\frac{1}{2}}/\sigma(k_N)\} .$$

Since $\sigma(k_N) \to \infty$ if follows from (4.2) that in order to verify Lyapunov's condition it suffices to show that $\{p_{k_N}(N)^2 \sigma^2(k_N)\}^{-1} = \{{p_N}^2 {\beta_N}^2 \sigma^2(k_N)\}^{-1} \to 0$, and this

follows directly from (4.1). (We would like to remark that the idea of the present proof is similar to a proof of a related theorem given by Farm (1971).)

To prove Theorem 3*, let $Z_i(N) = 0$ or 1 according as the ith cell is empty or occupied, after j balls have been distributed. Then $Z_i(N)$, $i = 1, \cdots, N$ are exchangeable, and $S_N(p, j) = \sum_{i=1}^{N} Z_i(N)$. Using this representation it is easy to obtain $ES_N(p, j) = N\{1 - (1 - p/N)^j\}$, and $\operatorname{Var} S_N(p, j) = N\{(1 - p/N)^j - (1 - 2p/N)^j\} - N^2\{(1 - p/N)^{2j} - (1 - 2p/N)^j\}$. (See also Park (1972).) Consider $V_N^2(j_N)$ defined in (1.2). If $V_N^2(j_N) \to \infty$ then one obtains, using $(1 - t) = \exp - (t + t^2/2 + \cdots)$ and some detailed analysis

$$\operatorname{Var} S_N(p_N, j_N) = V_N^2(j_N) + o(Ne^{-p_N j_N/N}) . \tag{4.3}$$

Similarly, some analysis yields $ES_N(p_N, j_N) = N(1 - e^{-p_N j_N/N}) + O(1)$. Thus

$$\lim P\left(\frac{S_N(p_N, j_N) - ES_N(p_N, j_N)}{(\operatorname{Var} S_N(p_N, j_N))^{\frac{1}{2}}} \leqq x\right) = \lim P\left(\frac{S_N(p_N, j_N) - N(1 - e^{-p_N j_N/N})}{V_N(j_N)} \leqq x\right). \tag{4.4}$$

Now by (1.1)

$$\begin{aligned} &P(\{S_N(p_N, j_N) - N(1 - e^{-p_N j_N/N})\} V_N^{-1}(j_N) \leqq x) \\ &\quad = P(\{W_N(p_N, k_N) - EW_N(p_N, k_N)\}\sigma^{-1}(k_N) \\ &\qquad \geqq \{j_N - EW_N(p_N, k_N)\}\sigma^{-1}(k_N)) , \end{aligned} \tag{4.5}$$

where we have put ($[y]$ denoting largest integer contained in y),

$$k_N = [xV_N(j_N) + N(1 - e^{-p_N j_N/N}) - 1] . \tag{4.6}$$

It is easy to establish that under the conditions of Theorem 3*, k_N of (4.6) satisfies the assumption of Theorem 3. Let $\Phi(x)$ denote the cumulative standard normal distribution function. Then the right-hand side of (4.5) tends to $\Phi(x)$ if and only if

$$\{j_N - EW_N(p_N, k_N)\}\sigma^{-1}(k_N) \to -x . \tag{4.7}$$

We shall show that (4.7) is correct. Theorem 3* then follows.

Set $k_N^* = N(1 - e^{-p_N j_N/N})$, and consider the value of $\sigma^2(k_N)$ given in the right-hand side of (4.1), when we allow also noninteger k_N. Since $V_N(j_N) = O([N(1 - e^{-p_N j_N/N})]^{\frac{1}{2}})$, it is easily established that for k_N^* as defined and k_N given in (4.6), $\sigma(k_N)/\sigma(k_N^*) \to 1$, and thus (4.7) follows if we show that $\{j_N - EW_N(p_N, k_N)\}\sigma^{-1}(k_N^*) \to -x$. Now $EW_N(p_N, k_N) = (N/p_N) \sum_{u=N-k_N}^{N} u^{-1}$ and simple integral estimation of the sum yields

$$-(N/p_N) \log \beta_N < EW_N(p_N, k_N) < -(N/p_N) \log (\beta_N - N^{-1}) , \tag{4.8}$$

and hence we shall use the left-hand side of (4.8) as an approximation of $EW_N(p_N, k_N)$. Thus, using (1.2) and (4.1), and omitting the square brackets

in (4.6),

$$\lim \frac{j_N - EW_N(p_N, k_N)}{\sigma(k_N{}^*)}$$
$$= \lim \frac{j_N + (N/p_N)\log\{e^{-p_N j_N/N} - xV_N(j_N)/N\}}{\sigma(k_N{}^*)}$$
$$= \lim \frac{(N/p_N)\log\{1 - e^{p_N j_N/(2N)}x\{1 - e^{-p_N j_N/N}(1 + p_N{}^2 j_N/N)\}^{\frac{1}{2}}N^{-\frac{1}{2}}\}}{(N^{\frac{1}{2}}/p_N)e^{p_N j_N/(2N)}\{1 - e^{-p_N j_N/N}(1 + p_N{}^2 j_N/N)\}^{\frac{1}{2}}} = -x\,.$$

5. Proof of Theorems 4* and 4. So far we have not used the exact distribution of $S_N(p, j)$, which may be of interest in its own right. By using the usual combinatorial formulas for the realization of exactly m among N events (see e.g. Feller (1957) page 96) it is easily found that

$$P(N - S_N(p, j) = m) = \sum_{r=m}^{N}(-1)^{r-m}\binom{r}{m}T_r$$
$$= \binom{N}{m}\sum_{t=0}^{N-m}(-1)^t\binom{N-m}{t}\left(1 - \frac{(t+m)p}{N}\right)^j, \tag{5.1}$$
$$m = 0, 1, \cdots, N,$$

where

$$T_r = T_r(p, j, N) = \binom{N}{r}\left(1 - \frac{pr}{N}\right)^j. \tag{5.2}$$

(See also Park (1972)). Our proof now follows the idea of the proof for $p \equiv 1$, given by Feller (1957) page 93. Since $(N - r)^r \leqq N(N-1)\cdots(N-r+1) \leqq N^r$ we have

$$N^r\left(1 - \frac{r}{N}\right)^r\left(1 - \frac{pr}{N}\right)^j \leqq r!\,T_r \leqq N^r\left(1 - \frac{pr}{N}\right)^j. \tag{5.3}$$

Using $\exp(-t/(1-t)) \leqq 1 - t$ on the left-hand side, and $1 - t \leqq \exp(-t)$ on the right-hand side of (5.3) yields $(1 - r/N)^r\{Ne^{-pj/(N-pr)}\}^r \leqq r!\,T_r \leqq \{Ne^{-pj/N}\}^r$. Now let $j = j_N$ and $p = p_N$ and suppose $Ne^{-p_N j_N/N} \to e^{-\rho} = \gamma \geqq 0$. Then for every fixed r both sides of the inequality tend to γ^r, i.e. $\lim(r!\,T_r) = \gamma^r$. Thus by (5.1) and the bounded convergence theorem we have $\lim P(N - S_N(p_N, j_N) = m) = (1/m!)\sum_{r=m}^{\infty}(-1)^{r-m}\gamma^r/(r-m)! = e^{-\gamma}\gamma^m/m!$, and Theorem 4* follows.

To obtain Theorem 4, let $y > 0$ be arbitrary. From (1.1) $P(\exp\{-W_N(p_N, N - b)p_N/N + \log(2N)\} > 2y) = P(N - S_N(p_N, [(N/p_N)\log(N/y)]) \leqq b - 1)$. Set $j_N = [(N/p_N)\log(N/y)]$. Then j_N satisfies the assumption of Theorem 4* with $\rho = -\log y$ and Theorem 4 now follows from (2.5).

REFERENCES

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

[2] FARM, ANTE (1971). Asymptotic normality in a capture-recapture problem, when catchability is affected by the tagging procedure. Unpublished.

[3] FELLER, W. (1957). *Probability Theory and its Applications*, **1**, 2nd ed. Wiley, New York.

[4] PARK, C. J. (1972). A note on the classical occupancy problem. *Ann. Math. Statist.* **43** 1698–1701.

[5] RÉNYI, A. (1962). Three new proofs and a generalization of a theorem of Irving Weiss. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **7** 203–214.

[6] SAMUEL-CAHN, E. (1973). Asymptotic distribution for the coupon collector's and sampling tagging problems, when tagging affects catchability. To appear.

DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
JERUSALEM, ISRAEL