PUBLICATIONS DE L'INSTITUT MATHÉMATIQUE
Nouvelle série tome 47 (61), 1990, 151–157

ON THE DISTRIBUTION OF WAITING TIME UNTIL k-TH REPETITION OF ANY EVENT

Dragan Banjević

Abstract. The random placing of balls continues until we find that one of boxes has been occupied k times, $k \geq 2$ ("birthday surprise"). The case of unlimited number of alternatives with unequal probabilities is discussed. Some exact and asymptotic formulas for the distribution of waiting time are given.

1. Introduction

In this paper we consider a classical model for waiting times which is also known as "birthday surprise". The random placing of balls continues until we find that one of boxes has been occupied k times, $k \geq 2$. The problem with equally likely alternatives is primarily discussed by Feller (1968), and considered by new methods by Newman (1960), Klamkin and Newman (1967) and Dwass (1969), where interesting results about expected waiting time are given. For the case $k = 2$, interesting asymptotic results are given by Arnold (1972). An overview on the problem is given by Johnson and Kotz (1977). Slightly different approach of the "birthday surprise" problem is considered by Saperstein (1972, 1975) and Naus (1974), with restrictions on the number of balls. Cerasoli (1983, 1984) and Buoncristiani and Cerasoli (1984) use the method given by Dwass (1969) (so called Poisson randomizaton method) and obtained some general results about occupancy problems with applications to "birthday" problem. Here we discuss the case of alternatives without assumption on equal probabilities. In this situation, the case $k = 2$ is considered by Banjević (1974).

Let p_i be the probability of placing one ball in the box number i, $p_i > 0$, $i = 1, 2, \dots, \sum p_i = 1$. Let $N_i(n)$ be the number of balls in the box number i after n independent placings, and $N = \min\{n : \text{for some } i,\ N_i(n) = k\}$.

AMS Subject Classification (1980): Primary 60 C 05, Secondary 60 J 10

Key words and phrases: waiting time, birthday surprise

Noting that $Q_k(n) = P(N > n) = P\big(\bigcap_i \{N_i(n) < k\}\big)$, and using Poisson randomization method introduced in Dwass (1969), we see that

$$g_k(t) = \sum_{n=0}^{\infty} Q_k(n)\frac{t^n}{n!} = \prod_{i=1}^{\infty}\sum_{j=0}^{k-1}\frac{(tp_i)^j}{j!},$$

but this formula is not so convenient for analysis, except in the equiprobable case. In this paper we give some explicit formulas for $Q_k(n)$, and also simple asymptotic formulas in the case $\max_i p_i \to 0$.

2. First formula for $Q_k(n)$

We see that

$$Q_k(n) = \sum_{j=1}^{\infty} \sum_{\substack{n_1+\cdots+n_j=n\\ 1\le n_i<k}} \frac{n!}{n_1!\ldots n_j!} \sum_{1\le i_1<\cdots<i_j} p_{i_1}^{n_1}\cdot\ldots\cdot p_{i_j}^{n_j}. \tag{1}$$

We need (1) in some finite form. In order to obtain this, let

$$p(n_1,\ldots,n_j) = \sum_{i_1,\ldots,i_j} p_{i_1}^{n_1}\cdot\ldots\cdot p_{i_j}^{n_j}, \quad 1\le n_1\le\ldots\le n_j, \tag{2}$$

where the sum is running over different integers $i_1,\ldots,i_j$. Let $p^{(i)} = \sum_j p_j^i$, $i = 1,2,\ldots$. Let $I \subset \{1,2,\ldots,j\}$, and for given $(n_1,n_2,\ldots,n_j)$, let $d(I) = \sum_{i\in I} n_i$, and $p^{[I]} = p^{(d(I))}$. Then we can express $p(n_1,n_2,\ldots,n_j)$ as the function of $p^{(1)}, p^{(2)},\ldots$.

Lemma 1. *We have*

$$p(n_1,\ldots,n_j) = \sum_{m=1}^{j}(-1)^{j-m}\sum_{(A)}(j_1-1)!\ldots(j_m-1)!\sum_{(B)}p^{[I_1]}\ldots p^{[I_m]},$$

where sum (A) is over $1\le j_1\le\ldots\le j_m\le j$, $j_1+\cdots+j_m=j$, *and sum (B) is over* $\{I_1,\ldots,I_m\}$, $I_1+\cdots+I_m=\{1,2,\ldots,j\}$, $|I_i| = j_i$ ($|I| = \mathrm{card}(I)$).

Proof. For $I = \{i_1,\ldots,i_s\}$ let us denote $p(I) = p(n_{i_1},\ldots,n_{i_s})$. From (2) we have

$$\begin{aligned}
p(n_1,\ldots,n_j,n_{j+1}) &= \sum_{i_1,\ldots,i_j} p_{i_1}^{n_1}\cdot\ldots\cdot p_{i_j}^{n_j}\sum_{i_{j+1}}p_{i_{j+1}}^{n_{j+1}}\\
&= \sum p_{i_1}^{n_1}\cdot\ldots\cdot p_{i_j}^{n_j}\big(p^{(n_{j+1})} - p_{i_1}^{n_{j+1}} - \cdots - p_{i_j}^{n_{j+1}}\big)\\
&= p^{(n_{j+1})}p(n_1,\ldots,n_j) - p(n_2,n_3,\ldots,n_j,n_1+n_{j+1})\\
&\quad - p(n_1,n_3,\ldots,n_j,n_2+n_{j+1}) - \cdots - p(n_1,n_2,\ldots,n_{j-1},n_j+n_{j+1}).
\end{aligned} \tag{4}$$

We proceed to derive the formula

$$p(n_1,\ldots,n_{j+1}) = \sum_{I\subset\{1,\ldots,j\}}(-1)^{|I'|-1}(|I'|-1)!p^{[I']}p(I), \tag{5}$$

$I + I' = \{1, 2, \ldots, j+1\}$.

From (4) it is evident that

$$p(n_1, \ldots, n_{j+1}) = \sum_{I \subset \{1, \ldots, j\}} a_j(I) p^{[I']} p(I),$$

for some coefficients $a_j(I)$. From (4) we have $a_j(\{1, \ldots, j\}) = 1$, and

$$\sum_{I \subset \{1, \ldots, j\}} a_j(I) p^{[I']} p(I) = p^{(n_{j+1})} p(n_1, \ldots, n_j) - \sum_{i=1}^{j} \sum_{I \subset \{1, \ldots, j\} \setminus \{i\}} a_{j-1}(I) p^{[I']} p(I).$$

Consider some fixed I, $|I| = r < j$. On the right-hand side, I is absent in exactly r sums and present in exactly $j - r = |I'| - 1$ sums, where $I + I' = \{1, 2, \ldots, j+1\}$, so that $a_j(I) = (-1)(j-r)a_{j-1}(I) = (-1)^{j-r}(j-r)!$, which gives (5). From (5) it is easy to obtain (3). $\square$

Example 1. $p(n_1) = p^{(n_1)}$, $p(n_1, n_2) = p^{(n_1)} p^{(n_2)} - p^{(n_1+n_2)}$, $p(n_1, n_2, n_3) = 2p^{(n_1+n_2+n_3)} - p^{(n_1)} p^{(n_2+n_3)} - p^{(n_2)} p^{(n_1+n_3)} - p^{(n_3)} p^{(n_1+n_2)} + p^{(n_1)} p^{(n_2)} p^{(n_3)}$. $\square$

From (1) and Lemma 1, we obtain

THEOREM 1. *We have*

$$Q_k(n) = \sum_{j=1}^{n} \sum_{(C)} \frac{n!}{n_1! \ldots n_j!} b(n_1, \ldots, n_j) p(n_1, \ldots, n_j), \tag{6}$$

where the sum (C) is over $n_1, \ldots, n_j$ *such that* $1 \le n_1 \le \cdots \le n_j < k$, $n_1 + \cdots + n_j = n$. $b(n_1, \ldots, n_j) = (a_1! \ldots a_s!)^{-1}$ *if* $n_1, \ldots, n_j$ *consists of* s *different groups with* a_i *members in* i*-th group,* $i = 1, 2, \ldots, s$, $a_1 + a_2 + \cdots + a_s = j$. $\square$

From (6) we see that $Q_k(n)$ is a function of $p^{(i)}$, $i = 1, 2, \ldots, n$.

Example 2. Let us denote $p(n_1, \ldots, n_j) = q(j)$ if $n_1 = \ldots = n_j = 1$. Then, from (6) we set

$$Q_2(n+1) = q(n+1) = \sum_{j=1}^{n} (-1)^{n-j} \frac{n!}{j!} p^{(n+1-j)} Q_2(j).$$

This formula was obtained by Banjević (1974). $\square$

3. Second formula for $Q_k(n)$

The formula in Theorem 1 is not convenient neither for large n, nor for approximations. Let

$$\binom{n}{j_1, \ldots, j_t} = \frac{n!}{j_1! \ldots j_t! (n - \sum j_i)!},$$

and

$$p_n(j_1,\dots,j_t)=\sum_{i_1,\dots,i_t} p_{i_1}^{j_1}\cdot\ldots\cdot p_{i_t}^{j_t}(1-p_{i_1}-\cdots-p_{i_t})^{n-\Sigma j_i}, \tag{7}$$

where $i_1,\dots,i_t$ are different integers. It is easy to see that

$$p_n(j_1,\dots,j_t)=\sum_{k_1,\dots,k_t}(-1)^{k_1+\cdots+k_t}\binom{n-\Sigma j_i}{k_1,\dots,k_t}p(j_1+k_1,\dots,j_t+k_t), \tag{8}$$

and $p_n(j_1,\dots,j_t)=p(j_1,\dots,j_t)$, if $n=\sum j_i$.

From the inclusion-exclusion principle, we set

$$R_k(n)=P(N\le n)=1-Q_k(n)=\sum_{1\le t\le n/k}(-1)^t P_t, \tag{9}$$

where

$$P_t=\sum_{(D)}\binom{n}{j_1,\dots,j_t}b(j_1,\dots,j_t)p_n(j_1,\dots,j_t) \tag{10}$$

and the sum (D) is over $j_1,\dots,j_t$, such that $k\le j_1\le\cdots\le j_t$, $j_1+\cdots+j_t\le n$.

Example 3. For $t=1$, $t=2$ from Example 1 and (8) and (9) we obtain $p_n(j)=\sum_{i=0}^{n-j}(-1)^i\binom{n-j}{i}p^{(j+i)}$, and

$$\begin{aligned}
P_1&=\sum_{j=k}^{n}\binom{n}{j}p_n(j)=\sum_{i=k}^{n}(-1)^{i-k}\binom{i-1}{k-1}\binom{n}{i}p^{(i)}\\
&=\binom{n}{k}p^{(k)}-k\binom{n}{k+1}p^{(k+1)}+\binom{k+1}{2}\binom{n}{k+2}p^{(k+2)}-\cdots,\\
P_2&=\frac{1}{2}\binom{n}{k,k}p^{(k)}p^{(k)}-k\binom{n}{k,k+1}p^{(k)}p^{(k+1)}+\cdots\\
&\quad-\frac{1}{2}\binom{n}{k,k}p^{(2k)}+k\binom{n}{k,k+1}p^{(2k+1)}+\cdots.
\end{aligned}$$

If $k\le n<2k$, $R_k(n)=P_1$, if $2k\le n<3k$, $R_k(n)=P_1-P_2$. In general, by Bonferoni's inequality, $P_1-P_2\le R_k(n)\le P_1$. □

Remark 1. The formula for P_t contains only terms of the form $p^{(i_1)}p^{(i_2)}\cdot\ldots\cdot p^{(i_s)}$, $s\le t$, $i_1+\cdots+i_s\ge kt$. Then the coefficient related to $p^{(r)}$, $k\le r<2k$, in $R_k(n)$, is the same as the corresponding one in P_1, as well as one for $p^{(r)}$, $2k\le r<3k$ in P_1-P_2, and one for $p^{(r)}p^{(i)}$, $r,i\ge k$, $2k\le r+i<3k$ in $-P_2$. □

Example 4. Let us consider directly the equiprobable case, i.e. $p_i=i/M$, $i=1,2,...,M$. Let for given k,M $Q_k(M,n)=P(N>n)$. Let $f(M,n,m)$ be the number of permutations of M objects, of the length n, such that any object may appear at most m times, $m\ge1$, $M\ge1$, $n\ge1$ (n — permutations with limited repetition, see Frucht (1966) and Mendelson (1981)). It is easy to see that $f(M,n,m) = M^n$ for $n \le m$, and $f(M,n,m) = 0$ for $n > Mm$, and that

$$f(M,n,m) = \sum_{i=0}^{m} \binom{n}{i} f(M-1, n-i, m). \tag{11}$$

From $Q_k(M,n) = f(M,n,k-1)/M^n$ and (11) we have the formula

$$Q_k(M,n) = \sum_{i=0}^{k-1} \binom{n}{i} \left(\frac{1}{M}\right)^i \left(1 - \frac{1}{M}\right)^{n-i} Q_k(M-1, n-i), \tag{12}$$

where $Q_k(M,n) = 1$, $n \le k-1$, $Q_k(M,n) = 0$, $n \ge (k-1)M$.

Mendelson (1981) gives another recursive formula for $f(M,n,m)$ which gives

$$Q_k(M,n+1) = Q_k(M,n) - \binom{n}{k-1} \left(\frac{1}{M}\right)^{k-1} \left(1 - \frac{1}{M}\right)^{n-k+1} Q_k(M-1, n-k+1). \tag{13}$$

Let $m(k) = \min\{n : Q_k(M,n) \le 0{,}5\}$ be median of the distribution. Using (13), calculation gives values for $m(k)$, expectation $E_k(N)$ and standard deviation $s_k(N)$ in Table 1. In the case $k = 2$, good approximations for $E(N)$ are given in McCabe (1970). □

4. Asymptotic formulas for $Q_k(n)$

Let p_i be ordered by magnitude, i.e. $1 > p_1 \ge p_2 \ge \cdots$. Then we have

THEOREM 2. *Let $p_1 \to 0$. Then*

$$R_k(n) = 1 - Q_k(n) \sim \sum_{i=k}^{m} (-1)^{i-k} \binom{i-1}{k-1} \binom{n}{i} p^{(i)}, \tag{14}$$

$m = \min\{n, 2k-2\}$, $k \ge 2$.

Proof. We see that $p^{(r)} < p^{(j)}$ for $r > j$, and $p^{(j_1)} \cdot \ldots \cdot p^{(j_t)} \le (p^{(k)})^t \le (p^{(k)})^2$ for $j_1, \ldots, j_t \ge k$, $t \ge 2$. Also $p^{(k)} \le p_1 p^{(k-1)} \le p_1^{k-1} \to 0$, and $p^{(s)}/p^{(j)} \le p_1^{s-j} \to 0$, $s > j$, if $p_1 \to 0$. We shall prove that $p^{(2k)} \le (p^{(k)})^2 \le p^{(2k-1)}$. The first inequality is obtained from $p(k,k) = p^{(k)}p^{(k)} - p^{(2k)} \ge 0$ (Example 1). In order to obtain the second, let the random variable X be such that $P(X = i) = p_i$, and $f(i) = p_i^{k-1}$. Then $p^{(2k-1)} = E(f(X))^2 \ge (Ef(X))^2 = (p^{(k)})^2$. We have $(p^{(k)})^2/p^{(2k-2)} \le p^{(2k-1)}/p^{(2k-2)} \le p_1 \to 0$, so that $p^{(s)} = o(p^{(2k-2)})$, $s > 2k-2$, and $p^{(j_1)} \cdot \ldots \cdot p^{(j_t)} = o(p^{(2k-2)})$, $j_1, \ldots, j_t \ge k$, $t \ge 2$. By Example 3 and Remark 1, we obtain the theorem. Notice that for $p_i = 1/M$, $i = 1, 2, \ldots, M$, $p^{(2k-1)} = (p^{(k)})^2$, so that formula (14) in the general case is the best formula which contains only "linear" terms $p^{(j)}$, $j \ge k$. □

THEOREM 3. *$Q_k(n) = \lim_{j\to\infty} H(j,n)$, where $H(j,n)$ satisfy the recursive equation*

$$H(j,n) = \sum_{i=0}^{k-1} p_j^i \binom{n}{i} H(j-1,n-i), \tag{15}$$

with $H(j,0) = 1$, $H(1,n) = p_1^n$, $n < k$, $H(1,n) = 0$, $n \geq k$, $p_1 \geq p_2 \geq \dots$.

Proof. Let $H(j,n) = \sum_{\substack{n_1+\dots+n_j=n \\ 0\leq n_i<k}} \frac{n!}{n_1!\cdot\ldots\cdot n_j!} p_1^{n_1}\cdot\ldots\cdot p_j^{n_j}$. Then it is easy to obtain (15). So

$$\begin{aligned} Q_k(n) &= \lim_{j\to\infty} P\left(\bigcap_{i=1}^{j}\{N_i(n) < k\}\right) \\ &= \lim_{j\to\infty} \sum_{m=0}^{n} \sum_{\substack{n_1+\dots+n_j=m \\ 0\leq n_i<k}} \binom{n}{n_1,\dots,n_j} p_1^{n_1}\cdot\ldots\cdot p_j^{n_j}(1-p_1-\dots-p_j)^{n-m} \\ &= \lim_{j\to\infty} \sum_{m=0}^{n} (1-p_1-\dots-p_j)^{n-m}\binom{n}{m} H(j,m) = \lim_{j\to\infty} H(j,n), \end{aligned}$$

from $0 \leq H(j,m) \leq 1$ and $(1-p_1-\dots-p_j)^{n-m} \to 0$, $j\to\infty$, for $0 \leq m < n$. □

We note that, in certain way, Theorem 3 is a generalization of (12).

The method used in Theorem 3 gives the possibility for the generalization of the model for waiting time. Let the box number i be fully occupied if it contains k_i balls, $k_i \geq 1$, $i = 1,2,\dots$ and let placing of balls continue until one of boxes has been occupied. Then we obtain

$$Q(n \mid k_1,k_2,\dots) = \lim_{j\to\infty} P\left(\bigcap_{i=1}^{j}\{N_i(n) < k_i\}\right) = \lim_{j\to\infty} H(j,n),$$

$$H(j,n) = \sum_{i=0}^{k_j-1} p_j^i \binom{n}{i} H(j-1,n-i),$$

$$H(j,0) = 1, \qquad H(1,n) = p_1^n, \quad n < k_1, \qquad H(1,n) = 0, \quad n \geq k_1.$$

Table 1. Median $m(k)$, expected waiting time $E_k(N)$ and standard deviation $s_k(N)$ for waiting of k repetitions for $M = 365$ birthdays.

k	2	3	4	5	6	7	8	9	10
$m(k)$	23	88	187	313	460	623	798	985	1181
$E_k(N)$	24,6	88,7	187,1	311,5	456,0	616,6	790,3	975,0	1168,7
$s_k(N)$	12,2	32,8	56,1	79,7	102,7	124,9	146,3	167,3	186,7

REFERENCES

[**1**] D. Banjević, *Generalization of waiting times model*, Mat. Vesnik **11** (**26**), (1974) 3–9.

[**2**] M. Cerasoli, *Poisson randomization in occupancy problems*, J. Math. Anal. Appl. **94** (1983), 150–165.

[**3**] M. Cerasoli, *Studio di leggi d'occupazione mediante misture poissoniane*, Boll. Un. Mat. Ital. A (**6**) **3** (1984), 289–295.

[**4**] J. H. Buoncristiani and M. Cerasoli, *Multidimensional occupancy problems with Poisson randomization*, J. Math. Anal. Appl. **104** (1984), 526–536.

[**5**] M. Dwass, *More birthday surprises*, J. Combin. Theory **7**, (1969), 258–261.

[**6**] W. Feller, *An Introduction to Probability Theory and its Applications, Vol. 1*, Wiley, New York, 1968.

[**7**] R. Frucht, *Permutations with limited repetitions*, J. Combin. Theory **1** (1966), 195–201.

[**8**] N. L. Johnson and S. Kotz, *Urn Models and Their Applications*, Wiley, New York, 1977.

[**9**] M. S. Klamkin and D. J. Newman, *Extension of the birthday surprise*, J. Combin. Theory **3** (1967), 279–282.

[**10**] B. McCabe, *Elementary problem E 2263*, Amer. Math. Monthly **77** (1970), 1008.

[**11**] H. Mendelson, *On permutations with limited repetition*, J. Combin. Theory, Ser. A **30** (1981), 351–353.

[**12**] J. Naus, *Probabilities for a generalized birthday problem*, J. Amer. Statist. Assoc. **69** (1974), 810–815.

[**13**] D. J. Newman, *The double dixie cup problem*, Amer. Math. Monthly **67** (1960), 58–61.

[**14**] B. Saperstein, *The generalized birthday problem*, J. Amer. Statist. Assoc. **67** (1972), 425–428.

[**15**] B. Saperstein, *Note on a clustering problem*, J. Appl. Probab. **12** (1975), 629–632.

[**16**] B. C. Arnold, *The waiting time until first duplication*, J. Appl. Prob. **9** (1972), 841–846.

Matematički fakultet
11001 Beograd, p.p. 550
Jugoslavija

(Received 13 09 1989)