This supplementary electronic material contains the Annexes of the paper: The coupon collector urn model with unequal probabilities in ecology and evolution. Zoroa, N., Lesigne, E., Fernández-Sáez, M.J., Zoroa, P. & Casas, J. Published in Journal of the Royal Society Interface.

A Study of Y_n

This annex includes the mathematical developments necessary to obtain the expressions for the probability mass function, distribution function, expectation, and variance of Y_n, given in Section 3.

By applying the general inclusion and exclusion principle, we find that, for any distinct elements $j_1, j_2, ..., j_k$ of H, and denoting $p_{j_1 j_2 ... j_k} = p_{j_1} + p_{j_2} + ... + p_{j_k}$,

$$P(\text{the set of parasitized hosts after } n \text{ draws is } \{j_1, j_2, ..., j_k\}) =$$

$$p^n_{j_1 j_2 ... j_k} - \sum_{\{i_1, i_2, ..., i_{k-1}\} \subset \{j_1, j_2, ..., j_k\}} p^n_{i_1 i_2 ... i_{k-1}} +$$

$$\sum_{\{i_1, i_2, ..., i_{k-2}\} \subset \{j_1, j_2, ..., j_k\}} p^n_{i_1 i_2 ... i_{k-2}} - ... + (-1)^{k-1} \sum_{i \in \{j_1, j_2, ..., j_k\}} p^n_i,$$

from which we deduce that

$$P(Y_n = k) =$$

$$\sum_{\{j_1, j_2, ..., j_k\} \subset \{1,2,...,N\}} p^n_{j_1 j_2 ... j_k} -$$

$$\binom{N-k+1}{N-k} \sum_{\{j_1, j_2, ..., j_{k-1}\} \subset \{1,2,...,N\}} p^n_{j_1 j_2 ... j_{k-1}} +$$

$$\binom{N-k+2}{N-k} \sum_{\{j_1, j_2, ..., j_{k-2}\} \subset \{1,2,...,N\}} p^n_{j_1 j_2 ... j_{k-2}} - ...$$

$$+(-1)^{k-1} \binom{N-1}{N-k} \sum_{\{j\} \subset \{1,2,...,N\}} p^n_j. \quad (1)$$

Using the notation $p_A = \sum_{i \in A} p_i$ for any $A \subset H$, this can be written in a more compact form

$$P(Y_n = k) = \sum_{A \subset H, |A| \leq k} (-1)^{k-|A|} \binom{N-|A|}{k-|A|} p^n_A, \text{ for } \min\{N, n\}, \quad (2)$$

where $|A|$ denotes the number of elements of the set A.

Let us now consider the distribution function of Y_n,

$$P(Y_n \leq k) = \sum_{j=1}^{k} P(Y_n = j) = \sum_{j=1}^{k} \sum_{A \subset H, |A| \leq j} (-1)^{j-|A|} \binom{N-|A|}{j-|A|} p_A^n =$$

$$\sum_{A \subset H, |A| \leq k} \left(\sum_{i=0}^{k-|A|} (-1)^i \binom{N-|A|}{i} \right) p_A^n. \tag{3}$$

As, for any integers K and $k \geq 0$ the equality

$$\sum_{i=0}^{k} (-1)^i \binom{K}{i} = (-1)^k \binom{K-1}{k}$$

holds, we obtain

$$P(Y_n \leq k) = \sum_{A \subset H, |A| \leq k} (-1)^{k-|A|} \binom{N-|A|-1}{k-|A|} p_A^n, \quad k = 1, 2, ..., N. \tag{4}$$

A similar expression can be seen in [4]. From (4) we can calculate the moments of Y_n. Let

$$m_k^{[n]} = \sum_{A \subset H, |A| = k} p_A^n,$$

for every $k \leq N$

$$\sum_{j=1}^{k} P(Y_n \leq j) = \sum_{l=1}^{k} (-1)^{k-l} \binom{N-l-2}{k-l} m_l^{[n]},$$

this gives, with $k = N-1$ and $k = N$,

$$\sum_{j=1}^{N-1} P(Y_n \leq j) = m_{N-1}^{[n]},$$

$$\sum_{j=1}^{N} P(Y_n \leq j) = m_N^{[n]} + m_{N-1}^{[n]} = 1 + m_{N-1}^{[n]}.$$

And, bearing in mind that

$$E(Y_n) = \sum_{j=1}^{N} P(Y_n \geq j) = 1 + \sum_{j=1}^{N} P(Y_n > j) = 1 + N - \sum_{j=1}^{N} P(Y_n \leq j),$$

we obtain the well known formula:

$$E(Y_n) = N - m_{N-1}^{[n]} = N - \sum_{i=1}^{N} (1-p_i)^n. \tag{5}$$

We were unable to find any expression for $E(Y_n^2)$ and the variance of Y_n, in previous studies. These two quantities can be obtained as follows. We compute, for $k \leq N$

$$\sum_{t=1}^{k}\sum_{j=1}^{t} P(Y_n \leq j) = \sum_{l=1}^{k}(-1)^{k-l}\binom{N-l-3}{k-l} m_l^{[n]},$$

then, for $k = N-2,\ N-1$ and N we obtain

$$\sum_{t=1}^{N-2}\sum_{j=1}^{t} P(Y_n \leq j) = \ m_{N-2}^{[n]},$$

$$\sum_{t=1}^{N-1}\sum_{j=1}^{t} P(Y_n \leq j) = \ m_{N-2}^{[n]} + m_{N-1}^{[n]},$$

and

$$\sum_{t=1}^{N}\sum_{j=1}^{t} P(Y_n \leq j) = \ m_{N-2}^{[n]} + 2m_{N-1}^{[n]} + m_N^{[n]}.$$

The last identity can be written:

$$\sum_{j=1}^{N} \frac{(N-j+1)(N-j+2)}{2} P(Y_n = j) = \ m_{N-2}^{[n]} + 2m_{N-1}^{[n]} + m_N^{[n]},$$

this gives

$$\frac{(N+1)(N+2)}{2} - \frac{2N+3}{2}E(Y_n) + \frac{1}{2}E(Y_n^2) = \ m_{N-2}^{[n]} + 2m_{N-1}^{[n]} + m_N^{[n]},$$

therefore

$$E(Y_n^2) = 2m_{N-2}^{[n]} - (2N-1)m_{N-1}^{[n]} + N^2 =$$

$$2\sum_{1\leq i<j\leq N}(1-p_i-p_j)^n - (2N-1)\sum_{i=1}^{N}(1-p_i)^n + N^2$$

and

$$Var(Y_n) = 2\sum_{1\leq i<j\leq N}(1-p_i-p_j)^n + \sum_{i=1}^{N}(1-p_i)^n\left(1-\sum_{i=1}^{N}(1-p_i)^n\right). \quad (6)$$

B Expectation of T_{k,N_1}

To prove Proposition 4.2 obtaining the expected value of T_{k,N_1}first we need to compute the probability of event $D_{i_1 i_2 \ldots i_k}$,

$$P(D_{i_1 i_2 \ldots i_k}) = P(\text{first host of } H_1 \text{ parasitized is } i_1)$$

P(second host of H_1 parasitized is i_2 | first host of H_1 parasitized was i_1) ...

$P(k-\text{th} \quad \text{host of } \mathrm{H_1} \text{ parasitized is} i_k | \text{ first } (k-1) \text{ hosts of } H_1 \text{ parasitized were } i_1, i_2, ..., i_{k-1})$.

Both in the case $q = 0$ ($H_1 = \{1, 2, ..., N\}$) and the case $q > 0$

$$P(\text{first parasitized host of } H_1 \text{ is } i_1) = \frac{p_{i_1}}{1-q},$$

where

$$q = \sum_{i \in H_2} p_i > 0.$$

For the rest of the factors

$P(h - \text{th parasitized host of } H_1 \text{ is } i_h | \text{ first parasitized hosts of } H_1 \text{ were } i_1, \ i_2, ..., i_{h-1}) =$

$$\sum_{r=0}^{\infty} p_{i_h} (q + \sum_{j=1}^{h-1} p_{i_j})^r = \frac{p_{i_h}}{1 - q - \sum_{j=1}^{h-1} p_{i_j}}, \ h = 1, 2, ..., k, \text{ where } q = \sum_{i \in H_2} p_i,$$

therefore,

$$P(D_{i_1 i_2 ... i_k}) = \frac{\prod_{j=1}^{k} p_{i_j}}{p(p - p_{i_1})(p - p_{i_1} - p_{i_2})...(p - \sum_{j=1}^{k-1} p_{i_j})}. \tag{7}$$

Let Π_k be the set of all k-permutations of 1, 2, ..., N_1. Then the events $D_{i_1 i_2 ... i_k}$ with $(i_1 i_2 ... i_k) \in \Pi_k$ constitute a partition of Ω, i.e. $D_{i_1 i_2 ... i_k} \cap D_{j_1 j_2 ... j_k} = \oslash$ if $i_1 i_2 ... i_k \neq j_1 j_2 ... j_k$ and

$$\sum_{(i_1 i_2 ... i_k) \in \Pi_k} P(D_{i_1 i_2 ... i_k}) = 1.$$

We can then write $E(T_{k,N_1})$ as follows,

$$E(T_{k,N_1}) = \sum_{(i_1 i_2 ... i_k) \in \Pi_k} E(T_{k,N_1} | D_{i_1 i_2 ... i_k}) P(D_{i_1 i_2 ... i_k}). \tag{8}$$

To compute the conditional expectations $E(T_{k,N_1} | D_{i_1 i_2 ... i_k})$, let us denote by X_h the random variable representing the number of draws elapsed after $h-1$ hosts of the set H_1 being parasitized and before a new host of the set H_1 is parasitized, $1 \leq h \leq k$. We can then write

$$T_{k,N_1} = X_1 + 1 + X_2 + 1 + ... + X_k + 1 = X_1 + X_2 + ... + X_k + k. \tag{9}$$

and therefore

$$E(T_{k,N_1}|D_{i_1 i_2 ... i_k}) = \sum_{h=1}^{k} E(X_h|D_{i_1 i_2 ... i_k}) + k \tag{10}$$

but

$$E(X_h|D_{i_1 i_2 ... i_k}) = E\left(X_h|\text{already parasitized hosts are those of } H_2 \text{ and } i_1 i_2 ... i_{h-1}\right) \tag{11}$$

One direct application of (4.1) would then be:

$$E(X_h|D_{i_1 i_2 ... i_k}) = \frac{q + \sum_{j=1}^{h-1} p_{i_j}}{p - \sum_{j=1}^{h-1} p_{i_j}} \quad \text{for } h = 1, 2, ..., k. \tag{12}$$

From (10) and (12)

$$E(T_{k,N_1}|D_{i_1 i_2 ... i_k}) = \left(\sum_{h=1}^{k} \frac{q + \sum_{j=1}^{h-1} p_{i_j}}{p - \sum_{j=1}^{h-1} p_{i_j}}\right) + k = \sum_{h=1}^{k} (\frac{q + \sum_{j=1}^{h-1} p_{i_j}}{p - \sum_{j=1}^{h-1} p_{i_j}} + 1) =$$

$$\frac{1}{p} + \frac{1}{p - p_{i_1}} + \frac{1}{p - p_{i_1} - p_{i_2}} + ... + \frac{1}{p - \sum_{j=1}^{k-1} p_{i_j}} =$$

$$\frac{1}{1-q} + \frac{1}{1 - q - p_{i_1}} + ... + \frac{1}{1 - q - \sum_{j=1}^{k-1} p_{i_j}}, \tag{13}$$

where

$$q = \sum_{i \in H_2} p_i = \sum_{i=N_1+1}^{N} p_i \quad \text{and} \quad p = \sum_{i \in H_1} p_i = \sum_{i=1}^{N_1} p_i.$$

Bearing in mind (8), (7) and (13) it follows that

$$E(T_{k,N_1}) = \sum_{(i_1 i_2 ... i_k) \in \Pi_k} \left(\frac{1}{p} + \frac{1}{p - p_{i_1}} + \frac{1}{p - p_{i_1} - p_{i_2}} + ... + \frac{1}{p - \sum_{j=1}^{k-1} p_{i_j}}\right)$$

$$\frac{\prod_{j=1}^{k} p_{i_j}}{p(p - p_{i_1})(p - \sum_{j=1}^{2} p_{i_j})...(p - \sum_{j=1}^{k-1} p_{i_j})}.$$

C Proof of Theorem 6.1

To obtain this proof first we state the following lemmas.

Lemma C.1. Let k and N be integers satisfying $1 < k \leq N-1$, then

$$\sum_{0\leq r\leq k-1} (-1)^{k-1-r} \binom{N-2}{r} \binom{N-2-r}{k-1-r} = 0.$$

Proof. For $x \in \mathbb{R}$ the equality

$$\sum_{0\leq r\leq n} \binom{a}{r} \binom{x}{n-r} = \binom{a+x}{n},$$

is satisfied, where $\binom{x}{h}$ is defined by (3.1) then

$$\sum_{0\leq r\leq n} \binom{a}{r} \binom{-b}{n-r} = \binom{a-b}{n},$$

and

$$\sum_{0\leq r\leq n} (-1)^{n-r} \binom{a}{r} \binom{b+n-r-1}{n-r} = \sum_{0\leq r\leq n} \binom{a}{r} \binom{-b}{n-r} = \binom{a-b}{n}.$$

Then, we obtain, with $a = N-2$, $n = k-1$, $b = N-k$

$$\sum_{0\leq r\leq k-1} (-1)^{k-1-r} \binom{N-2}{r} \binom{N-2-r}{k-1-r} = \binom{k-2}{k-1} = 0,$$

and the lemma follows.

Lemma C.2. Let q_1, q_2, ...,q_M be non-negative real numbers and $I = \{1, 2, ..., M\}$. For every $A \subset I$ let $q_A = \sum_{i\in A} q_i$. Then, for any integer $m \geq 0$,

$$\sum_{A\subset I, |A|\leq r} (-1)^{r-|A|} \binom{M-|A|}{r-|A|} q_A^m \geq 0.$$

Moreover, if $m \geq r$ and at least r of the values q_1, q_2, ...,q_M are greater than zero, then

$$\sum_{A\subset I, |A|\leq r} (-1)^{r-|A|} \binom{M-|A|}{r-|A|} q_A^m > 0.$$

Proof. If all the q_i are zero, there is nothing to prove. Let us suppose that $s = \sum_{i=1}^{M} q_i > 0$. Let $p_i = q_i/s$, $i = 1, 2, ..., M$. These values define the probability distribution $p = (p_1, p_2, ..., p_M)$ on I. From (3.3) it follows that

$$\sum_{A \subset I, |A| \le r} (-1)^{r-|A|} \binom{M-|A|}{r-|A|} q_A^m =$$

$$s^m \sum_{A \subset I, |A| \le r} (-1)^{r-|A|} \binom{M-|A|}{r-|A|} p_A^m = s^m P_p(Y_m = r),$$

which proves the lemma.

Proof of Theorem 6.1. Let $H' = \{3, 4, ..., N\}$. According to (3.4) we have:

$$P_p(Y_n \le k) = \sum_{A \subset H, |A| \le k} (-1)^{k-|A|} \binom{N-|A|-1}{k-|A|} p_A^n =$$

$$\sum_{A \subset H', |A| \le k} (-1)^{k-|A|} \binom{N-|A|-1}{k-|A|} p_A^n +$$

$$\sum_{A \subset H', |A| \le k-1} (-1)^{k-1-|A|} \binom{N-2-|A|}{k-1-|A|} (p_A + p_1)^n +$$

$$\sum_{A \subset H', |A| \le k-1} (-1)^{k-1-|A|} \binom{N-2-|A|}{k-1-|A|} (p_A + p_2)^n +$$

$$\sum_{A \subset H', |A| \le k-2} (-1)^{k-2-|A|)} \binom{N-3-|A|}{k-2-|A|} (p_A + p_1 + p_2)^n.$$

Then

$$P_p(Y_n \le k) - P_{p'}(Y_n \le k) =$$

$$\sum_{A \subset H', |A| \le k-1} (-1)^{k-1-|A|} \binom{N-2-|A|}{k-1-|A|} (p_A + p_1)^n +$$

$$\sum_{A \subset H', |A| \le k-1} (-1)^{k-1-|A|} \binom{N-2-|A|}{k-1-|A|} (p_A + p_2)^n -$$

$$\sum_{A \subset H', |A| \le k-1} (-1)^{k-1-|A|} \binom{N-2-|A|}{k-1-|A|} (p_A + p_1 + h)^n -$$

$$\sum_{A \subset H', |A| \le k-1} (-1)^{k-1-|A|} \binom{N-2-|A|}{k-1-|A|} (p_A + p_2 - h)^n.$$

Let f be the real function defined by

$$f(x) = \sum_{A \subset H', |A| \leq k-1} (-1)^{k-1-|A|} \binom{N-2-|A|}{k-1-|A|} (p_A + x)^n, \qquad x \in \mathbf{R}$$

This function is a polynomial of degree less than or equal to n. The coefficient of x^n is equal to

$$\sum_{A \subset H', |A| \leq k-1} (-1)^{k-1-|A|} \binom{N-2-|A|}{k-1-|A|} =$$

$$\sum_{0 \leq r \leq k-1} (-1)^{k-1-r} \binom{N-2}{r} \binom{N-2-r}{k-1-r}$$

and this is equal to 0 by Lemma C.1. The coefficient of x^{n-j} for $j = 1, 2, ..., n$ is

$$\binom{n}{j} \sum_{A \subset H', |A| \leq k-1} (-1)^{k-1-|A|} \binom{N-2-|A|}{k-1-|A|} p_A^j,$$

by the first part of Lemma C.2 with $I = H' = \{3, 4, ..., N\}$, $M = |I| = N-2$, $m = j$ and $r = k-1$, it follows that these coefficients are greater than or equal to zero. This polynomial function is then convex on $[0, +\infty)$, so that

$$f(p_1) + f(p_2) \geq f(\alpha p_1 + (1-\alpha)p_2) + f(\alpha p_2 + (1-\alpha)p_1) = f(p_1 + h) + f(p_2 - h),$$

$$f(p_1) + f(p_2) - f(p_1 + h) - f(p_2 - h) \geq 0.$$

However, this inequality is the same as

$$P_p(Y_n \leq k) - P_{p'}(Y_n \leq k) \geq 0,$$

which gives (6.2). Recalling the relationship between the random variables Y_i and the random variables T_j, we also obtain

$$P_p(T_{k+1} \leq n) \leq P_{p'}(T_{k+1} \leq n),$$

which is (6.3).

Moreover, from the second part of Lemma C.2. it follows that if at least $k-1$ of the values $p_3, p_4, ..., p_N$ are greater than zero and $n \geq k+1$, then the coefficient of x^{n-k+1} is greater than zero, where $n - k + 1 \geq 2$. So, at least one monomial of degree greater than or equal to 2 appears in the polynomial. The convexity is then strict, and we can write

$$f(p_1) + f(p_2) - f(p_1 + h) - f(p_2 - h) > 0,$$

and

$$P_p(Y_n \leq k) - P_{p'}(Y_n \leq k) > 0, \quad n = k+1, k+2, k+3...$$

which is equivalent to

$$P_p(Y_n \leq k) > P_{p'}(Y_n \leq k), \quad n = k+1, k+2, k+3...$$

and therefore to

$$P_p(T_{k+1} \leq n) < P_{p'}(T_{k+1} \leq n), \quad n = k+1, k+2, k+3...$$

This completes the proof.